3D Vision Technology in Precision Medical Robots

Ketong Yan

{cr906328449@gmail.com}

School of Automation,Beijing Institute of Technology,Beijing,China

Abstract. As minimally invasive surgery (MIS) has evolved towards intelligence and automation, the need for environment perception by medical robots has shifted from mere image transmission to deep 3D scene understanding. This article first reviews the evolution of 3D vision techniques in precision medical robotics, focusing on the shift from traditional geometric methods to emerging paradigms such as deep learning, Transformer architecture, and Neural Radiation Field (NeRF). The article then delves into the special visual challenges in endoscopic scenarios, and illustrates the latest breakthroughs in dynamic reconstruction and vision SLAM techniques in the context of cutting-edge literature such as EndoNeRF and 3D Gaussian Splatting. In addition, this paper also discusses the application of these techniques in real-world scenarios such as surgical instrument tracking, multimodal navigation, and haptic perception, aiming to provide a systematic reference to support future research and development of intelligent minimally invasive surgical robotic systems.

Keywords: 3D Vision, Precision Medical Robotics, Neural Radiance Fields, Dynamic 3D Reconstruction, Visual SLAM.

1 Introduction

The advent of minimally invasive surgery has revolutionized the treatment paradigm of modern medicine. Through tiny incisions, doctors utilize endoscopes and slender surgical instruments to operate deep inside the human body, greatly reducing the burden of trauma and recovery for patients. Assistive systems, represented by the da Vinci surgical robot, have taken this minimally invasive capability to new heights. However, despite the increasing sophistication of the hardware, existing surgical robotic systems are still essentially “teleoperation” tools. Doctors rely heavily on two-dimensional video images from endoscopes to make decisions, and this mode of operation results in a severe lack of depth information. Doctors need to go through a long learning curve before they can transform 2D images into 3D spatial cognition in their brains. This “hand-eye separation” limits the further improvement of surgical efficiency. In order to realize true precision medicine, i.e., to automate surgery, giving robots the ability to perceive 3D vision as well as or even better than human beings has become a core issue of concern for both academia and industry. Accurate 3D vision is not only a simple image

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364884

presentation, but also the cornerstone for realizing 3D reconstruction of surgical scenes, precise real-time tracking of instruments, and augmented reality navigation.

This paper will follow the logic of technological development to systematically sort out the status quo and change of 3D vision technology in precision medical robotics. The paper will first examine the foundation and hardware system of medical 3D vision, and analyze the rigorous requirements for imaging technology in the special environment of the body. On this basis, the article will delve into the generational evolution of depth estimation and 3D reconstruction techniques, especially analyzing why deep learning-based methods are gradually replacing traditional geometric matching algorithms, and how Neural Radiation Field (NeRF) and 3D Gaussian Sputtering (3DGS) techniques have brought new light to high-fidelity reconstruction. Subsequently, this article will focus on localization challenges in dynamic environments and discuss the adaptation of visual SLAM techniques for non-rigid surgical scenarios. Finally, the article will demonstrate the practical application value of 3D vision technology in instrument segmentation, multimodal alignment and cross-modal force haptic perception in the context of specific clinical needs.

The core of this paper aims to reveal how 3D vision technology can drive the evolution of medical robots from passive actuation tools to active and intelligent sensing subjects. By synthesizing and analyzing key literature including EndoNeRF, DefSLAM, and 3D Gaussian Splatting, this paper seeks to answer the key question of how to build robust and real-time 3D vision systems in surgical environments that lack Ground Truth and are full of dynamic deformations and complex illumination. This paper argues that data-driven deep learning approaches, especially the combination of self-supervised learning and neural rendering techniques, are redefining the perceptual boundaries of medical robotics and finding a new balance between real-time and high fidelity.

