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

Cross-view Exocentric and Egocentric Fusion for Robust Microsurgical Anastomosis Understanding

Yuxuan Liu, Yuyang Zhuge, Xinyao Zhou, Yating Luo, Yunfei Luan, Yao Guo, Guang-Zhong Yang

***Abstract*— Microsurgical anastomosis has become increasingly prevalent in surgical autonomy, requiring accurate and stable control of suturing needles and threads while enhancing the efficiency and safety of microsurgical operations. However, current systems predominantly employ top-down-view microscopes for intraoperative imaging, which are constrained by limited field-of-view and significant occlusion caused by instrument-tissue interactions. To address these challenges, we develop a dual-view vision system for microsurgical anastomosis, integrating both conventional top-down-view microscopes and eye-in-hand cameras mounted on surgical instrument tips. Our approach involves cross-view feature fusion through different schemes to improve microsurgical scene understanding, including surgical action recognition, gripper-object interaction prediction, and instrument pose estimation. Extensive anastomosis datasets are collected on our robotic platform and several experiments are conducted for detailed evaluation of the system performance. Quantitative and qualitative results demonstrate that our dual-view microsurgical system significantly outperforms single-view microscopes in terms of robust visual perception, and cross-view feature fusion improves both the accuracy and precision of anastomosis scene understanding.**

I. Introduction

Microsurgical robotics has undergone transformative advancements in minimally invasive surgery, enabling precise surgical control and submillimeter-scale operations through the integration of advanced robotic systems and high-resolution cross-modality visual perception [1]–[3]. Recent studies have focused on developing accurate and safe microsurgical robotic platforms, which have been successfully applied in surgeries such as flexible neural electrode implantation [4], [5], retinal vein cannulation [6], and intracranial virus injection [7]. Among these applications, microsurgical anastomosis represents one of the most difficult surgeries and is widely utilized in clinical practice, particularly in transplantation and flap transfer [8], [9]. This technique requires the precise reconnection of broken vessels involving dexterous manipulation of microvasculature and suturing instruments. Robot-assisted microsurgical anastomosis improves the efficiency and safety with sub-millimeter precision, demonstrating significant potential for clinical adoption.

For microsurgical robotics, exocentric (top-down-view) microscopes are commonly employed for visual perception [10], [11]. For instance, Nomura et al. [7] positioned the microscope above the surgical plane to capture high-resolution color images for vessel segmentation and intraoperative guidance. However, there exist several inherent challenges for the conventional top-down-view microscope systems. On the one hand, the field of view (FOV) is significantly restricted due to submillimeter-scale surgical manipulations [12]. The microscope is focused at a small area under high magnification, making the operator difficult to get a global scan of the surgical scene. Although existing works have attempted to expand the FOV using stereo or multiple microscopes, they primarily derive depth information while retaining the same exocentric perspective. On the other hand, frequent intraoperative instrument-tissue interactions induce severe self-occlusion in the top-down view. Self-occlusion makes the robot system unable to accurately understand scenarios and characterize interactions between instruments and tissues, leading to a decrease of surgical accuracy and safety.

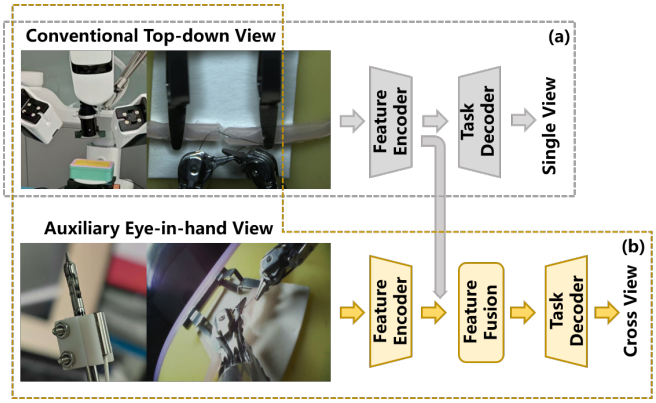

Fig. 1. Fusion of exocentric and egocentric images is designed for superior and robust performance on different tasks of microsurgical anastomosis understanding. (a) Conventional top-down-view microscope is implemented for microsurgical robots to capture global images. (b) Our established composite dual-view vision system with auxiliary eye-in-hand egocentric cameras mounted at the end of the instruments to capture detailed instrument-object interactions.

The integration of egocentric (eye-in-hand) cameras has emerged as a promising approach to enhance the accuracy and generalization capability of visual perception not only in natural scenarios [13]–[15] but also in surgical scenes [16]–[18]. The eye-in-hand camera is typically placed at the end of surgical instruments to provide a more detailed perception of instrument-object interactions, complementing the traditional exocentric microscope view. By using the images simultaneously captured by exocentric and egocentric cameras, cross-view fusion has been demonstrated to be

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

Y. Liu, Y. Zhuge, X. Zhou, Y. Luo, Y. Luan, Y. Guo and G.-Z. Yang are with the Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Institute of Medical Robotics, Shanghai Jiao Tong University, Shanghai, China. ({200009051yx, yao.guo, gzyang}@sjtu.edu.cn).

