Journal of Artificial Intelligence in Bioinformatics
http://dx.doi.org/10.62762/JAIB.2026.698089

REVIEW ARTICLE

Check for updates

Deep Learning-Based Segmentation, Action, and Phase Recognition in Laparoscopic Cholecystectomy: A Review

Hafsa Gulzar[1,†,*], **Muhammad Bilal**[2,†], **Muhammad Zubair Nawaz**[3], **Muhammad Yaqub**[4] and **Chayut Bunterngchite**[5]

[1] School of Computer Science and Technology, Donghua University, Shanghai 201620, China
[2] Department of Computer Science, Iowa State University, Ames 50011, United States
[3] Wuhan National Laboratory for Optoelectronics and School of Optical and Electronic Information, Huazhong University of Science and Technology, Wuhan 430074, China
[4] School of Automation, Central South University, Changsha 410083, China
[5] Division of Industrial and Logistics Engineering Technology, Faculty of Engineering and Technology, King Mongkut's University of Technology North Bangkok, Rayong Campus, Rayong 21120, Thailand

Abstract

The integration of artificial intelligence into biomedical image analysis represents a paradigm shift in clinical decision support and surgical informatics. Laparoscopic cholecystectomy (LC) is a widely performed minimally invasive procedure but remains technically challenging due to limited visibility, anatomical variations, and the risk of complications such as bile duct injury. Artificial intelligence (AI), particularly deep learning (DL), has shown great potential for improving intraoperative understanding through automated surgical video analysis. This review summarizes recent advances in DL-based LC video analysis, focusing on semantic segmentation and surgical action or phase recognition. Existing approaches are categorized into convolutional neural networks (CNNs), transformer-based architectures, and hybrid spatio-temporal models. For segmentation, U-Net variants, DeepLab architectures, and vision transformers are reviewed for identifying anatomical structures and surgical instruments. For action and phase recognition, CNN-LSTM frameworks, 3D CNNs, and video transformers are discussed for modeling surgical workflows and temporal dependencies. We also summarize commonly used datasets, including Cholec80 and CholecSeg8k, and evaluation metrics such as Dice score, IoU, accuracy, and F1-score. Major challenges include data scarcity, inter-surgeon variability, domain shift, and real-time computational constraints. Future research should emphasize self-supervised learning, multi-modal fusion, and clinically explainable AI to improve robustness, generalizability, and real-world surgical deployment.

Submitted: 06 June 2026
Revised: 10 July 2026
Accepted: 22 July 2026
Published: 22 September 2026

Vol. 2, **No.** 2, 2026.
10.62762/JAIB.2026.698089

***Corresponding author:**
Hafsa Gulzar
hafsa@mail.dhu.edu.cn

† These authors contributed equally to this work.

Citation
Gulzar, H., Bilal, M., Nawaz, M. Z., Yaqub, M., & Bunterngchite, C. (2026). Deep Learning-Based Segmentation, Action, and Phase Recognition in Laparoscopic Cholecystectomy: A Review. *Journal of Artificial Intelligence in Bioinformatics*, 2(2), 55–67.

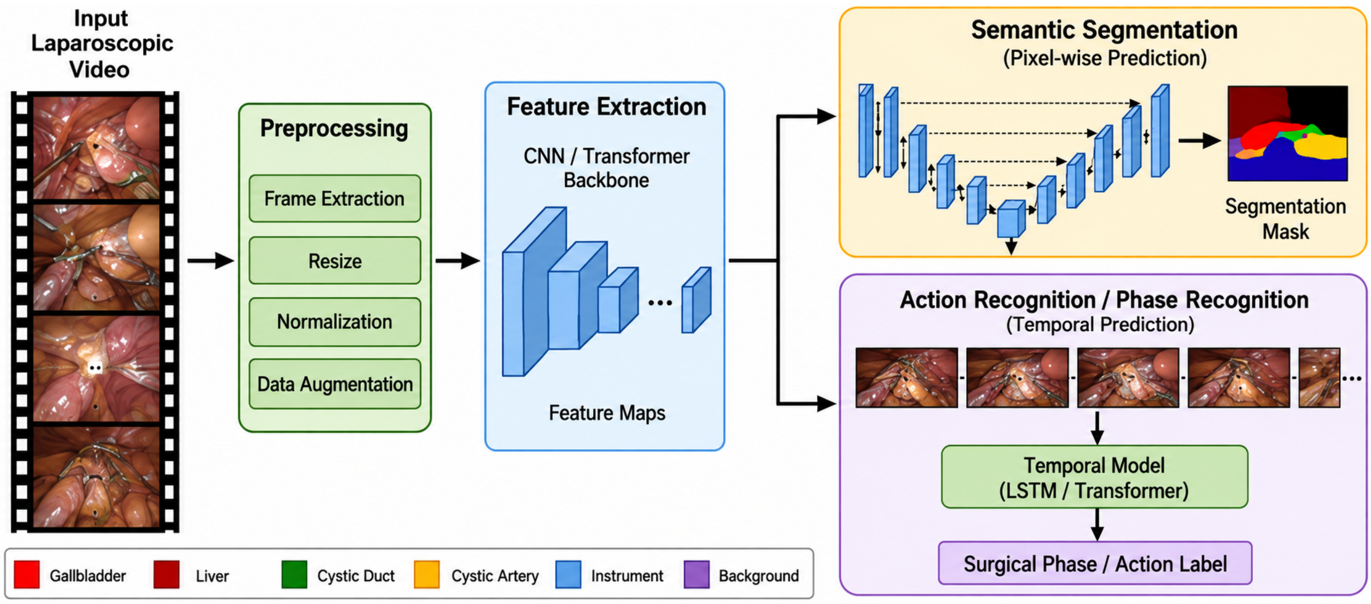

Figure 1. Overall framework of deep learning for LC video analysis.

Keywords: laparoscopic cholecystectomy, deep learning, semantic segmentation, surgical phase recognition, surgical action recognition.

