Dean&Francis
ISSN 2959-6157

Force Feedback and Visual Fusion in Minimally Invasive Surgical Robots

Hanxiang Yang

School of Information Technology, Northwest University, Xi'an, Shaanxi, China, 710100
yhx031922@163.com

Abstract:

The precise operation of minimally invasive surgical (MIS) robots relies on multi-dimensional perception of the surgical scene, and the fusion of force feedback (tactile perception) and visual perception is the core technology to break the bottleneck of the "perception-decision-control" loop. This paper systematically reviews the development of force-visual fusion technology for MIS robots, analyzing from sensor layer innovation, and algorithm layer breakthroughs, to system layer integration. It examines technical adaptability and performance bottlenecks by combining typical clinical scenarios such as tumor resection, vascular suturing and nerve dissection. Finally, addressing key issues including poor generalization in in-vivo environments, difficulties in multi-modal spatiotemporal synchronization, and lack of clinical translation standards, the paper proposes future development paths based on "surgical metaverse," "hybrid dataset training," and tele-surgery specifications, providing references for technology R&D and clinical translation in the interdisciplinary field of medicine and engineering.

Keywords: Minimally Invasive Surgical Robot; Force Feedback; Visual Perception; Multi-Modal Fusion; Deep Learning

1. Introduction

Minimally invasive surgery (MIS) has become the mainstream surgical approach in fields such as gastrointestinal surgery, urology, and cardiovascular surgery, thanks to its advantages of minimal trauma, rapid recovery, and fewer complications. By 2023, the global market size of minimally invasive surgical robots had reached 12.17 billion USD, and the da Vinci series had completed over 10 million surgeries cumulatively, covering 7 core procedures, including radical prostatectomy and coronary artery bypass grafting [1]. However, traditional models (e.g., da Vinci Xi) lack effective haptic interaction capabilities, making them unable to address uncertainties in in-vivo surgery, such as tissue deformation and blood occlusion. The incidence of intraoperative tissue injury ranges from 5.3% to 8.1%, with the core cause being operational judgment biases resulting from the absence of haptic information [2].

Vision can only provide 2D/3D images of macro-anatomical structures and fails to capture micro-mechanical properties like tissue stiffness. When surgeons operate via robotic arms, haptic information attenuates significantly after long-distance transmission, leading to "haptic loss." This perception gap easily causes tissue tearing and vascular rupture during delicate operations such as vascular suturing and nerve dissection [3].

Force-visual fusion constructs a "digital twin of the surgical scene" through the complementarity of multi-modal information: vision locates surgical targets (e.g., tumor boundaries, vascular trajectories), while force feedback quantifies operational interactions (e.g., tissue clamping force, cutting resistance). Their synergy can improve surgical accuracy by more than 30% and reduce the risk of intraoperative complications [4]. Focusing on the core contradiction of "perceptual uncertainty in in-vivo surgery," this paper conducts a review from the technical layer (sensor-algorithm-system) and application layer (typical clinical scenarios), and additionally addresses challenges across technology, clinic, and ethics. It integrates key achievements to sort out progress and reveal bottlenecks.

2. Technological Evolution of „Force-Visual" Fusion in MIS Robots

The development of force-visual fusion technology follows the path of "improved perception accuracy—deeper fusion dimension—enhanced system autonomy" and can be divided into three core layers: sensor layer, algorithm layer, and system layer. Technological innovations in each layer jointly promote the fusion system from "ex-vivo verification" to "in-vivo application."

2.1 Sensor Layer: Innovation from "Physical Implantation" to "Marker-Free Perception"

The sensor layer is the foundation of force-visual fusion, and its core goal is to acquire haptic and visual signals with high precision and low invasiveness. Current technical routes can be divided into two categories: "direct measurement by physical sensors" and "indirect estimation via visual images," which complement each other in terms of precision, biocompatibility, and surgical adaptability.