2 Medical 3D Vision Basics and Hardware System

In the early days of minimally invasive surgery, monocular endoscopes dominated. The simple structure and small diameter of the monocular lens allowed easy access to narrow body cavities. However, monocular vision does not provide direct depth information, and physicians can only rely on empirical cues such as shadows, texture gradients, or motion parallax to infer distances, which can easily lead to misjudgments in complex, high-risk operations. In order to solve this problem, binocular stereo vision technology has emerged. By simulating the principle of binocular parallax of the human eye, binocular endoscopes are able to directly calculate the depth information of the scene. Most of the mainstream surgical robotic systems are now equipped with a binocular vision system as standard. However, binocular vision is not perfect; it has strict requirements on the baseline distance between the two lenses, and traditional stereo matching algorithms tend to fail when facing smooth tissue surfaces in vivo that lack obvious texture [1].

Unlike open scenarios such as automated driving or industrial inspection, the internal environment of the human body has a very high degree of specificity and complexity. This specificity is mainly reflected in two aspects: lighting conditions and environmental dynamics. First of all, the in vivo environment completely relies on the point light source illumination that comes with the endoscope. This type of illumination leads to overexposure in the center of the image and darkness at the edges, resulting in an extremely uneven distribution of light. To make matters even more difficult, tissue surfaces are often covered with body fluids or mucus, which produces strong specular highlights when illuminated by a point light source. These reflective regions appear as highlighted white spots in the image, completely losing the original texture information and causing great interference to the feature-matching-based vision algorithms. Second, the in vivo environment is a typical non-rigid dynamic environment. Breathing movements, heartbeats, and the contact extrusion of surgical instruments cause soft tissues to be deformed at all times. This continuous non-rigid deformation breaks the “rigid scene assumption” commonly used in traditional computer vision algorithms, making 3D reconstruction and localization exponentially more difficult.

To overcome the limitations of vision-only solutions, researchers have attempted to introduce industrially proven depth sensors into the medical field.RGB-D cameras and Time-of-Flight (ToF) cameras have gained attention for their ability to directly acquire depth maps. However, in medical minimally invasive scenarios, these new sensors face insurmountable physical barriers. High-precision ToF sensors tend to be large and difficult to miniaturize to the point where they can fit through minimally invasive incisions (typically less than 12 mm). In addition, structured light sensors are prone to absorption or scattering of the infrared light they project in media filled with blood and smoke, resulting in a significant reduction in measurement accuracy. Therefore, the current industry consensus tends to exploit the potential of pure vision (RGB) algorithms to solve the problem by co-optimizing hardware and software. For example, a related study, RAFT-Stereo, demonstrated how advanced optical flow algorithms can be utilized to enhance binocular endoscope perception [2]. The study demonstrated that by utilizing Recurrent All-Pairs Field Transforms, the system can recover high-quality depth information by capturing minute pixel movements even inside narrow and sparsely textured cavities, which provides a strong technological support for the application of pure vision solutions in extreme environments.

3 Depth Estimation and 3D Reconstruction

3D reconstruction is more than just a visual display; it is a prerequisite for robots to understand the human anatomy. Only by constructing high-precision 3D organ models in the computer that contain geometric information can the robot perform effective obstacle avoidance path planning, accurately identify lesion targets, and provide the surgeon with intuitive surgical navigation.

3.1 Traditional 3D Reconstruction Methods

Before the popularization of deep learning, medical 3D reconstruction mainly relied on traditional multi-view geometry methods. The core of these methods lies in feature point matching, i.e., finding the same anatomical feature points in images from different viewpoints and calculating their spatial coordinates using the principle of triangulation. From the early sparse point cloud reconstruction, gradually evolving to the later dense reconstruction, the traditional methods perform reasonably well in texture-rich scenes. However, human soft tissue surfaces are often smooth and single-textured, resulting in difficult feature extraction and high matching error rates. In addition, traditional methods usually assume that the scene is static, which makes it difficult to cope with the real-time deformation of the tissue during surgery, and the reconstructed model often lags behind the changes of the real anatomical structure.