1 Introduction

Laparoscopic cholecystectomy (LC) is considered the gold standard surgical procedure for the treatment of gallbladder diseases due to its minimally invasive nature, reduced postoperative pain, shorter hospital stays and faster patient recovery [1, 2]. Although it is one of the most commonly performed surgeries worldwide, LC is still associated with significant intraoperative risks [3]. One of the most critical complications is injury to the common bile duct, which can lead to severe long-term morbidity and require complex corrective procedures. A major challenge in LC arises from the limited field of view, presence of smoke, occlusions caused by surgical instruments, and high variability in patient anatomy, all of which impede the accurate intraoperative identification of critical biliary structures and increase the risk of bile duct injury [4, 6]. These factors make it difficult for surgeons, especially trainees, to consistently identify critical anatomical structures such as the cystic duct, cystic artery, and gallbladder boundaries [6]. To address these challenges, computer-assisted surgical systems powered by artificial intelligence have gained increasing attention. With the rapid development of deep learning, particularly convolutional neural networks (CNNs) and transformer-based architectures, automated surgical video understanding has become feasible [8]. The rapid proliferation of deep learning research in this domain is evidenced by bibliometric analyses showing exponential growth in publications on surgical instrument segmentation, detection, and tracking [7]. These methods enable machines to learn meaningful representations directly from laparoscopic video data without requiring handcrafted features. In the context of LC, two fundamental computer vision tasks have emerged as the focus of research [9, 10]. Semantic segmentation involves pixel-wise classification of surgical images to identify and localize anatomical structures and surgical tools. Accurate segmentation is essential for providing visual guidance, improving anatomical awareness, and supporting intraoperative decision-making. The surgical action recognition (or Phase Recognition) task involves analyzing temporal sequences of surgical video frames to identify different phases of the operation, such as preparation, dissection of Calot's triangle, clipping and cutting, and gallbladder extraction [11]. Understanding surgical workflow progression is crucial for surgical training, skill assessment, and real-time assistance systems [12–14].

Early approaches in surgical video analysis relied on handcrafted features and traditional machine learning techniques, which struggled to generalize across different surgical settings [13]. In contrast, deep learning methods have demonstrated superior performance by learning hierarchical representations directly from data. Convolutional neural networks (CNNs) such as U-Net and DeepLab have become standard for segmentation tasks, while recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and more recently transformer-based video models have significantly advanced temporal modeling in action recognition [15]. Despite these advancements, several challenges still limit the deployment of deep learning systems in real clinical environments [8, 16].

Surgical data are often limited and expensive to annotate, leading to small and imbalanced datasets. Additionally, variations in surgical techniques across different surgeons and hospitals introduce domain shifts that reduce model generalization [17]. Real-time processing constraints further complicate clinical integration, as surgical assistance systems must operate with low latency and high reliability [18, 19]. This review aims to systematically analyze recent advances in deep learning-based segmentation and action recognition for laparoscopic cholecystectomy as illustrated in the overall framework presented in Figure 1. We categorize existing methods, compare their strengths and limitations, and discuss publicly available datasets and evaluation protocols [16, 20]. Finally, we identify key research gaps and outline future directions, including self-supervised learning, multi-modal fusion of video and instrument data, and the development of explainable AI systems to improve clinical trust and adoption [22]. Bridging the gap between technical advances and clinical adoption will require clinicians and engineers to share a common understanding of AI model capabilities and limitations in real surgical practice [21].

2 Deep Learning for Semantic Segmentation in Laparoscopic Cholecystectomy

Semantic segmentation in laparoscopic cholecystectomy (LC) plays a critical role in understanding the surgical scene at a pixel level. Unlike classification tasks that assign a single label to an entire image, semantic segmentation provides dense predictions, enabling precise localization of anatomical structures and surgical instruments [23]. This fine-grained understanding is essential in LC because surgeons operate in a highly constrained environment where accurate identification of tissues such as the gallbladder, cystic duct, liver, and surrounding vessels directly impacts surgical safety [24, 25]. The main challenges in LC segmentation include occlusions caused by surgical tools, presence of smoke and blood, specular reflections from laparoscopic cameras, and significant variability in anatomy across patients. These challenges make traditional computer vision approaches ineffective and have led to the adoption of deep learning-based methods that can learn robust hierarchical representations from data.

2.1 CNN-Based Semantic Segmentation Methods

Convolutional Neural Networks (CNNs) are the foundation of most medical image segmentation models. The most influential architecture in surgical vision is the U-Net, which introduced an encoder–decoder structure with skip connections [36]. Scoping reviews of artificial intelligence applied to invasive surgical video have confirmed U-Net and its variants as the dominant family of models across a broad range of surgical procedures [26]. The encoder progressively reduces spatial resolution while capturing semantic context, and the decoder reconstructs spatial details to produce pixel-wise segmentation masks. Skip connections preserve fine-grained spatial information, which is crucial in surgical scenes where anatomical boundaries are often subtle. Several variants of U-Net have been proposed to improve performance in LC segmentation tasks. R2U-Net combines recurrent and residual convolutional units within the U-Net encoder–decoder structure, enhancing feature reusability and gradient propagation across deeper networks [37]. Attention U-Net introduces attention gates that suppress irrelevant background features and focus on anatomically important regions. U-Net++ redesigns skip connections using nested and dense convolution pathways to reduce semantic gaps between encoder and decoder features [28]. In addition to U-Net variants, the DeepLab series (DeepLabv3 and DeepLabv3+) has been widely used in surgical segmentation. These models utilize atrous (dilated) convolutions to expand the receptive field without increasing computational cost. The Atrous Spatial Pyramid Pooling (ASPP) module allows the network to capture multi-scale contextual information, which is particularly useful in LC where organs appear at varying scales due to camera movement [38]. Despite their success, CNN-based methods are primarily spatial and process frames independently. This leads to temporal inconsistency when applied to surgical videos, where structures may appear unstable across consecutive frames [29].

2.2 Transformer-Based Segmentation Methods

To overcome the limitations of CNNs in capturing global context, transformer-based architectures have recently been introduced into surgical segmentation [3, 30]. Vision Transformers (ViTs) divide an image into patches and model relationships between all patches using self-attention mechanisms. This allows the network to capture long-range dependencies, which is especially beneficial in complex surgical scenes where anatomical structures may be partially occluded or visually similar. However, standard ViTs require large-scale datasets, which are often unavailable in medical domains. To address this, hierarchical transformer models such as the Swin Transformer have been proposed. Swin Transformers apply self-attention within local windows and progressively merge them across layers, making them computationally efficient and suitable for high-resolution medical images [39]. Comparative studies in laparoscopic surgery have demonstrated that Transformer-based models such as SegFormer achieve higher accuracy and better generalizability than CNN-based counterparts like DeepLabv3 in anatomical structure segmentation, though often at the cost of real-time processing capability [31].

