2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Procedure Recognition by Knowledge-Driven Segmentation in Robotic-Assisted Vitreoretinal Surgery

Zhen Li, Yawen Deng, Qiang Ye, Weihong Yu, Haoxiang Qi, Yaliang Liu, Zhangguo Yu, Gui-Bin Bian*

Abstract—**Internal limiting membrane (ILM) peeling is a vital vitreoretinal surgery procedure. However, due to the thickness of just 1-2 micrometers and the intricacies associated with its varying density and adhesion, the difficulty of manipulation exceeds the physiological limits of human perception and operation. Surgical robot is characterized by high precision and stability. However, navigating intricate intraocular environments and handling minuscule high-precision areas remain enormous challenges. These include issues of uneven lighting, field-of-view loss, and motion blur. This paper proposed a perception method named 'Multimodal Surgical Process Recognition based on Domain Knowledge and Segmentation (MSPR-DKS),' designed to address these challenges and provide input for the precise control of robots. Moreover, a comprehensive dataset focused on ILM peeling during macular hole surgeries was established. Experimental results underscore the efficacy of this approach, with segmentation accuracies exceeding 99.27% for instruments and macular holes and an average accuracy of 98.97% in recognizing surgical processes. This study paves the way for leveraging domain knowledge and image segmentation to improve robot-assisted manipulation of soft tissues in ophthalmology.**

This research is supported by the National Key Research and Development Program of China (Grant 2022YFB4702900), the National Natural Science Foundation of China (Grant 62027813, U20A20196), the Beijing Science Fund for Distinguished Young Scholars (JQ21016), the Excellent member of CAS Youth Innovation Promotion Association (Y2022054).

Zhen Li is with the School of Electronic and Information Engineering, Tongji University, 200092, Shanghai, China, and Institute of Automation, Chinese Academy of Sciences, 100190, Beijing, China (e-mail: zhen.li@ia.ac.cn).

Ya-Wen Deng is with the School of Mechatronic Engineering, Beijing Institute of Technology, Beijing, 100081, China and with the Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China (e-mail: 3120235412@bit.edu.cn).

Qiang Ye and Gui-Bin Bian are with the Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China (e-mail: qiang.ye@ia.ac.cn, guibin.bian@ia.ac.cn).

Wei-Hong Yu is with the Department of Ophthalmology, Peking Union Medical College Hospital, Chinese Academy of Medical Sciences and Peking Union Medical College, Beijing, 100005, China (e-mail: yuwh@pumch.cn).

Haoxiang Qi and Yaliang Liu are with the School of Mechatronic Engineering, Beijing Institute of Technology, Beijing, 100081, China (e-mail: 3120215098@bit.edu.cn, liuyaliang@bit.edu.cn).

Zhangguo Yu is with the School of Mechatronic Engineering, Beijing Institute of Technology, Beijing, 100081, China, with the Key Laboratory of Biomimetic Robots and Systems, Ministry of Education, Beijing 100081, China, with the International Joint Research laboratory of Biomimetic Robots and Systems, Ministry of Education, Beijing 100081, China, and also with the State Key Laboratory of Intelligent Control and Decision of Complex System, Beijing 100081, China (e-mail: yuzg@bit.edu.cn).

*Corresponding author: Gui-Bin Bian

I. Introduction

Macular hole is a common vitreoretinal disease, with a cumulative incidence of about 41.1 cases per 100,000 person-years [1], resulting in severe vision loss, vision distortion, central scotoma, and other symptoms. It significantly interferes with the patient's daily life, even workability. The primary surgical treatment for macular holes is vitrectomy combined with internal limiting membrane (ILM) peeling [2]. This procedure aims to sufficiently relieve the traction exerted by the vitreous and the ILM on the retinal tissue, facilitating the closure of the hole and restoring the retina's normal morphology and function. However, during the ILM peeling procedure, surgeons face challenges related to the ILM's minimal thickness of 1-2 micrometers, coupled with its varying density and adhesion. These factors render manual operation exceedingly difficult, often approaching the physiological limits of human precision [3]. The application of robot-assisted surgery offers higher precision and stability for anterior retinal membrane peeling procedures, including the ILM. This is particularly beneficial when dealing with delicate retinal areas and navigating complex intraocular environments [4].

In the current field of ophthalmology, although various advanced ophthalmic robotic systems such as MICRON [5], RAM!S [6], Preceyes [7], IRISS [8], and SHER [9] have been developed and extensively tested, they still face several key challenges when manipulating vitreoretinal diseases, such as macular holes. Accurately distinguishing focus like macular holes from adjacent tissues is challenging due to poor lighting and blurred boundaries. Secondly, delicate instrumentation manipulation within a complex surgical environment is difficult. Lastly, ensuring the continuity of the surgical procedure and achieving consistent treatment outcomes is essential. To overcome these challenges, the introduction of image-guided environmental perception technology [10], especially the segmentation of instruments and focus, as well as the automatic recognition of the surgical procedure, will greatly enhance the functionality and efficiency of robot-assisted macular hole treatment. This will enhance the precision manipulation of soft tissues in the vitreoretinal region of the eye.