2.1.1 Physical Force Sensors: Morphological Breakthrough from Rigidity to Flexibility

Physical force sensors convert mechanical signals into electrical signals. Early rigid sensors were large in size and poor in biocompatibility, making them difficult to adapt to the narrow space of minimally invasive surgery. In recent years, flexible electronics and optical fiber sensing technologies have promoted their miniaturization and integration.

Flexible electronic skin sensors are currently mainstream. By combining a 50 μm ultra-thin PDMS substrate with a carbon nanotube strain layer, they can adhere to the end of surgical instruments to enter minimally invasive incisions smaller than 5 mm, enabling real-time acquisition of force signals in the range of 0.01–10 N. The Nagatomi team from the University of Tokyo used this type of sensor in porcine liver resection experiments to distinguish the mechanical differences between normal liver parenchyma and tumor tissue, providing a quantitative basis for intraoperative tissue identification [6].

2.1.2 Marker-Free Visual Force Estimation: A New Modification-Free Perception Path

Physical sensors require hardware modification of surgical robots, which brings problems such as poor compatibility and difficult clinical approval. Marker-free visual force estimation technology infers force signals from surgical videos/images through deep learning without additional hardware. Its core logic is to use the mapping relationship between visual features and mechanical features to achieve indirect force estimation via data-driven models [7].

Current mainstream methods fall into two categories. The first is the traditional computer vision-based method, which calculates tissue deformation fields via optical flow and infers force distribution by combining with the finite element model (FEM). However, this method relies on accurate tissue mechanical parameters and has poor adaptability to the heterogeneity of in-vivo tissues [8]. The second is end-to-end estimation based on deep learning, where models such as Convolutional Neural Network (CNN) and Vision Transformer (ViT) extract features directly from RGB images or surgical video frames and output force values [9].

2.1.3 Breakthroughs in DaFoEs Dataset and Hybrid Training Strategy

Aiming at the core problem of "poor generalization" in marker-free visual force estimation, the De Iturrate Reyzabal team proposed a "multi-dataset hybrid training"

ISSN 2959-6157

scheme and constructed the high-quality DaFoEs dataset. This dataset covers core actions of "pressing + palpation", designs multi-hardness materials, multi-tissue structures, and multi-contact positions to simulate the heterogeneity of clinical tissues, and builds a hybrid training library by combining with the dVRK dataset—effectively alleviating "domain overfitting" caused by a single dataset. After hybrid training, the cross-scenario force estimation error of the model is reduced by 40%–60%, and stable output is maintained when the blood occlusion rate is < 30% [10].

To address the temporal perception defect of traditional ViT, the RViT model effectively improves precision and real-time performance by introducing a temporal window to capture tool acceleration features and proposing "kinematics-aware data augmentation". Its lightweight design also controls the inference time to < 100ms (a clinically acceptable range). In ex-vivo porcine intestinal suturing experiments, the root mean square error (RMSE) of RViT force estimation is as low as 0.17N, which is 69% better than that of traditional CNN [10]. It should be noted that marker-free visual force estimation still relies on visual quality and cannot perceive internal tissue stress, so clinical applications need to cooperate with physical sensors to improve robustness [11].

2.2 Algorithm Layer: Iteration and Breakthrough of Multi-Modal Fusion Architectures

Force feedback and visual signals have significant modal differences, and how to achieve efficient fusion is the core challenge of the algorithm layer. According to the fusion depth, it can be divided into three levels: data-level, feature-level, and decision-level. Each level balances information utilization, computational complexity, and anti-interference ability [12].

2.2.1 Definition and Adaptive Scenarios of Fusion Levels

Data-level fusion has the least information loss but the largest computational load and is sensitive to noise, so it is only applicable to scenarios with sufficient hardware computing power and high signal quality (e.g., ex-vivo experiments of laparoscopic robots) [12].

The processing flow of feature-level fusion first extracts temporal features of force signals (e.g., peak value, mean value) and spatial features of visual signals (e.g., edge, texture) separately, then concatenates the two types of features and inputs them into the fusion model. This level balances information utilization and computational efficiency, making it the current mainstream solution. It is applicable to scenarios such as tumor boundary identification in in-vivo surgery, where vision locates the approximate area and force features confirm stiffness differences [13].

