Surgical Tool Detection on CholecTrack20 Using Lightweight Deep Learning Models

Yuteng Zhao

Creative Computing Institute, University of the Arts London, London, SE5 8UF

y.zhao0820232@arts.ac.uk

Abstract. With the rapid development of minimally invasive surgery (MIS) and robot-assisted surgery, real-time, accurate, and robust surgical instrument detection and tracking has become a core research focus in medical AI. This review summarizes the current state of surgical tool detection and tracking, with a particular focus on the application of lightweight deep learning models on the CholecTrack20 dataset. Models such as MobileNetV2 and YOLOv8n demonstrate promising deployment potential and performance on embedded and low-computation platforms. We analyze their advantages and limitations in small-object detection, occlusion handling, multi-view tracking, and end-to-end real-time inference, and discuss potential improvements through multi-frame fusion, feature pyramid networks, lightweight attention modules, and data augmentation strategies. Furthermore, future research directions are outlined, including multimodal perception (vision + tactile/force feedback), explainable AI (XAI), and uncertainty estimation to ensure clinical safety and regulatory compliance. Overall, lightweight models offer practical deployment value for surgical tool detection and tracking, and provide a feasible pathway toward intelligent, multimodal surgical systems.

Keywords: Minimally Invasive Surgery, Surgical Tool Detection and Tracking, Lightweight Deep Learning.

1. Introduction

With technological advances, minimally invasive surgery (MIS) and robot-assisted surgery have become central trends in modern healthcare [1]. The evolution from traditional open surgery to endoscopic surgery, minimally invasive surgery (MIS), and ultimately robot-assisted surgery reflects a continuous drive toward greater precision, intelligence, and minimal invasiveness. The development of medical robotics and MIS has increased the demand for real-time visual assistance systems, which determine the operator's or robot's perception of anatomical structures and instrument states [2]. Such systems are fundamental for precise operation, navigation guidance, and immediate feedback.

Recent advances in deep learning–based surgical tool detection and tracking have progressed from single-frame recognition to multi-frame, temporally consistent tracking, moving from basic 2D detection toward 3D localization and pose estimation [3, 4]. Nevertheless, significant challenges remain: complex surgical environments involving multiple instruments, occlusions, and varying lighting conditions pose dual challenges to accuracy and real-time performance [5]. Additionally, there is a critical need for lightweight models suitable for clinical deployment, highlighting the importance of research on compact architectures and edge-device optimization to enhance performance in real-world surgical settings [6].

The CholecTrack20 dataset provides a dedicated video dataset for multi-class surgical instrument detection and tracking in laparoscopic surgery, providing high-quality, multi-view annotations of surgical tools, including spatial positions, classes, identities, operator information, surgical phases, and visual challenge labels [7]. It offers three distinct tracking perspectives to meet diverse clinical requirements, making it a valuable resource for research on surgical tool tracking and AI-assisted surgical systems.

In the following sections, this review initially presents the task of surgical instrument detection and tracking, detailing the required hardware and software support as well as the primary technical challenges. It then outlines the development of methods, from conventional image processing and hand-engineered features to contemporary deep learning techniques in the form of CNNs, RNNs, and

Transformers. Practical application demands within navigation, automation, and surgical training are then debated, followed by a review of the CholecTrack20 dataset and its benchmarking significance. Comparative experimentation of six lightweight models is then outlined, paying close attention to MobileNetV2 and YOLOv8n. The review concludes by noting current limitations and suggesting future research avenues, encompassing lightweight optimization, multimodal fusion, and explainability toward safe clinical translation.

This review focuses on the rapid development of MIS and robot-assisted surgery and examines how lightweight deep learning models, such as MobileNetV2 and YOLOv8n, can be effectively and accurately applied to CholecTrack20 for real-time surgical instrument detection and tracking. It also discusses model performance, deployment potential, and current research challenges.

2. Surgical Tool Detection and Tracking

2.1. Task Definition and Core Challenges

Artificial intelligence (AI) has been increasingly applied to surgical instrument detection and tracking, a core interdisciplinary research area combining computer vision, robotics, and medical engineering [8]. The goal is to achieve real-time, precise, and robust identification and localization of instruments during surgery, supporting surgical assistance systems, skill assessment, and safety monitoring. This integration is a cornerstone toward intelligent surgical ecosystems in which humans and machines cooperate.

