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

# One-Shot Autofocus via User-Adaptive Gaze Control for Robot-Assisted Microsurgery

Yunfei Luan, Yuxuan Liu, Yuyang Zhuge, Yating Luo,
Yao Guo, *Member*, *IEEE*, Guang-Zhong Yang, *Fellow*, *IEEE*

***Abstract*— Robot-assisted microsurgery (RAMS) is rapidly advancing with increasing levels of automation. Given the inherently shallow depth-of-field characteristics of surgical microscopes, integrating autofocus capabilities into RAMS has emerged as an urgent trend. Among existing solutions, gaze-induced autofocus has gained prominence due to its natural alignment with the surgeon's visual attention. However, gaze autofocus often relies on complex and non-intuitive triggering mechanisms, making it difficult to adapt for diverse users. Additionally, although the hill-climbing strategy is commonly employed to find the optimal focus plane, this process is inefficient for RMAS due to its slow convergence and inability to accommodate dynamic surgical scenarios. To address these limitations, we propose a novel gaze-controlled autofocus system featuring user-adaptive triggering and one-shot focusing. When a region is defocused and under the surgeon's gaze, our system rapidly achieves optimal focus with a single-step lens movement. Surgeons can easily adjust trigger sensitivity using a slider. Experiments validate the accuracy of our defocus estimation and triggering prediction algorithms. A user study demonstrates that the proposed system offers superior user-friendliness and operational efficiency compared to conventional systems.**

## I. Introduction

Robot-assisted microsurgery (RAMS) has become a burgeoning field of research in recent years [1], [2]. Robots enhance surgical precision and efficiency, offering significant benefits to the medical community. Currently, there is a growing trend towards automation for surgical robotics [3]–[5]. Automation alleviates surgeons from repetitive tasks, allowing them to focus on critical decision-making and complicated procedures.

In microsurgery, the surgical microscope serves as a primary observation tool for sub-millimeter operation fields [6], [7]. However, due to the trade-off of high resolution, its depth-of-field tends to be shallow, resulting in frequent defocus blur during the procedures. Conventionally, surgeons manually adjust the focus plane via foot pedals, which is time-consuming and distracting. With the advances in surgical automation, integrating autofocus capabilities into RAMS is an essential development goal.

To achieve this goal, an initial consideration is where to focus. Traditional methods set the focal plane for maximizing global sharpness or by tracking a predefined target [8]. However, the most likely focusing region is typically at the surgeon's visual attention, which is local and continuously shifting. For this reason, recent studies have introduced gaze-contingent focus control [9], [10], enabling the focus to adjust following the operator's point of regard (PoR), significantly improving user convenience.

A more challenging aspect is determining when to focus, i.e., focus trigger. Existing methods rely on predefined rules on gaze-scenario interaction [11] or specific gaze gestures [12], which are either complicated or stressful. In contrast, Cao *et al.* [13] proposes an attention-triggered mechanism, directly leveraging the pupil variation to determine whether to focus. While intuitive, the trigger becomes a distraction if the regarded region is already in focus. Therefore, integrating gaze and defocus information is necessary. Additionally, existing methods overlook the variability of triggering criteria across users due to factors like individual gaze behaviors and defocus tolerance. Adapting to individual operators requires time-consuming adjustment, hindering practical deployment. Therefore, improving user-adaptability is essential.

A further challenge lies in how to focus both accurately and rapidly. Deep-learning-based defocus estimation methods have been extensively explored on cell microscope datasets [14], [15]. Meanwhile, cellular specimens are commonly 2D, making these models difficult to generalize to 3D surgical environments. Large-scale depth estimation fundamental models, such as Depth-Anything-V2 [16], have also been applied to estimate pixel-wise depths, from which the axial out-of-focus distances can be derived by comparing against the fixed in-focus depth. Nevertheless, these models are error-prone for highly blurred and texture-sparse surgical images, which exhibit a domain gap from natural images. Recently, [17] proposes a 3D tracking framework for microsurgical autofocus. They employ a defocus-restoration-based tracker targeting the ROI (Region of Interest), and an order-constrained model predicting the ROI's defocus level. While this method achieves superior defocus prediction accuracy compared to prior works, it still requires multiple iterations to converge to the in-focus state, akin to traditional hill-climbing optimization strategies [18], [19]. In highly dynamic microsurgical procedures, rapid focus is critical. Therefore, it is necessary to advance from multi-shot optimization towards one-shot absolute out-of-focus distance prediction, which is expected to be effective across various microsurgical procedures with simple calibration.

In this work, we propose a gaze-controlled autofocus framework for robot-assisted microsurgery. The framework

This work was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX, and also in part supported by National Natural and Science Foundation of China under grant 62203296. (*Corresponding authors: Yao Guo, Guang-Zhong Yang*)

Y. Luan, Y. Liu, Y. Zhuge, Y. Luo, Y. Guo and G.-Z. Yang are with Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. ({yunfei.luan, yao.guo, gzyang} @sjtu.edu.cn).