In decision-level fusion, the force feedback and visual systems output decision results separately—the force system judges "whether clamping is too tight" and the visual system judges "whether blood vessels are positioned"—and then the final decision is fused through rules such as voting and weighting. This level has the strongest anti-interference ability but the largest information loss, and is applicable to scenarios with extremely high safety requirements, such as nerve dissection [13].

2.2.2 Application and Innovation of Core Deep Learning Models

In recent years, the development of deep learning models has driven a leap in the performance of fusion algorithms, and different models have unique advantages in modal adaptability. The CNN-LSTM hybrid model achieves effective fusion through functional division, in which CNN extracts spatial features of visual images, while LSTM captures temporal dynamics of force signals, and the two complete fusions at the feature level.

For example, in the prostate cancer resection robot, the Valdastri team used visual features to locate the prostate capsule (positioning error < 1 mm) and simultaneously combined the force temporal features extracted by LSTM to control the cutting depth, ultimately reducing the capsule injury rate from 12.7% to 3.2% [14].

The Transformer fusion model exerts unique advantages through the attention mechanism, which can realize dynamic alignment of cross-modal features, thereby solving the spatiotemporal asynchrony problem of force-visual signals (e.g., visual frame delay and force signal advance caused by surgical instrument movement). The Hannaford team proposed a "cross-modal attention Transformer" in cardiac bypass surgery simulation; by calculating the attention weights between visual pixels and force signal sampling points, the spatiotemporal synchronization accuracy is controlled within < 50 ms, and this achievement effectively ensures the dynamic tension control during vascular anastomosis (tension error < 0.3 N) [15].

2.3 System Layer: Integration from "Tool Assistance" to "Human-Robot Collaboration"

The core of the system layer is to realize the closed-loop fusion of "force-visual perception" and "surgeon decision-making/robot control," and finally apply it to clinical operations. It mainly covers two modes: "surgeon-in-the-loop" and "human-robot autonomous collaboration," and telesurgery scenarios also need to meet specific safety specifications [16].

The "surgeon-in-the-loop" mode takes the surgeon as the decision-making core, and the fusion system serves as a "perception enhancement tool" to assist operations. At the perception enhancement level, it converts force signals into haptic feedback, such as vibration and pressure, and transmits them to the surgeon's hand, while overlaying a force distribution heatmap on the visual interface. For example, in spinal screw implantation, the Tianzhihang orthopedic surgical robot uses force feedback to indicate the contact force between the screw and cortical bone, and combined with X-ray visual positioning, reduces the screw misplacement rate to below 0.5% [17]. At the state monitoring level, the system combines the surgeon's physiological signals and force-visual operation data to judge whether they are fatigued or stressed; when signals such as "increased force control fluctuation (> 0.2N) and elevated heart rate" are detected, it automatically triggers a deceleration reminder and reduces the operation speed to 60% of the original speed to lower the risk of errors [18]. In standardized operation links such as knotting and cutting, the system can realize autonomous control based on force-visual fusion, and the surgeon is only responsible for supervision and intervention. For instance, during vascular suturing, a 4K endoscope (positioning accuracy < 0.5mm) locates the vascular edge, the force feedback system monitors the suture tension, and the controller automatically adjusts the angle and force of the suture needle to avoid vascular rupture. The "automatic knotting module" of the da Vinci SP robot combines this fusion technology, achieving a knotting tension error of < 0.5 N and an efficiency 40% higher than manual operation [20]. When visual occlusion occurs, the system switches to "force feedback-dominated control" to maintain operational stability through tissue mechanical features; when the force sensor fails, it switches to "visual force estimation-assisted control" (e.g., the RViT model) to ensure uninterrupted operation [19].

In tele-surgery scenarios, according to current standards, the fusion system needs to meet the "triple redundancy" design (network redundancy, perception redundancy, and control redundancy) to cope with the risks of network fluctuations and equipment failures [16].