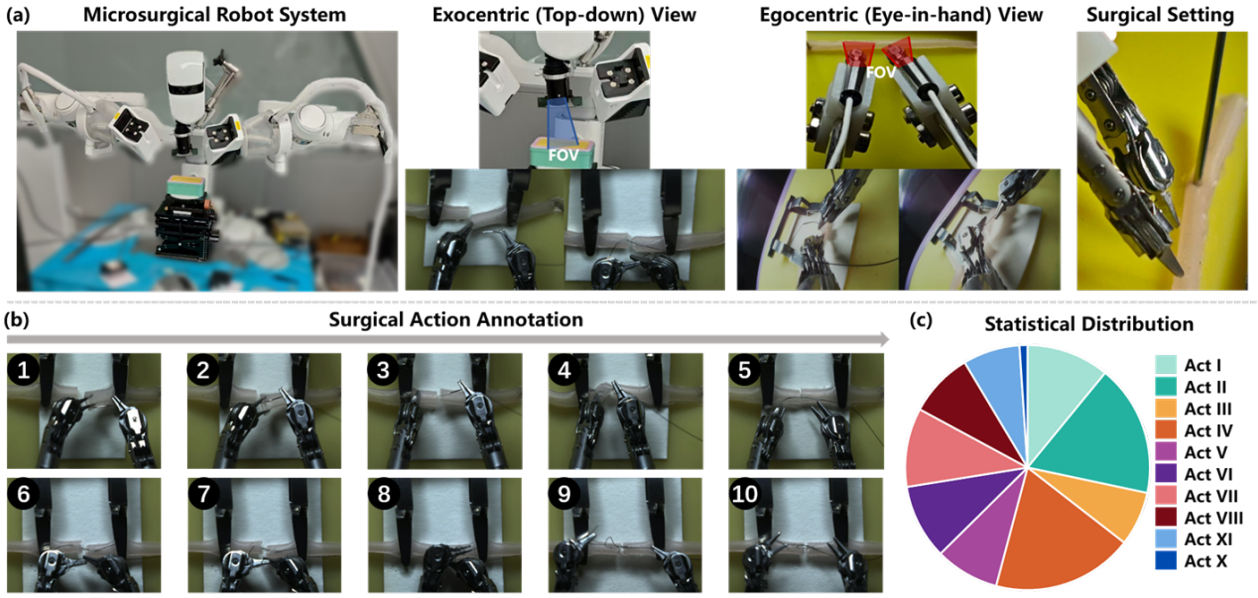

Fig. 2. Overview of our established microsurgical anastomosis robotic platform for data collection. (a) Our robot system contains exocentric (top-down-view) microscopes and egocentric (eye-in-hand) cameras placed at the end of surgical instruments. (b) The definition of the entire process of microsurgical anastomosis, including the process of suture making from action I to V and the process of knotting from action VI to X. (c) Statistical distribution of ten actions on our collected microsurgical dataset.

effective for surgical automatic operations [19]. For example, Kim et al. [17] used four cameras for surgical imitation learning, which includes two cameras on the top of the surgical plane and the other two fixed close to the grippers. However, existing works primarily apply them for end-to-end surgical task execution to merely validate the improvement of the success rate for specific operations, rather than systematically exploring the impact of dual-view systems on surgical scene understanding. The interpretability of the end-to-end framework is poor and can not show the superiority of the combination of two viewpoints. A comprehensive discussion and analysis of the dual-view system is needed to illustrate the benefits of cross-view fusion in microsurgery.

To address the aforementioned problems and challenges, in this work, we establish a dual-view vision system for microsurgical anastomosis and design networks for cross-view exocentric and egocentric (exo-ego) fusion for scene understanding. As shown in Fig. 1, the system integrates the conventional exocentric microscope with eye-in-hand egocentric cameras mounted at the end of surgical instruments, enabling submillimeter-scale precise anastomosis while providing complementary visual perceptions. To demonstrate the effectiveness of the dual-view vision system, we have designed networks to conduct cross-view fusion and incorporated it for surgical understanding, including surgical action recognition, gripper-object interaction prediction, and instrument pose estimation. Two distinct fusion schemes are implemented and discussed, i.e., frame-wise 2D and temporal 3D fusion, to systematically evaluate the benefits of incorporating egocentric visual features. Extensive experimental validation is conducted using a comprehensive dataset of intraoperative anastomosis procedures collected by our established robotic platform. Quantitative and qualitative results are reported and visualized to provide the evidence for superiority of dual-view systems in microsurgical anastomosis while offering insights into cross-view fusion strategies for surgical scene understanding.

The main contributions of this paper are as follows:

- We propose a dual-view vision system integrating both exocentric and egocentric perspectives to capture microsurgical robotic actions and instrument-tissue interactions, significantly advancing microsurgical scene understanding capabilities.
- We collect an extensive intraoperative dataset for microsurgical anastomosis under various surgical scenarios with annotated labels for multiple downstream tasks. The dataset will be made available upon request.
- Detailed experiments on our established robotic system are systematically validated to demonstrate the effectiveness of cross-view fusion compared with the conventional top-down-view microscope for robust microsurgical vision perception.

II. System Overview and Data Collection

To prove the effectiveness of our proposed framework for real surgical scenarios, we have established a microsurgical robot system for microsurgical anastomosis. The overview of our robot system is shown in Fig. 2-(a), which consists of three parts. The first part is the dual-arm robotic system with well-designed instruments for anastomosis, which can accurately and stably hold suturing needles and threads. The second part is the vision system. A high-resolution microscope is placed above the surgical plane to provide the top-down-view perception. Egocentric cameras are placed at the end of the instruments to deal with self-occlusions and perform exocentric-egocentric cross-view fusion. The last part is the surgical settings for anastomosis. We cut a