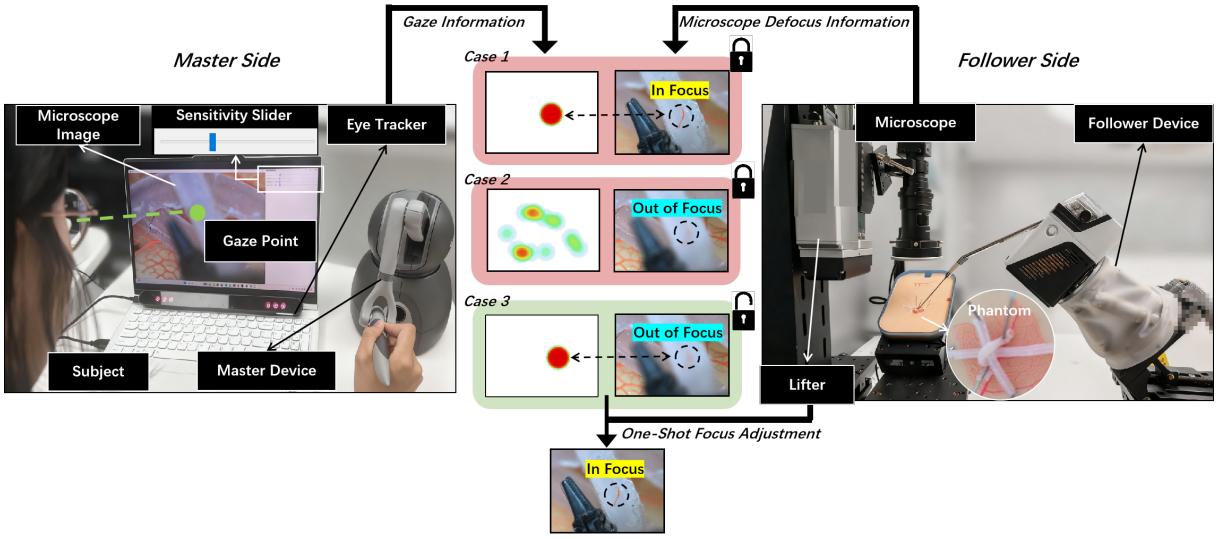


Fig. 1. Task Definition. During teleoperated surgery, the subject controls the master device while viewing the microscope video on the screen. An eye tracker captures PoR to determine the focus region. Autofocus triggers only if recent PoRs cluster near a defocused area. Trigger sensitivity is adjustable via an on-screen slider throughout the surgery. Once activated, the microscope lifter moves in one shot by the measured defocus distance for rapid focusing.

advances in user-adaptive triggering and one-shot focusing. As presented in Fig. 1, during the teleoperated microsurgery procedure, the operator employs the master device to conduct surgical operations and gazes at the front screen, which displays the real-time image stream from the microscope. The PoR on the screen captured by an eye tracker decides the to-focus region. The focus is only triggered when the recent PoRs are allocated near a region which is out of focus, i.e., when the operator observes a defocused region with high attention. The triggering sensitivity can be adjusted by dragging a slider on the screen intra-operatively. Once triggered, the lifter carrying the microscope is displaced by the absolute defocus distance of the region. The displacement is in one shot, guaranteeing rapidness. To realize user-adaptive triggering, a trigger prediction network is proposed. It integrates the gaze and defocus information from temporal and spatial perspectives, and outputs real-time trigger possibility. The user can set the possibility threshold based on their own experience. To realize one-shot focusing, a bi-directional defocus estimation network is introduced. With the expanded perception field and the additional sharpness cues, the network can distinguish near and far defocus, and estimate the absolute defocus distance map based on a single image. A key challenge common to both networks is obtaining high-quality labelled data. Thus, simple but efficient data collection and labelling methods are designed for the two networks, making our framework easily employed in practical RMAS procedures. Experiments on each network are performed to prove effectiveness and superiority. A user study is conducted to validate that our system is robust in practice, and obviously more user-adaptable and time-efficient than prior systems.

The main contribution of this paper is three-fold:

- A gaze-controlled autofocus framework is proposed for robot-assisted microsurgery, which features user-adaptive triggering and one-shot focusing, and is validated to be more user-friendly and time-efficient than the previous works.
- A triggering prediction network is proposed based on the integration of gaze and defocus information. The trigger sensitivity can be easily adjusted, facilitating high user-adaptability.
- A bi-directional defocus estimation network is proposed, which predicts the absolute defocus distance map based on a single image, enabling the one-shot focusing.

The remainder of this article is organized as follows: Section II illustrates the proposed framework and the detailed networks. Section III presents experiments evaluating the performance of each network and the whole system. Section IV concludes the article and discusses the future directions of our work.

## II. Methodology

### A. Framework Overview

To achieve one-shot and user-adaptive autofocus, our framework processes gaze information and microscope images, and outputs the decision on whether to focus and the moving distance of the microscope in real time. As shown in Fig. 2, it mainly consists of three modules, Gaze Information Processing Module (GIPM), Defocus Estimation Module (DEM), and Focus Trigger Prediction Module (FTPM).

*1) GIPM:* GIPM generates the spatial distribution and temporal features of recent PoRs. Given a sampling window with size $m$, it utilizes $m$ lastest data points from the eye tracker as the input, which includes the PoR coordinate relative to the screen $P = (p^x_{left}, p^y_{left}, p^x_{right}, p^y_{right}), p \in [0, 1]$ and the pupil diameters $Q = (q_{left}, q_{right}), q \in [0, 10]$, where $left$ and $right$ represent the left and right eye. To process the raw data sequence, first, blinking frames are selected by $q < 2$. In these frames, $p$ of the corresponding

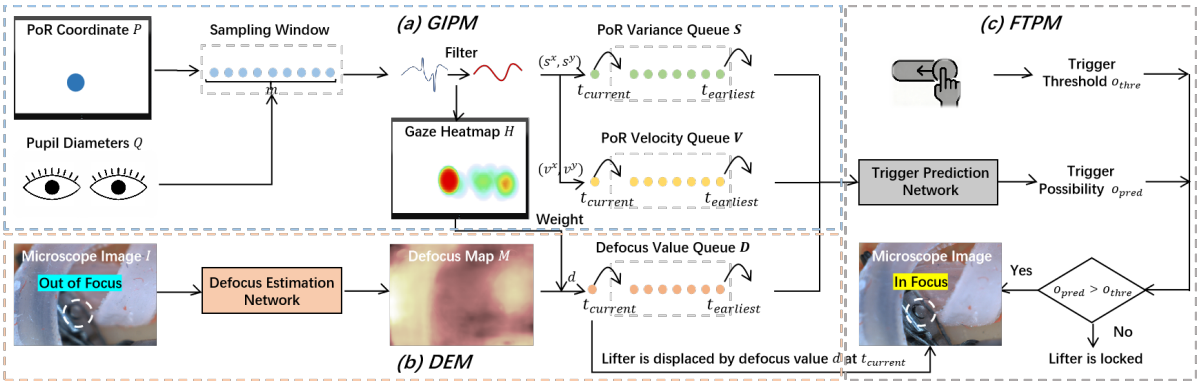


Fig. 2. Framework overview. The framework consists of three modules. (a) GIPM. The gaze information processing module extracts the spatial and temporal features of PoR according to recent raw PoR inputs and pupil diameters. (b) DEM. The defocus estimation module predicts physical defocus distances at the pixel level given the current microscope image, and the output is further weighted by the gaze heatmap to generate the average defocus value of the regarded region. (c) FTPM. The focus trigger prediction module predicts the current possibility of triggering based on gaze and defocus information. The trigger sensitivity can be adjusted by dragging the trigger threshold slider for different users. With the whole framework, the defocus region with high attention can be auto-focused in one shot. Otherwise, the lens remains locked to provide a stable view.