Hybrid architectures combining CNNs and transformers have shown strong performance in LC segmentation. In these models, CNNs are used to extract local spatial features such as edges and textures, while transformers model global dependencies between anatomical regions. This combination improves both boundary precision and contextual understanding. Overall, transformer-based segmentation models demonstrate superior generalization ability compared to pure CNNs, especially in challenging surgical environments with complex visual variability [32].

2.3 Spatio-Temporal Segmentation in Surgical Videos

Unlike natural images, laparoscopic videos contain strong temporal continuity. Consecutive frames are highly correlated, and ignoring temporal information often leads to inconsistent segmentation outputs. To address this, spatio-temporal segmentation models have been developed. One common approach is combining CNNs with recurrent neural networks such as Long Short-Term Memory (LSTM) networks. In this framework, CNNs extract spatial features from each frame, while LSTMs model temporal dependencies across frames. This helps maintain consistency in predicting anatomical structures over time. Another approach involves 3D CNNs, which extend convolution operations into the temporal dimension. These models process multiple frames simultaneously, capturing motion-aware representations. However, 3D CNNs are computationally expensive and require large memory resources. More recently, temporal attention mechanisms have been introduced into surgical video analysis. Originally developed for surgical phase recognition—where models such as OPERA [33] and Trans-SVNet [35] demonstrated that assigning differential importance weights to frames in a sequence allows the network to focus on clinically relevant moments while suppressing redundant or noisy frames—these attention strategies have since been adapted for spatio-temporal segmentation to improve temporal consistency of predicted masks across video sequences. A comprehensive summary of these segmentation methods, including their backbones, strengths, limitations, and applications, is provided in Table 1.

2.4 Key Challenges in LC Segmentation

Despite the remarkable progress achieved by deep learning-based segmentation methods in laparoscopic cholecystectomy (LC), several critical challenges still hinder their full performance and clinical translation, as demonstrated by comparative studies of segmentation networks on LC datasets [40, 41]. One of the major difficulties is occlusion and instrument interference, where surgical tools frequently enter the camera view and partially or completely block important anatomical structures such as the cystic duct, gallbladder, and surrounding tissues. This makes accurate pixel-level prediction significantly more complex, as the model must distinguish between instruments and overlapping organs in a highly dynamic environment. Another important challenge is the presence of smoke and blur artifacts, which are commonly produced during electrocautery and tissue manipulation. These visual disturbances reduce image clarity and introduce noise into the input frames, leading to degraded segmentation performance and unstable predictions across consecutive frames. In addition, class imbalance remains a persistent issue in surgical datasets, where small and less frequently visible structures such as the cystic duct or small vessels are heavily underrepresented compared to dominant background regions like liver tissue or empty surgical space. This imbalance often biases models toward larger structures and reduces sensitivity to clinically critical fine details [42].

Furthermore, inter-patient variability poses a significant limitation, as anatomical differences in tissue appearance, fat distribution, and pathological conditions vary widely across patients. As a result,

Table 1. Summary of segmentation methods.

Method	Backbone	Strength	Limitation	Application
U-Net [36]	CNN	Simple, effective	No temporal info	Baseline segmentation
R2U-Net [37]	CNN + Recurrent	Feature reusability via recurrent residual blocks	Higher complexity	Medical image segmentation
Attention U-Net [27]	CNN + Attention	Focus on key regions	Higher complexity	Surgical scenes
DeepLabv3+ [38]	CNN + ASPP	Multi-scale features	High computation	Complex anatomy
Swin Transformer [39]	Transformer	Global context	Data hungry	Advanced segmentation

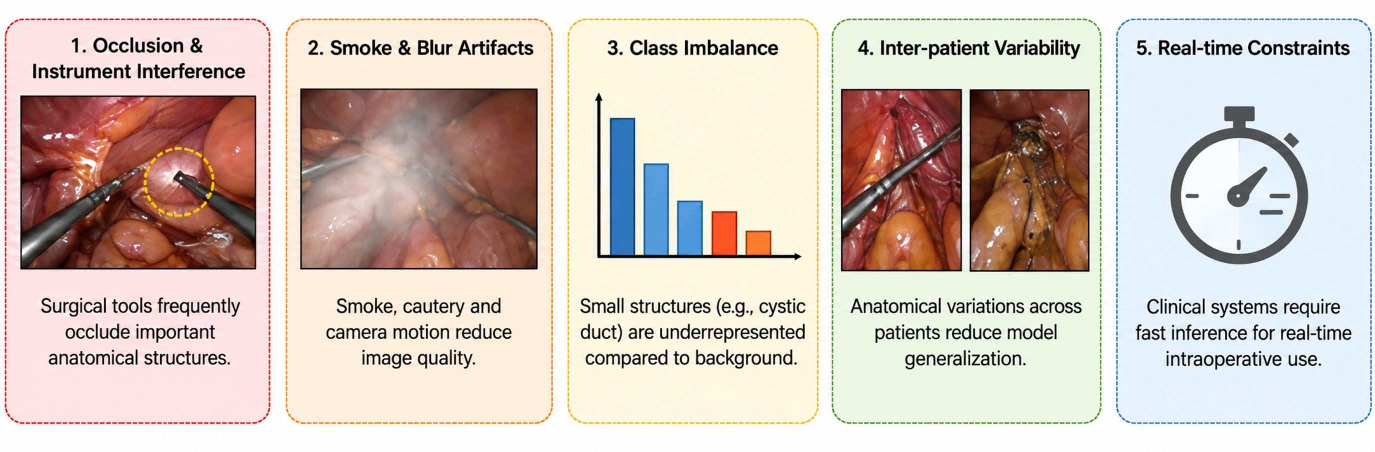

Figure 2. Key challenges in LC segmentation.

models trained on limited datasets often struggle to generalize effectively to unseen surgical cases from different hospitals or surgeons. Lastly, real-time computational constraints are a key barrier for clinical adoption, since intraoperative systems require fast and efficient inference to provide timely feedback without interrupting surgical workflow. Many high-performing deep learning models, particularly transformer-based or complex hybrid architectures, are computationally expensive and therefore difficult to deploy in real-time surgical environments [43]. Collectively, these challenges, which are visually summarized in Figure 2, highlight the need for more robust, efficient, and generalizable segmentation frameworks capable of handling noisy surgical conditions while maintaining high accuracy and real-time performance.