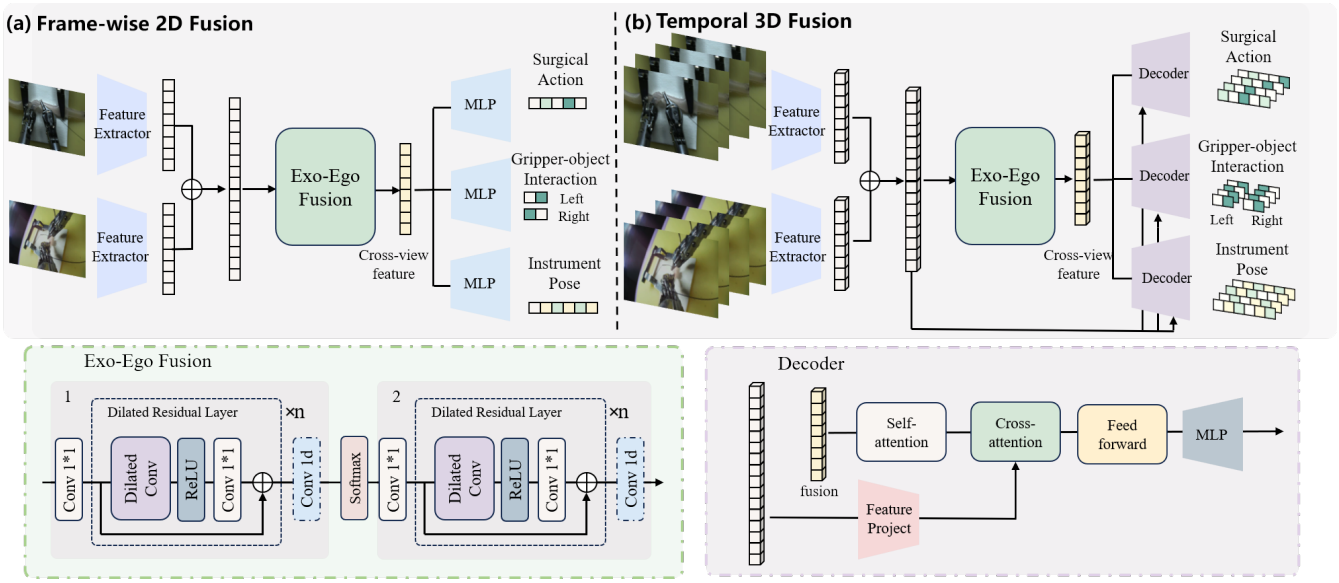

Fig. 3. Illustration of the cross-view exo-ego fusion scheme and three downstream tasks are evaluated for microsurgical anastomosis understanding. (a) Frame-wise 2D fusion is implemented with feature maps extracted by a single frame and no temporal constraints are used. (b) Temporal 3D fusion is designed by concatenating all feature maps derived by a sequence of images. The exo-ego fusion block is incorporated based on extracted features from both global and local images.

complete vessel into two separate parts to simulate intraoperative suturing operations. The suturing needle and thread are preoperatively assembled and placed on the grippers.

In Fig. 2-(b), the entire process of microsurgical anastomosis contains two procedures following the previous work [20]. The first process is to make a suture, which includes five sub-procedures. The operator inserts the left instrument inside the vessel and holds the suturing needle with the right instrument (Act-I), and passes the needle through the right part of the vessel (Act-II). Then the left instrument is extracted and pressed onto the other side (Act-III), and the operator passes the needle through the left part of the vessel (Act-IV). Finally, the operator resets the instrument and finishes the suture making (Act-V). The second process is knotting, which also includes five sub-procedures. The operator grabs the thread with the right instrument and conducts circular rotations around the left gripper twice (Act-VI & Act-VII). After that, the operator uses the left instrument to grab the end of the thread (Act-VIII) and withdraw it through the circular gap to tie a knot (Act-IX). Finally, the left and right instruments are reset and release the thread to complete the knotting (Act-X). Despite different surgical actions, we have also focused on the gripper-object interactions and instrument pose. The interactions of the grippers indicate whether the left and right grippers reach the target object, like suturing needles and threads. Consequently, the interaction state contains four elements (Left-0 & Left-1, Right-0 & Right-1), where 0 represents the closed state of the grippers and 1 represents the open state. The pose of instruments refers to six degrees of freedom for translation and rotation.

During the process of data collection, we have asked two operators to perform the entire process of anastomosis. Both two operators are well-trained on our established platform and can perform the entire process within four minutes. We collect exocentric images from the top microscope and the egocentric images from cameras placed on the instrument, and we have synchronized all the data. When needles and threads shed from the instruments, the operator manually moves them into the field of view and cuts the video part. Kinematic data is also recorded simultaneously when performing the surgery, and it contains the information for the state and pose of the grippers. In this work, surgical actions include ten different types and are annotated by operators watching the replay of videos. The distribution of ten classes is visualized in Fig. 2-(c), which represents the corresponding duration of different actions.

III. Methodology

A. Problem Statement

To demonstrate the effectiveness of our dual-view vision system for microsurgical anastomosis, we propose to implement exo-ego cross-view fusion on three tasks of scene understanding for evaluation with a multitask learning scheme, i.e., action recognition, interaction prediction, and pose estimation. Before introducing the detailed method for multitask learning, we define the problem and related notations. Let $\{(I^t_{exo}, I^t_{ego})|1 \leq t \leq T\} \in \mathcal{R}^{H\times W\times 3}$ represent the surgical image sequences with T frames collected by our established vision system, where H and W refer to the height and width of each image, respectively. The network $\mathcal{H}_\theta(\cdot)$ is designed to perform simultaneous estimation of surgical action A, interaction state S, and instrument pose P via a supervised multi-task learning scheme with parameters θ. In this work, the surgical action consists of ten sequential types during suture making and knotting, noted as $A = \{a^t\} \in \{0, 1, ..., 9\}$. The interaction state includes two terms indicating whether the left and right grippers reach the surgical objects, noted as $S = \{s^t_l, s^t_r\} \in \{0, 1\}$. The pose of the left and right instruments is represented as three rotation angles $P = \{p^t_l, p^t_r\}$, and it should be noted that we ignore

the translation terms due to cross-view variance and only preserve the rotation terms. The optimization problem can be described as:

$$\mathcal{H}_{\theta^*} = \arg\min_{\theta} \mathcal{L}(I_{exo}, I_{ego}, A, S, P),$$

where $\mathcal{L}(\cdot)$ represents the multi-task loss function during the training stage.

B. Network Design

The illustration of the multi-task learning network is visualized in Fig. 3, which consists of two parts, including the feature extraction process and cross-view feature fusion for surgical understanding. Three task-specific decoders are designed to perform surgical action, interaction state, and pose estimation with shared exocentric and egocentric fused features. We design two types of networks for evaluation based on different fusion dimensions, i.e., frame-wise 2D fusion and temporal 3D fusion. Experiments are conducted based on these two types for a comprehensive analysis.

