2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Scene-Aware Robotic Light Pipe Control for Vitreoretinal Surgery

Wenjun Lin, Wending Zhang, Chin-Boon Chng, Yong Jun Tan, and Chee Kong Chui

Abstract— **Surgical robotics have revolutionized medical procedures by offering enhanced precision and reduced complications. However, vitreoretinal surgery still relies heavily on manual techniques, where surgeons manage both a surgical tool and a light pipe, complicating operations and potentially affecting outcomes. To improve efficiency and outcomes while reducing workloads on surgeons, a novel vision-based robot-assisted system with advanced surgical scene understanding ability is proposed. The system automatically positions a light pipe held by a specialized surgical robot through optimization-based visual collaborative control. By identifying target areas for automatic illumination, the system allows surgeons to focus on surgical tasks and supports more complex surgeries such as three-arm procedures. Besides, the system enhances surgical safety by detecting surgical activities and dangerous areas and issuing alerts accordingly. Postoperatively, the system records tool trajectories and detected activities, providing data for surgical reports, skill evaluation, and training. Experiments prove the effectiveness of the control system, visual algorithm, and overall collaborative system.**

I. INTRODUCTION

Surgical robotics have been increasingly used in a wide range of surgical procedures, with a particular emphasis on Minimally Invasive Surgery (MIS) [1]. This trend is primarily driven by their capability to execute surgical tasks with remarkable precision, reduce the risk of complications, and shorten patient recovery times. Additionally, the use of robots in surgery can help reduce the physical strain on surgeons and improve overall efficiency. Advanced robotic systems, such as the da Vinci Surgical System, are now widely used in a variety of surgical procedures [2], [3], [4]. However, most intraocular procedures, including cataract and vitreoretinal surgeries, remain manual.

In vitreoretinal surgery, the intraocular surgical field is visualized through an ophthalmic microscope positioned above the patient. The surgeon operates a surgical tool with the dominant hand, while the non-dominant hand controls a light pipe to illuminate the area, as illustrated within the gray elliptical frame in Fig. 1. Several factors can affect the outcomes of the surgery, including the surgeon's skills and experience, physiological hand tremors, fatigue, and patient movement. Earlier studies have investigated and solved some of these challenges through various technical interventions. To mitigate the effects of hand tremors, robotic manipulators such as the Steady-Hand Eye Robot (SHER) [5] have been designed to manipulate tools or perform tasks with a high degree of precision and dexterity. In parallel, to improve surgical safety, advanced control strategies, such as adaptive controller [6] and active interventional control framework [7], have been proposed to constrain robotic motion within safe trajectories. Furthermore, to alleviate the cognitive and physical load on surgeons, an increasing number of studies have focused on automating sub-tasks such as tool positioning and navigation [8], [9], [10]. Despite these efforts, the bimanual nature of vitreoretinal surgery presents a ongoing challenge, requiring advanced skill and coordination due to its procedural complexity. To address the challenge, He et al. [11] introduced an automatic light pipe actuating system utilizing a force-sensing light pipe and a tool mounted on two SHERs. While this system enables coordinated motion, the light pipe's control strategy is based solely on aligning its orientation with the tool and lacks scene awareness. For truly intelligent assistance, the light pipe should autonomously adjust its position in response to dynamic surgical conditions.

To develop an intelligent surgical assistant, we introduce a vision-based robot-assisted vitreoretinal surgery system as illustrated in Fig. 1. Intraoperatively, an eye robot is introduced to autonomously adjust the light pipe based on real-time surgical scene understanding, thus freeing the surgeon's non-dominant hand and reducing both cognitive distraction and physical fatigue. Leveraging its scene understanding capabilities, the system is able to enhance safety by detecting surgical activities and issuing alerts for improper actions or tool contact with restricted areas. Postoperatively, the system can record surgical activities to automatically generate reports, thereby reducing the surgeon's documentation workload. It also tracks the tool tip position to reconstruct tool trajectories, which can be used for surgical skill assessment and training purposes. To enable these capabilities, we propose a surgical scene understanding model, DMNet, to detect surgical objects and activities. DMNet employs a "detect-and-match" strategy: it first detects all tools, tissues, and actions in current surgical scene, and subsequently constructs interaction quintuples through model-free post-processing. For intelligent robotic control, we further propose an optimization-based visual collaborative control method for automated light management. Together, these components establish a comprehensive framework for enhancing the precision, safety, and efficiency of vitreoretinal surgery through intelligent robotic assistance.

To evaluate the effectiveness of the proposed system, we created a phantom-based experimental environment to simulate the task of membrane peeling in vitreoretinal surgery.

*This work was supported in part by a Ministry of Education Academic Research Fund Tier 1 Grant (Robotic Seed Funding) (WBS: A-8003274-00-00)

Wenjun Lin, Wending Zhang, Chin-Boon Chng, Yong Jun Tan, and Chee Kong Chui are with Department of Mechanical Engineering, National University of Singapore, Singapore. Wenjun Lin and Wending Zhang are co-first authors. Wenjun Lin and Chee Kong Chui are corresponding authors: skyelin@nus.edu.sg, mpecck@nus.edu.sg

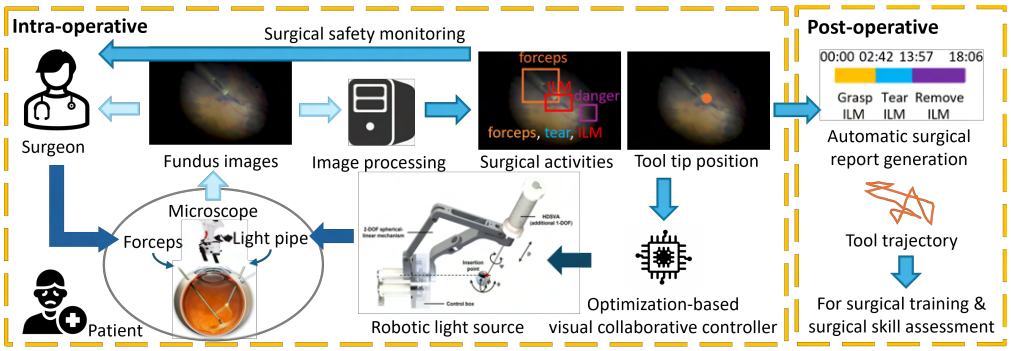

Fig. 1. Workflow of the proposed vision-based robot-assisted vitreoretinal surgery system.