3 Deep Learning for Action Recognition in Laparoscopic Cholecystectomy

Action recognition in laparoscopic cholecystectomy (LC) focuses on identifying and classifying surgical actions from video sequences, such as cutting, clipping, dissecting, suctioning, and grasping. Unlike static image analysis, action recognition requires understanding both spatial appearance and temporal dynamics across consecutive frames [43]. The overall pipeline for deep learning-based action recognition in laparoscopic cholecystectomy is depicted in Figure 3. This makes the task significantly more complex, as surgical actions are highly dependent on motion patterns, tool-tissue interaction, and procedural context. Accurate recognition of surgical actions is essential for workflow analysis, surgical skill assessment, and automated phase recognition systems. Beyond workflow monitoring, AI-based video analysis has also shown promise for the automated detection of intraoperative adverse events, enabling timely clinical intervention [34]. However, LC environments introduce several challenges, including occlusions caused by surgical instruments, rapid camera motion, smoke artifacts, and subtle visual differences between similar actions, which make robust action recognition difficult [44]. In the context of surgical video analysis, it is important to distinguish between two related but distinct tasks: surgical phase recognition and surgical action recognition. Surgical phase recognition refers to the classification of broad, sequential workflow steps of an operation (e.g., the seven standard phases in the Cholec80 dataset, such as preparation,

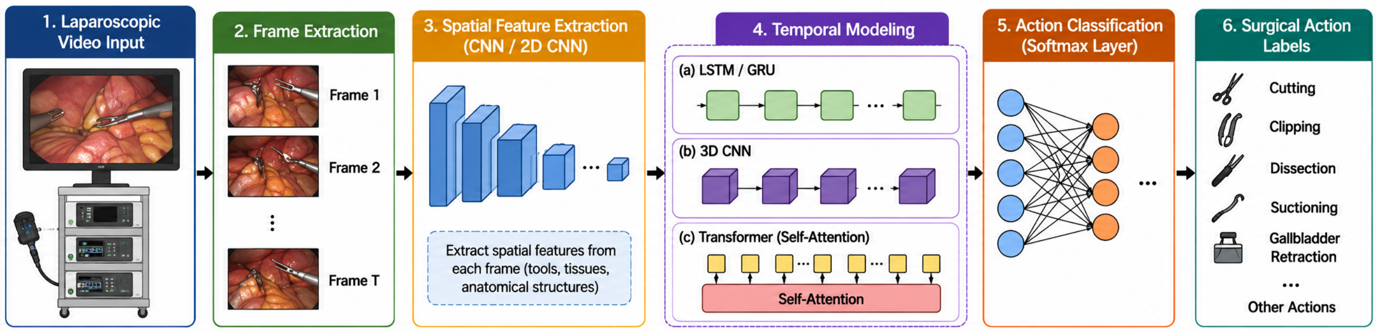

Figure 3. Overall pipeline of deep learning-based action recognition in laparoscopic cholecystectomy.

Calot's triangle dissection, and gallbladder extraction). In contrast, surgical action recognition involves identifying finer-grained, short-duration events or tool-tissue interactions, typically defined by a verb, an instrument, and a target (e.g., "grasping the gallbladder with forceps", as seen in the CholecT50 dataset). While phase recognition provides a macro-level understanding of the workflow, action recognition offers micro-level insights into specific surgical procedures. This section discusses deep learning approaches applicable to both temporal modeling tasks.

3.1 CNN-Based Action Recognition Methods

Convolutional Neural Networks (CNNs) form the foundation of many early and modern surgical action recognition models. Initially, 2D CNN architectures such as VGGNet, ResNet, and Inception networks were applied to extract spatial features from individual frames. These models effectively capture visual information such as surgical instruments, anatomical structures, and tissue appearance. However, since 2D CNNs process frames independently, they fail to model temporal dependencies, which are essential for distinguishing surgical actions that differ primarily in motion rather than appearance [45, 46]. To address this limitation, many approaches combine CNNs with temporal modeling techniques. A common framework is the CNN–LSTM architecture, where CNNs extract spatial features from each frame and Long Short-Term Memory (LSTM) networks model temporal dependencies across frame sequences. This hybrid approach has been widely used in LC action recognition to capture surgical workflow progression and improve temporal consistency in predictions. Variants such as Gated Recurrent Units (GRUs) have also been explored due to their simpler structure and faster computation. Despite their effectiveness, CNN–RNN based methods still struggle with long-range temporal dependencies and may fail to capture complex interactions between distant frames in long surgical videos [18].

3.2 3D CNN-Based Action Recognition Methods

To better capture spatio-temporal information, 3D Convolutional Neural Networks (3D CNNs) were introduced. Unlike 2D CNNs, 3D CNNs apply convolution operations across both spatial and temporal dimensions, enabling the model to learn motion-aware representations directly from video clips. Popular architectures such as C3D and I3D (Inflated 3D ConvNet) have been widely adopted in surgical action recognition tasks. In LC videos, 3D CNNs are particularly effective in recognizing dynamic actions such as dissection and clipping, where motion patterns are critical. These models learn joint spatial-temporal features, allowing them to better differentiate between visually similar actions. However, 3D CNNs are computationally expensive, require large memory resources, and depend heavily on large, annotated datasets, which are often limited in medical domains [22].

3.3 Transformer-Based Action Recognition Methods

Recently, transformer-based architectures have emerged as a powerful alternative for surgical action recognition. Vision Transformers (ViTs) and Video Transformers utilize self-attention mechanisms to model long-range dependencies across frames [24]. This allows the network to capture global temporal relationships, which is particularly useful in LC procedures where surgical actions may be temporally distributed and not strictly sequential. Transformer-based models divide video inputs into patches or tokens and compute relationships between them using attention mechanisms. This enables better modeling of complex temporal interactions compared to RNN-based methods. However, standard transformers require large-scale

