Medical Surgery Stream Segmentation to Detect and Track Robotic Tools

Syed Abdul Khader[*ø]
Massachusetts General Hospital, Harvard Medical School, Plaksha University.
syed.abdul@plaksha.edu.in

Vysakh Ramakrishnan[*ø]
Massachusetts General Hospital, Harvard Medical School, Plaksha University.
vysakh.r@plaksha.edu.in

Arian Mansur
Harvard Medical School.
arianmansur@hms.harvard.edu

Chi-Fu Jeffrey Yang
Massachusetts General Hospital.
cjyang@mgh.harvard.edu

Lana Schumacher
Massachusetts General Hospital.
lschumacher2@mgh.harvard.edu

Sandeep Manjanna
Plaksha University.
msandeep.sjce@gmail.com

***Abstract*—This paper presents a real-time image segmentation and tracking system that can operate on a continuous stream of endoscopic surgical videos. Such a system can find several valuable applications in the medical field, such as building real-time augmentations to assist with robotic surgeries, training medical residents, and summarizing surgery videos to generate reports having the whole understanding of the procedure. We have formulated a segmentation technique that requires minimal supervision and provides real-time tracking of the objects. The results from the evaluation of our approach indicate that even with minimal annotated data from surgeons, we can achieve good segmentation. This reduces the need for extensive and expensive data collection and annotation processes from robotic surgery. We evaluated our approach on two datasets, EndoVis 2017 and 2018, a dataset from the Robotic Instrument sub-challenge from MICCAI 2017 and 2018. Our results are on par with the state-of-the-art methods on EndoVis-17 and EndoVis-18 for binary segmentation, and in the case of multi-class, we do it with EndoVis-17. The main contribution of our paper is to give the segmentation results in the real-time streaming data.**

***Index Terms*—Robotic arm segmentation, Image segmentation, Endoscopic surgery video segmentation**

I. Introduction

Robot-assisted surgery has become increasingly adopted in recent years. It is one of the latest advancements in the field of surgery that has expanded the scope of treatment options for patients [1]. The robotic system translates a surgeon's hand movements onto robotic arms with improved agility, visualization, and precision than conventional open or laparoscopic approaches [2]. However, some challenges with adopting robotic surgery are the loss of haptic sensation and the steep learning curve. This is especially true for fields like thoracic surgery, where several reported cases of intraoperative catastrophes still occur and require conversion to an open thoracotomy [3]. Moreover, the conventional learning model is an apprenticeship model based on subjective observation by experienced surgeons. However, the increasing use of artificial intelligence in medicine has the potential to help build technical competence in trainees with automated feedback and improve patient outcomes. Such a system will require the ability to model gestures and complex actions with motion information. However, a prerequisite to this program requires the precise detection and localization of the robotic tools, especially to train large datasets. In this paper, we present a real-time image segmentation and tracking system that can operate on a continuous stream of endoscopic surgical videos. Such a tool can be very useful in providing the right feedback to the surgeon and also be used for training medical postgraduate trainees with more precise and effective assessment.

da Vinci Surgical System(dVSS) is used in most minimally invasive soft tissue operations, and the instruments' hands are almost similar across the surgeries. image segmentation of endoscopic robotic surgery videos and tool identification has been an active area of research both in robotics and computer vision communities. There have been several instrument segmentation challenges, such as EndoVis-17 [4] and EndoVis18 [5]. Even with continued research, there seems to be an increasing demand for an efficient algorithm to segment out the robotic tools from a continuous real-time input stream of endoscopic surgery video using minimal labeled data. Some of the methods [6]–[8] used various methods from UNet to unsupervised ways, but the results are not effective enough as the surgery videos are very different from natural videos.

Real-time segmentation and tracking the tools in the robotic surgery video streams can assist surgeons during the surgery, thus helping to reduce human errors. Such methods and models can also help in giving continuous and accurate feedback to trainees. Some of the existing models depend on a large datasets to improve the accuracy. On the other hand, there are models that do real-time segmentation, but with less precision. In this paper, we presents a new pipeline to detect and track robotic tools in real-time surgery video streams. This robust pipeline for video instance segmentation inherits the advantages of two popular segmentation models — YOLOv8-seg [9] and XMem [10].

Contributions. In this work, we propose a robust pipeline for real-time robotic arm segmentation during robotic surgery.