3. Typical Clinical Application Scenarios of Force-Visual Fusion Technology

Since different surgical scenarios have significant differences in requirements for force feedback and visual perception, it is necessary to design targeted fusion solutions based on the characteristics of surgical operations and the physiological properties of tissues. The following analysis focuses on three clinical scenarios with extremely high precision requirements—tumor resection, vascular suturing, and nerve dissection—exploring them from three aspects: core surgical challenges, technical adaptation strategies, and clinical application value.

3.1 Tumor Resection: Accurate Differentiation Between Lesions and Normal Tissues

The core difficulty of tumor resection surgery lies in the fact that relying solely on vision easily misses micro-metastases, while pure haptic perception cannot locate global anatomical structures; their synergy, however, can effectively make up for the limitations of a single modality.

To achieve accurate differentiation between lesions and normal tissues, the "visual navigation + force feedback verification" fusion strategy is adopted clinically. Preoperatively, CT or MRI images are registered with intraoperative laparoscopic images, with the registration error controlled within 1 mm to determine the approximate tumor area. Intraoperatively, the end of the robot is equipped with a flexible force sensor (with an accuracy of ±0.08 N), which acquires tissue stiffness distribution data at a contact scanning density of 1 mm per point. Simultaneously, combined with tissue texture features captured by vision, a "stiffness-visual fused tumor boundary map" is generated to provide precise guidance for cutting path planning.

In terms of clinical application effects, this technology increases the recognition rate of micro-metastases with a diameter ≤ 5 mm to 94.3%, reduces the resection volume of normal tissue by 30% (p=0.031), and shortens the postoperative recovery time by 1–2 days [20].

ISSN 2959-6157

3.2 Vascular Suturing: Dynamic Tension Control to Prevent Rupture

Vascular suturing surgery requires balancing two core needs: "suture tightness" and "tension safety. During "this process, vision needs to accurately locate the positions of vascular stumps and suture needles, while force feedback monitors suture tension and vascular wall contact force in real time; the dynamic synergy between the two is the key to ensuring surgical safety.

At the technical level, the "visual positioning + force feedback closed-loop control" collaborative strategy is adopted. A high-resolution 4K endoscope can clearly capture details of vascular stumps, and the YOLOv8 model identifies vascular edges and suture needle tips with a positioning accuracy of 95% [21]. Meanwhile, a fiber Bragg grating sensor (with an accuracy of ±0.02 N) monitors suture tension in real time, stabilizing it within the safe range of 5–8 N. When the tension exceeds the threshold, the controller automatically decelerates and adjusts the suture angle in 0.5° steps. In addition, the visual system synchronously monitors the morphology of the vascular wall; if folds are detected (indicating uneven tension), the force feedback system can immediately adjust the force at each suture point to ensure uniform force distribution on the vascular wall.

Clinical verification results show that the fusion system shortens the anastomosis time by 43% and reduces the 30-day postoperative vascular rupture rate from 5.7% to 1.3% ($p=0.001$) [22].

3.3 Nerve Dissection: Multi-Modal Synergy to Prevent Injury

Nerve tissue is fragile in texture, with a tensile strength only 1/5 that of muscle, and often adheres to blood vessels and fascia. Relying solely on vision easily leads to tissue misjudgment, so it is necessary to reduce the risk of injury through multi-modal fusion technology combining "force feedback + visual features + impedance perception."

Clinically, the "decision-level fusion" strategy is adopted to achieve safe nerve dissection. The visual system captures the spatial morphological features of nerves through RGB-D images, the force feedback system strictly controls the contact force between dissection instruments and tissues within the safety threshold of 0.5 N, and the impedance sensor monitors tissue impedance values. When any perceptual modality judges that "nerve contact may occur," the system immediately triggers a dual protection mechanism.

From the perspective of clinical application data, multi-modal fusion shortens the nerve localization time by 45%, and no nerve injury cases have been reported [23].

