Wrist Camera Pose-Guided Multi-View Fusion for Occlusion Reconstruction in Robot-Assisted Microsurgical Anastomosis

Xinyao Zhou[1,2†][0009−0008−9131−7738], Yuxuan Liu[1,2†][0000−0002−2431−2653], Yating Luo[1,2][0009−0006−6050−4106], Yunfei Luan[1,2][0009−0004−9810−6272], Guang-Zhong Yang[1,2][0000−0003−4060−4020], and Yao Guo[1,2*][0000−0001−8041−1245]

[1] Institute of Medical Robotics, School of Biomedical Engineering, Shanghai Jiao Tong University, Shanghai, China
{zhou_xinyao, 200009051yx, gzyang, yao.guo} @sjtu.edu.cn
[2] Shanghai Key Laboratory of Flexible Medical Robotics, Tongren Hospital, Shanghai Jiao Tong University, Shanghai, China

Abstract. Precise visual perception is a cornerstone of safety and efficacy in embodied robot-assisted surgery. However, the intricate operations of microsurgical anastomosis solely through the global vision system are often plagued by severe occlusions of frequent instrument-tissue interactions, which severely hinder the visibility of micro-scale structures like suturing threads. To tackle this challenge, we introduce the Pose-Guided Geometry-Aware Transformer, a novel framework for occlusion-aware reconstruction by leveraging a wrist-mounted camera configuration for surgical robots. Specifically, we first curate a multi-view dataset with well-calibrated wrist camera poses. By introducing an implicit camera pose encoding mechanism, our proposed method is able to effectively fuse geometric priors into the framework, facilitating robust cross-view alignment and feature aggregation. Such design exhibits superior robustness to diverse occlusion ratios and numbers of multi-view inputs, enabling the reconstruction of occluded regions with complementary wrist perspectives. Our method establishes a new multi-view benchmark for suturing thread recognition in microsurgical anastomosis and significantly outperforms the current state-of-the-art approaches. This work emphasizes the importance of the wrist-camera pose in multi-view fusion, providing a paradigm-shifting approach to occlusion recovery and fine-grained visual perception for surgical robotics.

Keywords: Multi-view Fusion · Occlusion Recovery · Surgical Robotics.

1 Introduction

Precise visual perception of fine anatomical and interventional structures is a prerequisite for safe and effective robot-assisted surgery [1, 11, 24]. In microsurgical

† These authors contributed equally to this work.
* Corresponding author.

anastomosis, the accurate segmentation of suturing threads is of particular importance. Unlike large, well-defined organs, surgical sutures are thin, highly deformable, and exhibit complex topological configurations. High-quality segmentation of such small structures is critical not only for awareness of intraoperative situation awareness [9] but also for enabling downstream capabilities such as autonomous suturing [13, 29], motion planning [22]. However current segmentation approaches, ranging from general large foundation models [15] to task-specific frameworks [8], are often inadequate for surgical sutures shown in Fig. 1-(b), as their slender morphology and complicate topology frequently lead to fragmented predictions or a complete failure to infer the geometry of occluded segments. With the rapid evolution of embodied artificial intelligence [10], surgical embodied robots have featured hardware configurations similar to general-purpose humanoid agents with dual-arm and primary global views [4, 6]. However, the strategic utilization of wrist-mounted perspectives remains fundamentally underexplored. In this work, we address this gap by focusing on the microsurgical anastomosis paradigm, where we harness complementary visual streams from wrist cameras to achieve robust perception and fine-grained segmentation of sutures under challenging occlusions. Leveraging multiple wrist views [2] makes the segments occluded in the primary view easily captured by them, thus enabling the reconstruction of self-occluded structures, shown in the Fig. 1-(c).