Hardware support is critical for ensuring system performance and reliability. Essential components include: (1) vision sensors, eIntel RealSense D435 depth camera for high-precision 3D localization and tracking, or AR HMD cameras integrated in the STTAR framework with reflective markers for real-time tracking and eye-motion-controlled cameras [9, 10]; (2) robotic arms and end-effectors, Senhance surgical robot system, combining vision guidance for accurate tool manipulation, supported by control platforms such as NVIDIA Isaac for Healthcare for integrating custom arms and sensors [11, 12]; (3) processing platforms and computational resources, e.g., NVIDIA Jetson embedded GPU platforms for real-time image processing and deep learning inference, and RTOS such as Linux-based RMP for multi-axis motion, vision, and AI integration [13, 14]; (4) localization and tracking systems, optical tracking in ROSA knee systems using reflective markers to precisely localize tools and patient anatomy [15]. These hardware and computational components form the cornerstone of AI-supported surgical platforms and thus ensure efficiency and reliability.

Surgical tool detection and tracking is inherently interdisciplinary, requiring real-time video analysis, precise robotic control, and adherence to clinical safety standards [16]. Core challenges include real-time processing, accuracy, robustness, and domain adaptation. Surgical videos are typically 1080p–4K at 30–60 fps, with tools moving rapidly, frequent occlusions, and articulated joints; models must maintain ≥25–30 fps while providing stable, low-jitter detection and tracking [17]. Articulated instruments pose precision challenges due to similar appearance, elongated shape, and fine details, necessitating pixel-level boundaries, keypoints, poses, and even 6D localization [18]. Real surgical environments include smoke, blood, reflections, high-frequency jitter, rapid deformations, and occlusions, with distribution shifts across surgical phases. Variations in hospitals, devices, procedures, and patient populations can significantly degrade cross-domain performance, highlighting the need for robust and adaptive models [19]. These variations in hospitals, devices, procedures, and patient populations significantly degrade cross-domain performance, underscoring the pressing need for robust, adaptive, and clinically validated AI models.

2.2. Technical Evolution

Surgical instrument detection and tracking has evolved from rule-based handcrafted feature methods (traditional image processing) to deep learning-driven, multi-task, end-to-end approaches, with technical development gradually moving toward higher accuracy, real-time performance, robustness, and cross-domain generalization [20]. Traditional methods rely on manually designed

image features and rule-based matching, such as edge detection, Hough transforms, and color segmentation, often combined with Kalman or particle filters for multi-frame tracking. While simple, these algorithms are highly sensitive to lighting, reflections, and occlusion, exhibiting poor generalization. Subsequent research further developed these approaches by employing classifiers such as Random Forests and SVMs in combination with handcrafted features like HOG, SIFT, and SURF for binary or multi-class detection, with exploration of multi-view fusion and structured prediction to improve generalization and accuracy [21]. Since 2016, deep learning methods have achieved significant progress: convolutional neural networks (CNNs) have been widely applied to instrument detection and segmentation tasks; recurrent neural networks (RNNs) and long short-term memory networks (LSTMs) have been used to process temporal sequences for instrument tracking and surgical phase recognition; and Transformer architectures have been introduced to enhance multi-task learning and cross-frame modeling. For instance, LapTool-Net utilizes an RCNN architecture to integrate spatial and temporal contextual information, enabling efficient multi-label detection of surgical instruments in laparoscopic videos, achieving end-to-end training and inference [22].

2.3. Practical Application Requirements

Real-time detection and high-precision tracking of surgical instruments play a crucial role in surgical navigation, robot-assisted automated procedures, and surgical skill assessment. However, different clinical applications impose distinct and often stringent requirements on algorithmic speed, accuracy, robustness, and latency. In surgical navigation, real-time detection and tracking of instrument tips must achieve precise localization to ensure accurate alignment with preoperative imaging (CT/MRI) or intraoperative 3D models [23]. Low-latency augmented reality (AR) integration is also required to enable surgeons to visualize instrument positions and key anatomical structures in real time on surgical displays or AR headsets. In automated or semi-automated surgery, instrument detection and tracking results need to be fed back to the robotic control system in real time to support visual servoing or assisted operation [24]. This demands high-frequency updates of instrument positions and contingency strategies to maintain safety in cases of error, latency, or tracking loss. Moreover, in surgical training and evaluation, instrument trajectories and video data can be analyzed automatically to quantify metrics such as movement paths, speed, and tremor frequency, supporting novice surgeons in correcting improper actions or planning subsequent steps [25].