3.2 Deep Learning-based 3D Reconstruction

The introduction of deep learning techniques has greatly improved the robustness of feature extraction and has become the key to solving the above challenges. However, one of the core pain points in the medical field is the lack of data, especially the lack of an accurate depth "Ground Truth". In real surgeries, doctors cannot use laser scanners to obtain the absolute depth of tissue as they can in the lab. To address this challenge, researchers have turned to synthetic data and transfer learning. Literature NeuralField-LDM proposes a generative approach based on Latent Diffusion Model (LDM), which is capable of generating high-quality 3D medical scene data [3]. This method not only expands the training dataset but also provides diverse anatomical structure samples for the model. Meanwhile, Sim-to-Real Transfer technology is dedicated to solving the domain divide between virtual and real data [4]. Through the Domain Adaptation algorithm, models trained in virtual environments can be successfully migrated to real surgical scenarios. This strategy effectively bypasses the challenge of real data labeling, allowing the model to maintain high prediction accuracy even when confronted with unseen real tissues. For completely unlabeled real surgical videos, self-supervised learning provides another solution. The literature Self-Supervised Monocular Depth Estimation proposes to utilize the Photometric Consistency (PC) between video frames as a supervised signal [5]. This study combined anatomical geometric constraints to allow the network to self-correct during training, and successfully realized accurate depth restoration in complex scenes such as the inner ear in the absence of truth values, demonstrating the great potential of data-driven methods in medical scenarios.

Convolutional Neural Networks (CNNs), while excelling in local feature extraction, have limitations in capturing long-range dependencies. In order to better understand the global anatomy of organs, the Transformer architecture has been introduced to medical image processing. The literature UNETR innovatively applies the Transformer to 3D medical image segmentation and reconstruction tasks. Through the Self-Attention (SA) mechanism, the model was able to establish global connections between pixel points, which significantly improved the understanding of complex organ morphology [6]. Further studies such as EDUS-Net, explored the advantages of multi-task learning. This study demonstrated that joint depth estimation and semantic segmentation can achieve complementarity between tasks [7]. Semantic information can help the depth estimation network to better distinguish object edges, while depth information can assist the segmentation network to understand the spatial hierarchy, and the two work in tandem to ultimately output a reconstruction result with clearer edges and more reasonable structure.

3.3 Neural Field-Based 3D Reconstruction

Although deep learning has improved feature extraction, traditional mesh or voxel representations are still inadequate when dealing with complex lighting and dynamic deformation. In recent years, the emergence of Neural Implicit Representation (NIR) has brought innovation. Literature EndoNeRF firstly introduces the Neural Radiation Field (NeRF) technique to endoscopic scenes [8].EndoNeRF successfully separates the static background from the dynamic deformation in the rendering process by constructing the Neural Implicit Field and modeling it in the time dimension. What's more, it can effectively simulate and deal with high light reflections common in surgery, generating highly realistic high-fidelity 3D models, breaking the bottleneck of traditional geometric methods that are difficult to deal with reflective surfaces. However, a significant drawback of NeRF is its slow inference speed, which makes it difficult to meet the real-time intraoperative demands. To address this pain point, the recent literature Deformable 3D Gaussian Splatting proposes an explicit representation method based on Gaussian spheres [9]. This technique employs 3D Gaussian Splatting (3DGS) to represent the scene, which retains the high accuracy of neural rendering and greatly improves the computational efficiency. It is shown that this method not only achieves sub-millimeter reconstruction accuracy, but also improves the rendering speed to the Real-time (RT) level. This signifies that high-fidelity dynamic 3D reconstruction technology finally has the possibility of practical implementation in clinical surgical navigation.

4 Visual SLAM and Dynamic Localization

Vision Simultaneous Localization and Mapping (SLAM) is at the heart of robot autonomy. In an unknown in vivo environment, SLAM gives robots the ability to "know where they are" and "what their surroundings look like". For precision medicine robots, SLAM is not only the basis for automatic cruise, but also a prerequisite for intraoperative repositioning, target memory, and multi-instrument cooperative operation. Without a stable localization system, the robot will not be able to maintain the consistency of operation during a long surgical procedure.