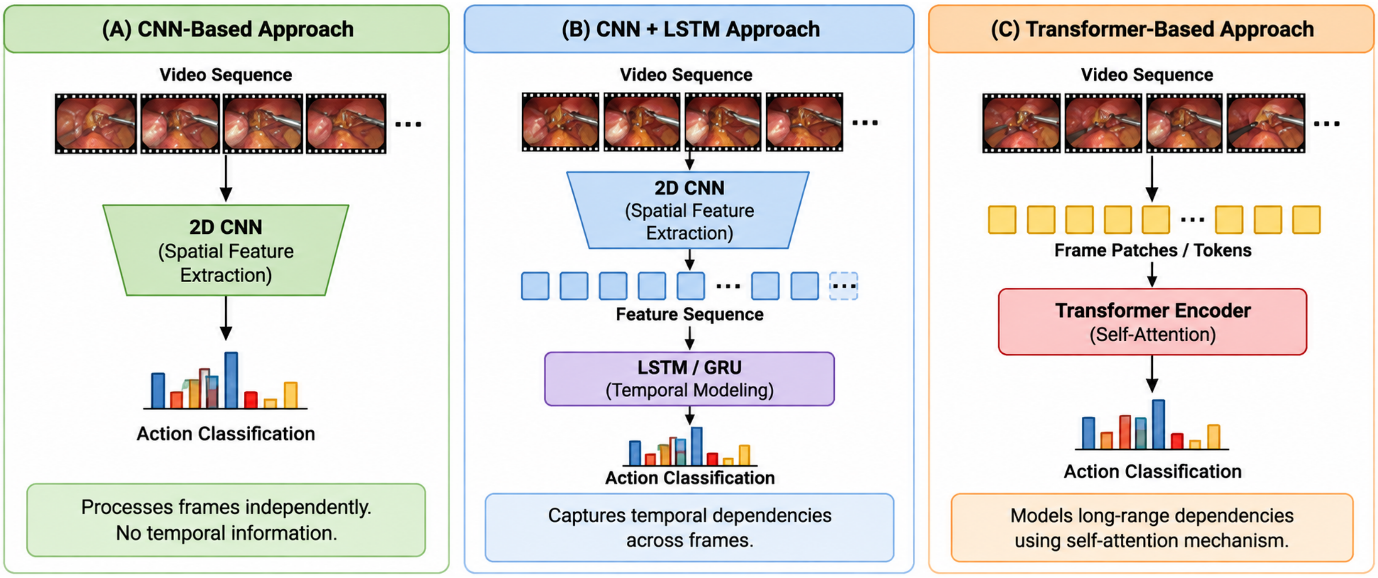

Figure 4. Comparison of deep learning architectures for action recognition in LC.

datasets for training, which limits their direct application in medical video analysis [47]. To address this, efficient variants such as TimeSformer and long-range video transformers specifically designed for surgical phase recognition [48], as well as hierarchical transformers like Swin Transformer-based video models, have been proposed, reducing computational complexity while maintaining strong performance in capturing both local and global temporal dependencies. The integration of such efficient transformer architectures into endoscopic and surgical AI systems represents a key trend in the broader field of surgical automation [49]. A visual comparison of these deep learning architectures for action recognition in LC is presented in Figure 4.

3.4 Hybrid Deep Learning Approaches

To leverage the strengths of different architectures, hybrid models combining CNNs, RNNs, and Transformers have been widely explored. In these frameworks, CNNs are used to extract spatial features, RNNs capture short-term temporal dynamics, and attention mechanisms or transformers model long-range dependencies. Such hybrid architectures are particularly effective in LC action recognition because they can simultaneously handle spatial complexity, temporal continuity, and global context. Some approaches also incorporate temporal attention mechanisms to selectively focus on clinically relevant frames while suppressing redundant or noisy information. Overall, hybrid methods provide a more balanced solution for surgical action recognition by combining spatial feature extraction with both short-term and long-term temporal modeling [50]. Table 2 provides a detailed summary of these deep learning-based action recognition methods in LC.

3.5 Key Challenges in LC Action Recognition

Despite significant advancements in deep learning-based action recognition for laparoscopic cholecystectomy, several challenges still limit performance and clinical deployment. One major challenge is visual ambiguity between actions, where different surgical steps may appear visually similar, especially when observed frame-by-frame without temporal context. This makes it difficult for models to accurately distinguish between actions such as cutting and dissecting [11]. Another critical issue is occlusion and tool interaction, where surgical instruments frequently obstruct anatomical structures, leading to incomplete or misleading visual information. Smoke, blur, and lighting variations further degrade video quality and introduce noise into the recognition process [56]. Class imbalance is also a significant challenge, as some surgical actions occur much more frequently than others, causing models to be biased toward dominant classes while underperforming on rare but clinically important actions. Additionally, inter-surgeon variability introduces inconsistencies in surgical style, making it difficult for models trained on limited datasets to generalize across different surgical environments [14]. Finally, real-time processing constraints remain a major barrier for clinical adoption. Many high-performing transformer-based or 3D CNN models are computationally expensive, making them difficult to deploy in real-time intraoperative systems

Table 2. Summary of deep learning-based action recognition methods in LC.

Method	Models	Key Idea	Strengths	Limitations
CNN + Temporal Memory [51]	ResNet + Non-local Temporal Bank	Long-range temporal relation modeling for surgical workflow	Robust long-term dependency modeling	Complex memory bank mechanism
CNN + Temporal Graph [52]	CNN + Position-Aware Graph Network	Spatial CNN features with graph-based temporal phase modeling	Structured temporal dependencies	Graph complexity, limited scalability
3D CNN-based [53]	C3D, I3D	Spatio-temporal convolution over video clips	Strong motion modeling	High computational cost
Multi-task Global Reasoning [54]	CNN + Global Reasoning Module	Joint multi-task learning for surgical scene understanding	Simultaneous multi-label recognition	Requires multi-task annotation
Hybrid models [55]	CNN + LSTM + Attention / Transformer hybrids	Combined spatial + temporal + attention	Balanced performance, robust	Complex architecture