* These authors contributed equally to this work.
ø Work done during internship at Thoracic Surgery Research Department, Massachusetts General Hospital.

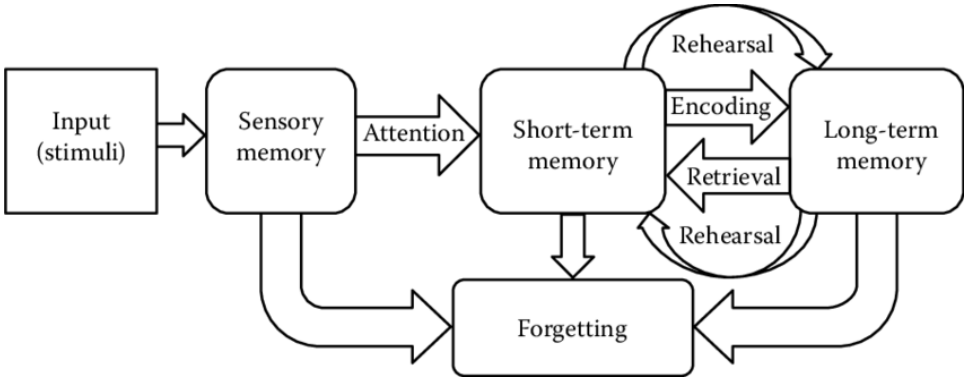

Fig. 1. Concept of Atkinson-Shiffrin memory model

The major contribution of this paper is a real-time segmentation and tracking pipeline to segment out robotic arms and tools from a endoscopic surgery video stream. We also experiment with reducing and eliminating the need for additional annotated data by using pre-trained models in some of the experiments. Evaluation shows that our approach is able to produce state-of-the-art segmentation results on an input video stream in real-time.

II. Related Work

Understanding of robot-assisted surgery (RAS) videos is an active research area. Robot-assisted surgery has witnessed a remarkable surge in growth, driven by the rapid advancements in robotics and imaging systems. These innovations have ushered in a new era of surgical precision characterized by enhanced visual capabilities, tactile feedback, and the dexterity of robotic arms. A fundamental cornerstone of this progress lies in real-time semantic segmentation, a vital process in robot-assisted surgery. This segmentation, marked by its accuracy and efficiency, not only aids in tracking surgical instruments but also provides invaluable context regarding the various tissues and instruments involved in the procedure. Below are some of the related events that happened in this space.

A. Medical Image Segmentation Using Transformer

Transformers have a stronger generalization ability and context extraction ability than convolutional neural network (CNN) structures, which have also been commonly applied in image segmentation. This is because the local receptive fields do not limit transformers as in CNNs and can, therefore, learn long-range dependencies between pixels. One of the specific advantages of transformers for image segmentation is that they can learn global context information, which can help improve the accuracy of segmentation masks. Also, they are not limited by the local connectivity of CNNs, which allows them to learn long-range dependencies between pixels. Zizu et al. [11] proposed TraSeTR, a novel track-to-segment transformer that dynamically integrates tracking cues to assist instance-level surgical instrument segmentation. The primary win for this technique is identity matching and contrastive query learning, which is carefully designed to track surgical instruments with large temporal variations. Yuqing Wang et al. [12] built a video instance segmentation framework built upon Transformers termed VisTR, which views the VIS task as a direct end-to-end parallel sequence decoding/prediction problem. There are also many transformer architecture techniques to do tracking of the surgical instruments. ViT [13] introduces the Transformer to image recognition and models an image as a sequence of patches, which attain excellent results compared to state-of-the-art convolutional networks.

B. Video Instance Segmentation

Recent advances in video instance segmentation have been made possible by the development of deep learning models. Most VIS methods employ a feature memory to store information available in the first frame and use it to segment any new frames. Deep learning models have been shown to be able to learn the complex relationships between objects in videos, which is essential for accurate instance segmentation. The state-of-the-art technique for VIS is MaskFreeVIS [14], achieving highly competitive VIS performance while only using bounding box annotations for the object state. Spatial-temporal graph neural networks(ST-GNNs) are a type of deep learning model that can be used to learn the spatial and temporal relationships between objects in videos. They are effective for various video understanding tasks, including video instance segmentation.