An artificial eyeball was fabricated to mimic the anatomy of a human right eye, with polydimethylsiloxane employed to create artificial membranes within the eyeball. A human operator is tasked to use grasping forceps to perform membrane peeling inside the simulated eyeball, while the procedure was recorded to generate a simulated retinal dataset for training and validating the visual model. Experiments on the simulated retinal dataset demonstrate the effectiveness of the proposed DMNet in understanding surgical scenes, achieving a mAP_{ITI} of 63.89% for surgical activity detection and an AP_{50} of 99.29% for object detection. Additionally, the optimization-based visual collaborative control method has been validated through point reaching experiments, yielding an average error of 2.5 mm, thereby confirming its effectiveness. Furthermore, the overall system has been evaluated for its capability in intelligent light positioning, showcasing its practical application in enhancing surgical procedures.

In summary, our main contributions include:

- We propose a novel vision-based robot-assisted vitreoretinal surgery system to improve surgical efficiency and outcomes while reducing the workload on surgeons.
- We propose a simple but effective surgical scene understanding model, DMNet, which adopts a "detect-and-match" strategy to detect surgical objects and activities.
- We introduce an optimization-based visual collaborative control method for vitreoretinal surgery that automatically adjusts lighting based on surgical activities.

II. Proposed System

The proposed intelligent vision-based robot-assisted vitreoretinal surgery system consists of three components: vision, mechanism, and control. These components work together to improve surgical precision and safety. During surgery, a microscope is positioned above the patient's eye to provide vision for the surgeon and the computer to understand the surgical scene. An eye robot then holds and automatically adjusts the light pipe based on the proposed optimization-based visual collaborative control method. Surgical activities and tool tip positions are detected using the proposed DMNet model. This information can be used to monitor surgical safety and support report generation, skill assessment, and training. The target illumination point, determined from the detected activities, is sent to the control system. By applying coordinate transformation, the target point is mapped from the image frame to the robot's workspace, thereby linking visual data with robotic motion control. Finally, the inverse kinematics is calculated by the optimization algorithm to generate the required motor angles, which are transmitted to the robot to adjust the light accordingly. The following sections describe these three components in detail.

A. Vision

To support intelligent robot-assisted vitreoretinal surgery, the visual component needs to perform the following tasks:

- Object detection: Detect all tools, tissues, and defined danger areas, regardless of whether they are involved in interactions.
- Instrument-tissue interaction detection [12]: Detect fine-grained surgical activities represented as ⟨tool class, tool bounding box, tissue class, tissue bounding box, action class⟩ quintuples in the surgical scene.

To solve the above tasks, we propose DMNet, a simple yet effective surgical scene understanding model tailored for intelligent robot-assisted vitreoretinal surgery. As shown in Fig. 2, the surgical scene understanding problem is reformulated as a detection and matching task. The "detection" involves detecting all tools, tissues, and actions in the scene, while the "matching" pairs these elements to form interaction quintuples. In DMNet, an action is represented by the smallest bounding box encompassing its corresponding tool box and tissue box, along with an action category. DMNet follows a "detect-and-match" strategy. After extracting features using a ResNet50 backbone [13], three parallel branches are used to detect tools, tissues, and actions independently. A model-free post-processing matching algorithm is then employed to match each detected action with its corresponding tool and tissue, completing the formation of interaction quintuples.

1) Detect: Each detection branch shares a similar structure based on Faster R-CNN [14] framework, incorporating a Region Proposal Network (RPN), a RoI Align layer, and a box head layer. RPN generates region proposals that are likely to contain objects and predicts objectness scores and bounding box offsets. These proposals and feature maps are then processed by the RoI Align layer to extract fixed-size features from each region, which are then passed to the box head for classification and bounding box regression.

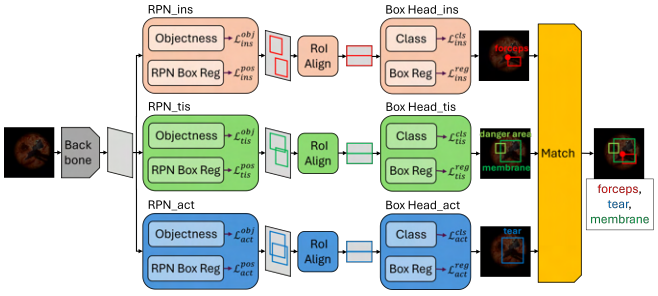

Fig. 2. Vision algorithm for surgical scene understanding.

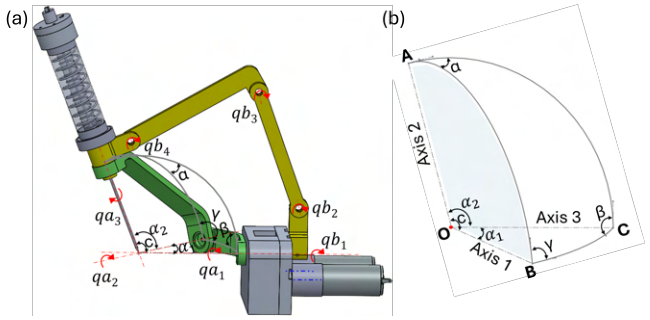

Fig. 3. Illustration of (a) Robot system; (b) Spherical geometry.