eye is regarded as invalid and reset as the latest valid $p$. Following that, a Kalman filter is applied to smooth $P$ to mitigate noise disturbance. Next, the aggregated PoR coordinate $P_{mid} = (p^x, p^y), p^x = (p^x_{left} + p^x_{right})/2, p^y = (p^y_{left} + p^y_{right})/2$ is calculated.

For the spatial distribution branch, given the screen width $w$ and height $h$, the gaze heatmap $H(x, y) \in \mathcal{R}^{h\times w}$ is yielded as

$$H(x,y) = \sum_{t=1}^{m} exp(-\frac{(x - p_t^x \times w)^2 + (y - p_t^y \times h)^2}{2\sigma^2}), \tag{1}$$

where $p_t^x$ and $p_t^y$ are the $p^x$ and $p^y$ of the $t$-th timestamp within the sampling window, and $\sigma$ is the gaussian standard deviation. After that, $H$ is normalized to guarantee that the sum of all pixel values equals 1. In this way, the weight map standing for spatial attention distribution is obtained.

For the temporal feature branch, the PoR variance $S = (s^x, s^y)$ is formulated as the variance of $p^x$ and $p^y$ among the whole sampling window, and the PoR velocity $V = (v^x, v^y)$ is formulated as the error of $p^x$ and $p^y$ between the last two timestamps. Subsequently, the PoR variance queue **S** and PoR velocity queue **V** are updated by the current $S$ and $V$, respectively. And the earliest $S$ and $V$ are dropped to stabilize the queue lengths. Following the above pipeline, the long-term and short-term temporal features are extracted and stored.

*2) DEM:* DEM estimates the defocus value among the user attention region. With the latest-captured RGB microscope image $I \in \mathcal{R}^{h\times w\times 3}$ as the input, DEM leverages a defocus estimation model, which features in direction-aware absolute defocus value measurement, to generate a defocus map $M \in \mathcal{R}^{h\times w}$. To integrate the defocus information across the entire attention field and reduce estimation error at any single gaze point, the output defocus value $d$ is generated as the heatmap-weighted sum of $M$, i.e.,

$$d = \sum_{x=0}^{w-1}\sum_{y=0}^{h-1} H(x,y) \times M(x,y). \tag{2}$$

After that, the current $d$ updates the defocus value queue **D** in the same way as **S** and **V** update. Finally, a bank of defocus values with inherent temporal and spatial information is obtained.

*3) FTPM:* FTPM predicts the likelihood of triggering autofocus at the current timestamp. A prediction network processes the input queues **S**, **V**, **D** and generates the trigger possibility $o_{pred} \in (0, 1)$, which represents a higher triggering necessity when it increases. If the score surpasses the user-defined trigger threshold $o_{thre} \in [0, 1]$, the autofocusing motor is enabled, and the microscope lens is directly displaced by the current defocus value $d$ along the optical axis. Through the selective triggering, for the defocused region where the user's attention targets, the lens takes a single step to focus on it. By contrast, for already in-focus or attention-dispersed circumstances, the lens is locked. Furthermore, to rapidly adapt to varied users, each user can increase the trigger threshold $o_{thre}$ when they feel the triggering is oversensitive or decrease it when they feel obviously delayed. Consequently, both in-time clear local observation and stable global viewing are guaranteed in an easily adapted manner.

### B. Defocus Estimation Network

*1) Data Collection and Labelling:* A practical labelling method for creating microscope defocus datasets is proposed. First, the microscope scans the object or scenario with a predefined step length $\delta$ and generates a stack of microimages $\{I_k \in \mathcal{R}^{h\times w\times 3} | k = 1, ..., l\}$, $k$ for image index and $l$ for the total image number. After that, leveraging the open-source focus-stack algorithm [20], the images are aligned to mitigate the off-axis shift of the lens during the scanning process. Subsequently, the most in-focus image index for each pixel is calculated, forming the in-focus index map $F \in \mathcal{R}^{h\times w}$. For each image $I_k$, the corresponding defocus map $M_k \in \mathcal{R}^{h\times w}$,

which reveals the axial offsets from the focus plane, can be yielded as

$$M_k(x, y) = (k - F(x, y)) \times \delta. \tag{3}$$

Following this pipeline, the observed scenario can be adjusted several times until sufficient stacks are collected. The image-label pairs generated from stacks can be used to train the single-image-based defocus estimation model, which directly predicts the physical defocus distance, thus guaranteeing the efficient one-shot autofocusing process.