In recent years, deep learning has advanced surgical image processing. Such advancements are evident in segmenting instruments and focus, as well as in recognizing surgical procedures [11]. Compared to traditional segmentation methods that rely on manual thresholds and features, segmentation accuracy has been significantly improved through deep learning, particularly using convolutional neural networks (CNN) [12]. U-Net [13], with its skip connections

and symmetrical encoder-decoder architecture, stands out for its ability to precisely capture details in medical images and efficiently distinguish between instruments and focus. Its variant, ResUnet [14], by incorporating residual structures, further enhances the network's depth and expressive capability, optimizing the performance of intraoperative segmentation tasks. Additionally, the vision transformer (ViT) [15], introduced by Dosovitskiy et al. as a transformer model specifically designed for visual tasks, processes images in a manner distinct from CNNs. By segmenting and serializing fixed-sized image patches, it can deeply capture global and long-range information within images, demonstrating a strong preference for shape features and exhibiting high robustness. Given these technological advancements, models that combine ResUnet and ViT, such as TransUnet [16], have emerged. Leveraging both spatial and attention mechanisms, TransUnet offers a more efficient and precise solution for intraoperative image segmentation. Moreover, its strong generalization capabilities for complex scenarios have been demonstrated in numerous experiments. In surgical procedure recognition, deep learning, especially methods combining CNN and long short-term memory (LSTM) [17], has dramatically advanced the accurate parsing of intraoperative video surgical procedures. Although early methods primarily relied on hidden Markov models to capture the temporal information of surgeries, their effectiveness and real-time performance were limited. However, with Recurrent neural networks (RNNs), especially LSTM, there has been an enhanced capability to capture the temporal information associated with various surgical procedures precisely. This automated detection method provides robots with rapid and accurate surgical procedure information, enabling more precise surgical decision-making.

However, much of the prior work on models for instrument and focus segmentation, as well as surgical procedure recognition, has been concentrated on ophthalmic cataract surgeries [18], general laparoscopic surgeries [19], and brain tumor resection [20]. Research on vitreoretinal surgery remains largely unexplored. Intraoperative image segmentation for treatment faces multiple challenges. Factors such as fundus reflection, obstructions from surgical instruments, magnification and focal length adjustments of the microscope, and dynamic disturbances during the surgery leading to uneven lighting, field-of-view loss, and motion blur significantly increase the complexity of image processing. While recognizing surgical procedures for can employ methods based on global and local features, similar to other surgeries, there still remain challenges. Despite being concise and annotations-free global feature methods may overlook intricate image details, making handling complex scenarios problematic. For local feature methods, while more effective for intricate surgical scenarios, necessitate additional annotation information, rendering the data acquisition and processing inefficient.

To address the aforementioned issues, the contributions of this paper are:

- An innovative perception approach for robot-assisted intraoperative ILM peeling in macular hole surgeries is proposed. This approach combines TransUnet with two segmentation tasks: one involving level set evolution (LSE) for high-precision macular hole lesion segmentation and another for intraoperative instrument segmentation. Furthermore, surgical procedure recognition during the surgery is achieved based on the segmented image features.
- A multimodal surgical procedure recognition method has been formulated, driven by domain knowledge. Visual representations capture overarching surgical changes, while key point distances quantify subtle spatial shifts. Harnessing multimodal information transformation and complementarity provides a comprehensive understanding of surgical focal points and instrument movements. This approach refines the precision in recognizing surgical procedures.
- A detailed dataset centered on ILM peeling in macular hole surgeries has been compiled. Spanning three types of data labels and six specific peeling actions provides substantial data backing for pertinent research. Experimental outcomes demonstrate a high segmentation accuracy of over 99.27% for both instruments and macular holes, and an impressive 98.97% average accuracy in surgical procedure recognition.

II. Proposed Method for Enhanced Robotic Surgery Perception

This study introduces a multimodal surgical procedure recognition algorithm based on domain knowledge and segmentation (MSPR-DKS) to enhance the perceptual capabilities of robot-assisted intraoperative ILM peeling during macular hole surgery. This approach explicitly targets intraoperative images during ILM peeling. During robotic surgical procedure recognition, it is expected to effectively addresses challenges such as uneven illumination, loss of field of view, and motion blur. The proposed MSPR-DKS architecture is illustrated in Fig. 1.