For each detection branch, the loss function includes both RPN loss and Fast R-CNN loss as described in the Faster RCNN model. The final multi-task loss is a combination of the aforementioned losses. Since each contributing loss behaves differently, the weight balancing parameters are not manually tuned but instead incorporated as learnable parameters during training, as introduced in Liebel et al. [15].

2) Match: During inference, Non-Maximum Suppression (NMS) is first applied separately for the outputs of each detection branch to produce the detection results for tools, tissues, and actions. Subsequently, a postprocessing matching strategy is employed to associate each detected action with its corresponding tool and target tissue, forming instrument-tissue interaction quintuple. Bounding boxes for actions, representing union boxes of the corresponding tools and tissues, provide positional information for matching. The position-based matching strategy can be summarized as follows:

(1) Pair the detected tools and tissues to create all possible tool-tissue pairs.
(2) Generate the union box for each tool-tissue pair.
(3) Calculate the Intersection over Union (IoU) between generated union boxes and detected action boxes.
(4) Match each detected action box (representing the detected tool and target tissue) with the unmatched union box that has the highest IoU, provided the IoU exceeds a set threshold (0.5).

In this way, the predicted tools, tissues, and actions are matched to form interaction quintuples, which describe the surgical activities. Tools and tissues that are not matched are identified as non-interacting objects, fulfilling the safety monitoring requirements of the surgical system. Based on the detected tool positions, the tool tip can be identified by locating the corner point of the tool bounding box that is closest to the center of the scene. This approach is reasonable as vitreoretinal surgery is a key-hole surgery, and the surgical tool can only enter the eyeball through a fixed incision. Therefore, tool tip position is restricted in the surgical scene.

Based on the detected information, the vision component will determine an illumination target point according to the designated lighting strategy (for example, the one described in Section III-E) and send it to the control component to calculate the control signal. Furthermore, by recording the tool tip positions and the surgical activities, the tool trajectories and the description of the surgical procedures in the surgical report can be automatically generated. The recorded trajectories and activities can be used for further surgical skill assessment and surgical training.

B. Mechanism

A robotic system is developed to automatically assist surgeons in performing surgical tasks. As shown in Fig. 3(a), the mechanism consists of a spherical (lower) linkage and a linear (upper) linkage, connected in parallel at their end to form a closed-loop chain. An end effector consisting of a tool actuator is mounted where both linkages meet. This system builds on the following spherical geometry to perform four DOF RCM constrained motion, including roll, pitch, yaw, and translation along the z axis of the end effector.

1) Spherical Geometry: As presented in Fig. 3(b), the slice of a spherical triangle is formed by the intersection of three great circles intersecting piecewise. The sphere's radius α_1 and α_2 are the parameters determined by mechanism design. In the context of spherical geometry, the location of the isocenter is equivalent to the practical use of RCM. In a general case where the designed arc length AC is not equal to BC, the spherical law of sines, which governs the ratio of the spherical arc length of unit radius to its corresponding spherical angle, can be stated as:

$$\sin\alpha_1/\sin\alpha = \sin\alpha_2/\sin\beta = \sin c/\sin\gamma \tag{1}$$

Then, spherical angle c and γ can be determined:

$$\begin{aligned}\cos c &= \cos\alpha_1\cos\alpha_2 + \sin\alpha_1\sin\alpha_2\cos\gamma \\ \cos\gamma &= -\cos\alpha\cos\beta + \sin\alpha\sin\beta\cos c\end{aligned} \tag{2}$$

2) Forward Kinematics Derivation: Table I describes the DH parameters of link A (the spherical linkage) and B (the linear linkage). In the table, qa_1, qa_2, qa_3 denote the joint angles of spherical linkage, while qb_1, qb_2, qb_3, qb_4 denote the joint angles of linear linkage as shown in Fig 3(a). The design parameters defining the dimensions of the linkage are selected according to the functional needs and the estimated surgical workspace. α in serial link A represents the spherical angle in Eq. 1 and Eq. 2, and d in serial link B represents the dimensional length of individual linear linkages.

With the listing of DH parameters, the symbolic representation of the transformation matrix is as follows:

$$\begin{aligned}A\_T_E^U &= A\_T_1^0 A\_T_2^1 A\_T_3^2 A\_T_E^3 \\ B\_T_E^U &= B\_T_1^0 B\_T_2^1 B\_T_3^2 B\_T_4^3 B\_T_E^4\end{aligned} \tag{3}$$

TABLE I
DH PARAMETER OF SERIAL LINK A AND B

LINK A				LINK B			
θ	r/mm	d	α	θ	r	d/mm	α
$qa_1+\pi$	155	0	0.79	qb_1	0	46	$\frac{\pi}{2}$
qa_2	0	0	-1.84	$qb_2-0.25$	0	118	0
$qa_3+\frac{\pi}{2}$	-116.36	0	0	$qb_3+2.50$	0	133.22	0
				$qb_4+1.19$	0	31	$-\frac{\pi}{2}$

TABLE II
PARAMETERS FOR FORWARD KINEMATICS

σ_{11}	$-0.68c_3-0.18c_2c_3-0.71s_2s_3$
σ_{21}	$0.5s_3(2*c_1c_2-1.41s_1s_2)-c_3(0.26c_1s_2-0.68s_1+0.18c_2s_1)$
σ_{31}	$0.5s_3(1.41c_1s_2+2*c_2s_1)-c_3(0.68c_1-0.18c_1c_2+0.26s_1s_2)$
σ_{12}	$0.68s_3+0.18c_2s_3-0.71c_3s_2$
σ_{22}	$s_3(0.26c_1s_2-0.68s_1+0.18c_2s_1)+0.5c_3(2c_1c_2-1.41s_1s_2)$
σ_{32}	$s_3(0.68c_1-0.18c_1c_2+0.26s_1s_2)$
σ_{13}	$0.68c_2-0.18$
σ_{23}	$0.18s_1+0.97c_1s_2+0.68c_2s_1$
σ_{33}	$0.97s_1s_2-0.68c_1c_2-0.18c_1$
P_x	$ToolLen(0.68c_2-0.18)+155$
P_y	$ToolLen(0.18s_1+0.97c_1s_2+0.68c_2s_1)$
P_z	$ToolLen(0.18c_1+0.68c_1c_2-0.97s_1s_2)$

$c_1=\cos(qa_1)$, $c_2=\cos(qa_2)$, $c_3=\cos(qa_3)$, $s_1=\sin(qa_1)$, $s_2=\sin(qa_2)$, $s_3=\sin(qa_3)$, and $ToolLen$ is the tool length exceeding the RCM point along the direction of the end effector.