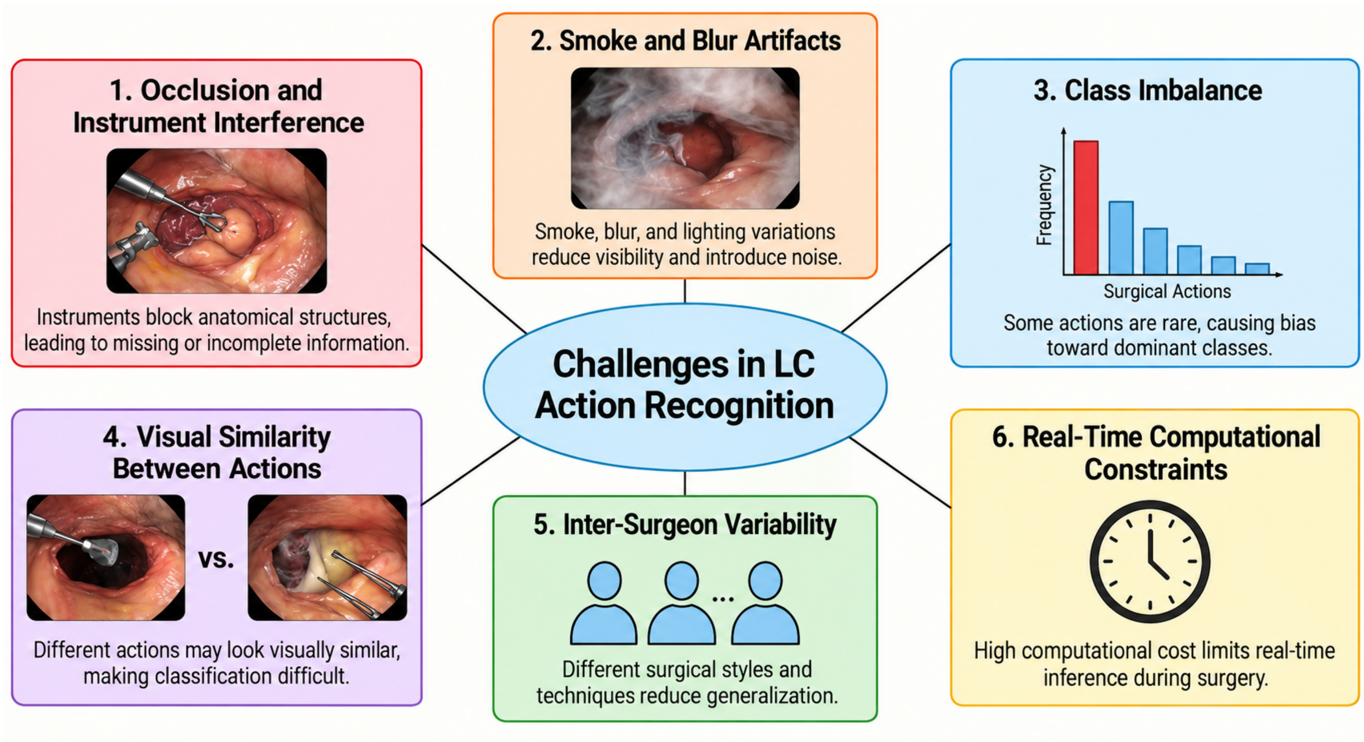

Figure 5. Key challenges in LC action recognition.

where fast and reliable predictions are required [57]. These key challenges in LC action recognition are further highlighted in Figure 5.

4 Public Datasets and Evaluation Metrics

The advancement of deep learning models in laparoscopic cholecystectomy (LC) is heavily dependent on the availability of high-quality, annotated datasets and standardized evaluation protocols. Public datasets have played a pivotal role in benchmarking algorithms, fostering reproducible research, and driving the development of robust surgical AI systems. Several public datasets have been established to support research in surgical segmentation and action recognition [58]. Cholec80 is one of the most widely used datasets for surgical phase recognition, consisting of eighty full-length LC videos with temporal annotations for seven distinct surgical phases, serving as the primary benchmark for action and workflow recognition [10]. Derived from this, CholecSeg8k provides pixel-wise semantic segmentation annotations for 8,080 keyframes, including thirteen classes that cover critical anatomical structures and surgical instruments, making it a standard benchmark for LC segmentation [59]; more recently, CholecInstanceSeg has extended this line with tool instance segmentation annotations [60]. Additionally, CholecT50 comprises fifty videos focusing on fine-grained surgical action triplets (instrument, verb, target) alongside tool and phase labels [46]. The HeiChole benchmark offers a comprehensive multi-task resource containing videos annotated for actions, phases, tools, and instrument presence, which is highly valuable for developing multi-task and hybrid models [14]. Furthermore, the annual Endoscopic Vision (EndoVis) challenges hosted at the MICCAI conference provide specialized datasets for tasks such as instrument segmentation, tracking, and surgical scene parsing [61].

To ensure fair comparison across different models, standardized evaluation metrics are employed for both segmentation and action recognition tasks. For semantic segmentation, the Dice Similarity Coefficient (DSC), or F1-score, is the most common metric, measuring the overlap between the predicted segmentation mask and the ground truth, which is particularly important for evaluating small structures. The Intersection over Union (IoU), or Jaccard Index, calculates the ratio of the intersection to the union of the predicted and ground truth masks, providing a strict measure of localization accuracy [62]. Pixel accuracy is also utilized, though it can be misleading in cases of severe class imbalance, which is common in surgical scenes. For action and phase recognition, frame-wise accuracy measures the percentage of frames where the predicted action or phase matches the ground truth label. Mean Average Precision (mAP) is often utilized for tool detection and multi-label action recognition to evaluate performance across different classes. Additionally, temporal metrics such as the Edit Score and Frame-wise Jaccard are used to evaluate how well a model predicts the boundaries and sequence of surgical phases, specifically penalizing over-segmentation and fragmented predictions [63].

5 Future Research Directions

Despite the remarkable progress in applying deep learning to LC video analysis, several critical gaps remain, and addressing these challenges will require innovative approaches to data efficiency, model interpretability, and clinical integration. One promising direction is the adoption of self-supervised and unsupervised learning. The reliance on large-scale, manually annotated datasets remains a significant bottleneck in surgical AI. Self-supervised learning offers a viable solution by allowing models to learn robust feature representations from vast amounts of unannotated surgical videos. By designing pretext tasks, such as predicting the temporal order of video clips or reconstructing masked video patches, self-supervised learning can significantly reduce the dependency on expensive pixel-wise and temporal annotations, enabling models to generalize better across different surgical domains [64].

Another crucial area for future research is multi-modal fusion for context-aware systems. Current models predominantly rely on visual data alone, yet the operating room is a highly multi-modal environment in which diverse data streams—including video, depth, instrument kinematics, and audio—co-occur simultaneously. Semantic scene graph frameworks developed for holistic OR modeling, such as 4D-OR [65], demonstrate the feasibility and value of integrating these modalities and offer a generalizable blueprint for multi-modal fusion in laparoscopic cholecystectomy analysis. Integrating visual data with other modalities, such as surgical instrument kinematics from robotic systems, tool usage logs, audio, and preoperative patient data, can provide a more holistic understanding of the surgical scene [45]. This fusion can improve the accuracy of action recognition and enable prescriptive AI systems that anticipate the surgeon's next steps and dynamically assess surgical safety metrics like the Critical View of Safety [66**?**]. Furthermore, the development of clinically explainable AI (XAI) is essential for the adoption of these systems in high-stakes clinical environments. Deep learning models often operate as black boxes, and developing XAI techniques tailored for surgical video, such as generating attention maps that highlight influential anatomical regions or providing natural language explanations for phase transitions, will bridge the gap between algorithmic output and clinical trust [67].