The core design principle of DKS-MSPR is to achieve accurate surgical procedure recognition based on efficient image segmentation. In the first phase of MSPR-DKS, by integrating TransUnet and LSE techniques, obfuscation due to surgical instruments has been optimized, accomplishing precise segmentation tasks for both macular hole focus and intraoperative tools. In the second phase of MSPR-DKS, a multimodal surgical procedure recognition technique is designed based on domain knowledge and segmented image features. On the one hand, serializing the segmented images of focus and instruments and then utilizing the temporal convolutional network (TCN) combined with the BiLSTM method. On the other hand, considering the operational characteristics of the macular hole ILM peeling surgery, key points of the instruments and the distances between them are extracted, and the CNN combined with the BiLSTM algorithm is employed, enriching the image modality data. Ultimately, precise intraoperative surgical procedure recognition is achieved by merging these two methodologies and capitalizing on multimodal information transformation and complementarity.

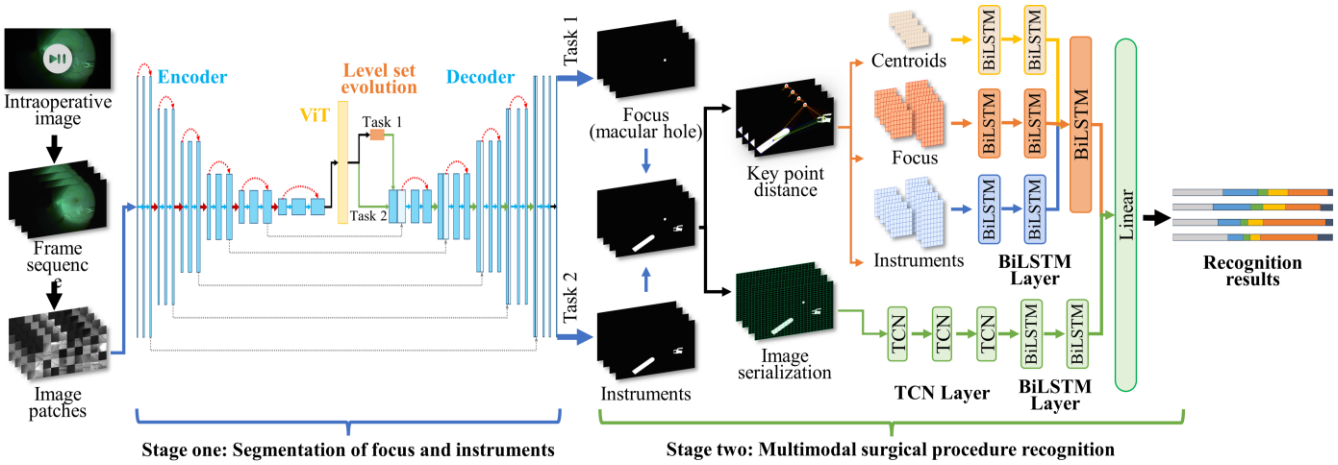

Fig. 1. The MSPR-DKS Framework for enhancing the perceptual capabilities of robot-assisted intraoperative ILM peeling. The design incorporates TransUnet and level-set evolution techniques for image segmentation in stage one and through the synergy of TCN and BiLSTM to recognize the surgical procedure in stage two.

A. Stage One: High-Precision Segmentation Using Level Set Evolution Layer and TransUnet

In this Stage, a combination of TransUnet and the LSE layer is employed to achieve high-precision segmentation of macular hole focus and intraoperative instruments.

(1) Data Preprocessing Workflow

Continuous intraoperative videos are first transformed into a frame sequence format during the preprocessing phase. Following this, each frame undergoes a series of modifications: it is converted from RGB to grayscale to highlight essential details, standardized to ensure consistent input scale, and then enhanced using methods like Adaptive Histogram Equalization (CLAHE) and gamma correction for improved clarity. The data values of each frame are then normalized to the range [0, 1]. For segmentation purposes, the data is labeled accordingly for two specific tasks. Each image is further segmented into multiple patches, but any patch with its center outside the visible field of view is discarded. Once this preprocessing is complete, the dataset is divided into training and testing sets, preparing it for the subsequent model training and testing procedures.

(2) Model Implementation

The six-layer deep TransUnet model is applied to the macular hole and intraoperative instrument segmentation. However, during the bridging stage, an additional LSE layer is introduced for the macular hole segmentation task to optimize the segmentation outcome.

Advantages of TransUnet: TransUnet merges the self-attention mechanism of the Transformer with the encoder-decoder structure of U-Net. This synergy gives it an exceptional capability to capture global features, making it highly suitable for segmenting focus and instruments in macular hole surgeries. By expanding to a depth of six layers, TransUnet can more the stratified features present in macular hole surgery profoundly discern. A deeper architecture provides a more extensive receptive field, capturing richer contextual information. This ensures accuracy and stability in differentiating complex ocular backgrounds from instruments with blurred edges.

Role of the Level Set Evolution Layer: To enhance the detection of the intricate structures of the macular hole, the LSE layer is integrated within the bridging layer of TransUnet. Given the potential motion blur in macular hole surgery images caused by rapid focus movements or minute microscope shifts, capturing the boundaries of small focus areas like the macular hole can be challenging. Here, the LSE layer ensures flexibility and high segmentation accuracy with its natural handling of topological changes, sub-pixel accuracy, robustness to complex shapes, and capability to maintain holistic area information.