1) Frame-wise 2D exo-ego Fusion: The network architecture of frame-wise 2D exo-ego fusion type is illustrated in Fig. 3-(a). Given the input exocentric and egocentric images, two feature encoders $\{\mathcal{H}^{enc}_{exo}(\cdot), \mathcal{H}^{enc}_{ego}(\cdot)\}$ are used to obtain exocentric and egocentric features of size 2048, respectively. The feature extractor is built upon ResNet50 without the last linear layer, and the two networks share the same architecture but different weights. The input of the exo-ego fusion block $\mathcal{H}^{f}(\cdot)$ is the concatenation of the extracted frame-wise exocentric and egocentric features, and several dilated residual layers are implemented to conduct cross-view fusion. Three task-specific decoders $\{\mathcal{H}^{dec}_{act}(\cdot), \mathcal{H}^{dec}_{sta}(\cdot), \mathcal{H}^{dec}_{pos}(\cdot)\}$ are designed with fully connected layers to predict surgical action, state, and pose with 10, 4, and 6 parameters for each task. Since the prediction is based on frame-wise image without temporal correlations, it is defined as a 2D exo-ego fusion scheme.

2) Temporal 3D exo-ego Fusion: Different from frame-wise fusion scheme, the second type incorporates temporal cues for feature fusion as shown in Fig. 3-(b). Given the image sequence with T frames, we first use ResNet50 to extract frame-wise features for both viewpoints of size $T\times2048$. It should be noted that the frame-wise feature extractor shares the same architecture as mentioned before. Different from 2D feature fusion, we implement temporal dilated residual networks for 3D exo-ego feature fusion and the size of fused feature maps is $T\times32$. Similarly, three task-specific decoders are implemented to predict results of size $T\times10$, $T\times4$ and $T\times6$ parameters. The difference of frame-wise and temporal 3D exo-ego fusion is the input feature maps, where the former type uses single-frame exocentric and egocentric feature and the latter incorporates sequence-level feature maps. With these two types of exo-ego fusion scheme, we can efficiently demonstrate the effectiveness of our proposed exo-ego vision system compared with a conventional monocular microscope.

C. Training Details

Given the input exocentric image I^t_{exo} and egocentric image I^t_{ego}, the frame-wise feature maps are acquired by feature encoders and concatenated for the downstream tasks:

$$f^t_{exo} = \mathcal{H}^{enc}_{exo}(I^t_{exo}),\ f^t_{ego} = \mathcal{H}^{enc}_{ego}(I^t_{ego}),$$
$$f^t = \mathcal{H}^{f}(f^t_{exo} \oplus f^t_{ego}).$$

When training the frame-wise exo-ego feature fusion network, for surgical action recognition, our designed action branch predicts the probability for each action and the cross-entropy loss function is utilized for classification with ten classes:

$$\mathcal{L}_{act} = \mathrm{CE}(a^t, a^t_{gt}),\ a^t = \mathcal{H}^{dec}_{act}(f^t).$$

For surgical state recognition, we also use cross-entropy loss function for both left and right grippers to determine whether they reach the target object with two classes for each side:

$$\mathcal{L}_{sta} = \mathrm{CE}(s^t_l, s^t_{l,gt}),\ s^t_l = \mathcal{H}^{dec}_{sta}(f^t).$$

For instrument pose estimation, we implement L1 distance loss function for the predicted rotation angles and ground truth for both left and right instruments:

$$\mathcal{L}_{pos} = 0.5 \times (||p^t_l - p^t_{l,gt}||_1 + ||p^t_r - p^t_{r,gt}||_1).$$

The training loss function $\mathcal{L}$ is calculated as the weighted sum of three terms:

$$\mathcal{L}^{2D} = \omega_{act}\mathcal{L}_{act} + \omega_{sta}\mathcal{L}_{sta} + \omega_{pos}\mathcal{L}_{pos},$$

where $\{\omega_{act}, \omega_{sta}, \omega_{pos}\}$ represent the coefficients for different downstream tasks and are selected by parameter ablation studies for the best performance.

Different from the frame-wise exo-ego feature fusion method, temporal 3D fusion takes the feature sequence as input and implements a temporal convolutional network to obtain fused features:

$$f = \{f^t\}^T_{t=0} = \mathcal{H}^{f}(\{f^t_{exo} \oplus f^t_{ego}\}^T_{t=1}).$$

For three tasks, the loss function is calculated based on the average of frame-wise supervision:

$$\mathcal{L}^{3D} = \frac{\sum^T_{t=0} \mathcal{L}^{2D}(f, a^t_{gt}, s^t_{gt}, p^t_{gt})}{T}.$$

IV. Experiment

A. Data Preprocessing

The data used in this work is collected by our established microsurgical robot system with both exocentric and egocentric cameras, where the exocentric microscope is placed with a top viewpoint and the egocentric camera is fixed at the end of the instrument. As mentioned before, we have collected images for two different procedures of microsurgical anastomosis, including suture making and knotting. About 80 video sequences are collected for suture making and 60 sequences are collected for knotting, which constitute the entire dataset for training and evaluation with about 66k frames. We have

TABLE I

COMPARISON RESULTS BY FRAME-WISE 2D EXO-EGO FUSION SCHEME WITH DIFFERENT METHODS

Data Amount	Approaches	Action (%)				Interaction (%)				Pose	
		Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1	MAE(°)	<10°(%)
η=1.0	Ours	**91.74**	**89.11**	**92.22**	**90.06**	**96.96**	**95.30**	**95.93**	**95.61**	**8.27**	**74.56**
	Exo	90.92	88.19	90.81	89.11	95.93	93.59	94.63	94.10	8.61	71.34
	Ego	90.03	87.37	90.30	88.11	96.10	94.13	94.63	94.38	8.37	73.94
η=0.5	Ours	**87.89**	**85.10**	**87.68**	**85.63**	96.77	94.31	96.32	95.27	**9.16**	**66.20**
	Exo	83.28	79.56	78.72	78.04	**97.23**	**95.68**	**96.34**	**96.00**	9.40	65.37
	Ego	85.02	78.28	85.90	78.62	94.26	90.62	92.63	91.57	9.07	66.96
η=0.25	Ours	75.24	**72.07**	**73.28**	66.20	92.32	87.07	90.30	88.54	10.18	56.30
	Exo	**77.08**	70.57	70.78	**67.07**	**94.54**	**91.61**	**92.58**	**92.09**	**10.02**	**57.55**
	Ego	75.47	67.91	72.19	65.93	90.41	82.36	88.99	85.04	10.40	55.25