Table 1. Performance Requirements for Surgical Instrument Detection and Tracking Across Clinical Scenarios

Clinical Scenario	Speed / Throughput	Accuracy (Localization / Overlay)	End-to-End Latency (E2E)	Notes
Neurosurgery / Spinal Navigation	30–60 fps	≈1–2 mm	<100 ms	High-risk; millimeter accuracy and low latency critical; AR overlays must be stable.
Laparoscopy (Human–Robot / Semi-Automation)	30–60 fps; 4K	2–5 mm	<100–200 ms	Real-time instrument feedback required; 4K increases compute load; <100 ms latency improves human–robot interaction.
Teleoperation / Remote Surgery	≥30 fps (ITU>50 fps)	Task-safe	<100 ms ideal; <200 ms acceptable; ≤320 ms feasible	Latency >200 ms affects control and workload; optimize communication.
Microsurgery / Ophthalmology Training & Assessment	25–30 fps (higher offline)	≈1–2 mm or stricter	<200 ms intraoperative feedback; offline unrestricted	Focus on tip trajectory reconstruction and fine-grained skill assessment; offline analysis can use high-precision computation.

Table 1 shows that different clinical scenarios impose markedly distinct requirements on speed, accuracy, and end-to-end (E2E) latency: high-precision navigation demands millimeter-level localization with extremely low latency (<100ms), whereas remote and semi-automated surgeries emphasize stable real-time performance and high frame rates. Microsurgical training and postoperative assessment prioritize high-precision analysis, often allowing offline processing. This diversity of requirements highlights the critical importance of lightweight models and efficient

inference frameworks to balance real-time responsiveness, accuracy, and computational resource constraints.

3. Datasets and Benchmarking

3.1 CholecTrack20 Dataset Overview

Figure 1 illustrates the multi-view surgical instrument tracking and multidimensional annotations from the CholecTrack20 dataset. This dataset targets multi-class, multi-instrument tracking (MCMOT) in laparoscopic cholecystectomy and provides three trajectory perspectives: visible (in-screen FoV), intracorporeal, and intraoperative (spanning the entire procedure). Annotations include spatial and identity information: instrument bounding boxes (bbox), instrument categories (7 classes: grasper, bipolar, hook, scissors, clipper, irrigator, bag), trajectory IDs (provided for each perspective), and operator identity (Operator ID). Contextual labels cover surgical phase, out-of-screen/in-screen (OOCV), out-of-body/in-body (OOB), and various visual challenge tags such as occlusion, crowding, bleeding, smoke, lens contamination, reflections, blur, and trocar-view limitations. The dataset consists of 20 real surgical videos (~15 hours), with ~35k annotated frames (sampled at 1 fps) and ~65k instrument instances [26].