C. Multi-scale feature fusion

Multi-scale feature aggregation is essential for robust performance in instance segmentation due to the wide range of scale variation of objects in these images. Objects in instance segmentation images can vary significantly in size, making it difficult for a single-scale model to segment all objects in the image accurately. In a recent paper by Liu et al. (2023) [15], researchers addressed the challenge of fusing multi-modal medical images, which require accurate feature extraction at various scales due to their intricate and detailed nature. Conventional convolutional neural networks (CNNs) struggle with this task. To tackle this problem, they introduced a novel CNN architecture for multi-scale feature fusion to

enhance the quality of fused multi-modal medical images. The network consists of two trunks, three branches, and fusion modules (FMs) to efficiently combine multi-scale features, resulting in the generation of fused images.

III. Approach

In this work, we propose a robust pipeline Figure 2for real-time Video Instance Segmentation (VIS) that leverages the strengths of two prominent models, YOLOv8-seg [9] for segmentation and XMem [10], to process streaming video data efficiently. The pipeline is delineated into two primary phases: initial segmentation and continuous segmentation propagation.

A. Initial Segmentation

Given the absence of ground truth masks in real-time video streams, the first frame of the video stream is processed using YOLOv8-seg, an advanced segmentation model known for its speed and accuracy. YOLOv8-seg generates a predicted mask that serves as a surrogate for the ground truth mask required by the XMem model [10]. This initial segmentation phase provides the foundation for the subsequent segmentation propagation phase.

B. Segmentation Propagation

The XMem [16] model, inspired by the Atkinson-Shiffrin memory model Figure 1, facilitates long-term video object segmentation through a unified feature memory store. It comprises three types of memory: sensory memory for current image processing, working memory to retain past frames, and long-term memory to assimilate old frames when the working memory reaches capacity. XMem utilizes the predicted mask obtained from YOLOv8-seg in the initial segmentation phase to propagate segmentation across subsequent frames. This mechanism ensures that segmentation is continuously refined and propagated through the video stream, leveraging the memory-based architecture of XMem to accommodate the dynamic nature of video data.

C. Integration and Optimization

The integration of YOLOv8-seg and XMem is orchestrated such that YOLOv8-seg remains operational until a frame containing a robotic arm is detected. Upon detection, the predicted mask and the RGB image are relayed to initialize the XMem model, transitioning the segmentation process to XMem for all subsequent frames. In scenarios where the initial frame lacks a robotic arm, there are no predictions from YOLO, resulting in no object to track for XMem and not having the required mask in future frames when a robotic arm appears. This integration ensures a seamless transition and accurate segmentation throughout the video stream.

An additional optimization is implemented whereby if the segmentation area delineated by YOLO's prediction is less than 3% of the image, the mask is not forwarded to XMem. Instead, YOLO awaits the subsequent frame, persisting until a frame with the robotic arms occupying at least 3% of the image area is encountered. This optimization ensures that only frames with a significant presence of the robotic arm are used to initialize the XMem model to allow the model a more precise understanding of the robotic arm in the video, enhancing the pipeline's efficiency and precision.

IV. Experimental Setup

Datasets. Our study utilizes surgical imagery from the EndoVis challenge datasets of 2017 [4] and 2018 [5]. These datasets consist of endoscopic video frames collected during robot-assisted surgery using the da Vinci Xi surgical system.

The first of the datasets used in this paper was provided by the MICCAI 2017 EndoVis Robotic Instrument Segmentation sub-challenge. The dataset consists of 10 sequences of the abdominal porcine procedures recorded using da Vinci Xi systems. Each sequence consists of 300 frames sampled at a frequency of 1 Hz and has images from two RGB stereo from left and right cameras. For every left camera image, there is a corresponding ground truth mask. We converted the labels in two ways: for binary, we made the arms with the label '1' and everything else as background with the label '0'. For the multi-class segmentation task, we labeled the classes from 1 to 6 for each of the six robotic arms (Bipolar Forceps (BC), Prograsp Forceps (PF), Large Needle Driver (LND), Vessel Sealer (VS), Monopolar Curved Scissors (MCS), Grasping Retractor (GR)) in the sequence. Each image was of the size 1920x1080 but was cropped to a size of 1280x1024 to remove the black borders on both sides. The data is already split into train and test by having the initial 225 frames of the first eight sequences, totaling 1800 frames as the training set, and the remaining 75 frames of each sequence are part of the test. For the last two sequences, the entire 300 frames are used as a test set to check for generalization. In total, we have 1200 frames in the test set.