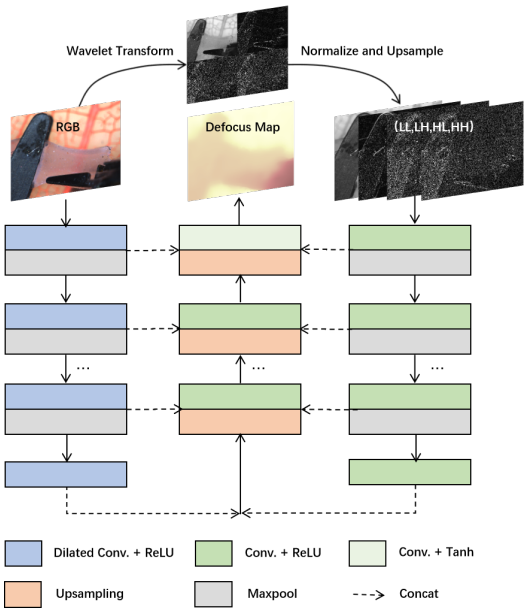


Fig. 3. The structure of the bi-directional defocus estimation network. Based on U-Net, the network uses dilated convolutions to encode the RGB image to provide an enlarged field of view. And the sharpness cues are extracted by wavelet transform and explicitly put into an additional encoding branch. A shared decoding branch and the skip connection mechanism are adopted. In this way, the network can clearly distinguish between near and far defocus and make more accurate predictions on texture-less regions.

*2) Network Structure:* A bi-directional defocus estimation network is proposed for one-shot autofocus, which features in accurately distinguishing the near and far defocus, and estimating defocus at the texture-less regions. Considering the crucial requirement of real-time inference, the light-weighted U-Net structure is adopted as the baseline network. As presented in Fig. 3, the improvement mainly focuses on two perspectives. First, the texture of the microsurgical scenario tends to be more sparse than natural scenarios, making the standard depth perception field rather confined and thus misleading the judgment of defocus direction. To mitigate this effect, the convolutions in the encoding stage are replaced by dilated convolutions [21] for an enlarged perception field. Additionally, sharpness cues are explicitly extracted and put into an additional encoding branch for facilitating defocus value estimation. This information provides straightforward guidance on the extent of bluriness, and it also marks the edges and corners of the original image, which are the confidential regions for estimating defocus. To achieve the extraction of sharpness, 2D discrete wavelet transform (DWT) is applied to the RGB image $I \in \mathcal{R}^{h \times w \times 3}$ to generate the approximation subband $LL \in \mathcal{R}^{\frac{h}{2} \times \frac{w}{2}}$, the horizontal detail subband $LH \in \mathcal{R}^{\frac{h}{2} \times \frac{w}{2}}$, the vertical detail subband $HL \in \mathcal{R}^{\frac{h}{2} \times \frac{w}{2}}$, and the diagonal detail subband $HH \in \mathcal{R}^{\frac{h}{2} \times \frac{w}{2}}$. After that, the four subbands are normalised and upsampled to the original size before they are put into the network. With RGB and sharpness cues prepared, the dual-branch U-Net encoder extracts their features respectively, followed by a feature concatenation unit. After feature aggregation, the shared decoder outputs the depth map, with additional information from the skip connections of the corresponding layers of both encoders. The loss function is a combination of Mean Squared Error (MSE) and Structure Similarity Index Measure (SSIM) loss, with a carefully adjusted ratio. Consequently, the proposed network boosts the defocus direction awareness and facilitates a more accurate defocus estimation, especially in the smooth areas.

### C. Trigger Prediction Network

*1) Data Collection and Labelling:* To capture the inherent features of gaze and defocus on different triggering necessities, four volunteers are recruited to conduct teleoperation microsurgery on our platform for 10 minutes each, with the eye tracker recording their gaze information. The GIPM and DEM modules run in real time, so that the queues **S**, **V**, **D** are stored. The users are provided with a foot pedal to trigger focus. In detail, once they feel the observed region out of focus, they push the pedal. In this way, the microscope focuses on the region by a one-time displacement. The state of the foot pedal is also recorded as the reference for triggering necessity. After data collection, the queues are paired with the pedal state at the same timestamp. For the moment of pressing the pedal, the queues are labelled as 1. Otherwise, for the time span between two pressing events, after downsampling timestamps to avoid data repeats and maintain a balance between classes, the queues at sampled timestamps are labelled as 0. Note that timestamps closer to any pressing event than a defined threshold are dropped, as they are ambiguous to label as 0 or 1. Following this pipeline, the input queues and output hard labels are obtained.

*2) Network Structure:* A long short-term memory (LSTM) network is adopted for triggering prediction since it features temporal memory and is lightweight. Given the queue length $l_q$, $\mathbf{S} \in \mathcal{R}^{l_q \times 2}$, $\mathbf{V} \in \mathcal{R}^{l_q \times 2}$, and $\mathbf{D} \in \mathcal{R}^{l_q \times 1}$ are concatenated on the last channel and put into the LSTM network with the input size 5 and hidden size $s_h$. The last hidden state is followed by a multi-layer perception (MLP) block to regress the possibility of triggering. For the loss function, due to the class imbalance, the focal loss [22] is leveraged to stress the precise prediction of positive samples. Let the output possibility $p$, the loss $\mathcal{L}_f$ is defined as

$$\begin{aligned} \mathcal{L}_f = &- \sum_{i=1}^{n_p} \alpha \times (1 - p^i)^\gamma \times \log(p^i) \\ &- \sum_{i=1}^{n_n} (1 - \alpha) \times (p^i)^\gamma \times \log(1 - p^i), \end{aligned} \tag{4}$$

where $n_p$ and $n_n$ are the numbers of positive and negative samples, $\alpha \in [0, 1]$ and $\gamma \in [0, 5]$ are factors determining the attention allocation of the loss. For emphasising positive samples, high $\alpha$ and $\gamma$ values are adopted. Subsequently, the model enables real-time triggering possibility prediction based on gaze and defocus. Without defining a triggering rule explicitly, the model learns about the inherent features of human attention and integrates these features into a single possibility value. In this way, the adjustment of triggering sensitivity is much more convenient by just resetting the possibility threshold, increasing the user-adaptability.