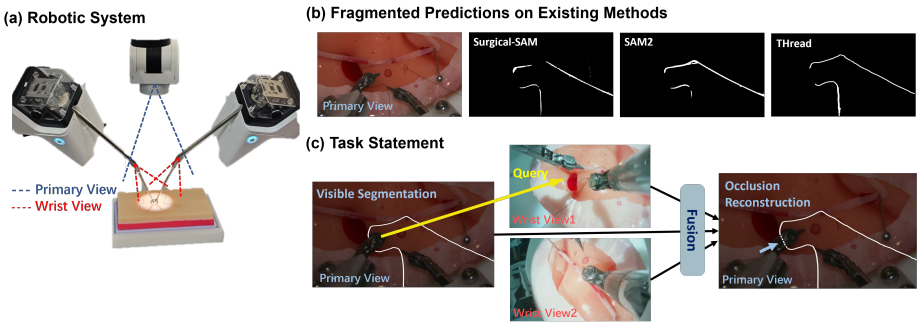

Fig. 1. (a): The surgical robot configurations with one primary camera and two wrist cameras. (b): Performances with methods Surgical-SAM [8], SAM2 [15], THread [30]. (c): The overview of our task to reconstruct the occluded segments.

Existing occlusion-handling strategies can be broadly divided into single-view completion and multi-view fusion. Single-view approaches, whether leveraging geometric priors [21, 30] or generative inpainting [25, 26], remain fundamentally limited by the information available in one perspective and degrade rapidly under large-scale occlusions. Multi-view reconstruction offers a principled alternative by aggregating complementary information across viewpoints [17]. However, unlike the fixed camera arrays in autonomous driving [7,28], wrist-mounted cameras on surgical robots undergo arbitrary 6-DoF motions, introducing extreme pose

diversity that renders naïve feature alignment ineffective. A pose-aware fusion mechanism is therefore essential.

In this paper, we investigate the potential of integrating camera pose encoding within transformer architectures to facilitate robust wrist-view fusion. We propose cross-view alignment module, leveraging camera pose priors to simultaneously inject 3D perspective cues and 2D geometric information into each patch embedding. This enables the network to learn cross-view correspondences, thus significantly enhancing the performance of occlusion reconstruction. We curate the first multiple wrist-view occlusion dataset and evaluate our proposed method across different occlusion ratios and perspective counts, where our method demonstrates leading performance. To the best of our knowledge, this work represents the first attempt to explore multi-view information fusion specifically from the wrist-mounted perspectives, showing potential in wider applications for surgical robotic manipulation and perception tasks.

2 Methods

The proposed model accepts an input set of $\{(\mathbf{V}_i, \mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i)\}_{i=1}^{M}$, where $\mathbf{V_i} \in \mathbb{R}^{H\times W\times 3}$ denotes the input wrist-camera image, $\mathbf{K}_i \in \mathbb{R}^{3\times 3}$ and $[\mathbf{R}_i|\mathbf{t}_i] \in \mathbb{R}^{3\times 4}$ represent the associated camera intrinsics and extrinsics for M views, respectively. In order to simulate the occlusion in the primary view, we implement Proactive Occlusion Masking to manually impose masking. Our objective is to leverage $M-1$ complementary wrist perspectives to perform surgical suture segmentation and reconstruct occluded structures within the primary view. As shown in Fig. 2, the framework is structured into a two-stage pipeline: feature-level alignment pre-training, followed by fine-tuning for the downstream task. We employ DINOv3 [19] as the backbone, capitalizing on its pre-trained fine-grained perceptual capabilities for feature extraction. Within the Cross-view Alignment Module, both camera intrinsic and extrinsic parameters are encoded and injected into the Transformer architecture alongside Rotary Positional Embedding (RoPE) [20], facilitating the implicit learning of geometric feature alignment. This cross-view alignment capability provides significant potential for generalization across segmentation and other visual perception tasks.

2.1 Proactive Occlusion Masking

In practical surgical scenarios, instrument-induced self-occlusion is typically centrally localized within the operative field and varies in scale, posing an inherent paradox for ground-truth acquisition. Since occluded regions are unobservable by the hardware setup, obtaining accurate training labels is fundamentally impossible. To circumvent this challenge, we propose a Proactive Occlusion Masking strategy, where we apply deliberate masks to occlusion-free target images and force the network to reconstruct features within these regions. It ensures absolute label fidelity by allowing us to annotate raw images before simulating occlusion, and it also enables us to manipulate various mask ratios to make