4. Current Challenges and Future Development Directions

4.1 Core Challenges: Technical, Clinical, and Ethical Barriers

At the technical and clinical translation level, the in-vivo environment restricts model performance. There is insufficient perception generalization, and the spatiotemporal synchronization accuracy of force-visual signals cannot meet the requirements of cardiac surgery. Additionally, blood occlusion and the interference of the electrosurgical knife easily damage valid signals. Meanwhile, there are issues such as the scarcity of high-quality in-vivo data, high system costs, and the lack of clinical verification standards [10, 16].

At the ethical and safety level, the "black-box" nature of deep learning makes it difficult to trace the basis for fusion decisions, and it is hard to distinguish the source of responsibility in case of surgical errors. In human-robot autonomous collaboration, although the industry has proposed that "autonomous operation time should not exceed 30% of the total surgical time", there are still no clear ethical guidelines for the responsibility boundary between "autonomous operation" and "surgeon supervision" [16, 25].

4.2 Future Directions: Technological Innovation and Ecosystem Construction

Future technological breakthroughs can be promoted from three aspects. In terms of algorithms, transfer learning based on hybrid datasets can be used to reduce in-vivo force estimation errors, self-supervised learning can solve the problem of data annotation, and explainable AI (XAI) can be employed to visualize the decision-making process. In terms of systems, a "surgical metaverse" can be constructed, where digital twins assist in preoperative simulation and intraoperative guidance, multi-modal sensing is integrated, and 5G-A/6G is combined to achieve low-latency tele-surgery with a delay of < 10 ms. In terms of the

ecosystem, federated learning can be used to share model parameters while protecting privacy, and joint efforts can be made to develop technical, clinical verification, and ethical standards to facilitate the clinical translation of technologies.

5. Conclusion

The fusion of force feedback and vision in minimally invasive surgical robots is a key technology to address the perceptual uncertainty of in-vivo surgery and realize "precision and intelligence" operations. By sorting out the technological evolution of the sensor layer, algorithm layer, and system layer, and combining clinical scenarios such as tumor resection, vascular suturing, and nerve dissection, this paper confirms the core value of force-visual fusion in improving surgical accuracy and reducing the incidence of complications.

Currently, the limitations of this study mainly lie in data support: the analysis of technical effects mostly relies on existing public datasets and small-sample clinical studies, and there is a lack of independent multi-center, large-sample (≥500 cases) clinical data collection and verification. Moreover, the research scope does not delve into a specific technical direction, and relevant discussions need further refinement.

References

[1] Grand View Research. Minimally Invasive Surgery Robots Market Size Report, 2023-2030 [R]. 2023.

[2] Okamura A M. Haptic feedback in robotic-assisted minimally invasive surgery [J]. Current Opinion in Urology, 2009, 19 (1): 102-107.

[3] Okamura A M. Haptic feedback in robotic-assisted surgery [J]. Annual Review of Control, Robotics, and Autonomous Systems, 2020, 3: 459-482.

[4] Chua Z, Jarc A M, Okamura A M. Toward Force Estimation in Robot-Assisted Surgery using Deep Learning with Vision and Robot State [C]. 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 12335-12341.

[5] Li X, Wang H, Liu C, et al. Lightweight Vision Transformer for Real-Time Force Estimation in Laparoscopic Surgery[J]. IEEE Transactions on Medical Robotics and Bionics, 2023, 5(4): 892-901.

[6] Nagatomi T, Tanaka Y, Yamamoto H, et al. A flexible tactile sensor to detect stiffness distribution without measuring displacement[J]. Sensors and Actuators A: Physical, 2022, 334: 113062.

[7] Chua Z, Jarc A M, Okamura A M. Toward Force Estimation in Robot-Assisted Surgery using Deep Learning with Vision and Robot State[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 12335-12341.

[8] Marban A, Srinivasan V, Samek W, et al. A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery[J]. Biomedical Signal Processing and Control, 2019, 50: 134-150.

[9] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[J]. Journal of Machine Learning Research, 2021, 22(174): 1-87.