In the LSE process, the encoder output initializes the level set function, from which gradients and magnitudes are computed across all spatial points. This is followed by gradient normalization, divergence calculation for curvature determination, and an update of the level set function using the curvature information.

Firstly, compute the discrete gradients of the level set function $\varnothing$ in the x and y directions.

$$\varnothing_x = \frac{\varnothing(x+1,y)-\varnothing(x-1,y)}{2} \tag{1}$$

$$\varnothing_y = \frac{\varnothing(x,y+1)-\varnothing(x,y-1)}{2} \tag{2}$$

Next, based on the gradients obtained from the previous step, calculate their magnitude.

$$|\nabla\varnothing| = \sqrt{\varnothing_x^2+\varnothing_y^2} \tag{3}$$

Third, normalize the gradients.

$$\varnothing_{x,norm} = \frac{\varnothing_x}{|\nabla\varnothing|} \tag{4}$$

$$\varnothing_{y,norm} = \frac{\varnothing_y}{|\nabla\varnothing|} \tag{5}$$

Fourth, calculate the curvature. The curvature is the divergence of the normalized gradients.

$$k = \frac{\varnothing_{x,norm}(x+1,y) - \varnothing_{x,norm}(x-1,y)}{2} + \frac{\varnothing_{y,norm}(x,y+1) - \varnothing_{y,norm}(x,y-1)}{2} \quad (6)$$

Fifth, update the level set function. Using the calculated curvature, the level set function is updated.

$$\varnothing_{next} = \varnothing - \Delta t \times k \quad (7)$$

Where, Δt is a time step size, representing the rate of evolution. $\varnothing_{next}$ is the level set function for the next time step.

Incorporating such an LSE layer further bolsters TransUnet's capability to segment macular holes in complex backgrounds.

By the end of stage one, we can independently produce results for both focus (macular hole) and instruments (fiber optic and peeling forceps), which will provide crucial input for intraoperative procedure recognition in stage two.

B. Stage Two: Multi-Modal Surgical Procedure Recognition Based on Segmented Image Features and Domain Knowledge

In intraoperative imaging of the ILM retinal membrane peeling procedure, the utilization of untreated images can hamper the accurate identification of surgical procedures due to inherent lighting and field-of-view challenges. Specifically, during the ILM peeling, relying solely on fiber-optic illumination can lead to uneven illumination in the visuals, as the light is susceptible to disturbances from changes in the fiber's angle or position. Moreover, as surgical instruments approach or overlay the macular hole during the peeling, the focus within the microscope's field of view could be obscured. Hence, prioritizing the segmentation of focus and instruments, followed by procedure recognition, can effectively simplify image characteristics and counteract the redundancies and misjudgments caused by lighting issues while accurately pinpointing the obscured macular hole location.

A domain-knowledge-based multi-modal method is the stage two approach. The images segmented in the first stage and the extracted key point distances are employed in a dual-modality to achieve precise intraoperative procedure recognition. The focus (macular holes) and surgical instruments (fiber optic and peeling forceps) derived from intraoperative images are fed into the procedure recognition model. Based on the image features from the first stage, key contours and centroids of focus and instruments are extracted using domain knowledge, subsequently yielding the key point distance modality, supplementing the original image modality. The model employs TCN and BiLSTM units to extract, aggregate, and interpret features from the dual modality data respectively. Ultimately, both modalities are fused at a linear layer, with the final precise procedure recognition outcome obtained through an argmax layer.

(1) Feature extraction of image modalities using TCN combined with BiLSTM

Combining a three-layer Temporal Convolutional Network (TCN) and a two-layer BiLSTM demonstrates superior performance for the identification procedure in ILM peeling surgery. After merging the segmented macular hole and the instrument regions, the images are serialized. Subsequently, through the three-layer TCN with its convolutional structure, long-term dependencies within the temporal data of ILM peeling surgery images are effectively captured, showcasing impressive parallelism attributes. This ensures that pivotal early events are accurately identified, even in extended sequences. The subsequent BiLSTM layers, through their bidirectional configuration, grasp the sequence information from both forward and backward perspectives, further augmenting comprehension of intricate contextual relationships within the sequence. In summation, integrating TCN and BiLSTM offers a model fortified with manifold advantages for sequential image recognition during surgical procedures. Such an amalgamation comprehensively seizes the anterior-posterior correlation of surgical steps while parallel processing capability and stable gradient propagation boost training efficiency and stability.