The spherical linkage A is chosen to derive the forward kinematics while combining the restriction from linear linkage B. The function of qa_2 and qa_3 represented by qa_1 and qb_1 can be derived based on the spherical trigonometry. First, initialize with an input variable to serve as the crank angle:

$$\beta=\pi-qa_1-qb_1 \tag{4}$$

Substitute parameters from Table I into Eq.1 and Eq. 2, the functional representation of qa_3 and qa_2 are as follows:

$$qa_3=\sin^{-1}[\frac{\sin(\beta)\sin(\alpha_1)}{\sin(\alpha_2)}]$$
$$qa_2=-\text{sign}(qa_3)\text{sign}(\pi-\beta)\cos^{-1}[\frac{N}{D}]$$

where :

$$N=\sin(qa_3)\sin(\beta)\cos(\alpha_1)\cos(\alpha_2)-\cos(qa_3)\cos(\beta)$$
$$D=1-\sin(qa_3)\sin(\beta)\sin(\alpha_1)\sin(\alpha_2) \tag{5}$$

The full solution of forward kinematics is then given as:

$$Unified\_T_E^U=\begin{bmatrix}\sigma_{11}&\sigma_{12}&\sigma_{13}&P_x\\ \sigma_{21}&\sigma_{22}&\sigma_{23}&P_y\\ \sigma_{31}&\sigma_{32}&\sigma_{33}&P_z\\ 0&0&0&1\end{bmatrix} \tag{6}$$

where $Unified\_T_E^U$ represents the transformation from the base joint of the spherical link to the end effector joint and the parameters are listed in Table II.

C. Control

Establishing an optimized, collaborative, vision-based control system requires an integration framework that bridges robotics, vision, and control. In this component, hand-eye calibration links the visual data to robotic motion control. The optimization algorithm will then ensure an optimal inverse kinematics solution.

1) Hand-Eye Calibration: The calibration of the camera is conducted first to obtain the transformation function between pixel coordinates and camera coordinates, as defined by the following equation:

$$\begin{bmatrix}X\\Y\\Z\end{bmatrix}=ZK^{-1}\begin{bmatrix}u\\v\\1\end{bmatrix} \tag{7}$$

where (u,v) is the target point under pixel coordinates, Z is the depth information, K is the camera intrinsic matrix, and (X,Y,Z) is the 3D camera coordinate. In vitreoretinal surgical scenarios, depth variation is limited due to the constrained eyeball workspace. Hence, a constant depth value is used, obtained by measuring the distance between the camera coordinate origin and the main working plane in the eyeball. The transformation of the target point P from camera coordinates to the base coordinate of the spherical link is as follows:

$$P=T_{cam}^{base}\begin{bmatrix}X\\Y\\Z\\1\end{bmatrix} \tag{8}$$

where T_{cam}^{base} is the homogeneous transformation matrix between base and camera coordinates given by measurement.

2) Optimization-based Control via Nonlinear Programming (NLP): With the transformation determined, an optimization problem based on constrained nonlinear programming can be formulated to calculate the inverse kinematics, and the output will be the two joint angles of the two motors qa_1 and qb_1:

$$\begin{aligned}\min_{\theta_1,\theta_2}\ &\|PQ\times Z_{\text{axis}}(\theta)\|+\lambda_{\text{smooth}}\left\|\theta-\theta_{\text{prev}}\right\|^2\\ \text{s.t.}\ &lowbound_1\le\theta_1\le upbound_1,\\ &lowbound_2\le\theta_2\le upbound_2\end{aligned} \tag{9}$$

where PQ is the vector that passes the origin of the coordinate of the end effector Q and the target point P, Z_{axis} is the direction vector of the end effector, λ_{smooth} is the weight coefficient, θ and θ_{prev} represent the base joint angles at the current step and the last step respectively, θ_1 and θ_2 represent qa_1 and qb_1, and $lowbound$ and $upbound$ determine the limit of the range of two base joint angles.

This NLP problem aims to minimize the distance between the direction vector Z_{axis} and the illumination target point P to provide optimal lighting for surgical operations. In addition, minimizing changes in joint motion using a weighted coefficient helps prevent excessive displacement of the two links, ensuring smooth and stable light positioning. In membrane peeling, smooth surgical manipulation is essential, and the robot's joint motions should exhibit minimal variability accordingly. Therefore, the objective function aims to minimize the distance and displacement of the joint to provide a safe, precise, and rapid response to the surgeon's movement.

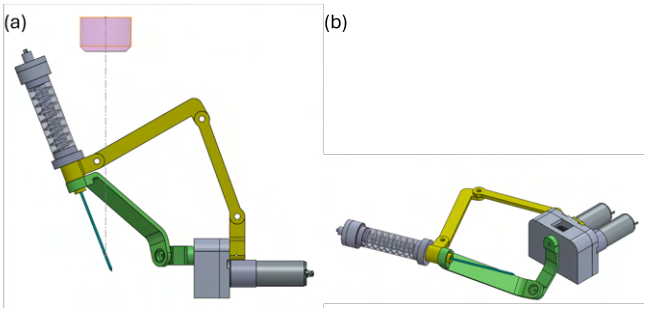

Fig. 4. Unsafe conditions: (a) Blocking camera; (b) Hurting patient's face.