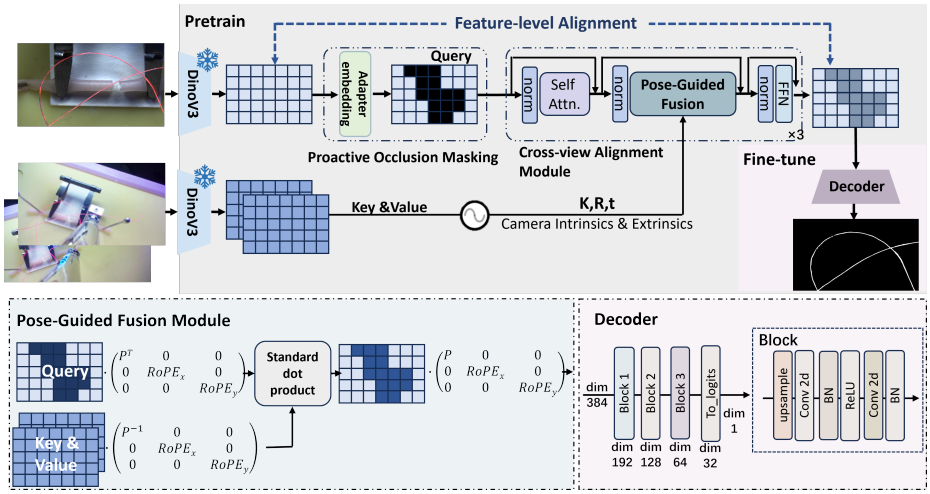

Fig. 2. Overview of the model architecture. During the pre-training phase, features of the primary view via proactive occlusion masking serve as the Query, while concatenated features from wrist views constitute the Key and Value. Cross-view Alignment Module incorporates encoded camera intrinsics and extrinsics to achieve occlusion reconstruction. The final high-fidelity segmentation masks are generated during the downstream fine-tuning stage.

the network learn robust correspondences across various viewpoints. Proactive masking can be implemented at either the image or feature level. Given that we utilize DINOv3 pre-trained weights, imposing rigid image-level masks would cause a distribution shift from natural images and introduce undesirable artifacts at feature boundaries. Thus we adopt patch-level masking scheme within the feature space where, unlike the discrete and scattered patterns of masked image modeling [3, 14, 16], our generated masks are spatially contiguous and centrally clustered to emulate the surgical instrument interference.

2.2 Revisiting the Camera Parameters

In contrast to conventional multi-view fusion and occlusion completion methods that rely on fixed viewpoints, the relative spatial relationships between our 6-DoF wrist perspectives vary arbitrarily across diverse poses hence relative camera pose guidance is indispensable for robust perception. Both intrinsics and extrinsics of each image are represented as the projection matrix $\mathbf{P}_i = \mathbf{K}_i[\mathbf{R}_i|\mathbf{t}_i] \in \mathbb{R}^{3\times4}$. For notational convenience, these 3×4 projection matrices can be made invertible by lifting to $\tilde{\mathbf{P}}_i \in \mathbb{R}^{4\times4}$ with the standard basis vector. It can be used to compute 2D image coordinates from world coordinates $\left[\tilde{\mathbf{x}}_i\ 1\right]^T \propto \tilde{\mathbf{P}}_i\tilde{\mathbf{X}}_{\text{world}}$, where $\tilde{\mathbf{x}}_i \in \mathbb{R}^3$ and $\tilde{\mathbf{X}}_{\text{world}} \in \mathbb{R}^4$ are homogeneous coordinates in the image and world respectively. Consequently, the geometric correspondence between the image coordinates in two distinct viewpoints can be easily