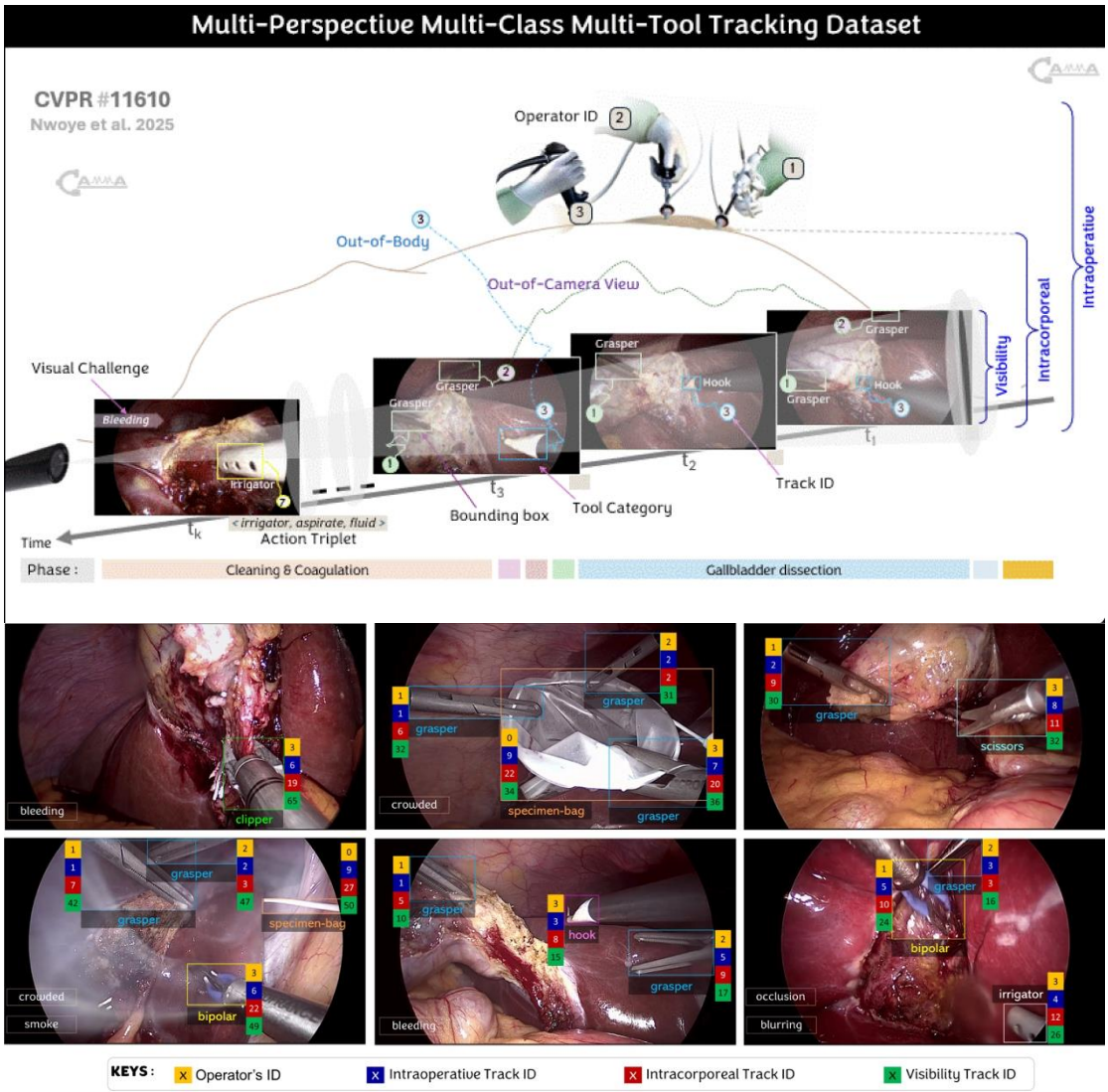

Figure 1. Schematic diagram of CholecTrack20 multi-angle surgical tool tracking and multi-dimensional annotation [26]

The high-quality, diverse annotations facilitate surgical instrument detection, tracking, and deployment of lightweight models. Surgical phase and visual challenge labels enable models to learn robust features under complex conditions. Models can be trained jointly on single-frame detection and multi-frame tracking tasks, reducing train/test distribution bias and promoting generalization for end-to-end lightweight architectures. The three perspectives and multi-scene coverage allow lightweight models to learn multi-scale features efficiently while maintaining detection accuracy across different viewpoints, making the dataset suitable for embedded systems or surgical robots. The moderate video frame rate and annotation density balance sufficient training samples with manageable inference speed and memory usage for lightweight models, supporting deployment on resource-constrained platforms.

3.2 Benchmark Results and Analysis

Table 2 presents a comparison of six lightweight models for real-time surgical tool detection in terms of architecture, efficiency, and deployment. The evaluation metrics include: Params (model parameters in millions), FLOPs (forward inference computation in Giga FLOPs), Input (input image resolution), Speed (inference frame rate, GPU/edge device reference), Deployment (suitability for embedded or lightweight platforms), and Notes (key advantages or application scenarios).