Another dataset used in this study was from the same challenge but from MICCAI 2018, where they had 19 sequences, of which 15 were combined in the training set, and the remaining four were used for testing. Like EndoVis17, each sequence was sampled at 1Hz and consisted of 300 frames each. In this dataset, the annotations are done differently, as they also have anatomies annotated along with arms, but for this study, the anatomies were excluded, and only the robotic arms were considered for the binary segmentation task. The dataset also doesn't have instrument-level annotations for the robotic arms. Hence, no multi-class segmentation experiments were conducted on this dataset.

Training. The training for both the YOLO and XMem was performed in PyTroch using the official implementation. The YOLO version 8 was used, which also introduces segmentation apart from its original ability for object detection. The model used was Yolov8x-seg, the largest model available with 74M parameters. The training was done for 100 epochs with early-stop on to stop the training if there was no significant improvement after five epochs. The batch size used was 16 with an image size of 640x512, and the training had the usual mosaic augmentations where it uses mix-up and other augmentations. Since each frame doesn't have more than five arms at once,

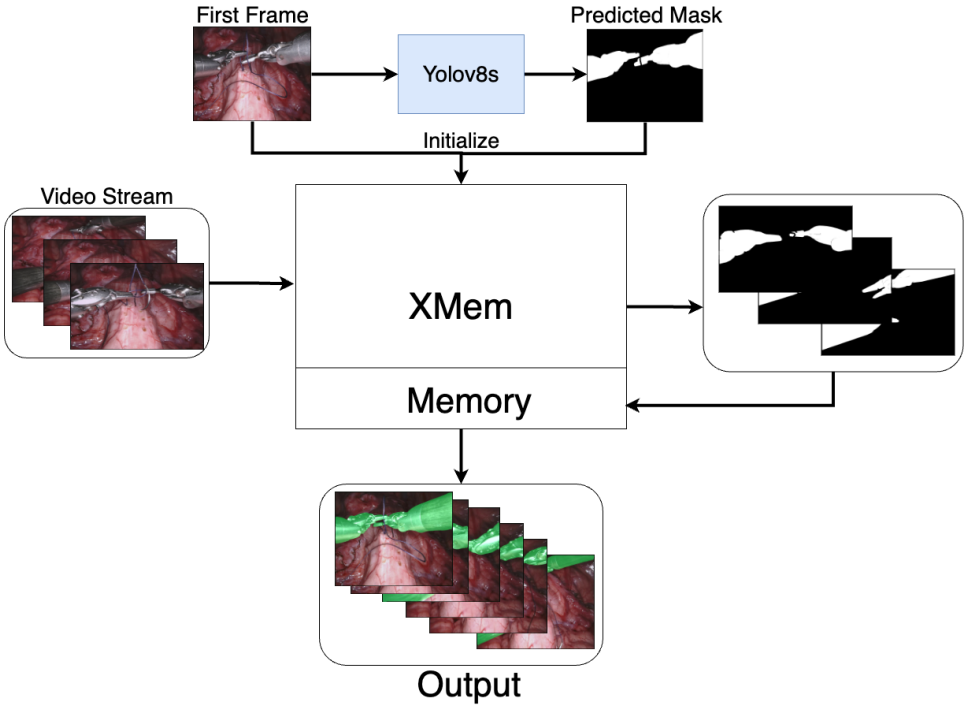

Fig. 2. Architecture of the Pipeline

the max_det variable, which limits the maximum number of instances of an object, is set to 10. Towards the end, for the last ten epochs, the mosaic augmentations are disabled to keep the input to the model as close to the original input and reduce the under-fitting of the model in the final stage, which was proved to work better in the YOLOv8 work by Jocher et al [17].

The standard XMem training consists of 4 stages, namely 0,1,2 and 3 for static images, blender images, main long training, and short main training, respectively. Since our dataset is much smaller than natural image datasets like DAVIS [18] and YouTubeVOS [19], we will individually use the XMem pre-trained weights as a starting point and do stage 3, which is shorter main training on EndoVis 17 and 18. The memory values for update (k), key, value, and hidden dimensions are set to default XMem at 5, 64, 512, and 64, respectively.