## III. Experiments and Results

### A. Implementation Details

*1) Hardware setup:* As displayed in Fig. 1, the system can be divided into three units, including the teleoperation unit, the eye tracking unit, and the autofocusing unit. For teleoperation, on the master side, a Geomagic Touch master hand sends the user-operated 6D pose of the surgical tool. On the follower side, a Physik Instrumente 6-axis micro-motor receives poses and repeats the trajectory. Its end is connected to a grasper, which can manipulate the blood vessel phantom flexibly. For eye tracking, a 7invensun A3 desktop eye tracker is placed at the bottom of the screen, where the real-time microsurgery image is displayed. After gaze calibration, the tracker can capture the PoR coordinate on the screen and the pupil diameters at 60 $Hz$. For autofocusing, a HIKROBOT MV-CH120-11UC microscope with 1× eyepiece and 2× objective lens is used to capture real-time video streams at $20Hz$. It is fixed on a ZengGuang motorised lifter, whose repeat accuracy is $0.01mm$ and maximum control frequency is $5Hz$. At the defocus stack collection process in II-B.1, the lifter step $\delta$ is set as $0.1mm$ and the total scanning distance is $15mm$.The inter-device communication is established via the Robot Operating System (ROS).

*2) Algorithm implementation:* At a workstation with an NVIDIA RTXTM A6000, the proposed framework is implemented under a Python and PyTorch environment. For inference, the framework runs at 20 $Hz$. The sampling window length $m$ of gaze formation is 20, the queue length $l_q$ for history data storage is 60, and the gaussian standard deviation $\sigma$ for heat map generation is 20. For training the defocus estimation model, a SGD optimiser is applied with the learning rate 5e-5, the momentum terms 0.9 and the weight decay 5e-4. The batch size is 8. The dilation rate of the RGB encoder is 4. Data augmentation methods including random flipping and cropping are conducted for increasing robustness. The weight ratio of MSE and SSIM loss is adjusted to 5:1. The Early Stop strategy is used with patience 20. For training the trigger prediction model, an Adam optimiser is leveraged with the learning rate 5e-4. The batch size is 32. The hidden size $s_h$ is 128. The loss parameter $\alpha$ and $\gamma$ are set as 0.75 and 2. The Early Stop strategy is used with patience 50.

*3) Dataset:* For defocus data, we collected 120 stacks, with 150 images in each stack. The image size is 480 × 640 × 3. The stacks are randomly divided into the train, validation, and test sets with a ratio of 8:1:1. For trigger data, we collected 1487 queue-label pairs with positive samples 165 and negative samples 1322. They are also split into the train, validation, and test sets with a ratio of 8:1:1 randomly.

### B. Evaluation on Defocus Estimation Network

TABLE I
Quantitative comparison results on defocus estimation

| Method | MAE($mm$) | Acc | Time($s$) |
|---|---|---|---|
| U-Net | 0.876 | 0.655 | **0.018** |
| Dilated U-Net | <u>0.793</u> | <u>0.689</u> | **0.018** |
| Ours | **0.671** | **0.754** | <u>0.033</u> |

The best result is in **bold**, and the second best result is <u>underlined</u>. The same annotations are used below.

*1) Evaluation metrics:* Three evaluation metrics are leveraged, including Mean Absolute Error (MAE), Accuracy (Acc), and Inference Time (Time). Given $n$ test images, MAE is defined as

$$MAE = \frac{1}{n \times h \times w} \sum_{i=1}^{n} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} |M_i(x, y) - D_i(x, y)|, \tag{5}$$

where $D_i$ and $M_i$ are the predicted and ground-truth defocus maps of the $i$-th image, respectively. Given that the tolerable defocus distance of the microscope is [-1.0$mm$, 0.8$mm$] investigated on 4 volunteers, Acc is defined as

$$Acc = \frac{1}{n \times h \times w} \times \sum_{i=1}^{n} \sum_{x=0}^{w-1} \sum_{y=0}^{h-1} l(-1.0 < M_i(x, y) - D_i(x, y) < 0.8), \tag{6}$$

where $l(\cdot)$ equals 1 if the input is True, otherwise 0. Finally, Time refers to the average inference time of a single frame.

*2) Results:* Our method is compared with two methods, the standard U-Net and the U-Net with dilated convolutions (Dilated U-net). In Fig. 4, the standard U-Net is prone to the wrong prediction of defocus direction, especially in peripheral regions, which is quite explainable since the limited perception field seldomly affects local bluriness estimation, but makes it hard to handle the relative depth between the region and its surroundings. The misdirection phenomenon largely threatens accurate autofocus, with the potential that heavier defocus happens after the reverse adjustment. Changing conventional convolutions to dilated ones effectively mitigates this phenomenon. Meanwhile, without the additional cues of sharpness, the prediction accuracy of the Dilated U-Net is still inferior to our method, suggesting the effect of introducing sharpness information. In Table I, compared with U-Net and Dilated U-Net, our method has a decreased MAE by 23.4% and 15.4%, respectively, and an increased Acc by 15.1% and 9.4%, respectively. Despite a higher cost of inference time, it still satisfies the real-time inference given the $20Hz$ framerate of image acquisition for our system.

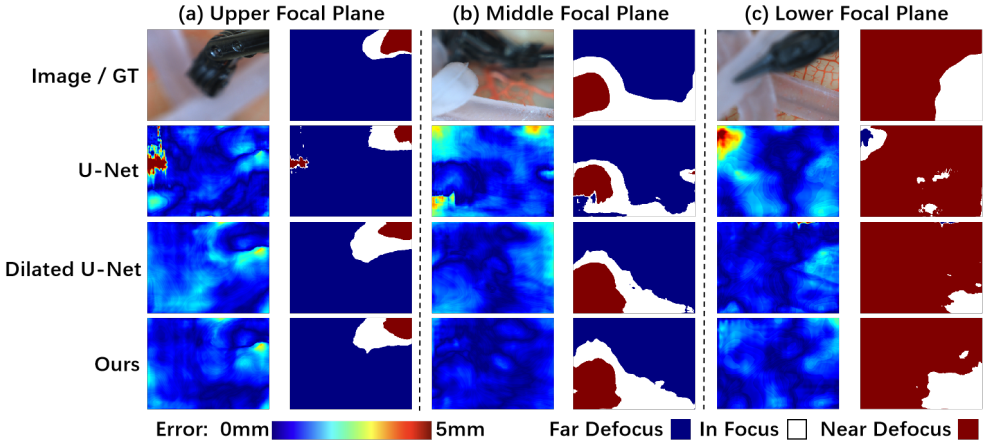


Fig. 4. Qualitative comparison results on defocus estimation. The evaluation is under three circumstances, including (a) the focal plane is at the top of the scenario, (b) the focal plane is in the middle of the scenario, and (c) the focal plane is at the bottom of the scenario. The first row presents the original images and the ground truth defocus direction maps. And the other three rows present the prediction error map and the predicted defocus direction map of each method. Far defocus: $< -1.0mm$. In focus: $[-1.0mm, 0.8mm]$. Near focus: $> 0.8mm$.