Table 2. Lightweight Model Comparison: Architecture, Efficiency, and Deployment for Real-Time Surgical Tool Detection

Model	Architecture / Type	Params (M)	FLOPs (G)	Input (px)	Speed (fps)	Deployment	Notes
MobileNetV2	Depthwise-separable CNN	3.4	0.3	224×224	50–100	TFLite, Edge	Lightweight, easy deployment
EfficientNet-B0	Compound scaling CNN	5.3	0.39	224×224	30–60	TFLite, ONNX	Good accuracy-efficiency tradeoff
YOLOv5s	CSP-Darknet + PANet	7	17	640×640	140	PyTorch, TensorRT	End-to-end detection, real-time capable
YOLOv8n	Nano variant, CNN + PANet	4.5	7.5	640×640	160	PyTorch, TensorRT	Fastest YOLO variant, suitable for embedded
Tiny-YOLOv4	CSPDarknet53-Tiny	6	6.9	416×416	220	TensorRT, Jetson Nano/Orin	Ultra-lightweight, high fps
ShuffleNetV2	Channel split + pointwise	1.0	0.14	224×224	120	Mobile/Edge	Extremely lightweight, low-power deployment

From the perspective of network complexity and computational efficiency, ultra-lightweight models such as ShuffleNetV2 and MobileNetV2 feature extremely low parameter counts (<4M) and FLOPs (<0.5G), making them suitable for resource-constrained platforms such as handheld devices or embedded ARM boards. These models are ideal for real-time scenarios or low-power edge devices but may lack precision in complex backgrounds or small-target detection. Medium-lightweight models (EfficientNet-B0) achieve a favorable balance between parameters and performance, though their inference speed is slightly lower than YOLO series models. YOLOv8n has higher computational demands (~7.5G FLOPs) but, when combined with optimized inference frameworks such as TensorRT, can achieve up to 160 fps, providing an effective trade-off between speed and accuracy. Lightweight high-performance models such as YOLOv5s and Tiny-YOLOv4 demonstrate robust real-time performance; Tiny-YOLOv4 can reach 220 fps at low resolution (416×416), though its accuracy is slightly below YOLOv8n.

Regarding deployment suitability, MobileNetV2 and ShuffleNetV2 support TFLite/edge deployment and are thus ideal for low-compute devices. Tiny-YOLOv4 and YOLOv8n are optimized for NVIDIA Jetson Nano/Orin or TensorRT platforms, offering flexible deployment. YOLOv5s and YOLOv8n run smoothly at high-resolution inputs (640×640), making them suitable for high-precision cloud inference or GPU server deployment.

From the perspective of CholecTrack20, MobileNetV2 is particularly well-suited: it has a small parameter count (3.4M), high inference speed (50–100 fps), and is ideal for embedded or resource-constrained deployment. Given the 1 fps annotation of CholecTrack20 frames, MobileNetV2 can

support end-to-end lightweight deployment. Potential improvements include integrating a NAS-FPN (Neural Architecture Search–Feature Pyramid Network) as proposed in, which can enhance multi-scale feature fusion while maintaining a lightweight design, thereby improving detection of small instruments and occluded scenarios [27].

YOLOv8n supports end-to-end object detection with high inference speed and PyTorch/TensorRT deployment, making it suitable for real-time tracking on high-resolution CholecTrack20 videos (640×640). Potential enhancements include incorporating SELSA (Sequence-Level Semantics Aggregation) to aggregate key-frame features across sequences, improving robustness under occlusion and deformation while maintaining tool ID consistency [28]. Additionally, lightweight data augmentation techniques such as MixUp allow linear mixing of samples with negligible computational overhead, enhancing generalization and robustness to visual challenges such as occlusion, smoke, and reflections [29].

Existing real-time surgical instrument detection work demonstrates that lightweight models combined with multi-scale and position-sensitive modules can achieve practical accuracy while maintaining real-time performance. For example, proposed an anchor-free transformer-based architecture for laparoscopic instrument detection, leveraging transformer layers, multi-scale positional encoding, and contrastive learning [30]. This approach achieved significant improvements: mAP increased by ~4% with an inference speed gain of ~113%, and an approximate 7% improvement over baseline models. Here, mAP (mean Average Precision) serves as a standard metric to evaluate detection accuracy across multiple categories, with higher mAP values indicating better overall precision.