TABLE II

COMPARISON RESULTS BY TEMPORAL 3D EXO-EGO FUSION SCHEME WITH DIFFERENT METHODS

Data Amount	Approaches	Action (%)				Interaction (%)				Pose	
		Accuracy	Recall	Precision	F1	Accuracy	Recall	Precision	F1	MAE (°)	<10°(%)
η=1.0	Ours	**93.76**	**91.30**	**93.96**	**92.35**	**98.66**	**98.01**	**98.15**	**98.08**	**6.38**	91.55
	Exo	92.09	90.73	89.86	90.08	98.32	97.60	97.58	97.59	6.97	89.28
	Ego	91.57	90.16	90.72	90.34	97.83	97.25	96.59	96.92	6.39	**91.66**
η=0.5	Ours	**93.02**	**90.34**	**92.70**	**91.23**	**97.89**	96.60	**97.33**	**96.96**	**6.87**	**88.78**
	Exo	90.52	88.35	91.03	89.43	97.83	**97.29**	96.55	96.92	7.02	87.26
	Ego	90.06	86.49	89.53	87.43	96.16	93.34	95.50	94.37	7.00	87.99
η=0.25	Ours	**87.76**	**86.09**	**88.30**	**86.72**	96.98	**96.23**	95.23	95.72	**7.26**	**83.88**
	Exo	85.96	82.82	87.81	84.22	**97.08**	96.05	**95.61**	**95.83**	7.47	81.76
	Ego	84.72	81.36	83.51	82.00	94.70	90.89	93.62	92.16	7.61	82.13

asigned 70 sequences with about 33k frames for training different networks and the left 70 sequences for evaluation. The ground truth labels for surgical action are annotated by experts who conduct the data collection of the robot system. Gripper-object interaction state indicates the contact of instruments and objects and is derived by the sensor placed on the gripper, whose value more than the threshold of 0.05 refers to the positive contact. The kinematics of the surgical robot are obtained to compose the pose of instruments.

B. Implementation Details

The proposed method is implemented by PyTorch and is trained on NVIDIA TITAN RTX 3090 GPU with 24 GB memory. Deep networks are all optimized by Adam optimizer and 20 epochs are used for the training stage with the learning rate set as 1e-4. All images are downsampled to 320×240 and normalized to [0,1] by dividing the maximum value. During the training process of the temporal exo-ego fusion network, we use the pretrained encoder for frame-wise feature extraction to make the training process converge fast by freezing the parameters to directly extract feature maps, and we only optimize the weights of temporal fusion decoders. The coefficients in loss functions are set as $\omega_{act} = 1.0$, $\omega_{sta} = 1.0$ and $\omega_{pos} = 10.0$.

C. Evaluation Metrics

For the task of surgical action and state recognition, we report four metrics commonly used for classification, i.e., accuracy(↑), F1 score(↑), recall(↑) and precision(↑). It should be noted that the metrics for interaction recognition are calculated by the average of left and right grippers. For the pose estimation, we report Mean Absolute Error (MAE↓) of three rotation angles and the percentage error lower than 10 degrees (<10°(%)↑). All metrics are calculated between the predicted results and the ground truth.

V. COMPARISON RESULTS AND DISCUSSION

In the subsequent sections, without further clarify, the **bold** value in the table refer to the best performance for each experiment, respectively. The coefficient η represents the used images for training according to the predefined training split. For example, $\eta = 0.5$ refers to half of the amount of training data. We compare our proposed cross-view fusion scheme noted as ours with each individual viewpoint noted as exo and ego, which only feeds the exocentric or the egocentric images as input to our network for the fair comparison. The quantitative and qualitative results are evaluated on three tasks, i.e., surgical action recognition (Action), gripper-object interaction recognition (Interaction) and instrument pose estimation (Pose).

A. Quantitative Results

The quantitative results of the frame-wise exo-ego fusion scheme compared with two single viewpoints are reported in Table I for three downstream tasks. Four metrics are reported for the classification tasks and two metrics are reported for pose regression, and experiments with different amounts of training data are conducted. It can be found that our method achieves the best performance on all metrics when using the entire training split. Compared with the conventional

exocentric viewpoints, our method achieves the improvement of 1.1% and 1.6% on F1 score for action and interaction recognition, and the decrease of 4.1% for average pose error. With half of the training data, our proposed method achieves the best performance on surgical action recognition and pose estimation, but with poor performance compared with single exocentric viewpoint. A significant improvement of 9.7% on F1 score and decrease of 2.6% on MAE is achieved compared with single viewpoint on action recognition and pose estimation. When the network is trained with one fourth data, exo-ego fusion scheme does not achieve superior performance compare with single exocentric or egocentric viewpoint. It means that the frame-wise fusion scheme has more robust performance with respect to larger amount of training data making the fusion network completely convergence and avoid overfitting, which reflects the constraints of fusion scheme only based on single image without temporal cues.

The results of the temporal 3D fusion scheme are reported in Table II with the same metrics mentioned above. It can be found that with one hundred percent of training data, our method achieves the best performance on all metrics except the percentage lower than 10 degrees. For this metric, our proposed method achieves the similar performance with egocentric view and much better performance than that of exocentric view. Compared with exocentric network, our method achieves the improvement of 2.5% and 2.6% on F1 score and $<10°(\%)$ for action recognition and pose estimation, respectively. With half of the training data, our method also achieves the best performance on all metrics except the recall for state recognition. A significant improvement of 2.0% and 1.7% on F1 score and $<10°(\%)$ is gained by cross-view fusion compared with single exocentric viewpoint. When the training data is decreased to 25%, our method still achieves the best overall performance except three metrics. From the above results, it can be found that adding temporal information greatly contributes to better performance compared with frame-wise 2D feature fusion. Different from the frame-wise exo-ego fusion scheme, temporal fusion makes our method relatively more robust to the data amount. With a small amount of training data, the temporal 3D exo-ego scheme still benefits from cross-view feature fusion.