Qiu and Ren proposed a low-cost transoral endoscopic navigation system, which is the first application of visual SLAM for this type of procedure [10]. It generates endoscopic trajectories with 3D point clouds of the oral cavity in real time and visualizes them with contour enhancement. The system aligns intraoperative SLAM reconstruction with preoperative CT data with high precision, solves the scaling problem, and provides the surgeon with deep lesion information and precise navigation. Phantom experiments confirm that the trajectory and point cloud alignment errors are as low as 0.614 mm and 0.836 mm, respectively, significantly improving surgical safety and accuracy. However, most of the traditional vision SLAM systems (e.g., the famous ORB-SLAM) are based on the "rigid environment assumption", i.e., objects in the environment are considered to be stationary. This assumption fails completely inside the human body. Tissue deformations caused by breathing and pulses can cause feature points to shift, leading conventional SLAM systems to believe that the camera is moving, resulting in large localization errors. In order to solve this problem, a classical geometric approach to non-rigid environments has been proposed in the literature DefSLAM [11]. The study introduces the concept of "Deformation Graph" to model the tissue surface as a deformable network of nodes. By simultaneously estimating the rigid motion of the camera and the non-rigid deformation of the tissue, DefSLAM successfully achieves stable localization in dynamic environments. In addition to solving the deformation problem, it is equally important to understand the semantic information of the environment. A purely geometric map is just a bunch of point clouds and lacks medical significance. Literature Real-time Semantic SLAM proposes to fuse semantic segmentation information while building a map [12]. This semantic SLAM system allows robots to not only see "obstacles", but also identify 'tumors', "blood vessels", or "nerves". ". This semantic level of understanding provides critical support for intelligent decision making and enables the robot to actively avoid risky areas.

5 Application Scenarios of 3D Vision Technology in Precision Medical Robots

5.1 Surgical Instrument Segmentation and Tracking

In robot-assisted surgery, the prerequisite for precise control of the robotic arm is to grasp the exact position of the end effector (such as a scalpel and forceps) in three-dimensional space in real time. Traditional visual tracking is often affected by motion blur caused by the rapid movement of instruments. Daneshgar Rahbar et al. proposed an innovative intraoperative visual intelligence system, which aims to enhance the monitoring and situational awareness of open surgery through real-time and accurate instrument segmentation [13]. Its core highlight is the use of a convolutional neural network called Enhanced U-Net with GridMask (EUGNet), which significantly improves the accuracy and real-time performance of instrument segmentation. The system achieved a high accuracy of 85.5% in the simulation environment, and the wireless transmission delay was only 200 milliseconds, which is sufficient to support real-time applications. This work provides key technical support for improving surgeons' real-time mastery of surgical instruments, improving surgical efficiency, and assisting in training.

5.2 Preoperative-Intraoperative Multimodal Registration

One important direction of precision medicine is augmented reality (AR) navigation, which involves overlaying high-resolution CT or MRI image models acquired before surgery with real-time intraoperative endoscopic videos. This technology gives doctors "X-ray vision," enabling them to see blood vessels and nerves hidden beneath the surface of tissues. The key to achieving this function lies in multimodal registration. Zhang et al. first applied deep hashing (DH) technology to the registration of preoperative CT and intraoperative video images in laparoscopic liver surgery, aiming to provide automatic and robust initial alignment for augmented reality (AR) [14]. This study used a content retrieval (CBIR) framework to pre-render a large number of liver views from CT and extract contour features, and then used a DH network trained with a triplet scheme to encode them into low-dimensional hash codes. During the operation, the system performed contour segmentation and encoding on the endoscopic video images, and quickly found matching hash codes through nearest neighbor search, thereby achieving accurate registration. This method effectively replaced time-consuming manual initialization (average 4.5 seconds vs 2 minutes 30 seconds) and achieved clinically relevant alignment in seven out of eight patients, significantly enhancing the clinical translational potential of AR in liver surgery.