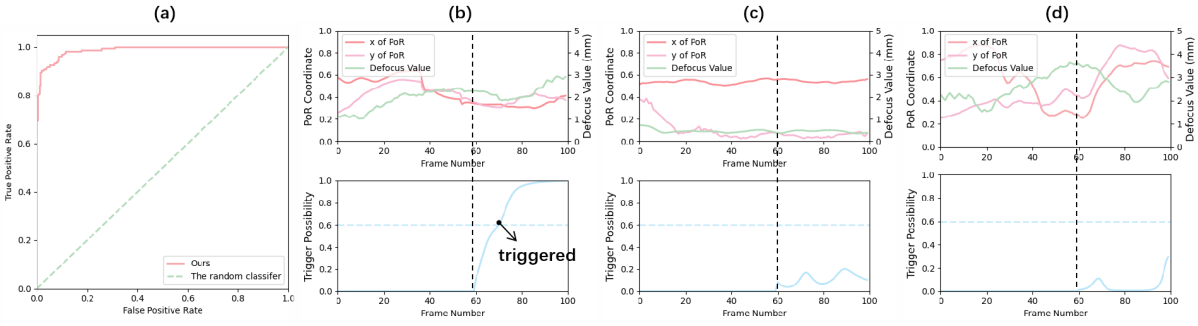


Fig. 5. Visualisation of triggering prediction network evaluation. (a) The ROC curve on the test dataset. (b)-(d) The variation of predicted trigger probability with real-time PoR and defocus value on an unseen user. The black dashed lines mean the start of prediction when the input queues are full. The blue dashed lines show the user-defined threshold. And the trigger activation is highlighted by a black dot.

### C. Evaluation on Trigger Prediction Network

The network is first evaluated on the test dataset. As presented in Fig. 5(a), the ROC curve is steep with an AUC of 0.981. At the optimal threshold by closest-to-(0,1) criteria, the network achieves a 92.1% true positive rate (TPR) and a 4.1% false positive rate (FPR). These results demonstrate the model's strong capability in distinguishing the scarce positive samples. Following that, for a user unseen during training who sets the trigger threshold as 0.6, we visualise the predicted trigger probability against real-time PoR and defocus values in Fig. 5 (b)-(d). In (b), with stable PoR and high defocus in the regarded area, the trigger probability consistently rises above the threshold, activating autofocus. In (c), despite the nearly fixed PoR, the defocus remains below 1 $mm$, suggesting the attention spot is in focus. The model correspondingly outputs a low probability, locking the autofocus. In (d), varying PoR reflects dispersed attention, leading again to low trigger probabilities and a fixed microscope state. Under the three circumstances, the network exhibits robust user adaptability.

### D. Evaluation on Framework

*1) Comparison methods:* We compare four focusing frameworks, including *Pedal*, *Gaze + Pedal*, *Rule-Based Gaze*, and *Ours*. *Pedal* refers to pressing a pair of pedals to adjust the focus plane up or down step by step. *Gaze + Pedal* means the user should press a pedal to trigger the lifter, and it is displaced by the defocus of PoR in one shot. In *Rule-Based Gaze* and *Ours* modes, the system automatically focuses on the defocused and regarded region in one shot, otherwise, it is locked. However, in terms of triggering prediction, in *Rule-Based Gaze*, an explicit rule is made, including the limitation of defocus value, staring time, and PoR variance. The users need to adjust these parameters to adapt. By contrast, *Ours* uses a model to output the trigger possibility, thus the users can directly adjust the possibility threshold to adapt. All of the adjustable parameters are displayed in the form of draggable sliders on the computer screen, which remain accessible throughout the surgical procedure.

*2) User study settings:* To evaluate the user performance and preference among different focus methods, 12 subjects aged from 22 to 28 (9 males, 3 females), all with a biomedical engieering background, were recruited to conduct

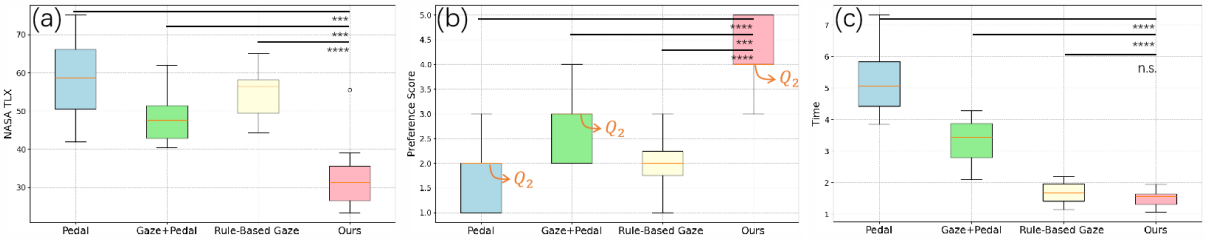


Fig. 6. Overall performance comparison among four focusing modes. The performances are evaluated by boxplots of (a) NASA TLX, (b) preference score, and (c) completion time. The pair-wise t-test was performed between Ours and each of the other three modes.