Statistical visualization of quantitative results are shown in Fig. 4 for surgical action and interaction recognition, and Fig. 5 shows results for instrument pose estimation. In Fig. 4-(a) and Fig. 4-(b), it visualizes the comparison between six approaches on surgical action recognition and gripper-object interaction prediction, respectively. The three rows refers to $\eta = 1.0$, $\eta = 0.5$, and $\eta = 0.25$. It can be found that the temporal 3D fusion performs significantly better than frame-wise 2D feature fusion. In Fig. 5, it shows the distribution of angle errors of instrument pose estimation with the left on X axis and the right on Y axis. The upper part shows the cases for temporal 3D exo-ego fusion and the lower part shows the frame-wise cases. Each column represents different amount of training data η similar to that mentioned above. The red line visualizes the angular error within ten

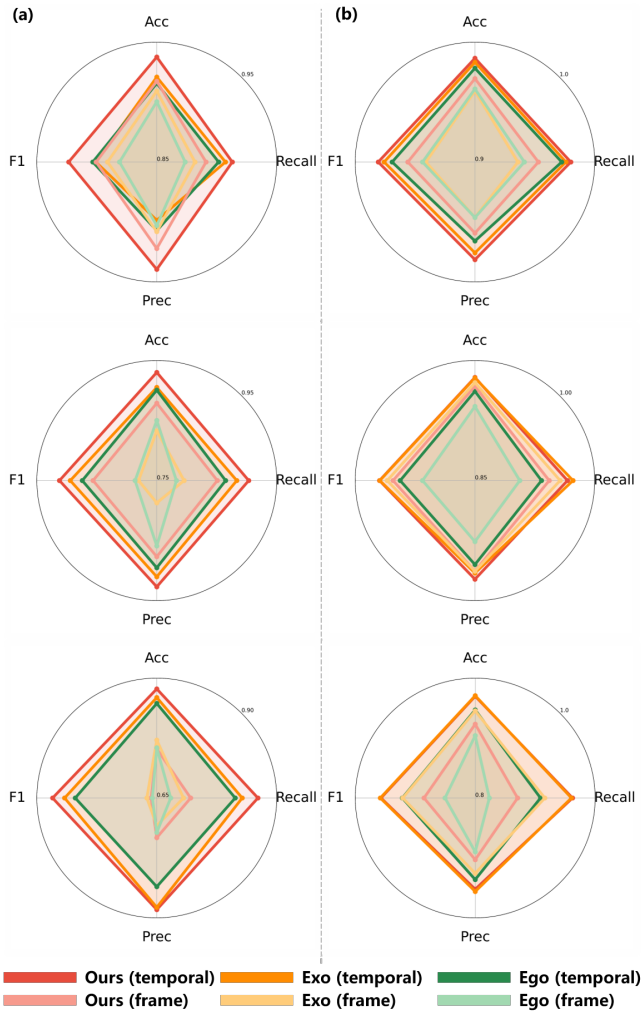

Fig. 4. Visualization of comparison results on four metrics for (a) action recognition and (b) interaction prediction. Three rows refer to models with different training data $\eta = 1.0$, $\eta = 0.5$, and $\eta = 0.25$.

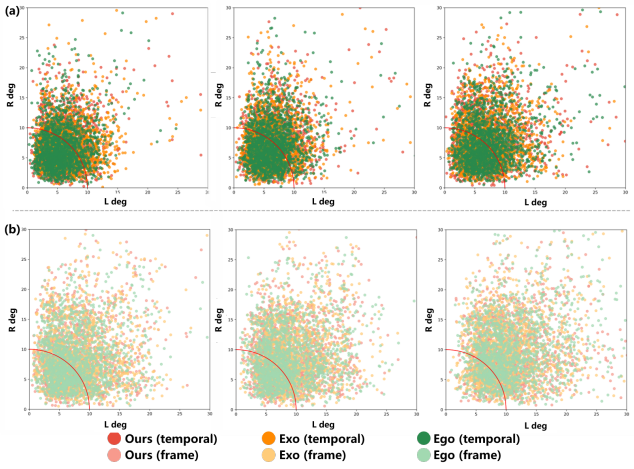

Fig. 5. Visualization of comparison results for instrument pose estimation with (a) temporal 3D fusion and (b) frame-wise 2D fusion. Three columns refer to models with different training data $\eta = 1.0$, $\eta = 0.5$, and $\eta = 0.25$.