Each image was Randomly Resized and cropped at 384x384 with a random scale between 0.36-1 during every iteration. The batch size used for training is 8, and the number of frames to consider for each sequence during training is also 8, per the max GPU memory availability. The GPU being used is A10 with 24GB GPU vRAM. The learning rate was set to 1e-5, and a cosine scheduler was set with warm-up starting at 200 iterations and ending at 700 iterations. The total number of iterations to train was 3000, and finally, a fine-tuned setup was done where the augmentations were reduced for another 500 iterations.

Inference. The inference pipeline is designed to process streaming video data sequentially using the YOLOv8-seg model and the XMem module. Initially, the YOLOv8-seg model analyzes the video frame-by-frame to detect a robotic arm within a frame, with the condition that the detection covers at least 3% of the frame area. Once such a frame is identified, YOLOv8-seg sends the frame and the corresponding predicted mask to XMem to initialize the segmentation process. Following this initialization, the stream of frames is directed to XMem for further segmentation. XMem operates at a speed of 20 frames per second (FPS) to perform segmentation inference. Additionally, it stores the results of past inferences in its working memory every five frames. This operation continues until the working memory reaches a capacity of 6GB of virtual RAM (vRAM), after which XMem consolidates the data and transfers the older frames to Long Term Memory. This structured approach ensures a streamlined operation for real-time segmentation and storage, facilitating the primary objective of robotic arm detection and segmentation in streaming video data.

Metrics. Specifically, we employed a widely recognized metric, the Jaccard Index, also popularly known as intersection-over-union(IoU). It quantifies how well the model can distinguish objects and compares the predicted masks with the hand-labeled ground truth mask. The IoU score lies between 0 and 1, where 1 means there is a perfect overlap and the predicted mask is as good as the ground truth, and an IoU of 0 indicates no overlap at all, where the predicted mask has no intersection with the ground truth mask.

$$IoU = \frac{\sum_i (m_i \cdot n_i)}{\sum_i (m_i + n_i - m_i \cdot n_i)}$$

The mean IoU over all the test patient sequences P is given

by

$$mIoU = \frac{1}{P}\sum_{i=1}^{P}(IoU_i)$$

For multi-class segmentation, the mean intersection over union was done by first averaging the class over all the patient sequences and finally averaging over all the multiple classes, which can be described as:

$$mIoU_{multi} = \frac{1}{C}\sum_{j=1}^{C}(\frac{1}{P}\sum_{i=1}^{P}IoU_i)_j$$

V. Results and Discussion

In the rigorous evaluation for binary segmentation on the EndoVis17 dataset, several models, including UNet [20], TernausNet-16 [21], U-Net++ [22], and TMA-Net [23], among others, were methodically assessed to understand their efficacies in the context of Intersection over Union (IoU) scores. Our proposed model notably attained the highest IoU of 92.26%, demonstrating a subtle yet crucial enhancement in segmentation accuracy over the state-of-the-art TMA-Net, which previously topped the performance charts with an IoU of 91.6%, as shown in Table I. This score of IoU is the mean score averaged over all the ten sequence test sets.

TABLE I
Comparison Results of Various Models on EndoVis2017 Binary Segmentation Dataset

Method	mIoU
ISINet [24]	65.18
Yolov8-seg [17]	77.60
UNet [20]	79.44
TernausNet-16 [21]	83.60
UNet++ [22]	87.21
Nested UNet [25]	87.21
TMA-Net [23]	91.60
Ours	**92.26**

We also evaluated the pipeline's individual components to assess each's performance and efficiency. As shown in Table II, only using Yolo gives a mean IoU of 77.6%, and using only pre-trained XMem gives mIoU of 82.5%. After finetuning the XMem and providing the first ground truth to start the tracking, we get the mean IoU of 92.6%, which is the benchmark. Now, in practice, since we won't have the first frame mask during streaming, connecting the Yolo model predicting the first mask and providing it to XMem achieves the same result with a mean IoU of 92.26%.