(2) The key point distances are extracted from the segmented images of the focus and instruments

In the ILM peeling procedure of macular hole surgeries, ascertaining the precise distance between the focus and the surgical instruments is crucial for optimizing surgical outcomes and safety, and given the unique characteristics of this operation, this distance can also serve as a data modality for effectively identifying the surgical procedure. From the perspective of the surgical procedure, after the ILM is stained, specialized ILM forceps are utilized to grasp the ILM at an avascular zone approximately 1DD (disc diameter, used as a reference for describing the size of retinal lesions) away from the periphery of the macular hole. Subsequently, its detachment from the retinal nerve fiber layer is executed. Following this, the peeling is carried out along the hole. For holes measuring a diameter of 400 μm or less, the ILM is generally peeled within a 3DD radius centered on the hole. Moreover, when contemplating the significance of image processing in this scenario, an in-depth examination of the segmented binary images becomes exceedingly pivotal. In particular, by exploring the characteristics of image contours, one can gain comprehensive insights into the shapes, sizes, and relative positions of objects present in the image. This is attributed to the fact that objects' contours in images often accurately represent their actual condition. Thus, by extracting key points from the contours of the focus (macula hole) and instruments (fiber optic and peeling forceps), then calculating the distances between the centroids of these three contours and their corresponding key points, invaluable insights can be garnered for the identification of the surgical procedure.

The extraction of key point distances involves four stages. Initially, centroids in an image, reflecting average pixel positions, are derived from white pixels' x and y coordinates. Next, analyzing these centroids allows for segmenting the image into left and right, distinguishing between fiber optic and peeling forceps. This method eschews detailed segmentation by focusing on x-coordinate differences. The Douglas-Peucker algorithm then simplifies contours, and finally, the Euclidean formula gauges distances between centroids or contours. This analyzes the relative positions and relationships between various structures or features in the image. The resultant distance schematic can be seen in Fig. 2.

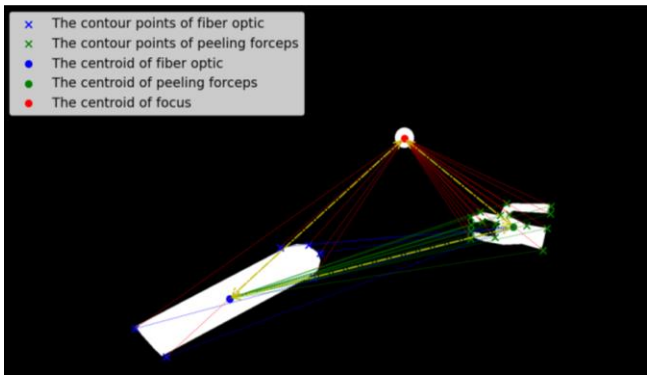

Fig. 2. Key point distances between the focus (macular hole) and the instruments (fiber optic and peeling forceps).

The resultant distance schematic can be observed in Fig. 2. There are three types of distance data. Specifically, these are the distances between the centroid of the focus and the centroids of the two instruments, the distances from the centroid of the focus to the key points on the instrument contour, and the distances from one instrument contour to the centroid of the other instrument.

(3) Multi-Feature extraction of distance modalities using BiLSTM

It effectively extracts temporal correlations among various types of distances within the same surgical procedure. In Step 1, each data type is fed into BiLSTM Layers 1 and 2, capturing the temporal correlations of each time point with data of the same type. Step 2 integrates the temporal correlations of the three types of distance data in BiLSTM Layer 3, inputting all hidden state data from the final node of the bidirectional sequence into the Linear Layer.

Lastly, the image and distance modalities are fused in a 1:1 ratio in their respective Linear Layers. This facilitates the interpretation of the features, deriving the recognition weights for the surgical procedure, and the final recognition result is outputted through the argmax layer.

III. Materials and Experiment

A. Surgical Procedure Types and Data Collection

This study primarily focuses on video data obtained during the ILM peeling phase of macular hole surgery. This data involves the focus (macular hole) and two instruments (the fiber optic and the peeling forceps). Within the ILM peeling phase, six surgical procedure categories have been defined, including localization of the peeling site, touching the ILM, grasping the ILM, extracting the ILM, tearing the ILM, and discarding the ILM at the boundaries. These categories correspond to the two segmentation types between the focus and instruments with their movements across consecutive frames. Fig. 3 presents examples of the surgical procedure categories examined in this study.

B. Data Collection and Annotation

The dataset comprises genuine ILM peeling surgery videos with a 1920×1080 and 96 dpi resolution, constituting 1,399 frames, sourced from four distinct patients with macular holes. This study was conducted under the ethical review of the Beijing Union Hospital of the Chinese Academy of Medical Sciences. Each frame has been manually annotated,

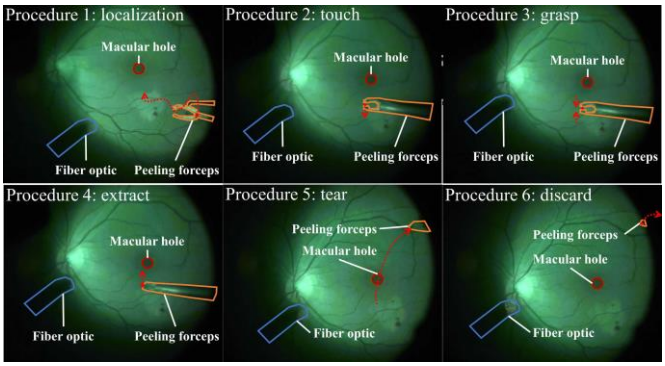

Fig. 3. Six Examples of Surgical Procedures.

detailing the contours of the macular hole, fiber optic, and peeling forceps. A sequence is formed by stacking 13 consecutive frames, with six frames preceding and following a central frame. The surgical procedure category is annotated based on the central frame within this sequence, facilitating accuracy evaluation in subsequent automated surgical procedure recognition models.