5.3 Cross-modal Sensing

Most existing commercial surgical robots lack tactile feedback, and doctors cannot perceive the reaction force of tissues during operation, which can easily lead to tissue tearing due to excessive force. Since force sensors are difficult to deploy on minimally invasive instruments, researchers have begun to explore vision-based force perception solutions. The "Deep Tactile Guidance" (DHG) system proposed by Fekri et al. was the first to apply deep learning to the transfer of skills in bone drilling surgery[15]. This system innovatively integrates heterogeneous data from multiple sources, including physical characteristics such as drill position, penetration depth, simulated drill temperature, and bone layer type, rather than relying solely on single force feedback data. By adopting an improved LSTM architecture, DHG can effectively learn the dynamic behavior of expert surgeons under these multimodal inputs and predict accurate tactile force feedback. This sensor fusion mechanism significantly improves the accuracy and robustness of guidance signals, providing novice doctors with a more comprehensive and realistic training experience, thereby promoting the remote transfer and learning of surgical skills under special restrictions such as the epidemic. Masui et al. aimed to address the challenge of organ damage caused by the lack of tactile feedback in robot-assisted laparoscopic surgery and proposed a deep learning-based visual method to estimate instrument manipulation forces in real time [16]. This study was the first to integrate a triaxial force sensor into a real surgical forceps and used a clinical endoscopy system to record manipulation images and synchronized force data of isolated pig kidneys. They demonstrated that limiting the region of interest (ROI) of the image to the tip of the instrument (detected by the YOLO algorithm) could significantly improve the estimation accuracy when performing force estimation. Compared with the “non-clipping model” using the complete surgical scene, the “clipping model” showed a significantly improved hyperforce detection classification accuracy (CA) and the estimated force waveform was closer to the actual measured value. Although there is still room for improvement in the mean absolute error (MAE) of force amplitude estimation, the method achieved real-time hyperforce alarm within a computational delay of 20 milliseconds, laying the foundation for the development of surgical safety control devices and helping to clarify “pseudo-tactile feedback” and apply it to surgeon training.

6 Conclusion

This article provides a comprehensive overview of the latest advancements and transformations in 3D vision technology within the field of precision medical robotics. From early monocular and binocular vision foundations to today's deep learning-enabled 3D reconstruction, the field is undergoing a profound shift from "geometry-driven" to "data-driven." In particular, the emergence of neural rendering technologies such as EndoNeRF and 3D Gaussian Splatting has provided revolutionary solutions to the long-standing problems of glare interference and non-rigid deformation in surgical scenarios. Simultaneously, the optimization of training strategies such as self-supervised learning and Sim-to-Real has effectively alleviated the bottleneck of data annotation difficulties in the medical field, making it possible to implement algorithms in real-world clinical environments.

Despite significant technological breakthroughs, numerous challenges remain on the road to clinical application. First, the conflict between real-time performance and computational power persists. High-fidelity neural rendering algorithms typically require expensive GPU computing resources, which contradicts the low-power and integrated requirements of mobile medical devices. Deploying lightweight models on edge computing devices is a key area for future research. Second, robustness in extreme scenarios needs improvement. Existing visual algorithms often fail in extreme environments such as severe bleeding, dense smoke, or heavily contaminated lenses. Future research should further explore multi-sensor fusion schemes to enhance the system's perception capabilities in harsh conditions.

3D vision technology forms the "eyes" and "brain" of precision medical robots. With the deep integration of the global perception capabilities of Transformer and the semantic understanding capabilities of EndoGSLAM, future medical robots will no longer be merely extensions of doctors' hands. They will possess the ability to understand anatomical structures, predict tissue deformation, and make intelligent decisions. This technological advancement will propel medical robots from "machine-assisted surgery" to "intelligent autonomous surgery," ultimately achieving safer, more precise, and more efficient medical services.