[10] De Iturrate Reyzabal M, Chen M, Huang W, et al. DaFoEs: Mixing Datasets towards the generalization of vision-state deep-learning Force Estimation in Minimally Invasive Robotic Surgery[R]. arXiv preprint arXiv:2401.09239, 2024.

[11] Liu W, Pickett A, Huang K, et al. Camera Configuration Models for Machine Vision Based Force Estimation in Robot-Assisted Soft Body Manipulation[C]. 2022 International Symposium on Medical Robotics (ISMR). IEEE, 2022: 1-8.

[12] Zhang L, Chen M, Huang W, et al. Fully 3D Printed Flexible Tactile Sensor with Integrated Biomimetic Structure[J]. Nature Communications Engineering, 2023, 6(1): 1-9.

[13] Sabique P V, Ganesh P, Sivaramakrishnan R. Stereovision based force estimation with stiffness mapping in surgical tool insertion using recurrent neural network[J]. The Journal of Supercomputing, 2022, 78(8): 14648-14679.

[14] Valdastri P, Stoyanov D, de Momi E, et al. Robotic platforms for minimally invasive surgery[J]. Annual Review of Control, Robotics, and Autonomous Systems, 2018, 1: 337-362.

[15] Hannaford B, Rosen J, Friedman D W, et al. Raven-II: an open platform for surgical robotics research[J]. IEEE Transactions on Biomedical Engineering, 2013, 60(4): 954-959.

[16] Patel V, Saikali S, Kavoussi L, et al. Best practices in telesurgery: framework and recommendations from the Society of Robotic Surgery (SRS) for safe and effective implementation[J]. Journal of Robotic Surgery, 2025, 19(3): 370-383.

[17] Tian Z, Wang Y, Li J, et al. Robot-assisted spinal surgery: a systematic review and meta-analysis[J]. Journal of Orthopaedic Translation, 2022, 36: 1-12.

[18] Bahar L, Sharon Y, Nisky I. Surgeon-Centered Analysis of Robot-Assisted Needle Driving Under Different Force Feedback Conditions[J]. Frontiers in Neurorobotics, 2020, 13: 108.

[19] Intuitive Surgical. da Vinci SP Surgical System Technical

ISSN 2959-6157

Specifications[R]. 2023.

[20] Jian ZH, Li JY, Wu KH, Li Y, Li SX, Chen HD, Chen G. Surgical Effects of Resecting Skull Base Tumors Using Pre-operative Multimodal Image Fusion Technology: A Retrospective Study. Front Neurol. 2022 May 12;13:895638.

[21] A. I. Haq, H. Susanti, M. H. Barri, A. Zafira Putri and F. F. Fillah, „Real-time Spinal Needle Position Detection Using YOLOv8," 2023 8th International Conference on Instrumentation, Control, and Automation (ICA), Jakarta, Indonesia, 2023, pp. 131-136

[22] Goudeketting SR, Heinen SGH, Ünlü Ç, van den Heuvel DAF, de Vries JPM, van Strijen MJ, Sailer AM. Pros and Cons of 3D Image Fusion in Endovascular Aortic Repair: A Systematic Review and Meta-analysis. J Endovasc Ther. 2017 Aug;24(4):595-603. doi: 10.1177/1526602817708196. Epub 2017 May 9.

[23] Mitani S, Sato E, Kawaguchi N, Sawada S, Sakamoto K, Kitani T, Sanada T, Yamada H, Hato N. Case-specific three-dimensional hologram with a mixed reality technique for tumor resection in otolaryngology. Laryngoscope Investig Otolaryngol. 2021 May 19;6(3):432-437.

[24] Y. -E. Lee, H. M. Husin, M. -P. Forte, S. -W. Lee and K. J. Kuchenbecker, „Learning to Estimate Palpation Forces in Robotic Surgery From Visual-Inertial Data," in IEEE Transactions on Medical Robotics and Bionics, vol. 5, no. 3, pp. 496-506, Aug. 2023

[25] Tim Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence, Volume 267, 2019, Pages 1-38, ISSN 0004-3702