IV. Result and Analysis

A. Segmentation Results of Focus and Instruments

TransUnet, in conjunction with LSE layers neural network, was utilized for image segmentation across two tasks. The collected intraoperative ILM peeling frame sequences were trained and tested, allocating 70% of the dataset for training and 30% for testing. To identify the optimal parameters, the training set underwent a five-fold cross-validation. The training was executed on an NVIDIA GeForce RTX 3080-Ti graphics processor, employing the PyTorch framework. The Adam optimization algorithm was adopted with a learning rate set at 0.0005 and a batch size of 64. In the task of macular hole segmentation, the level set was updated three times. Five commonly used image segmentation metrics were employed: Accuracy (Acc), Precision (PC), Dice coefficient (Dice), Specificity (SP), and Sensitivity (SE). Comparisons were made against four conventional algorithms: a 5-layer deep TransUnet, ResUnet, and AUnet. Additionally, for the macular hole segmentation task, a comparison was made with a 6-layer deep TransUnet algorithm, which did not utilize the level set.

TABLE I
Comparison of Segmentation Results of Four Algorithms

Types	Algorithm	Acc	Dice	PC	SE	SP
Macular hole	AUnet [21]	99.84%	33.02%	68.40%	21.76%	**99.98%**
	6-layer deep ResUnet [14]	**99.96%**	89.26%	87.36%	**91.25%**	99.97%
	5-layer deep TransUnet+LSE [16]	99.95%	86.97%	91.13%	83.17%	**99.98%**
	Our algorithm without LSE	99.95%	88.46%	90.83%	86.21%	**99.98%**
	Our algorithm	**99.96%**	**89.68%**	**91.37%**	88.05%	**99.98%**
Instruments	AUnet [21]	96.96%	52.26%	86.26%	37.49%	99.72%
	6-layer deep ResUnet [14]	98.35%	78.60%	92.90%	68.12%	99.75%
	5-layer deep TransUnet [16]	97.39%	61.16%	90.02%	46.31%	**99.76%**
	Our algorithm	**98.57%**	**82.33%**	**92.98%**	**73.39%**	99.74%

For the instruments, the metrics were: ACC at 98.57% and Dice at 82.33%. For the macular hole, they were ACC at 99.96% and Dice at 89.68%. The recognition of these two

tasks outperformed other algorithms in terms of comprehensive performance across five metrics. Specifically, macular hole segmentation showed the best results in Acc, Dice, PC, and SP, while instrument segmentation excelled in Acc, Dice, PC and SE. The experimental outcomes are presented in Table I. Fig. 4 showcased the segmented results for the two tasks. The results showed that a six-layer deep TransUnet more effectively segmented instruments and focus. Moreover, incorporating the LSE layer has enhanced the focus segmentation model's performance.

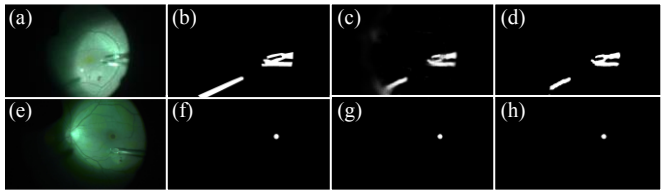

Fig. 4. Example of segmentation results for instruments and macular hole. (a), (b), (c), and (d) represent the instrument‘s original image, label, predicted probability map, and predicted binary image, respectively; (e), (f), (g), and (h) represent the macular hole’s original image, label, predicted probability map, and predicted binary image, respectively.

B. Recognition Results of the Surgical Procedure

The MSPR-DKS was employed for multi-modal surgical procedure recognition. As with segmentation, the preprocessed intraoperative ILM peeling frame sequences were used for training and testing, allocating 70% of the dataset for training and 30% for testing. To determine the optimal parameters, a five-fold cross-validation was conducted on the training set. The training was performed on an NVIDIA GeForce RTX 3080-Ti graphics processor using the PyTorch framework. The Adam optimization algorithm was adopted with a learning rate set at 0.0002 and 0.00005, weight decay at 1e-8, and a batch size of 64. In the classification module, image modality and distance modality were fused at a ratio of 1:1. Four commonly used action classification metrics were employed: Accuracy (Acc), Precision (PC), Recall (R), and F1 score (F1).