established as: $\left[\tilde{\mathbf{x}}_i\ 1\right]^T = (\tilde{\mathbf{P}}_i\tilde{\mathbf{P}}_j^{-1})\left[\tilde{\mathbf{x}}_j\ 1\right]^T = \mathbf{M}_{i,j}\left[\tilde{\mathbf{x}}_j\ 1\right]^T$, where $\mathbf{M}_{i,j} \in \mathbb{R}^{4\times4}$ denotes the relative projection matrix. It can be easily proved that $\mathbf{M}_{i,j}$ remains independent of the choice of world coordinate, as it exclusively encapsulates the relative geometric relationships. Given the inherent permutation invariance of the Transformer architecture [23], positional embeddings like RoPE are essential for the model to perceive the 2D spatial arrangement of image patches. Inspired by [5], we propose to integrate the relative camera pose $\mathbf{M}_{i,j}$ with RoPE, thereby injecting 3D spatial-geometric priors into 2D patch embeddings. Standard dot product attention is $\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{Q^TK}{\sqrt{d}}\right)V$, where $Q \in \mathbb{R}^{s\times d}$, K and $V \in \mathbb{R}^{(M-1)s\times d}$, $M-1$ denotes the number of wrist views. In this work, we formulate a projection matrix $\mathbf{D}_i$ by integrating the normalized camera projection matrix $\mathbf{P}$ and the RoPE parameters: $\mathbf{D} = \begin{bmatrix} \mathbf{I}_{d/8}\otimes\mathbf{P} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & RoPE_x & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & RoPE_y \end{bmatrix}_{d\times d}$, where $RoPE_x, RoPE_y \in \mathbb{R}^{\frac{d}{4}\times\frac{d}{4}}, \mathbf{I}_{d/8}\otimes\mathbf{P} \in \mathbb{R}^{\frac{d}{2}\times\frac{d}{2}}$. We denote the batched matrix-vector product as $\circledcirc$, the Kronecker product as $\otimes$, and the identity matrix of size k as $\mathbf{I}_k \in \mathbb{R}^{k\times k}$. The total dimensionality of $\mathbf{D}$ is defined to align with the feature dimension d. $\mathbf{D}_i$ is subsequently injected into the Query, Key, and Value representations within the attention layers and further incorporated into the final output for numerical stability:

$$\begin{aligned}\mathrm{Attn}^r(Q,K,V) &= \mathbf{D}_w \circledcirc \mathrm{Attn}(\mathbf{D}_o \circledcirc Q,\ \mathbf{D}_w^{-1}\circledcirc K,\ \mathbf{D}_w^{-1}\circledcirc V) \\ &= \mathbf{D}_w(\mathrm{softmax}\left(\frac{Q^T(\mathbf{D}_o^T\mathbf{D}_w^{-1})K}{\sqrt{d}}\right)\mathbf{D}_w^{-1}V),\end{aligned} \tag{1}$$

where $\mathbf{D}_o, \mathbf{D}_w$ denote the matrix for the occluded perspective and wrist perspectives, respectively. $(\mathbf{D}_o^T\mathbf{D}_w^{-1})_{\frac{d}{2}} = \mathbf{I}_{d/8}\otimes\mathbf{M}_{o,w}$ plays a crucial role in relative pose guided fusion.

2.3 Training Framework and Loss Functions

Our training procedure follows a two-stage training strategy, including multi-view feature alignment pre-training and task-specific fine-tuning for the downstream segmentation. Throughout the entire training process, the DINOv3 backbone weights are kept frozen. During pre-training we optimize our framework using a composite loss function containing Cosine Similarity Loss ($\mathcal{L}_{\cos}$), InfoNCE Loss ($\mathcal{L}_{\mathrm{nce}}$) [12] and Gradient Constraint Loss ($\mathcal{L}_{\mathrm{grad}}$). $\mathcal{L}_{\cos}$ is employed to enforce directional consistency between the reconstructed feature vectors and the pristine features extracted from occlusion-free target images. $\mathcal{L}_{\mathrm{nce}}$ prevents feature collapse by contrasting masked patches against a dictionary of global context. $\mathcal{L}_{\mathrm{grad}}$ explicitly penalizes inconsistencies in structural details and edge gradients, ensuring the network maintains sharp boundaries for fine-grained suture segmentation. During the fine-tune stage, the pretrained encoder is frozen for one epoch to allow for initial convergence of the decoder module, after which