References

[1] Y. Gu, J. Lv, J. Bo, B. Zhao, Y. Chen, J. Tao, Y. Qin, W. Wang, J. Liang, Joint dense 3D reconstruction method for endoscopic images of weak texture scenes. IEEE Access 9, 138254–138266 (2021)

[2] L. Lipson, Z. Teed, J. Deng, RAFT-Stereo: Multilevel recurrent field transforms for stereo matching. arXiv preprint arXiv:2109.07547 (2021)

[3] J.A. Barragan, J. Zhang, H. Zhou, A. Munawar, P. Kazanzides, Realistic data generation for 6D pose estimation of surgical instruments, in Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 13347–13353 (2024)

[4] J. Rothfuss, B. Sukhija, L. Treven, F. Dörfler, S. Coros, A. Krause, Bridging the sim-to-real gap with Bayesian inference. arXiv preprint arXiv:2402.0xxxx (2024)

[5] W. Li, Y. Hayashi, M. Oda, T. Kitasaka, K. Misawa, K. Mori, EndoSelf: Self-supervised monocular 3D scene reconstruction of deformable tissues with neural radiance fields on endoscopic videos, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2024, Lecture Notes in Computer Science, vol. 15006 (Springer Nature Switzerland, Cham, 2024), pp. 241–251

[6] Hatamizadeh, Y. Tang, V. Nath, D. Yang, A. Myronenko, B. Landman, H.R. Roth, D. Xu, UNETR: Transformers for 3D medical image segmentation, in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 574–584 (2022)

[7] Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, D. Xu, Swin UNETR: Swin Transformers for semantic segmentation of brain tumors in MRI images. arXiv preprint arXiv:2201.01266 (2022)

[8] Y. Wang, Y. Long, S.H. Fan, Q. Dou, Neural rendering for stereo 3D reconstruction of deformable tissues in robotic surgery, in Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 (Springer Nature Switzerland, Cham, 2022), pp. 431–441

[9] Y. Liu, C. Li, C. Yang, Y. Yuan, EndoGaussian: Real-time Gaussian splatting for dynamic endoscopic scene reconstruction. (2024)

[10] L. Qiu, H. Ren, Endoscope navigation with SLAM-based registration to computed tomography for transoral surgery. Int. J. Intell. Robot. Appl. 4, 252–263 (2020)

[11] J. Lamarca, S. Parashar, A. Bartoli, J.M.M. Montiel, DefSLAM: Tracking and mapping of deforming scenes from monocular sequences. IEEE Trans. Robot. 37, 291–303 (2021)

[12] K. Wang, C. Yang, Y. Wang, S. Li, Y. Wang, Q. Dou, X. Yang, W. Shen, EndoGSLAM: Real-time dense reconstruction and tracking in endoscopic surgeries using Gaussian splatting. (2024)

[13] M. Daneshgar Rahbar, G. Pappas, N. Jaber, Toward intraoperative visual intelligence: Real-time surgical instrument segmentation for enhanced surgical monitoring. Healthcare 12, 1112 (2024)

[14] H. Zhang, S. Bulathsinhala, B.R. Davidson, M.J. Clarkson, J. Ramalhinho, Deep hashing for global registration of preoperative CT and video images for laparoscopic liver surgery. Int. J. Comput. Assist. Radiol. Surg. 20, 1461–1469 (2025)

[15] P. Fekri, J. Dargahi, M. Zadeh, Deep learning-based haptic guidance for surgical skills transfer. Front. Robot. AI 7 (2021)

[16] K. Masui, N. Kume, M. Nakao, T. Magaribuchi, A. Hamada, T. Kobayashi, A. Sawada, Vision-based estimation of manipulation forces by deep learning of laparoscopic surgical images obtained in a porcine excised kidney experiment. Sci. Rep. 14, 9686 (2024)