TABLE II
COMPARISON OF PROCEDURE RECOGNITION RESULTS OF FOUR ALGORITHMS

Procedure labels	Evaluation metrics	Our Algorithm	Only Image Mode	Only Distance Mode	No Distance Distinction
Procedure 1	PC	1.00	0.90	1.00	0.95
	R	1.00	1.00	1.00	1.00
	F1	1.00	0.95	1.00	0.97
Procedure 2	PC	0.96	0.96	0.96	0.96
	R	1.00	1.00	1.00	0.96
	F1	0.98	0.98	0.98	0.96
Procedure 3	PC	1.00	1.00	1.00	0.00
	R	0.67	0.67	0.67	0.00
	F1	0.80	0.80	0.80	0.00
Procedure 4	PC	1.00	1.00	1.00	0.78
	R	1.00	1.00	1.00	1.00
	F1	1.00	1.00	1.00	0.88
Procedure 5	PC	1.00	1.00	1.00	0.91
	R	1.00	0.95	0.98	1.00
	F1	1.00	0.97	0.99	0.95
Procedure 6	PC	1.00	1.00	0.80	0.00
	R	1.00	1.00	1.00	0.00
	F1	1.00	1.00	0.89	0.00
Overall Acc		**98.97%**	96.91%	97.93%	91.75%

The surgical procedure recognition results are shown in Fig. 5, with an average Acc of 98.97%. Specifically, the recognition F1 for Procedure 1, 4, 5 and 6 peaked at 100% due to its distinctive duration and motion characteristics. In contrast, the recognition for Procedure 3 was the lowest at 80% because it had the shortest duration, and surgeons did not deliberately differentiate between it and the adjacent Procedure 2 during manual operations. The proposed algorithm was compared with four other algorithms under multi-modal and single-modal scenarios, as shown in Table II. The recognition results for all surgical procedures using the proposed algorithm significantly outperformed other algorithms, indicating the efficacy of the modality cross-fusion architecture adopted by the proposed algorithm.

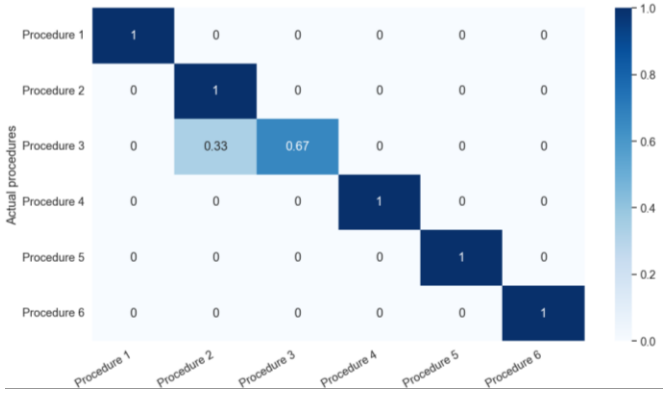

Fig. 5. Procedures recognition results by MSPR-DKS.

C. Discussion

Undeniably, there are limitations concerning real-time surgical procedure recognition and data annotation. Firstly, this study experiences a 200 ms delay during real-time surgical procedure recognition. This is attributed to the necessity of combining data that arises after capturing sequences with a sliding window. Future research will incorporate a surgical procedure prediction model to alleviate the system's latency. Secondly, despite the notable accuracy achieved in segmentation and procedure recognition, the current study needs to extensively explore the inherent intent behind these actions, including specific targeting objectives or anticipated peeling trajectories. Such nuances, crucial for a comprehensive understanding of the procedure, will be the focus of subsequent research to provide deeper insights into surgical procedure recognition.

V. CONCLUSION

MSPR-DKS is a deep learning architecture designed to enhance environmental perception for robots during IML peeling in macular hole surgeries. It strongly emphasizes precise image segmentation of the surgical area and tools, addressing challenges like uneven illumination and motion blur encountered in intraoperative ILM images. Accurate procedure recognition is achieved by leveraging the complementary nature of visual and distance information. A comprehensive dataset regarding ILM peeling in macular hole surgeries was also established, covering three data labels and six specific peeling actions. Experimental results indicate that the proposed image segmentation algorithm yields a Dice coefficient of up to 86.01%, and the surgical procedure recognition has a high accuracy rate of 98.97%. This research is expected to enhance the perception ability and manipulation accuracy for surgical robot, rendering robot-assisted ophthalmic surgeries safer and more intelligent.

REFERENCES

[1] F. S. Ali, J. D. Stein, T. S. Blachley, S. Ackley, and J. M. Stewart, "Incidence of and risk factors for developing idiopathic macular hole among a diverse group of patients throughout the United States," *JAMA Ophthalmology*, vol. 135, no. 4, p. 299, Apr. 2017.

[2] I. Chatziralli, G. Machairoudia, D. Kazantzis, G. Theodossiadis, and P. Theodossiadis, "Inverted internal limiting membrane flap technique for myopic macular hole: A meta-analysis," *Survey of Ophthalmology*, vol. 66, no. 5, pp. 771–780, Jun. 2021.