4. Discussion and Future Directions

4.1 Current Limitations

On surgical instrument detection and tracking datasets such as CholecTrack20, which feature small-scale instruments and real intraoperative challenges including occlusions, tool in/out of frame, smoke, bleeding, and high specular reflections, lightweight backbones (MobileNetV2) and lightweight detectors (YOLOv8n) are deployment-friendly for embedded systems but face performance bottlenecks due to limited model generalization and dataset scale, particularly in small-object recall, ID consistency, and robustness under challenging visual conditions. G. Loza reports that MobileNetV2-based lightweight detectors (MobileNetV2+SSDLite) exhibit insufficient recall and localization accuracy for small objects [31]. MobileNetV2 employs depthwise separable convolutions and inverted residual bottlenecks to significantly reduce parameter count and computational cost, but this comes at the expense of feature representation capacity, especially in abstracting high-level semantic information. Under the strong visual noise and challenging conditions annotated in the CholecTrack20 dataset, MobileNetV2 struggles to maintain effective feature extraction. Although the YOLO series (including the latest YOLOv8) demonstrates strong performance in many medical imaging and endoscopic applications, notes that narrow training data distribution and single-device acquisition can substantially limit generalization to other clinical environments—a limitation frequently reported in validation studies [32]. While YOLOv8n benefits from transfer learning and pretraining, its performance is constrained in real surgical videos, which are highly heterogeneous and lack large-scale annotated datasets, further highlights that data collection and annotation for real surgical instrument images are costly, and privacy considerations as well as complex intraoperative scenes restrict dataset scale and diversity, thereby limiting model generalization [33].

4.2 Future Research Directions

Lightweight backbones, such as MobileNetV2, offer inherent advantages for edge deployment but lag behind large Transformer- or CNN-based architectures in feature representation and small-object detection. Two complementary strategies can be pursued to address this gap: first, designing

"lightweight ViT/hybrid CNN–ViT" architectures as suggested in to enhance global context modeling [34]. The second focuses on applying knowledge distillation (KD) as described in to transfer knowledge from a large, high-capacity teacher model (a large ViT or strong detector) to a smaller student model, thereby significantly improving performance while maintaining low computational cost [35]. Real surgical environments often include occlusions, smoke, and fluid disturbances, which can degrade the reliability of purely vision-based methods. In robotic-assisted surgery, multi-modal tactile feedback systems can provide effective solutions [36]. For example, integrating tactile sensors into robotic manipulators and fusing visual, tactile, and force information can provide complementary cues for instrument recognition, contact detection, and fine-motion estimation (sudden force changes signaling grasp/release events), improving detection and tracking stability under occlusion or degraded visual conditions.

Clinical deployment requires AI systems to be not only accurate but also interpretable, verifiable, and compliant with regulatory and safety standards, as noted in [37]. Future research should integrate explainable AI (XAI) techniques, such as saliency visualization, feature attribution, and model confidence/uncertainty quantification, with system-level safety mechanisms, including anomaly detection, failure recovery strategies, and deterministic latency guarantees [38]. Research directions include local interpretability (frame-level heatmaps, keypoint contribution), temporal consistency explanation (why an instrument ID is lost at a specific time), and regulatory-oriented approaches (clinical validation, risk assessment, user-readable reports). Methodologically, combining uncertainty estimation (temperature scaling, Bayesian approximation) with XAI can enhance clinical acceptability and facilitate regulatory review.

5. Conclusion

With the rapid evolution of minimally invasive and robot-assisted surgery, the demand for real-time, accurate, and robust surgical tool detection and tracking has grown substantially. These capabilities are essential for improving surgical precision, enhancing intraoperative navigation, and supporting advanced functionalities such as semi-autonomous assistance and decision support in the operating room. Lightweight deep learning models, including MobileNetV2 and YOLOv8n, have demonstrated considerable promise for deployment on resource-constrained or embedded platforms. When evaluated on the CholecTrack20 dataset, these models achieved competitive performance delivering reliable real-time detection and tracking while maintaining high inference speed and low computational overhead, making them particularly well-suited for edge computing devices and surgical robotic systems.