Fig. 4 illustrates two highly unsafe conditions that may occur during eye surgery. In Fig. 4(a), only qa_1 rotates to 90° while qb_1 remains in its initial condition. This configuration will block the camera's view and increase the risk of surgical errors. Fig. 4(b) occurs when qb_1 rotates to 90° while qa_1 stays in its initial position. This configuration poses a risk of contact with the patient's face, thus compromising safety. Therefore, to avoid these unsafe scenarios and based on the workspace analysis in Section III-D.1, the joint angle limits for qa_1 and qb_1 are established as follows: $lowbound_1 = -\frac{\pi}{2}$, $lowbound_2 = -\frac{\pi}{3}$, $upbound_1 = \frac{\pi}{2}$, $upbound_2 = \frac{\pi}{3}$.

To solve this NLP problem, the sequential least squares programming (SLSQP) is applied for its high speed and accuracy [16] to calculate the optimal base joint angles qa_1 and qb_1. Subsequently, these values will be transmitted to the low-level controller as the reference signal to control the robot.

III. Experiments and Results

A. Experimental Setup

A phantom-based vitreoretinal surgery environment was established for experimentation, with a focus on the membrane peeling task, which is a representative and commonly performed vitreoretinal surgical procedure. An artificial eyeball was designed to replicate the anatomy of a human right eye. It featured a 3D-printed polylactic acid shell, lined with white clay to ensure a smooth interior surface. In addition, the fundus was represented by a hand-painted image in acrylic. Polydimethylsiloxane was used to create the artificial membranes placed in the simulated eyeball. Artificial membrane peeling was performed using grasping forceps (JARIT 460-230) with a tip width of 1mm, a tip length of 8mm, and a working length of 250mm.

As shown in Fig. 5, a CMOS camera (Toshiba USB3, V2406MCFBU) was mounted on a vertically adjustable platform and paired with a telescopic lens (Computer, TEC-M55) to simulate a surgical microscope system. The camera was elevated to provide a clearance height of approximately 15cm. A Python-based User Interface was used to capture videos for image analysis. The light pipe was simulated using a 5mm diameter straight hollow tube equipped with a miniature LED light source (Nichia, NSPW315DS), which illuminated the interior of the simulated eyeball prototype with a luminous intensity of approximately 2.5 to 3.4 cd.

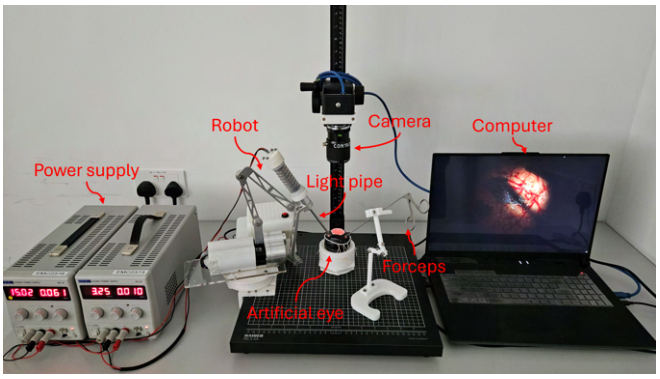

Fig. 5. Illustration of experiment setup.

For the robotic system, an Arduino Mega ATmega2560 was used as the microcontroller unit to implement PID position control and PI speed control for the motors. Communication between the high-level control system on the PC and the Arduino board was managed via ROS Noetic. Desired motor angles were transmitted to the Arduino board through ROS topics, which then autonomously controlled the motors.

The data flow of simulated system can be summarized as:

1) Surgical scene understanding: Detect instruments, tissues, defined danger areas, and surgical activities in the surgical scene based on the proposed DMNet model.
2) Optimization-based visual collaborative control: Calculate reference control signal via nonlinear programming with the target points from the visual algorithm.
3) Arduino board execution: Receive reference control signals via ROS serial port and execute control based on PID position control and PI speed control.

B. Data Collection

Before system testing, simulated surgical video data needs to be collected to train the model. A total of 36 simulated surgery videos, each with a resolution of 512×512, were recorded. The lighting in these videos was manually controlled by a human operator. These videos were downsampled to 3fps and randomly split into a training set of 29 videos, a validation set of 3 videos, and a test set of 4 videos, with 1,699 image frames in the training set, 178 frames in the validation set, and 188 frames in the test set. Each video was annotated with object categories, locations, and types of actions performed. The details are provided in Table III.

TABLE III
Instrument, tissue, action label classes and interaction combinations of the retinal dataset.

Instrument	Tissue	Action	Interaction combination
forceps	membrane danger area	grasp tear remove	forceps-grasp-membrane forceps-tear-membrane forceps-remove-membrane

C. Visual Algorithm Testing

1) Evaluation Metrics: For visual algorithm testing, Mean Average Precision (mAP) was used as the evaluation metric

for both object detection and surgical activity (instrument-tissue interaction) detection. For object detection, we followed the mAP calculation method used in the COCO dataset [17]. Following works for instrument-tissue interaction [12], we calculated and denoted the mAP for instrument-tissue interaction detection as mAP_{ITI}. In addition, accuracy is used to evaluate the performance of the surgical activity category.

To evaluate the correctness of the detected tool tip, Point-to-point Error for Landmark (PEL) was used. In addition, following evaluation metrics for tool tracking, tool tip point was extended into a tip area by generating a bounding box of 50 pixels side length centered at the tip point. mAP was then computed between predicted and ground-truth tip bounding boxes. This metric is denoted as 50-AP_{50} in this paper.