We also trained our model on the multi-class segmentation task on the same EndoVis17 dataset, and Table IV details the performance of various models compared to our proposed model. The task at hand involves segmenting six distinct classes, namely Bipolar Forceps (BF), Prograsp Forceps (PF), Large Needle Driver (LND), Vessel Sealer (VS), Grasping Retractor (GR), and Monopolar Curved Scissors (MCS), with the performance on each class being evaluated using the Intersection over Union (IoU) metric. The mIoU, representing the mean IoU across all the classes, serves as a robust indicator of a model's overall segmentation accuracy and consistency across different class types.

TABLE II
Comparision Results of Yolo, PreTrained(PT) and FineTuned(FT) XMem on EndoVis17 Binary Segmentation

XMem	Yolov8-seg	mIoU
PT	✗	82.50
FT	✗	92.26
✗	✓	77.60
FT	✓	**92.26**

TABLE III
Comparison Results of Various Models on EndoVis2018 Binary Segmentation Dataset

Method	mIoU
Yolov8-seg [17]	66.70
UNC [26]	66.77
IRCAD [27]	69.06
OTH [27]	70.43
LBDT [26]	71.9
MTTR [28]	72.2
VIS-Net [29]	74.2
Ours	**82.40**

Our proposed model demonstrates a remarkable superiority over the contemporary state-of-the-art models with a mIoU of 74.43. This showcases not only the model's ability to segment different classes accurately but also its consistency in maintaining a high level of performance across all classes. The nearest competitor, S3Net(+MaskRCNN), trails with a mIoU of 46.55, highlighting a significant gap in the performance. Our model showcases a compelling lead in individual class IoU values, particularly in the BF, PF, LND, VS, and MCS classes, with IoU values of 88.6, 85.12, 93.3, 91.4, and 88.2, respectively. This solid performance across individual classes substantiates the model's robustness and high adaptability to different class types within the dataset.

Our proposed model significantly outperformed existing state-of-the-art models in another set of results on the EndoVis18 Robotic Arm Binary Segmentation task. Our model achieved a remarkable IoU score of 82.4, a substantial improvement over the preceding state-of-the-art model, VisNet, which secured an IoU of 74.2. Other notable models, like MTTR and LBDT, exhibited an IoU of 72.2 and 71.9, respectively, indicating the superior efficacy of our model in delineating the robotic arm from the background.

Also, when doing a similar pipeline components study as did with EndoVis17, we find that Yolov8-seg severely underperforms. Due to the lower performance of the Yolo model, the first mask supplied to the XMem is inferior in precision compared to the ground truth mask. Hence, we can see a clear drop in performance when a FineTuned XMem with a ground truth Mask is compared to a FineTuned XMem with Yolo predictions with a clear drop of 5.1% in mIoU.

TABLE IV
COMPARISON RESULTS OF VARIOUS MODELS ON ENDOVIS2017 MULTI-CLASS SEGMENTATION DATASET

Method	BF	PF	LND	VS	GR	MCS	mIoU
TernausNet-11 [21]	13.45	12.39	20.51	5.97	1.08	1	10.17
MF-TAPNET [30]	16.39	14.11	19.01	8.11	0.31	4.09	10.77
ISINet [24]	38.70	38.50	50.09	27.43	2.01	28.72	28.96
TraSeTR [11]	45.2	56.7	55.8	38.9	11.4	31.3	36.79
S3Net(+Mask2former) [31]	49.48	29.91	70.61	32.98	19.53	18.35	38.13
S3Net(+MaskRCNN) [31]	75.08	54.32	61.84	35.5	**27.47**	43.23	46.55
Ours	**88.6**	**85.12**	**93.3**	**91.4**	0	**88.2**	**74.43**

TABLE V
COMPARISION RESULTS OF YOLO, PRETRAINED(PT) AND FINETUNED(FT) XMEM ON ENDOVIS18 BINARY SEGMENTATION

XMem	Yolov8-seg	mIoU
PT	✗	72.40
FT	✗	**87.50**
✗	✓	66.70
FT	✓	82.40

VI. LIMITATIONS

Initial Mask Precision: The accuracy of the initial mask forwarded to XMem is paramount for effective object tracking. In scenarios where the initial mask is imprecise, XMem encounters challenges in reliably tracking the object across subsequent frames.

Robotic Arm Detection: The performance of the pipeline is contingent on the accurate detection of the robotic arm by YOLO within the mask. Should YOLO fail to detect one of the robotic arms, XMem, in turn, will be unable to track that particular arm, thereby diminishing the overall performance.