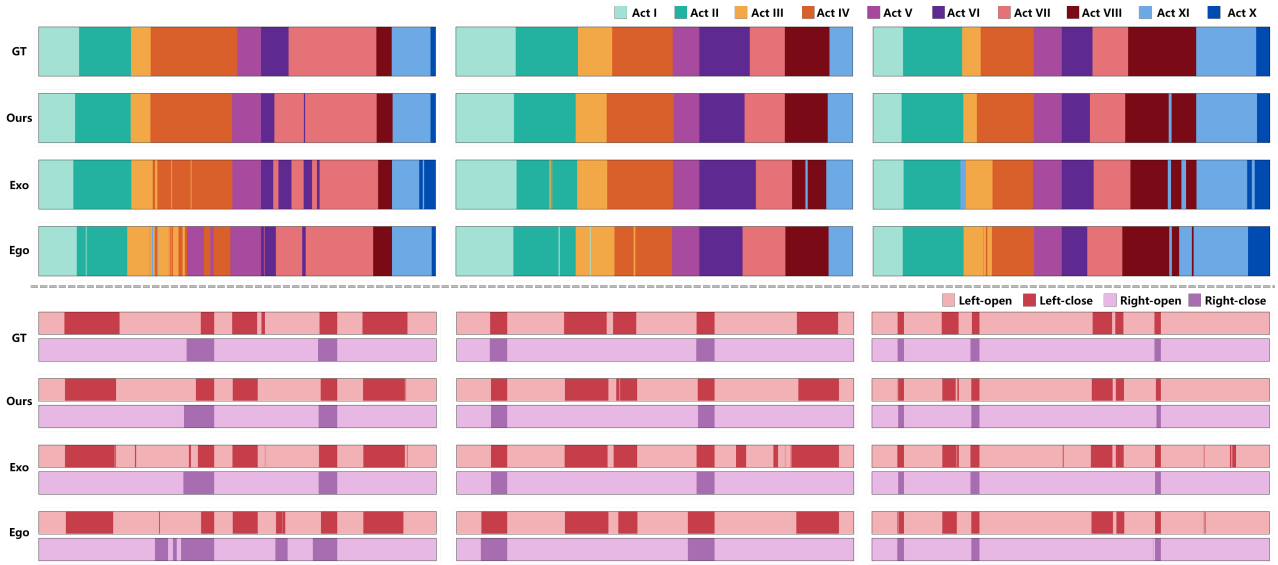

Fig. 6. Qualitative results of anastomosis action recognition and gripper-object interaction prediction. The upper part shows the results for action recognition, with different colors representing different actions. The lower part shows the results for the left and right grippers separately. The first row refers to the ground truth labels, and the second to fourth rows visualize different methods.

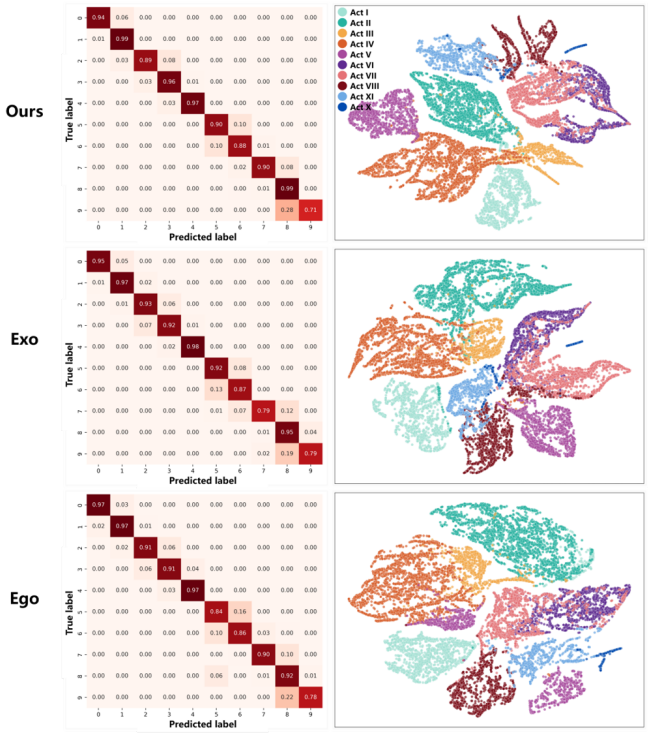

Fig. 7. Confusion matrix and t-SNE visualization results of different methods on anastomosis action recognition. The first row refers to our method and the second to third rows refer to single viewpoints.

degrees as the metric $<10°(\%)$, and more scattered points under this region refers to more accurate pose estimation. Compared with the second row for frame-wise feautre fusion, temporal 3D exo-ego fusion performs better for both left and right instrument pose estimation with more points falling into the left-lower part of the red line. It can also be found that our proposed exo-ego fusion scheme achieves the more accurate pose estimation compared with single viewpoints, and the

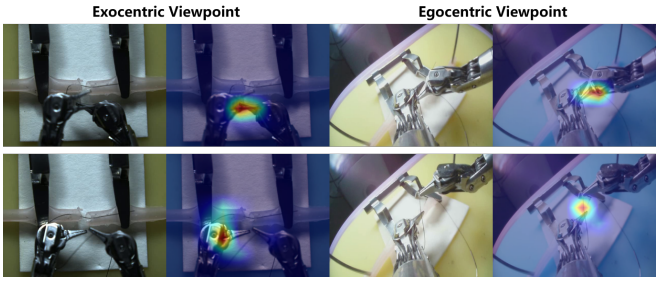

Fig. 8. Qualitative CAM visualization of cross-view feature fusion by our method for action recognition. The left part shows the global perception of exocentric viewpoint and the right part shows the local egocentric viewpoint.

distribution of scattered points is more compact.

B. Qualitative Results

The qualitative results of surgical action and interaction recognition are shown in Fig. 6 and all results are acquired by the temporal 3D fusion model with the entire training dataset. The upper part of the figure visualizes three examples of our proposed fusion scheme compared with the single exocentric or egocentric viewpoint on the task of ten surgical actions recognition. Different actions are represented with different colors. It can be found that our method achieves the most accurate and smooth prediction without significant temporal discontinuity. For surgical long sequences, our method effectively utilizes the cross-view features and improves the precision and stability through temporal fusion. The lower part of Fig. 6 shows three examples on the task of gripper state estimation, and the first line refers to the left gripper and the second line refers to the right gripper. During our established microsurgical scenarios, the left gripper has more operations than the right part, with more frequent state changes of close and open. It is obvious that the cross-view fusion scheme contributes to a more accurate perception of instrument interaction and can capture detailed state changes

according to a long anastomosis sequence.

More qualitative results on the task of surgical action recognition are provided in Fig. 7 and all results are estimated by temporal 3D fusion scheme with full training data. The left part visualizes the confusion matrix and the right part shows the t-SNE feature clustering [21]. Note that features used for t-SNE analysis are extracted by the frame-wise feature extractor. From the results of classification, compared with the individual global exocentric viewpoint and local egocentric viewpoint, our proposed exo-ego fusion scheme achieves the most accurate performance of action recognition, with only errors acceptably occurring near the true labels. From the feature space, the cross-view fusion scheme makes the class-wise feature robust to the distribution of training labels and form clusters perfectly.

We also provide two examples of visualizations of feature representations by class activation mapping [22] in Fig. 8, which is commonly implemented to show the focused area of classification. It can be found that our proposed cross-view fusion scheme concentrates on instrument-tissue interactions a lot, especially for the self-occluded regions. It can prove that the exo-ego cross-view system contributes to more robust visual perception of self-occlusions caused by frequent instrument-tissue interactions.