the encoder except backbone is fully unfrozen. To ensure the topological connectivity and width consistency of the fine-grained suture structures, we employ the clDice loss [18], which implements a cross-constraint between the soft-skeleton and the volume. The topology precision T_{prec} measures the fraction of the predicted skeleton that aligns with the ground-truth region, whereas the topology sensitivity T_{sens} assesses how well the ground-truth skeleton is captured by the predicted volume. By focusing on centerline alignment, clDice effectively enhances the reconstruction of slender sutures under challenging occlusions. The combined loss functions are formulated as the weighted sum of individual terms:

$$\mathcal{L}_{\text{pretrain}} = \lambda_{\cos}\mathcal{L}_{\cos} + \lambda_{\text{nce}}\mathcal{L}_{\text{nce}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}},\ \mathcal{L}_{seg} = 1 - \frac{2 \cdot T_{prec} \cdot T_{sens}}{T_{prec} + T_{sens}}. \quad (2)$$

3 Experiments and Results

3.1 Dataset and Settings

Given the absence of specialized datasets for fine-grained surgical sutures occlusion recovery, we have meticulously curated and annotated a novel Multiple Wrist-view occlusion dataset from a robotic suturing platform, featuring fine-grained suture segmentation masks and well-calibrated camera poses. The dataset comprises 12 video sequences corresponding to 12 static top views and approximately 17,000 wrist-camera images at a resolution of 1280 × 720. This collection is partitioned into validation and testing sets, each consisting of two sequences and totaling roughly 3,000 images per set. Camera poses were rigorously calibrated using either calibration patterns or robotic kinematics, ensuring the reprojection error less than 0.5 pixels across all sequences. The dataset is accessible at `https://github.com/Zoeric80/Multiple-Wrist-view-Microsurgical-Anastomosis-Occlusion-Dataset`.

While our architecture is capable of processing high-resolution images, we resize the input images to 640×352 to balance computational efficiency with the architectural of DINOv3. Based on our empirical evaluations, we set the weight $\lambda_{\cos} = \lambda_{\text{grad}} = 1, \lambda_{\text{nce}} = 0.1$. We pretrained on the Realestate dataset with a batch size of 4 for 35 epochs. All experiments are implemented on an NVIDIA RTX A6000 GPU. For the segmentation, we report five metrics, including the Dice coefficient (Dice↑), Peak Signal-to-Noise Ratio (PSNR↑), Intersection over Union (IoU↑), L1 distance (L1↓), and centerlineDice (clDice↑).

3.2 Comparison Results

Table. 1 compares the performance of proposed method with several state-of-the-art baselines under various mask ratios. Among these, AOT-GAN [26] and CRFill [27] represent inpainting-based approaches, while THread [30] leverages spatial context and geometric topology. Our method consistently demonstrates superior performance across all occlusion levels. At lower masking ratios, THread

achieve favorable results, particularly in the clDice metric. However, their efficacy degrades rapidly as the occlusion area expands. Conversely, while AOT-GAN maintain higher image similarity under heavy occlusion, they often yield inaccurate geometric structures. While our framework ensures higher geometric fidelity, the predictive confidence for fine-grained details naturally decreases as the masking ratio increases. Qualitative results are visualized in Fig. 3. As occlusion levels increase, inpainting-based approaches tend to produce significantly more erroneous geometric structures, but our method demonstrates remarkable robustness in maintaining topological fidelity.