Additionally, real-time optimization and edge deployment must be prioritized. Intraoperative AI systems must operate with ultra-low latency to provide actionable feedback without disrupting the surgical workflow. Future research should focus on model compression techniques to deploy heavy transformer-based models on edge computing devices within the operating room. Knowledge distillation, which transfers learned representations from a large teacher model into a compact student network and has been demonstrated to be effective for model compression in endoscopic image analysis [5], alongside network pruning and quantization, offers a principled pathway for reducing the inference cost of surgical AI models without sacrificing predictive accuracy. Finally, the emergence of foundation models and surgical vision-language models presents a paradigm shift. Adapting large-scale foundation models to the surgical domain through domain-specific fine-tuning could enable zero-shot or few-shot segmentation and action recognition, drastically reducing the need for task-specific model training and handling the high variability and long-tail distribution of rare surgical events [68].

6 Conclusion

Deep learning has revolutionized the analysis of laparoscopic cholecystectomy videos, offering unprecedented capabilities in semantic segmentation and action recognition. By automating the identification of critical anatomical structures and surgical workflow phases, these technologies hold immense potential for enhancing intraoperative safety, improving surgical training, and enabling real-time computer-assisted interventions. This review has systematically categorized the evolution of these methods, from foundational convolutional neural networks and U-Net variants to advanced spatio-temporal models and transformer-based architectures. While significant strides have been made in model performance, the transition from retrospective research to prospective clinical deployment is hindered by challenges such as data scarcity, domain shift, visual artifacts, and real-time computational constraints. Overcoming these barriers will require a concerted shift toward data-efficient learning paradigms like self-supervised learning, the integration of multi-modal clinical data, and the development of explainable, lightweight AI systems. As the field progresses toward the integration of surgical foundation models and context-aware prescriptive AI, the ultimate goal of highly assisted, intelligent operating rooms is becoming increasingly attainable. Continued collaboration between computer scientists, clinical surgeons, and regulatory bodies will be essential to ensure these technologies are safe, robust, and ultimately beneficial for patient outcomes.

Data Availability Statement

Data sharing is not applicable to this article as no new datasets were generated or analyzed; all datasets discussed are publicly available from their respective original sources.

Funding

This work was supported without any funding.

Conflicts of Interest

The authors declare no conflicts of interest.

AI Use Statement

The authors declare that no generative AI was used in the preparation of this manuscript.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] Fernicola, A., Palomba, G., Capuano, M., De Palma, G. D., & Aprea, G. (2024). Artificial intelligence applied to laparoscopic cholecystectomy: what is the next step? A narrative review. *Updates in Surgery, 76*(5), 1655-1667. [CrossRef]

[2] Liao, W., Zhu, Y., Zhang, H., Wang, D., Zhang, L., Chen, T., ... & Ye, Z. (2025). Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: a systematic review. *Frontiers in Surgery, 12*, 1551838. [CrossRef]

[3] Zhou, R., Wang, D., Zhang, H., Zhu, Y., Zhang, L., Chen, T., ... & Ye, Z. (2025). Vision techniques for anatomical structures in laparoscopic surgery: a comprehensive review. *Frontiers in Surgery, 12*, 1557153. [CrossRef]

[4] Corallino, D., Balla, A., Coletta, D., Pacella, D., Podda, M., Pronio, A., ... & Aldrighetti, L. (2025). Systematic review on the use of artificial intelligence to identify anatomical structures during laparoscopic cholecystectomy: a tool towards the future. *Langenbeck's archives of surgery, 410*(1), 101. [CrossRef]

[5] Gammulle, H., Chen, Y., Sridharan, S., Klein, T., & Fookes, C. (2023). Learning through guidance: knowledge distillation for endoscopic image classification. *arXiv preprint arXiv:2308.08731*. [CrossRef]

[6] Anteby, R., Horesh, N., Soffer, S., Zager, Y., Barash, Y., Amiel, I., ... & Klang, E. (2021). Deep learning visual analysis in laparoscopic surgery: a systematic review and diagnostic test accuracy meta-analysis. *Surgical endoscopy, 35*(4), 1521-1533. [CrossRef]

[7] Yousef, M., Aly, K. E., Ahmed, M., Ahmed, F. A., Al Jalham, K., & Balakrishnan, S. (2026). Bibliometric analysis of deep learning for surgical instrument segmentation, detection and tracking in minimally invasive surgery. *Frontiers in Digital Health, 8*, 1633888. [CrossRef]

[8] Maier-Hein, L., Vedula, S. S., Speidel, S., Navab, N., Kikinis, R., Park, A., ... & Jannin, P. (2017). Surgical data science for next-generation interventions. *Nature Biomedical Engineering, 1*(9), 691-696. [CrossRef]

[9] Levendovics, R., Levendovics, T., Kronreif, G., & Haidegger, T. (2024). Surgical data science: Emerging trends and future pathways. *Recent Advances in Intelligent Engineering: Volume Dedicated to Imre J. Rudas' Seventy-Fifth Birthday*, 65-84. [CrossRef]

[10] Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., & Padoy, N. (2016). Endonet: a deep architecture for recognition tasks on laparoscopic

videos. *IEEE transactions on medical imaging, 36*(1), 86-97. [CrossRef]

[11] Mascagni, P., Alapatt, D., Sestini, L., Altieri, M. S., Madani, A., Watanabe, Y., ... & Hashimoto, D. A. (2022). Computer vision in surgery: from potential to clinical value. *npj Digital Medicine, 5*(1), 163. [CrossRef]

[12] Pescio, M., Marzola, F., Distefano, G., Leoncini, P., Ammirati, C. A., Barontini, F., ... & Arezzo, A. (2026). Artificial intelligence in minimally invasive and robotic gastrointestinal surgery: Major applications and recent advances. *Journal of Personalized Medicine, 16*(2), 71. [CrossRef]

[13] Kitaguchi, D., Takeshita, N., Hasegawa, H., & Ito, M. (2022). Artificial intelligence-based computer vision in surgery: Recent advances and future perspectives. *Annals of gastroenterological surgery, 6*(1), 29-36. [CrossRef]