[3] M. R. Romano, T. Rossi, A. Borgia, F. Catania, T. Sorrentino, and M. Ferrara, "Management of refractory and recurrent macular holes: A comprehensive review," *Survey of Ophthalmology*, vol. 67, no. 4, pp. 908–931, Jan. 2022.

[4] A. Ebrahimi, S. Sefati, P. Gehlbach, R. H. Taylor, and I. I. Iordachita, "Simultaneous online registration-independent stiffness identification and tip localization of surgical instruments in robot-assisted eye surgery," *IEEE Trans. on Robotics*, vol. 39, no. 2, pp. 1373–1387, Sep. 2023.

[5] S. Yang, R. A. MacLachlan, J. N. Martel, L. A. Lobes, and C. N. Riviere, "Comparative evaluation of handheld robot-aided intraocular laser surgery," *IEEE Trans. on Robotics*, vol. 32, no. 1, pp. 246–251, Mar. 2016.

[6] M. A. Nasseri, M. Eder, S. Nair, E. C. Dean, M. Maier, and D. Zapp et al., "The introduction of a new robot for assistance in Ophthalmic Surgery," in *2013 Proc. 35th Annual International Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, 2013, pp. 5682–5.

[7] R. Ladha, T. Meenink, J. Smit, and M. D. de Smet, "Advantages of robotic assistance over a manual approach in simulated subretinal injections and its relevance for gene therapy," *Gene Therapy*, vol. 30, no. 3–4, pp. 264–270, May 2021.

[8] C. Shin, M. J Gerber, Y. Lee, M. Rodriguez, S. A. Pedram, and J. Hubschman et al., "Semi-automated extraction of lens fragments via a surgical robot using semantic segmentation of OCT images with deep learning - experimental results in ex vivo animal model," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5261–5268, Apr. 2021.

[9] M. Zhou, J. Wu, A. Ebrahimi, N. Patel, Y. Liu, and N. Navab et al., "Spotlight-based 3D instrument guidance for autonomous task in robot-assisted retinal surgery," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 7750–7757, Jul. 2021.

[10] M. J. Gerber, J.-P. Hubschman, and T.-C. Tsao, "Automated retinal vein cannulation on silicone phantoms using optical-coherence-tomography-guided robotic manipulations," *IEEE/ASME Trans. on Mechatronics*, vol. 26, no. 5, pp. 2758–2769, Oct. 2021.

[11] X. Chen, X. Wang, K. Zhang, K.-M. Fung, T. C. Thai, and K. Moore et al., "Recent advances and clinical applications of deep learning in medical image analysis," *Medical Image Analysis*, vol. 79, p. 102444, Apr. 2022.

[12] X. Luo, G. Wang, T. Song, J. Zhang, M. Aertsen, and J. Deprest et al., "Mideepseg: Minimally Interactive segmentation of unseen objects from medical images using Deep Learning," *Medical Image Analysis*, vol. 72, p. 102102, May 2021.

[13] O. Ronneberger, P. Fischer and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation", in *2015 Proc. 18th International Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, Munich, 2015, pp. 234-241.

[14] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749–753, May 2018.

[15] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, and T. Unterthiner et al., "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *2021 Proc. the International Conf. on Learning Representations (ICLR)*, Vienna, 2021.

[16] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, and Y. Wang et al., "TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation," *arXiv preprint*, Feb. 2021.

[17] C. Lin, F. Lee, L. Xie, J. Cai, H. Chen, and Li Liu et al., "Scene recognition using multiple representation network," *Applied Soft Computing*, vol. 118, p. 108530, Mar. 2022.

[18] Z.-L. Ni, X.-H. Zhou, G.-A. W, W.-Q. Yue, Z. Li, and G.-B. Bian et al., "Surginet: Pyramid attention aggregation and class-wise self-distillation for surgical instrument segmentation," *Medical Image Analysis*, vol. 76, p. 102310, Feb. 2022.

[19] D. Kitaguchi, Y. Lee, K. Hayashi, K. Nakajima, S. Kojima, and H. Hasegawaet al., "Development and validation of a model for laparoscopic colorectal surgical instrument recognition using convolutional neural network–based instance segmentation and videos of laparoscopic procedures," *JAMA Network Open*, vol. 5, no. 8, Aug. 2022.

[20] Y.-W. Luo, H.-Y. Chen, Z. Li, W.-P. Liu, K. Wang, and L. Zhang et al., "Fast instruments and tissues segmentation of micro-neurosurgical scene using high correlative non-local network," *Computers in Biology and Medicine*, vol. 153, p. 106531, Feb. 2023.

[21] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, and Kazunari Misawa et al., "Attention U-Net: Learning Where to Look for the Pancreas," in *2018 Proc. Medical Imaging with Deep Learning*, Amsterdam, 2018.