TABLE II

TRIGGERING AND FOCUSING PERFORMANCE COMPARISON BETWEEN RULE-BASED GAZE MODE AND OURS

| User | Rule-Based Gaze | | | | | Ours | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | F1 | F-Acc | Trial | Precision | Recall | F1 | F-Acc | Trial |
| User1 | 0.82 | 0.53 | 0.64 | 0.78 | 8 | 0.81 | 0.88 | 0.84 | 0.81 | 1 |
| User2 | 0.84 | 0.94 | 0.89 | 0.81 | 5 | 0.92 | 0.92 | 0.92 | 0.85 | 3 |
| User3 | 0.50 | 0.33 | 0.40 | 0.80 | 3 | 0.90 | 0.82 | 0.86 | 0.67 | 3 |
| User4 | 0.80 | 0.67 | 0.73 | 0.75 | 3 | 0.83 | 0.91 | 0.87 | 0.70 | 2 |
| User5 | 0.67 | 0.86 | 0.75 | 0.83 | 3 | 0.91 | 0.91 | 0.91 | 0.60 | 1 |
| User6 | 0.75 | 0.27 | 0.40 | 0.67 | 8 | 0.91 | 0.83 | 0.87 | 0.80 | 0 |
| User7 | 0.63 | 1.00 | 0.77 | 0.60 | 5 | 1.00 | 0.86 | 0.92 | 0.67 | 1 |
| User8 | 0.73 | 0.80 | 0.76 | 0.88 | 7 | 0.88 | 0.83 | 0.86 | 0.67 | 0 |
| User9 | 0.78 | 0.88 | 0.82 | 0.71 | 6 | 0.82 | 0.93 | 0.88 | 0.79 | 2 |
| User10 | 0.71 | 0.83 | 0.77 | 0.60 | 5 | 0.93 | 0.87 | 0.90 | 0.92 | 0 |
| User11 | 0.78 | 0.70 | 0.74 | 0.71 | 2 | 1.00 | 0.77 | 0.87 | 0.70 | 0 |
| User12 | 0.67 | 0.75 | 0.71 | 0.67 | 10 | 0.88 | 0.88 | 0.88 | 0.93 | 1 |
| Average | 0.72 | 0.71 | 0.70 | 0.73 | 5.42 | **0.90** | **0.87** | **0.88** | **0.76** | **1.17** |
| Diff | - | - | - | - | - | 0.18 *** | 0.16 * | 0.18 ** | 0.03 $^{n.s.}$ | -4.25 *** |

a user study. Before the study, all subjects were trained on how to use the teleoperation system, the gaze tracker, the assistive pedals, and the GUI of draggable slides. After the training, the users are required to complete four groups of teleoperation surgeries on the blood vessel phantom in a random order. The groups use the four focusing modes respectively, and other settings are identical. The experiment lasts for 5 minutes per group. The whole operation process during an experiment, including the surgical images, the PoRs, the triggering states, and the lifter heights at each timestamp, is recorded. During two experiments, a ten-minute break is set, during which the users need to fill in a NASA Task Load Index (NASA TLX) [23] form. After finishing the four experiments, the users are required to give a preference score ranging from 0 to 5 to each of the focusing techniques. The higher score represents a higher preference. Additionally, they are shown the experimental recordings of *Rule-Based Gaze* and *Ours* in a random order. And they are required to mark the three types of events, including successful trigger, wrong trigger, and missed trigger. Success trigger means the intended trigger is realised; wrong trigger means the unintended trigger happens; and missed trigger means the intended trigger is missed. They also need to point out whether the regarded region is focused, i.e., successful focus, on each success trigger. Lastly, an expert calculated the completion time of every focus event in the four experiments. The time starts when a defocused region is regarded, and ends when the lifter's displacement is completed.

*3) Evaluation metrics:* Three general metrics are used to evaluate the overall performance of each of the four focusing modes. The metrics include the NASA TLX score indicating the cognitive load during the task, the preference score, and the average focus completion time of the task. Another five metrics are introduced for evaluating the triggering and focusing performance in detail between *Rule-Based Gaze* and *Ours*. Given a user's successful trigger times $n_s$, wrong trigger times $n_w$, missed trigger times $n_m$, and successful focus times $n_f$ in one experiment, the Precision is defined as $\frac{n_s}{n_s+n_w}$; the Recall is defined as $\frac{n_s}{n_s+n_m}$; the F1 score is defined as $\frac{2\times Precision\times Recall}{Precision+Recall}$; the focus success rate $F-Acc$ is defined as $\frac{n_f}{n_s}$. Lastly, Trial means the times of parameter adjustment during adaptation to the system.

*4) Results:* In Fig. 6, *Ours* reduces the NASA TLX and increases the preference score compared to the other three modes by a large margin, suggesting its clear advantage in enhancing user experience and reducing operation load. Additionally, since no extra labour is needed to trigger the lifter, the focus time is also obviously shorter than *Pedal* and *Gaze + Pedal* modes. For *Rule-Based Gaze* mode, the users tend to spend much time adjusting parameters at the beginning, since it is challenging to optimize the three parameters simultaneously. For this reason, this mode only has a moderate performance on the NASA TLX and preference score. But the completion time has no significant

difference from *Ours*, showing that it is a time-saving method with low adaptability. *Gaze + Pedal* is a relatively well-received method among users due to its simplicity. Meanwhile, in the teleoperation, many other pedals are also used for controlling the manipulator. Thus, users need to switch their feet frequently. This explains the relatively long time cost of this mode. Finally, the *Pedal* mode receives the highest levels of stress and dissatisfaction. This mode costs lots of labour to press a pair of pedals step by step, and it is also the most time-consuming. Compared *Pedal* with *Gaze + Pedal*, the superiority of the proposed defocus estimation and one-shot focusing is clearly evident.

Furthermore, the triggering and focusing performance of *Ours* is explored in comparison to *Rule-Based Gaze* in Table II. *Ours* surpasses *Rule-Based Gaze* in Precision, Recall, and F1 by a large margin, and needs significantly fewer trials for parameter adjustment, indicating that it is a highly user-adaptable method. After fewer trials on possibility threshold optimisation, the trigger prediction model fits the user better. Additionally, the focus success rate exceeds 70% in practice in both modes, indicating that the system can meet the focusing requirement in one shot in most cases.