2) Implementation Details: DMNet was trained for 50 epochs using AdamW optimizer and the initial learning rate is set to 0.0001 with 0.1 decayed at the 40th epochs. Models were tested on one GeForce RTX 3060 GPU.

TABLE IV

EXPERIMENTAL RESULTS ON THE RETINAL TEST SET.

Object Detection (%)			Activity (%)		Tip Localization	
AP_{50}	AP_{75}	AP	mAP_{ITI}	Acc	PEL	50-AP_{50}
99.29	78.37	67.11	63.89	68.62	6.63 pixels	73.85%

3) Experimental Results: The experimental results on the retinal test set are listed in Table IV. For tool and tissue detection, the proposed DMNet achieves 99.29% AP_{50}, 78.37% AP_{75}, and overall 67.11% AP, proving its object detection capability. For activity detection, DMNet achieves 63.89% mAP_{ITI}, demonstrating its ability to detect instrument-tissue interactions. When evaluating the performance of activity categories, DMNet achieves an accuracy of 68.62%. For tool tip localization, the PEL is 6.63 pixels, which is a relatively small error compared to the image size of 512 pixels. Besides, the 50-AP_{50} is 73.85%, indicating that DMNet can localize the tool tip accurately within the tolerance range.

Fig. 6 presents some visualization results from the retinal test set, which illustrates the model's ability to understand the surgical scene and its robustness under various lighting conditions. As shown in the third column, even when a tissue is not actively interacted with, the model still recognizes and includes it in the output alongside other interaction pairs.

The example results of the generated tool trajectory and surgical activity timeline are presented in Fig. 7. This automatically generated information could be useful for surgical assessment, surgical training, and surgical report generation.

To evaluate the performance of DMNet against existing methods, experiments were conducted on two instrument-tissue interaction detection benchmark: the PhacoQ Dataset and the CholecQ Dataset [18]. The experimental results, summarized in Table V, show that DMNet achieves comparable mAP_{ITI} to state-of-the-art models on both datasets. Notably, DMNet provides significant advantages in terms of reduced inference time and a smaller model size, making it more efficient for practical deployment. Moreover, compared to end-to-end models QPIC and AIPNet, DMNet can identify instruments and tissues that are not involved in interactions, better aligning with the requirements of the proposed system.

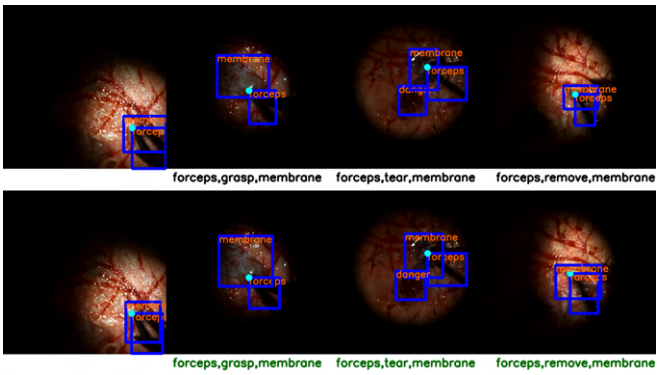

Fig. 6. Example results on the retinal test set. The first row shows the ground truth, while the second row shows the results from the proposed DMNet. The ground-truth or detection bounding boxes of tools and tissues are marked in blue. Correct interaction detection results are marked in green. The detected tool tip points are marked in light blue.

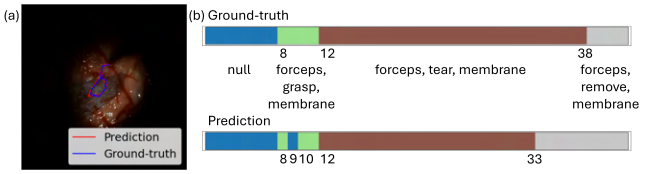

Fig. 7. Example results of the generated (a) tool trajectory and (b) activity timeline for the surgical report.

TABLE V

COMPARISON RESULTS ON PHACOQ AND CHOLECQ DATASETS.

Methods	Param	FPS	Time	PhacoQ	CholecQ
				mAP_{ITI}	mAP_{ITI}
Baseline [14]	94M	7.56	0.1323s	33.75%	28.69%
iCAN [19]	123.8M	7.03	0.1423s	34.13%	31.73%
SCG [20]	95.4M	6.47	0.1545s	34.63%	35.27%
QDNet [12]	131.6M	0.66	1.5071s	34.89%	36.83%
ITIDNet [18]	148.74M	4.31	0.2321s	36.82%	39.01%
QPIC [21]	36.88M	20.45	0.0489s	21.52%	24.07%
AIPNet [22]	261.79M	9.92	0.1008s	37.35%	39.42%
DMNet	70.5M	13.41	0.0746s	33.96%	36.77%

Param denotes model parameters, FPS denotes frames per second, and Time denotes the processing time per frame.

D. Control System Testing

1) Constrained Workspace Analysis: To ensure the constrained workspace meets operational requirements, it is essential to assess whether it covers the entire target area within the eyeball. Fig. 8(a) illustrates the simulated constrained workspace of the end-effector tool tip. To evaluate whether the robot can reach all working areas, the workspace is projected to a fixed main working plane at z=-45mm according to the base coordinate of robot linear linkage. As shown in Fig. 8(b), the red area represents the reachable workspace on the plane, while the blue circle represents the required workspace. It is evident that, despite the constraints, the robot's workspace fully covers the required area.

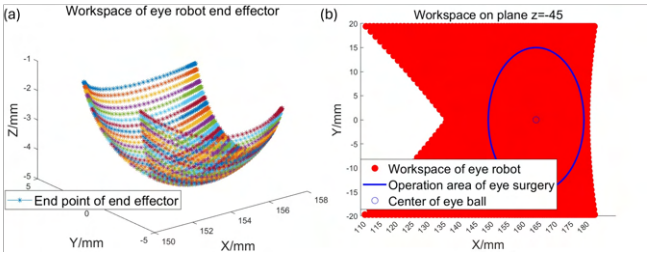

Fig. 8. (a) Constrained workspace; (b) Workspace on plane z=-45mm.