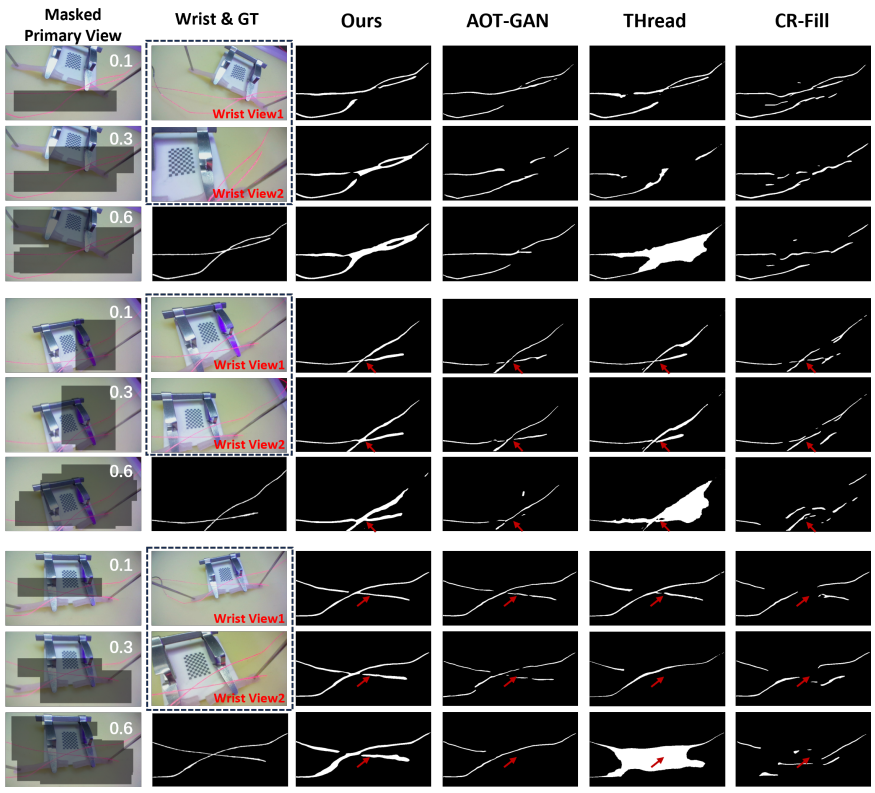

Fig. 3. Qualitative comparison results of our proposed method and other state-of-the-art baselines under various masking ratios. We utilize a dual-view fusion setup, which achieves the optimal performance.

3.3 Ablation Studies

As illustrated in Fig. 4, two sets of ablation experiments were conducted to validate the impact of view count and the camera pose encoding module. The

Table 1. Quantitative Comparisons of the Proposed Method With State-of-the-Art Models on the Dataset. The **bold** and underline results reported in the tables indicate the best and the second best performances. η denotes the mask ratios.

Methods	η	PSNR ↑	Dice ↑	IOU ↑	L_1 ↓	clDice ↑
AOT [26]	10%	19.08±1.30	0.68±0.12	0.52±0.12	0.64±0.24	0.76±0.15
	30%	18.16±1.37	0.60±0.14	0.44±0.14	0.80±0.29	0.66±0.18
	60%	16.74±1.26	0.43±0.15	0.28±0.13	1.08±0.31	0.46±0.17
CR-Fill [27]	10%	18.15±1.54	0.61±0.13	0.45±0.13	0.80±0.30	0.69±0.16
	30%	16.87±1.43	0.49±0.13	0.34±0.12	1.06±0.33	0.55±0.16
	60%	15.47±1.16	0.32±0.11	0.19±0.08	1.43±0.35	0.35±0.12
THread [30]	10%	19.80±0.88	0.77±0.04	0.63±0.04	0.52±0.12	**0.87±0.03**
	30%	17.83±0.84	0.67±0.07	0.51±0.08	0.82±0.20	**0.79±0.06**
	60%	11.19±1.04	0.33±0.06	0.19±0.04	3.86±1.30	0.55±0.07
Ours	10%	**20.15±0.32**	**0.77±0.03**	**0.64±0.06**	**0.46±0.06**	0.86±0.03
	30%	**19.16±0.45**	**0.71±0.03**	**0.55±0.04**	**0.58±0.07**	0.78±0.05
	60%	**17.15±0.56**	**0.55±0.06**	**0.38±0.05**	**0.92±0.10**	**0.59±0.07**

vanilla baseline adopts the same architecture and two-stage training strategy, with only the pose-guided encoding removed (RoPE retained). Results show that integrating pose encoding consistently outperforms the vanilla baseline across all view configurations, with peak performance at two wrist views. Insufficient views fail to capture occluded targets, while excessive views introduce noise, leading to coarser predictions. Furthermore, a dual-view setup aligns with mainstream embodied robotic hardware configurations. Conversely, the vanilla model's performance declines as view count increases due to its inability to learn view-dependent geometric relationships.