Introduction of New Arms: The pipeline exhibits a limitation concerning the introduction of new arms in future frames. Since XMem relies on the initial mask predicted by YOLO for object tracking, adding a new arm not present in the initial mask results in XMem's inability to recognize and track the newly introduced arm.

Grasping Retractor Identification: In the multi-class scenario, the dataset encompasses only 27 instances of the Grasping Retractor class, posing a significant challenge for the model in identifying this particular class. Consequently, this scarcity of instances culminates in a 0% Intersection over Union (IoU) score for the Grasping Retractor, reflecting a substantial limitation in the model's capacity in such scenarios.

VII. CONCLUSION AND FUTURE WORK

We presented a real-time object segmentation framework of robotic tools during surgery. Presented pipeline generates state-of-the-art results on the standard datasets, as illustrated in tables I, III, and IV. This framework can be easily deployed on real-time streaming video since it does n0t require any manual annotation of the initial mask. It can directly take the RGB frame as input and output the mask, which can be integrated into the da Vinci system to create an augmented visual feedback.

Another use for the system would be in annotation, where the annotator would have access to the datasets of previously annotated robotic instruments, which only need to be reviewed and tweaked as needed instead of doing it from scratch, which would significantly decrease overall annotation time and increases the total data available for future training, which is one of the main bottlenecks, especially in surgery-related research.

Future work can be extended to solve some of the above-mentioned limitations. One of the major concerns in the medical domain is the lack of data, especially in surgical AI; currently, the data needs to be annotated and reviewed by medical professionals, which is time-consuming and costly. To overcome this, one can use foundational models, which are trained using Self-Supervised Learning [32], on top of which we can use other models. For example, here in XMem [10], we use ResNet50 and ResNet18 for feature extraction, which is trained on ImageNet, but if we have foundational models on surgical data, we could use those models in XMem [10], increasing the performance. Another direction for the data would be data synthesis, where artifacts like blood, smoke on the screen, and occlusion of arms are not present in the current datasets but can be artificially created to increase the quantity and diversity of the datasets.

An additional limitation of not being able to detect the new robotic arms in frames during XMem [10] inference can be overcome by using Yolo in parallel to XMem [10]. If Yolo identifies any new arm, it can reinitialize the XMem [10] with a new predicted mask along with the new arm, or save the new arm in the memory of XMem [10] to make the model aware and propagate it in future frames.

VIII. ACKNOWLEDGEMENTS

This work was supported by GPU resources from the Lambda Labs funded by Massachusetts General Hospital, Harvard Medical School, and partly also by GPUs resources at Plaksha University, India.

REFERENCES

[1] M. A. Mederos, R. L. Jacob, R. Ward, R. Shenoy, M. M. Gibbons, M. D. Girgis, D. Kansagara, D. Hynes, P. G. Shekelle, and K. Kondo, "Trends in robot-assisted procedures for general surgery in the veterans health administration," *Journal of Surgical Research*, vol. 279, pp. 788–795, 2022.

[2] A. R. Lanfranco, A. E. Castellanos, J. P. Desai, and W. C. Meyers, "Robotic surgery: a current perspective," *Annals of surgery*, vol. 239, no. 1, p. 14, 2004.

[3] E. L. Servais, D. L. Miller, D. Thibault, M. G. Hartwig, A. S. Kosinski, C. T. Stock, T. Price, S. M. Quadri, R. S. D'Agostino, and W. R. Burfeind, "Conversion to thoracotomy during thoracoscopic vs robotic lobectomy: predictors and outcomes," *The Annals of Thoracic Surgery*, vol. 114, no. 2, pp. 409–417, 2022.

[4] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, S. Bodenstedt *et al.*, "2017 robotic instrument segmentation challenge," *arXiv preprint arXiv:1902.06426*, 2019.

[5] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen *et al.*, "2018 robotic scene segmentation challenge," *arXiv preprint arXiv:2001.11190*, 2020.

[6] M. Islam, D. A. Atputharuban, R. Ramesh, and H. Ren, "Real-time instrument segmentation in robotic surgery using auxiliary supervised deep adversarial learning," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2188–2195, 2019.

[7] C. da Costa Rocha, N. Padoy, and B. Rosa, "Self-supervised surgical tool segmentation using kinematic information," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 8720–8726.