Nevertheless, lightweight models face inherent challenges in terms of generalisation, small-object detection, and robustness under highly complex surgical conditions. Scenarios characterised by occlusions, smoke, blood contamination, and specular reflections can severely degrade visual clarity and reduce detection accuracy. To mitigate these issues, techniques such as multi-frame or temporal feature aggregation, lightweight attention modules, and feature pyramid networks have shown promise in enhancing semantic representation, improving small-object detection, and stabilising ID consistency across dynamic and visually challenging surgical scenes.

Future research should also prioritise multimodal sensing and fusion strategies, combining visual, haptic, and force feedback data to provide richer contextual information and enable fine-grained motion estimation of surgical instruments. In addition, incorporating explainable AI techniques and uncertainty quantification is crucial for improving clinical interpretability, fostering trust among practitioners, and ensuring compliance with regulatory and safety standards. Emerging approaches such as knowledge distillation and CNN–Transformer hybrid architectures further offer opportunities to strengthen feature expressiveness while maintaining the efficiency and deployability of lightweight frameworks.

In summary, lightweight deep learning models should not be viewed reduced alternatives to heavier networks, but rather as a strategic cornerstone for clinically deployable surgical AI. On

CholecTrack20, they show how compact architectures can trade off speed and accuracy, providing viable solutions for real-time deployment on embedded and robotic systems. Their actual value is in facilitating scalable, cost-effective, and safety-driven AI systems with the ability to support a variety of surgical tasks in navigation, automation, and training. The essential challenge ahead is developing lightweight frameworks into multimodal, intelligent perception systems that combine accuracy, computational efficiency, robustness, and interpretability. Striking this balance will be imperative for delivering autonomous, reliable, and clinically validated surgical assistance technologies in real-world operating rooms.

6. References

[1] Abdalla Osman, E. I., Mubarak Ismail, M. M. E., Hassan Mukhtar, M. A., Babiker Ahmed, A. U., Abd Elfrag Mohamed, N. A., & Alamin Ibrahim, A. A. Cureus, 17 (3), e81339 (2025).

[2] S. W. Wong and P. Crowe, Journal of Robotic Surgery 17, 1873 (2023).

[3] B. Ghanekar, L. R. Johnson, J. L. Laughlin, M. K. O'Malley, and A. Veeraraghavan, International Symposium on Biomedical Imaging (ISBI) (Pp. 1-5). IEEE. 22 (2025).

[4] W. Guo, J. Wu, Z. Chen, Q. Zhao, M. Xu, Z. Lei, and H. Liu, in Lecture Notes in Computer Science (2025), pp. 168–177.

[5] J. C. Á. Cerón, G. O. Ruiz, L. Chang, and S. Ali, Medical Image Analysis 81, 102569 (2022).

[6] N. P. P. Gala, May 2025 International Research Journal on Advanced Engineering and Management (IRJAEM) 3 (05): 1657-1665 (2025).

[7] C. I. Nwoye, K. Elgohary, A. Srinivas, F. Zaid, J. L. Lavanchy, and N. Padoy, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 8942 (2025).

[8] R. Alrasheed, O. A. Waraga, M. A. Talib and M. A. Moufti,2024 Global Digital Health Knowledge Exchange & Empowerment Conference (gDigiHealth.KEE), pp. 1-7 (2024).

[9] H. Aoki, & N. Fujita in Seventeenth International Conference on Quality Control by Artificial Vision Vol. 13737, pp. 227-234 (2025).

[10] A. Martin-Gomez, H. Li, T. Song, S. Yang, G. Wang, H. Ding, N. Navab, Z. Zhao, and M. Armand, IEEE Transactions on Visualization and Computer Graphics 30, 3578 (2023).

[11] T. Allyne, MD & M. Luca, MD, A SAGES Technology and Value Assessment. (2018).

[12] T. Mostafa, D.P Andres, M.Masoud, NVIDIA Technical Blog (2025).

[13] V. Schorp, F. Giraud, G. Pargätzi, M. Wäspe, L. Von Ritter-Zahony, M. Wegmann, N. A. Cavalcanti, J. G. Henao, N. Bünger, D. Cachin, S. Caprara, P. Fürnstahl, and F. Carrillo, 17th Hamlyn Symposium on Medical Robotics (2025).