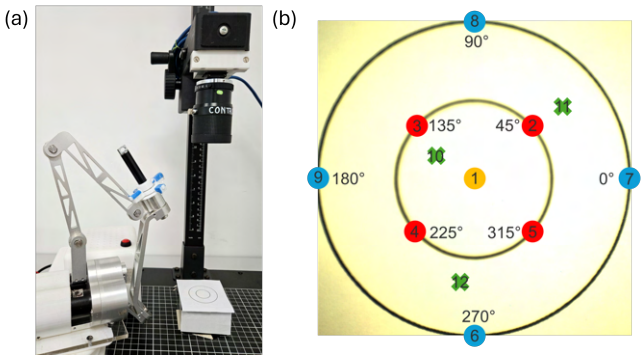

Fig. 9. Illustration of (a) test scenario and (b) target test point locations.

2) Control System Accuracy Validation: To evaluate the accuracy of the control system, the light pipe was replaced with a laser pointer, and the artificial eyeball was replaced by a piece of white paper positioned at the same position, as shown in Fig. 9(a). This paper represents the surgical operation plane at z=-45mm. A concentric circle representing the maximum operational area of the eyeball visible to the camera was printed on the paper. The outer circle denotes the full operational area and the inner circle (with half the radius) indicates the main operational area. Twelve points were selected to validate the positioning accuracy. Four points on the outer circle were placed along the vertical and horizontal axes, and four points on the inner circle were located along the diagonals at 45° and -45°. One point corresponds to the center of the concentric circles, while the remaining three were randomly chosen within the operational area.

This experiment was specifically designed to evaluate the accuracy of the control system alone, excluding the influence of the visual algorithm. Each target point was tested with 10 repeated trials. The Euclidean distance between the target point and the laser point on the concentric circle testing plane was manually measured and recorded as the positioning error of the control system. As shown in Fig. 10, the maximum observed error was 6 mm, while the minimum was 0.5 mm, resulting in an average error of 2.5 mm. The primary sources of error are mechanical and calibration inaccuracies. Mechanical factors include joint tremors and motor angle sensing errors. For example, point nine, located on the left side at 180°, exhibited the lowest accuracy. This is due to mechanical errors caused by friction and sensing errors when a large base joint angle of the spherical linkage is needed. Calibration errors arise from uncertainties in the hand-eye

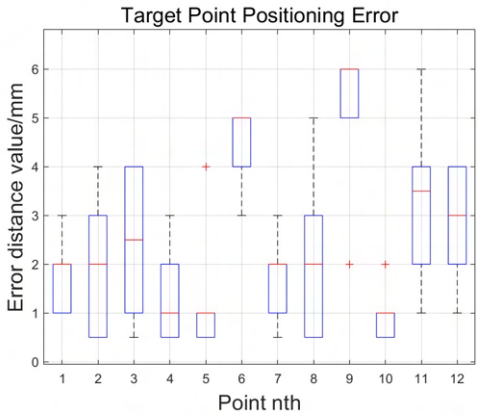

Fig. 10. Control system accuracy validation.

transformation matrix parameters and slight positional shifts during experimentation. Given the robot's relatively small operational workspace, minor differences between mathematical equations and the physical setup can result in notable deviations in performance. Although such errors may be unavoidable, the light spot from a typical light pipe has a significantly wider cone angle than a focused laser spot, making the current error level tolerable for practical use. Consequently, these inaccuracies have minimal impact on the effectiveness of the system to illuminate the intended surgical area.

E. Collaborative Software and Hardware Simulation

The proposed surgical system requires intricate collaboration between software and hardware components. To test the system's capability for automatic light control, experiments were conducted in a simulated environment. The experimental results are demonstrated in the supplementary video.

As the tool tip can indicate where surgical activity occurs and is often used as a tracking target, we evaluated the system's ability to follow it by testing whether the robot could automatically adjust the lighting to track a forceps moving randomly within the simulated eye. As shown in the experiment video, the system can track the tooltip within the simulated eyeball.

In addition, we simulated the membrane peeling procedure, with the system autonomously controlling the lighting. This simulation aims to assess the effectiveness of the system during active surgical procedures. In this experiment, the automatic light control strategy was configured as follows:

1) If no objects are detected, the illumination target point is set to the center of the image.
2) If only tissue is detected, the illumination target point is set to the center of the tissue bounding box.
3) If the tool is detected and there is no tool-tissue interaction, the illumination point is set to the tool tip.
4) If tool-tissue interaction is detected, the illumination target point is set to the center of the intersection between the tool and tissue bounding boxes.

Fig. 11 presents frames from the experimental video,

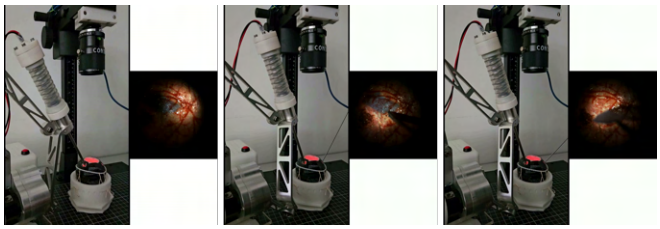

Fig. 11. Example experiment video frames to present the system's ability to track the illumination target point during active surgical procedures.

demonstrating the system's ability to track the illumination target point during active surgical procedures.