[8] A. Wang, M. Islam, M. Xu, and H. Ren, "Rethinking surgical instrument segmentation: A background image can be all you need," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 355–364.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.

[10] H. K. Cheng and A. G. Schwing, "Xmem: Long-term video object segmentation with an atkinson-shiffrin memory model," in *European Conference on Computer Vision*. Springer, 2022, pp. 640–658.

[11] Z. Zhao, Y. Jin, and P.-A. Heng, "Trasetr: track-to-segment transformer with contrastive query for instance-level instrument segmentation in robotic surgery," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 11 186–11 193.

[12] Y. Wang, Z. Xu, X. Wang, C. Shen, B. Cheng, H. Shen, and H. Xia, "End-to-end video instance segmentation with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8741–8750.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[14] L. Ke, M. Danelljan, H. Ding, Y.-W. Tai, C.-K. Tang, and F. Yu, "Mask-free video instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 857–22 866.

[15] J. Liu, L. Zhang, A. Guo, Y. Gao, and Y. Zheng, "Multi-scale feature fusion convolutional neural network for multi-modal medical image fusion," in *Proceedings of the 2023 4th International Conference on Computing, Networks and Internet of Things*, ser. CNIOT '23. New York, NY, USA: Association for Computing Machinery, 2023, p. 913–917.

[16] R. C. Atkinson and R. M. Shiffrin, "Human memory: A proposed system and its control processes," in *Psychology of learning and motivation*. Elsevier, 1968, vol. 2, pp. 89–195.

[17] G. Jocher, A. Chaurasia, and J. Qiu, "Yolo by ultralytics (version 8.0. 0)[computer software]," 2023.

[18] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 724–732.

[19] N. Xu, L. Yang, Y. Fan, J. Yang, D. Yue, Y. Liang, B. Price, S. Cohen, and T. Huang, "Youtube-vos: Sequence-to-sequence video object segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 585–601.

[20] 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: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[21] A. A. Shvets, A. Rakhlin, A. A. Kalinin, and V. I. Iglovikov, "Automatic instrument segmentation in robot-assisted surgery using deep learning," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*. IEEE, 2018, pp. 624–628.

[22] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE transactions on medical imaging*, vol. 39, no. 6, pp. 1856–1867, 2019.

[23] L. Yang, H. Wang, Y. Gu, G. Bian, Y. Liu, and H. Yu, "Tma-net: A transformer-based multi-scale attention network for surgical instrument segmentation," *IEEE Transactions on Medical Robotics and Bionics*, 2023.

[24] C. González, L. Bravo-Sánchez, and P. Arbelaez, "Isinet: an instance-based approach for surgical instrument segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2020, pp. 595–605.

[25] Y. Xia, S. Wang, and Z. Kan, "A nested u-structure for instrument segmentation in robotic surgery," in *2023 International Conference on Advanced Robotics and Mechatronics (ICARM)*. IEEE, 2023, pp. 994–999.

[26] Z. Ding, T. Hui, J. Huang, X. Wei, J. Han, and S. Liu, "Language-bridged spatial-temporal interaction for referring video object segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4964–4973.

[27] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[28] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4985–4995.

[29] H. Wang, L. Zhu, G. Yang, Y. Guo, S. Zhang, B. Xu, and Y. Jin, "Video-instrument synergistic network for referring video instrument segmentation in robotic surgery," *arXiv preprint arXiv:2308.09475*, 2023.

[30] Y. Jin, K. Cheng, Q. Dou, and P.-A. Heng, "Incorporating temporal prior from motion flow for instrument segmentation in minimally invasive surgery video," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part V 22*. Springer, 2019, pp. 440–448.

[31] B. Baby, D. Thapar, M. Chasmai, T. Banerjee, K. Dargan, A. Suri, S. Banerjee, and C. Arora, "From forks to forceps: A new framework for instance segmentation of surgical instruments," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 6191–6201.

[32] S. Ramesh, V. Srivastav, D. Alapatt, T. Yu, A. Murali, L. Sestini, C. I. Nwoye, I. Hamoud, S. Sharma, A. Fleurentin *et al.*, "Dissecting self-supervised learning methods for surgical computer vision," *Medical Image Analysis*, vol. 88, p. 102844, 2023.