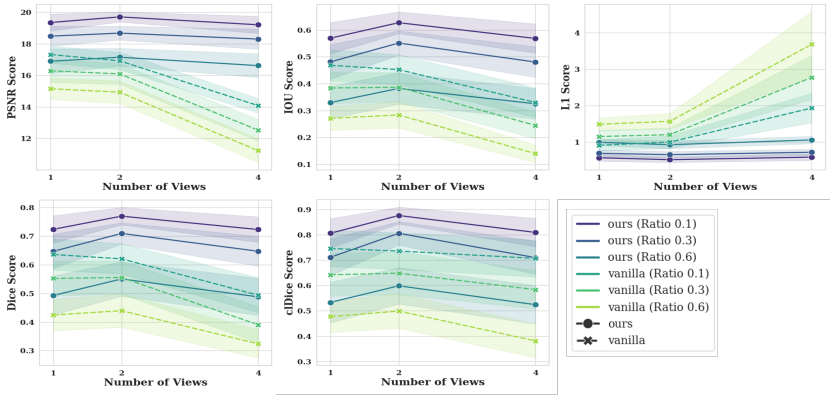

Fig. 4. Ablation study metrics for our proposed method and the vanilla baseline across varying masking ratios and numbers of viewpoints.

4 Conclusion

We propose a camera pose-guided multiple wrist-view fusion framework for the occlusion reconstruction and segmentation of surgical sutures. To support this fine-grained task, we meticulously curated a novel dataset, establishing the first benchmark for micro-scale flexible object reconstruction under multi-view occlusion. Our approach robustly demonstrates that wrist-mounted perspectives significantly enhance fine-grained perception, underscoring the transformative potential of wrist cameras in microsurgical anastomosis procedures.

Acknowledgments. This work was supported by Shanghai Municipal Science and Technology Major Project 2021SHZDZX. This work was supported by National Natural and Science Foundation of China under grant 62203296.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Cao, S., Fan, J., Shao, L., Sun, Q., Xu, T., Ai, D., fu, T., Xiao, D., Song, H., Zhang, T., Yang, J.: Robot-assisted osteotomy and reconstruction with ar guidance in maxillofacial reconstructive surgery. Cyborg and Bionic Systems **7** (05 2026). https://doi.org/10.34133/cbsystems.0590
2. Cuevas-Velasquez, H., Li, N., Tylecek, R., Saval-Calvo, M., Fisher, R.B.: Hybrid multi-camera visual servoing to moving target. In: 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1132–1137. IEEE (2018)
3. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16000–16009 (2022)
4. Kim, J.W., Chen, J.T., Hansen, P., Shi, L.X., Goldenberg, A., Schmidgall, S., Scheikl, P.M., Deguet, A., White, B.M., Tsai, D.R., et al.: Srt-h: A hierarchical framework for autonomous surgery via language-conditioned imitation learning. Science robotics **10**(104), eadt5254 (2025)
5. Li, R., Yi, B., Liu, J., Gao, H., Ma, Y., Kanazawa, A.: Cameras as relative positional encoding. arXiv preprint arXiv:2507.10496 (2025)
6. Li, S., Zhou, J., Jia, Z., Yeung, D.Y., Mason, M.T.: Learning accurate objectness instance segmentation from photorealistic rendering for robotic manipulation. In: International Symposium on Experimental Robotics. pp. 245–255. Springer (2018)
7. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q., Dai, J.: Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(3), 2020–2036 (2024)
8. Liu, H., Zhang, E., Wu, J., Hong, M., Jin, Y.: Surgical sam 2: Real-time segment anything in surgical video by efficient frame pruning. arXiv preprint arXiv:2408.07931 (2024)
9. Liu, Y., Zhou, X., Luo, Y., Luan, Y., Xiao, Z., Guo, Y., Yang, G.Z.: Simultaneous surgical stereo depth and motion estimation via brightness-aware self-supervised learning. Pattern Recognition **179**, 113510 (2026). https://doi.org/https://doi.org/10.1016/j.patcog.2026.113510