VI. Conclusions

This work presents a cross-view vision system combining exocentric and egocentric perspectives to enhance visual perception in robotic microsurgical anastomosis. Our system integrates a conventional top-down-view microscope for global scene perception with eye-in-hand cameras mounted on surgical instruments to capture localized instrument-tissue interactions. To demonstrate the efficacy of the proposed exo-ego fusion scheme, we design two distinct fusion networks, i.e., the frame-wise 2D and temporal 3D fusion, and evaluate on three critical visual perception tasks, including surgical action recognition, gripper-object interaction prediction, and instrument pose estimation. A large amount of intraoperative anastomosis data with annotated labels is collected and diverse experiments are conducted on it. The superior performance across all three tasks demonstrates that our exo-ego cross-view system not only enhances the quality of visual perception but also provides more accurate and interpretable feature representations for microsurgical understanding. Future work will focus on enlarging our dataset with different intraoperative scenes and extending our results to in vivo cases to demonstrate the potential application for clinical use.

References

[1] P. E. Dupont, "Medical robots learn to be autonomous," p. eadz8291, 2025.

[2] A. Gao, R. R. Murphy, W. Chen, G. Dagnino, P. Fischer, M. G. Gutierrez, D. Kundrat, B. J. Nelson, N. Shamsudhin, H. Su *et al.*, "Progress in robotics for combating infectious diseases," *Science Robotics*, vol. 6, no. 52, p. eabf1462, 2021.

[3] Y. Liu, J. Zhou, Y. Luo, S.-L. Chen, Y. Guo, and G.-Z. Yang, "Fpm-r2net: Fused photoacoustic and operating microscopic imaging with cross-modality representation and registration network," *Medical Image Analysis*, p. 103698, 2025.

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

[5] Y. Liu, Y. Luo, Y. Luan, X. Zhou, J. Yang, Y. Guo, and G.-Z. Yang, "Towards accurate brain electrode implantation via cross-modality fusion of white-light and photoacoustic microscopy," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025, pp. 7726–7732.

[6] 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 Transactions on Mechatronics*, vol. 26, no. 5, pp. 2758–2769, 2020.

[7] S. Nomura, S.-I. Terada, T. Ebina, M. Uemura, Y. Masamizu, K. Ohki, and M. Matsuzaki, "Arvis: a bleed-free multi-site automated injection robot for accurate, fast, and dense delivery of virus to mouse and marmoset cerebral cortex," *Nature Communications*, vol. 15, no. 1, p. 7633, 2024.

[8] H. Saeidi, J. D. Opfermann, M. Kam, S. Wei, S. Léonard, M. H. Hsieh, J. U. Kang, and A. Krieger, "Autonomous robotic laparoscopic surgery for intestinal anastomosis," *Science robotics*, vol. 7, no. 62, p. eabj2908, 2022.

[9] J. M. Lasso and R. P. Cano, "Practical solutions for lymphaticovenous anastomosis," *Journal of reconstructive microsurgery*, vol. 29, no. 01, pp. 001–004, 2013.

[10] Y. Luan, Y. Luo, Y. Liu, M. Zhang, Y. An, J. Yang, Y. Guo, and G.-Z. Yang, "Autofocusing with 3-d tracking for robot-assisted microsurgery," *IEEE/ASME Transactions on Mechatronics*, 2025.

[11] Y. Jiang, C. Zhong, C. Qin, Z. Sun, and S. Liu, "Synchronous rotation-based knot tying on mini-incisions using dual-arm nanorobot," *IEEE Transactions on Biomedical Engineering*, 2025.

[12] G. Malzone, G. Menichini, M. Innocenti, and A. Ballestín, "Microsurgical robotic system enables the performance of microvascular anastomoses: a randomized in vivo preclinical trial," *Scientific Reports*, vol. 13, no. 1, p. 14003, 2023.

[13] Y. Liu, J. Yang, X. Gu, Y. Guo, and G.-Z. Yang, "Egohmr: Egocentric human mesh recovery via hierarchical latent diffusion model," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9807–9813.

[14] K. Grauman, A. Westbury, E. Byrne, Z. Chavis, A. Furnari, R. Girdhar, J. Hamburger, H. Jiang, M. Liu, X. Liu *et al.*, "Ego4d: Around the world in 3,000 hours of egocentric video," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 995–19 012.

[15] Y. Liu, J. Yang, X. Gu, Y. Guo, and G.-Z. Yang, "Ego+ x: An egocentric vision system for global 3d human pose estimation and social interaction characterization," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 5271–5277.

[16] J. W. Kim, T. Z. Zhao, S. Schmidgall, A. Deguet, M. Kobilarov, C. Finn, and A. Krieger, "Surgical robot transformer (srt): Imitation learning for surgical tasks," in *Proceedings of The 8th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 270, 06–09 Nov 2025, pp. 130–144.

[17] J. W. Kim, J.-T. Chen, P. Hansen, L. X. Shi, A. Goldenberg, S. Schmidgall, P. M. Scheikl, A. Deguet, B. M. White, D. R. Tsai *et al.*, "Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning," *Science robotics*, vol. 10, no. 104, p. eadt5254, 2025.

[18] K. Hsu, M. J. Kim, R. Rafailov, J. Wu, and C. Finn, "Vision-based manipulators need to also see from their hands," in *International Conference on Learning Representations*.

[19] E. Özsoy, A. Mamur, F. Tristram, C. Pellegrini, M. Wysocki, B. Busam, and N. Navab, "Egoexor: An ego-exo-centric operating room dataset for surgical activity understanding," *arXiv preprint arXiv:2505.24287*, 2025.

[20] A. Huaulmé, D. Sarikaya, K. Le Mut, F. Despinoy, Y. Long, Q. Dou, C.-B. Chng, W. Lin, S. Kondo, L. Bravo-Sánchez *et al.*, "Micro-surgical anastomose workflow recognition challenge report," *Computer Methods and Programs in Biomedicine*, vol. 212, p. 106452, 2021.

[21] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.