[14] W.-L. Chuang, M.-H. Yeh, and Y.-L. Yeh, Actuators 10, 141 (2021).

[15] Anon, Zimmer Biomet: Warsaw, IN, USA, (2023).

[16] G Loza, P. Valdastri, S Ali Healthcare Technology Letters, 11 (2-3), 48-58. (2024)

[17] X. Du, T. Kurmann, P.-L. Chang, M. Allan, S. Ourselin, R. Sznitman, J. D. Kelly, and D. Stoyanov, IEEE Transactions on Medical Imaging 37, 1276 (2018).

[18] A. Qayyum, H. Ali, M. Caputo, H. Vohra, T. Akinosho, S. Abioye, I. Berrou, P. Capik, J. Qadir, and M. Bilal, Scientific Reports 15, (2025).

[19] J. Liu, X. Guo, and Y. Yuan, IEEE Transactions on Medical Imaging 41, 715 (2021).

[20] D. Bouget, M. Allan, D. Stoyanov, P. Jannin, Medical image analysis, 35, 633-654 (2017).

[21] R. A. Rizal, J. S. Sihotang, R. Gultom In 2019 International Conference of Computer Science and Information Technology (ICoSNIKOM) (pp. 1-6). IEEE (2019).

[22] B. Namazi, G. Sankaranarayanan, and V. Devarajan, Surgical Endoscopy 36, 679 (2021).

[23] D. S. Yanni, B. M. Ozgur, R. G. Louis, Y. Shekhtman, R. R. Iyer, V. Boddapati, A. Iyer, P. D. Patel, R. Jani, M. Cummock, A. Herur-Raman, P. Dang, I. M. Goldstein, M. Brant-Zawadzki, T. Steineke, and L. G. Lenke, Neurosurgical FOCUS 51, E11 (2021).

[24] L. Qiu, C. Li, and H. Ren, Healthcare Technology Letters 6, 159 (2019).

[25] A. Zia, Y. Sharma, V. Bettadapura, E. L. Sarin, and I. Essa, International Journal of Computer Assisted Radiology and Surgery 13, 443 (2018).

[26] C. I. Nwoye, K. Elgohary, A. Srinivas, F. Zaid, J. L. Lavanchy, & N. Padoy, arXiv preprint arXiv: 2312.07352 (2023).

[27] G. Ghiasi, T.-Y. Lin, and Q. V. Le, 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019).

[28] H. Wu, Y. Chen, N. Wang, and Z.-X. Zhang, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) (2019).

[29] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, arXiv (Cornell University) (2017).

[30] G. Loza, P. Valdastri, and S. Ali, Healthcare Technology Letters 11, 48 (2023).

[31] L. Li, B. Li, and H. Zhou, PeerJ Computer Science 8, e1145 (2022).

[32] H. D. Viet, T. T. Nguyen, H. N. Lam, B. P. Nguyen, T. Q. Vu, H. M. Nguyen, V. T. Pho, H. H. Dang, D. V. Sang, and T. T. Nguyen, Journal of Medical Artificial Intelligence 0, 0 (2023).

[33] L. Wiese, L. Hinz, E. Reithmeier, P. Korn, and M. Neuhaus, Computers 14, 69 (2025).

[34] B. Zhao, R. Song, and J. Liang, 2021 IEEE/CVF International Conference on Computer Vision (ICCV) 6123 (2023).

[35] A. Moslemi, A. Briskina, Z. Dang, and J. Li, Machine Learning with Applications 18, 100605 (2024).

[36] A. Abiri, J. Pensa, A. Tao, J. Ma, Y.-Y. Juo, S. J. Askari, J. Bisley, J. Rosen, E. P. Dutson, and W. S. Grundfest, Scientific Reports 9, (2019).

[37] L. Farah, J. M. Murris, I. Borget, A. Guilloux, N. Martelli M. & S. I. M. Katsahian, Mayo Clinic proceedings. Digital health, 1 (2), 120–138. (2023).

[38] Z. Sadeghi, R. Alizadehsani, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladík, S. Nahavandi, & P. M Pardalos, Computers & Electrical Engineering, 118, 109370–109370. (2024).