10. Long, Y., Lin, A., Kwok, D.H.C., Zhang, L., Yang, Z., Shi, K., Song, L., Fu, J., Lin, H., Wei, W., et al.: Surgical embodied intelligence for generalized task autonomy in laparoscopic robot-assisted surgery. Science Robotics **10**(104), eadt3093 (2025)
11. Luan, Y., Luo, Y., Liu, Y., Zhang, M., An, Y., Yang, J., Guo, Y., Yang, G.Z.: Autofocusing with 3-d tracking for robot-assisted microsurgery. IEEE/ASME Transactions on Mechatronics **31**(1), 220–231 (2026). https://doi.org/10.1109/TMECH.2025.3585533
12. Oord, A.v.d., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748 (2018)
13. Ostrander, B.T., Massillon, D., Meller, L., Chiu, Z.Y., Yip, M., Orosco, R.K.: The current state of autonomous suturing: a systematic review. Surgical Endoscopy **38**(5), 2383–2397 (2024)
14. Pan, W., Xu, Z., Yan, J., Wu, Z., Tong, R.K.y., Li, X., Yao, J.: Semi-supervised semantic segmentation meets masked modeling: Fine-grained locality learning matters in consistency regularization. Pattern Recognition p. 112586 (2025)
15. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
16. Seo, Y., Kim, J., James, S., Lee, K., Shin, J., Abbeel, P.: Multi-view masked world models for visual robotic manipulation. In: International Conference on Machine Learning. pp. 30613–30632. PMLR (2023)
17. Shen, W., Wang, Y., Liu, M., Wang, J., Ding, R., Zhang, Z., Meijering, E.: Multi-view integration network for multitask robotic surgical scene analysis. IEEE Transactions on Instrumentation and Measurement (2025)
18. Shit, S., Paetzold, J.C., Sekuboyina, A., Ezhov, I., Unger, A., Zhylka, A., Pluim, J.P., Bauer, U., Menze, B.H.: cldice-a novel topology-preserving loss function for tubular structure segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16560–16569 (2021)
19. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
20. Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)
21. Tran, M., Vo, K., Yamazaki, K., Fernandes, A., Kidd, M., Le, N.: Aisformer: Amodal instance segmentation with transformer. arXiv preprint arXiv:2210.06323 (2022)
22. Wang, X., Guo, S., Xu, Z., Zhang, Z., Sun, Z., Xu, Y.: A robotic teleoperation system enhanced by augmented reality for natural human-robot interaction. Cyborg and Bionic Systems **5** (04 2024). https://doi.org/10.34133/cbsystems.0098
23. Xu, H., Xiang, L., Ye, H., Yao, D., Chu, P., Li, B.: Permutation equivariance of transformers and its applications. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5987–5996 (2024)
24. Yang, G.Z., Bellingham, J., Dupont, P.E., Fischer, P., Floridi, L., Full, R., Jacobstein, N., Kumar, V., McNutt, M., Merrifield, R., et al.: The grand challenges of science robotics. Science Robotics **3**(14), eaar7650 (2018)
25. Yu, Y., Zeng, Z., Zheng, H., Luo, J.: Omnipaint: Mastering object-oriented editing via disentangled insertion-removal inpainting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17324–17334 (2025)
26. Zeng, Y., Fu, J., Chao, H., Guo, B.: Aggregated contextual transformations for high-resolution image inpainting. IEEE transactions on visualization and computer graphics **29**(7), 3266–3280 (2022)

27. Zeng, Y., Lin, Z., Lu, H., Patel, V.M.: Cr-fill: Generative image inpainting with auxiliary contextual reconstruction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 14164–14173 (2021)
28. Zhou, B., Krähenbühl, P.: Cross-view transformers for real-time map-view semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13760–13769 (2022)
29. Zhou, X.Y., Guo, Y., Shen, M., Yang, G.Z.: Application of artificial intelligence in surgery. Frontiers of medicine **14**(4), 417–430 (2020)
30. Zhou, X., Liu, Y., Zhang, M., Li, J., Guo, Y., Yang, G.Z.: Deep coarse-to-fine networks for robust segmentation and pose estimation of surgical suturing threads. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 520–526. IEEE (2025)