## IV. Conclusions

In conclusion, the inherent shallow depth-of-field of surgical microscopes makes autofocus essential in robot-assisted microsurgery (RAMS). Gaze-driven autofocus offers a natural alignment with the surgeon's visual attention and has emerged as a promising solution. However, current gaze-based methods often lack user adaptability in focus triggering. Furthermore, searching for the optimal focus via the widely used hill-climbing strategy is less compatible with the highly dynamic RAMS.

To address these limitations, we developed a gaze-controlled autofocus framework that incorporates user-adaptive triggering and one-shot focusing. A triggering prediction network is proposed leveraging gaze and defocus information. The predicted possibility grows as the user pays attention to a defocused region. Only when the possibility exceeds the user-defined threshold can autofocus be triggered. Thus, the trigger sensitivity can adapt among different users. Additionally, a bi-directional defocus estimation network is introduced, which can predict the absolute defocus distance map via a single image, enabling the rapid one-shot focusing. With sharpness cues introduced and the perception field enlarged, the network improves prediction accuracy by 15.1% and reduces the prediction error by 23.4% compared to the baseline. A user study with 12 subjects indicates that our autofocus system needs the least workload and completion time compared to the other three common prior focus systems, including manual, half-automated, and automated with a rule-based trigger. To adapt to individual users, our framework only needs an average of 1.17 times of adjustment, far less than 5.42 times by the rule-based method. The average triggering precision and recall are over 85%, and the success rate of one-shot focusing is over 70% in practice, suggesting the robustness of our system. Future work will focus on improving the accuracy of defocus estimation and evaluating system performance in more complicated RAMS tasks.

## References


[1] T. J. van Mulken, R. M. Schols, A. M. Scharmga *et al.*, "First-in-human robotic supermicrosurgery using a dedicated microsurgical robot for treating breast cancer-related lymphedema: a randomized pilot trial," *Nat. Commun.*, vol. 11, no. 1, p. 757, 2020.

[2] D. Zhang, W. Si, W. Fan *et al.*, "From teleoperation to autonomous robot-assisted microsurgery: A survey," *Mach. Intell. Res.*, vol. 19, no. 4, pp. 288–306, 2022.

[3] G.-Z. Yang, J. Cambias, K. Cleary *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," *Sci. Robot.*, vol. 2, no. 4, p. eaam8638, 2017.

[4] J. W. Kim, J.-T. Chen, P. Hansen *et al.*, "Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Sci. Robot.*, vol. 10, no. 104, p. eadt5254, 2025.

[5] Y. Long, A. Lin, D. H. C. Kwok *et al.*, "Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery," *Sci. Robot.*, vol. 10, no. 104, p. eadt3093, 2025.

[6] Y. Liu, Y. Luo, Y. Luan *et al.*, "Towards accurate brain electrode implantation via cross-modality fusion of white-light and photoacoustic microscopy," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2025, pp. 7726–7732.

[7] Y. An, J. Yang, B. He *et al.*, "A microscopic vision-based robotic system for floating electrode assembly," *IEEE/ASME Trans. Mechatron.*, vol. 29, no. 5, pp. 3810–3820, 2024.

[8] L. Ma and B. Fei, "Comprehensive review of surgical microscopes: technology development and medical applications," *J. Biomed. Opt.*, vol. 26, no. 1, pp. 010 901–010 901, 2021.

[9] S. K. Tedla, S. MacKenzie, and M. Brown, "Looktofocus: Image focus via eye tracking," in *Proc. Symp. Eye Track. Res. Appl.*, 2024, pp. 1–7.

[10] Y. Bi, Y. Su, N. Navab *et al.*, "Gaze-guided robotic vascular ultrasound leveraging human intention estimation," *IEEE Robot. Autom. Lett.*, 2025.

[11] J. Zhang, B. Wang, Z. Pan *et al.*, "Gazescope: A framework of gaze attention-based automatic field-of-view adjustment for laparoscopic robots," *IEEE Robot. Autom. Lett.*, 2025.

[12] K. Fujii, G. Gras, A. Salerno *et al.*, "Gaze gesture based human robot interaction for laparoscopic surgery," *Med. Image Anal.*, vol. 44, pp. 196–214, 2018.

[13] Y. Cao, S. Miura, Y. Kobayashi *et al.*, "Pupil variation applied to the eye tracking control of an endoscopic manipulator," *IEEE Robot. Autom. Lett.*, vol. 1, no. 1, pp. 531–538, 2016.

[14] C. Zhang, Y. Gu, J. Yang *et al.*, "Diversity-aware label distribution learning for microscopy auto focusing," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 1942–1949, 2021.

[15] T. Albuquerque, A. Moreira, and J. S. Cardoso, "Deep ordinal focus assessment for whole slide images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 657–663.

[16] L. Yang, B. Kang, Z. Huang *et al.*, "Depth anything v2," *Proc. Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 21 875–21 911, 2024.

[17] Y. Luan, Y. Luo, Y. Liu *et al.*, "Autofocusing with 3-d tracking for robot-assisted microsurgery," *IEEE/ASME Trans. Mechatron.*, 2025.

[18] C. Herrmann, R. S. Bowen, N. Wadhwa *et al.*, "Learning to autofocus," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2020, pp. 2230–2239.

[19] A. Abuolaim, A. Punnappurath, and M. S. Brown, "Revisiting autofocus for smartphone cameras," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 523–537.

[20] B. Forster, D. Van De Ville, J. Berent *et al.*, "Complex wavelets for extended depth-of-field: A new method for the fusion of multichannel microscopy images," *Microsc. Res. Tech.*, vol. 65, no. 1-2, pp. 33–42, 2004.

[21] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 472–480.

[22] T.-Y. Lin, P. Goyal, R. Girshick *et al.*, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[23] S. G. Hart and L. E. Staveland, "Development of nasa-tlx (task load index): Results of empirical and theoretical research," in *Adv. Psychol.* Elsevier, 1988, vol. 52, pp. 139–183.