IV. Conclusion

In conclusion, we propose an intelligent vision-based robot-assisted system for vitreoretinal surgery with the surgical scene understanding ability to automatically assist surgeons in specific tasks. The proposed system has the capacity to improve surgical outcomes by providing intelligent control over lighting based on the surgical context, enhancing safety by monitoring surgical activities, and reducing the workload of surgeons by automating certain aspects of the surgical procedure. To achieve this, we propose a simple but efficient surgical scene understanding model, DMNet, to detect the surgical objects and activities in the given scenes. Furthermore, an optimization-based visual collaborative control method is developed to automatically position the light. The results of the experimental and simulation studies confirm the effectiveness of the proposed system in the workspace.

While the proposed system shows promising results, several limitations remain. The simulated eyeball is approximately three times larger than a standard adult eyeball, simplifying the setup by eliminating the need for an aspheric lens but limiting realism. Future work will involve conducting ex-vivo experiments on porcine models. Additionally, despite using thrust and axial bearings at linkage joints, friction is observed on the contact surfaces of the spherical mechanism. Occasional shaft misalignments lead to undesirable jerking and stalling at certain configurations. However, these issues can be addressed through further mechanical optimization.

Currently, this robot is set to control the light source. However, it can be adapted for other surgical tasks, such as needle insertion and drug delivery. This system has the potential to evolve into a comprehensive robotic-assisted system for vitreoretinal surgery.

References

[1] B. S. Peters, P. R. Armijo, C. Krause, S. A. Choudhury, and D. Oleynikov, "Review of emerging surgical robotic technology," *Surgical endoscopy*, vol. 32, pp. 1636–1655, 2018.

[2] S. Maeso, M. Reza, J. A. Mayol, J. A. Blasco, M. Guerra, E. Andradas, and M. N. Plana, "Efficacy of the da vinci surgical system in abdominal surgery compared with that of laparoscopy: a systematic review and meta-analysis," *Annals of surgery*, vol. 252, no. 2, pp. 254–262, 2010.

[3] H. Rafii-Tari, C. J. Payne, and G.-Z. Yang, "Current and emerging robot-assisted endovascular catheterization technologies: a review," *Annals of biomedical engineering*, vol. 42, pp. 697–715, 2014.

[4] P. J. Johnson, C. M. R. Serrano, M. Castro, R. Kuenzler, H. Choset, S. Tully, and U. Duvvuri, "Demonstration of transoral surgery in cadaveric specimens with the medrobotics flex system," *The Laryngoscope*, vol. 123, no. 5, pp. 1168–1172, 2013.

[5] R. Taylor, P. Jensen, L. Whitcomb, A. Barnes, R. Kumar, D. Stoianovici, P. Gupta, Z. Wang, E. Dejuan, and L. Kavoussi, "A steady-hand robotic system for microsurgical augmentation," *The International Journal of Robotics Research*, vol. 18, no. 12, pp. 1201–1210, 1999.

[6] A. Ebrahimi, N. Patel, C. He, P. Gehlbach, M. Kobilarov, and I. Iordachita, "Adaptive control of sclera force and insertion depth for safe robot-assisted retinal surgery," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9073–9079.

[7] C. He, N. Patel, M. Shahbazi, Y. Yang, P. Gehlbach, M. Kobilarov, and I. Iordachita, "Toward safe retinal microsurgery: Development and evaluation of an rnn-based active interventional control framework," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 4, pp. 966–977, 2019.

[8] J. W. Kim, C. He, M. Urias, P. Gehlbach, G. D. Hager, I. Iordachita, and M. Kobilarov, "Autonomously navigating a surgical tool inside the eye by learning from demonstration," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 7351–7357.

[9] P. Zhang, J. W. Kim, and M. Kobilarov, "Towards safer retinal surgery through chance constraint optimization and real-time geometry estimation," in *2021 60th IEEE Conference on Decision and Control (CDC)*. IEEE, 2021, pp. 5175–5180.

[10] Y. Koyama, M. M. Marinho, M. Mitsuishi, and K. Harada, "Autonomous coordinated control of the light guide for positioning in vitreoretinal surgery," *IEEE Transactions on Medical Robotics and Bionics*, vol. 4, no. 1, pp. 156–171, 2022.

[11] C. He, E. Yang, N. Patel, A. Ebrahimi, M. Shahbazi, P. Gehlbach, and I. Iordachita, "Automatic light pipe actuating system for bimanual robot-assisted retinal surgery," *IEEE/ASME Transactions on Mechatronics*, vol. 25, no. 6, pp. 2846–2857, 2020.

[12] W. Lin, Y. Hu, L. Hao, D. Zhou, M. Yang, H. Fu, C. Chui, and J. Liu, "Instrument-tissue interaction quintuple detection in surgery videos," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 399–409.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[15] L. Liebel and M. Körner, "Auxiliary tasks in multi-task learning," *arXiv preprint arXiv:1805.06334*, 2018.

[16] D. Kraft, "A software package for sequential quadratic programming," *Forschungsbericht- Deutsche Forschungs- und Versuchsanstalt fur Luft- und Raumfahrt*, 1988.

[17] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[18] W. Lin, Y. Hu, H. Fu, M. Yang, C.-B. Chng, R. Kawasaki, C. Chui, and J. Liu, "Instrument-tissue interaction detection framework for surgical video understanding," *IEEE Transactions on Medical Imaging*, 2024.

[19] C. Gao, Y. Zou, and J.-B. Huang, "ican: Instance-centric attention network for human-object interaction detection," in *British Machine Vision Conference*, 2018, p. 41.

[20] F. Z. Zhang, D. Campbell, and S. Gould, "Spatially conditioned graphs for detecting human-object interactions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 319–13 327.

[21] M. Tamura, H. Ohashi, and T. Yoshinaga, "Qpic: Query-based pairwise human-object interaction detection with image-wide contextual information," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 10 405–10 414.

[22] W. Lin, Y. Hu, L. Hao, H. Fu, C. Chui, and J. Liu, "Aipnet: Action-instance progressive learning network for instrument-tissue interaction detection," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–11, 2025.