[14] Wagner, M., Müller-Stich, B. P., Kisilenko, A., Tran, D., Heger, P., Mündermann, L., ... & Bodenstedt, S. (2023). Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the HeiChole benchmark. *Medical image analysis, 86*, 102770. [CrossRef]

[15] Jin, Y., Li, H., Dou, Q., Chen, H., Qin, J., Fu, C. W., & Heng, P. A. (2020). Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical image analysis, 59*, 101572. [CrossRef]

[16] Demir, K. C., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., & Yang, S. H. (2023). Deep learning in surgical workflow analysis: a review of phase and step recognition. *IEEE Journal of Biomedical and Health Informatics, 27*(11), 5405-5417. [CrossRef]

[17] Ali, S. (2022). Where do we stand in AI for endoscopic image analysis? Deciphering gaps and future directions. *npj Digital Medicine, 5*(1), 184. [CrossRef]

[18] Garrow, C. R., Kowalewski, K. F., Li, L., Wagner, M., Schmidt, M. W., Engelhardt, S., ... & Nickel, F. (2021). Machine learning for surgical phase recognition: a systematic review. *Annals of surgery, 273*(4), 684-693. [CrossRef]

[19] Hegde, S. R., Namazi, B., Iyengar, N., Cao, S., Desir, A., Marques, C., ... & Sankaranarayanan, G. (2024). Automated segmentation of phases, steps, and tasks in laparoscopic cholecystectomy using deep learning. *Surgical endoscopy, 38*(1), 158-170. [CrossRef]

[20] Siddique, N., Paheding, S., Elkin, C. P., & Devabhaktuni, V. (2021). U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE access, 9*, 82031-82057. [CrossRef]

[21] Limon, D., Satish, V., Raghavan, N., Nguyen, P., & Rajesh, A. (2026). Artificial intelligence in surgery revisited: a 2025 guide to understanding and applying AI models in clinical practice. *The American Surgeon™, 92*(3), 687-697. [CrossRef]

[22] Hussain, S. M., Brunetti, A., Lucarelli, G., Memeo, R., Bevilacqua, V., & Buongiorno, D. (2022). Deep learning based image processing for robot assisted surgery: a systematic literature survey. *IEEE Access, 10*, 122627-122657. [CrossRef]

[23] Dick, L., Boyle, C. P., Skipworth, R. J., Smink, D. S., Tallentire, V. R., & Yule, S. (2024). Automated analysis of operative video in surgical training: scoping review. *BJS open, 8*(5), zrae124. [CrossRef]

[24] Li, Y., Zhao, Z., Li, R., & Li, F. (2024). Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends. *Artificial Intelligence Review, 57*(11), 291. [CrossRef]

[25] Tokuyasu, T., Iwashita, Y., Matsunobu, Y., Kamiyama, T., Ishikake, M., Sakaguchi, S., ... & Inomata, M. (2021). Development of an artificial intelligence system using deep learning to indicate anatomical landmarks during laparoscopic cholecystectomy. *Surgical endoscopy, 35*(4), 1651-1658. [CrossRef]

[26] King, A., Fowler, G. E., Macefield, R. C., Walker, H., Thomas, C., Markar, S., ... & Blencowe, N. S. (2025). Use of artificial intelligence in the analysis of digital videos of invasive surgical procedures: scoping review. *BJS open, 9*(4), zraf073. [CrossRef]

[27] Oktay, O., Schlemper, J., Folgoc, L. L., Lee, M., Heinrich, M., Misawa, K., ... & Rueckert, D. (2018). Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*. [CrossRef]

[28] Zhou, Z., Siddiquee, M. M. R., Tajbakhsh, N., & Liang, J. (2018). UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support* (pp. 3–11). Springer, Cham. [CrossRef]

[29] Grammatikopoulou, M., Sanchez-Matilla, R., Bragman, F., Owen, D., Culshaw, L., Kerr, K., Stoyanov, D., & Luengo, I. (2024). A spatio-temporal network for video semantic segmentation in surgical videos. *International Journal of Computer Assisted Radiology and Surgery, 19*(2), 375–382. [CrossRef]

[30] Bravo, A., Razzaq, S., Murari, M., Ahuja, A., Birchett, D., & Schumacher, L. (2026). Transformers in surgical artificial intelligence: A domain-stratified, study-level narrative review. *JTCVS open*, 101597. [CrossRef]

[31] Kolbinger, F. R., Rinner, F. M., Jenke, A. C., Carstens, M., Krell, S., Leger, S., Distler, M., Weitz, J., Speidel, S., & Bodenstedt, S. (2023). Anatomy segmentation in laparoscopic surgery: comparison of machine learning and human expertise – an experimental study. *International Journal of Surgery, 109*(10), 2962–2974. [CrossRef]

[32] El Boghdady, M., Temori, S., Khaireldin, D., Ewalds-Kvist, B., Ghazanfar, M. A., & Aroori, S. (2025). The role of artificial intelligence in enhancing safety assessment of laparoscopic cholecystectomy: a systematic review. *HPB*. [CrossRef]

[33] Czempiel, T., Paschali, M., Ostler, D., Kim, S. T., Busam, B., & Navab, N. (2021, September). Opera: Attention-regularized transformers for surgical phase recognition. In *International conference on medical image computing and computer-assisted intervention* (pp. 604-614). Cham: Springer International Publishing. [CrossRef]

[34] Eppler, M. B., Sayegh, A. S., Maas, M., Venkat, A., Hemal, S., Desai, M. M., ... & Goldenberg, M. G. (2023). Automated capture of intraoperative adverse events using artificial intelligence: a systematic review and meta-analysis. *Journal of Clinical Medicine, 12*(4), 1687. [CrossRef]

[35] Gao, X., Jin, Y., Long, Y., Dou, Q., & Heng, P. A. (2021, September). Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *International conference on medical image computing and computer-assisted intervention* (pp. 593-603). Cham: Springer International Publishing. [CrossRef]

[36] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention* (pp. 234-241). Cham: Springer international publishing. [CrossRef]

[37] Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., & Asari, V. K. (2019). Recurrent residual U-Net for medical image segmentation. *Journal of medical imaging, 6*(1), 014006-014006. [CrossRef]

[38] Chen, L. C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018, September). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision* (pp. 833-851). Cham: Springer International Publishing. [CrossRef]

[39] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B. (2021, October). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF international conference on computer vision (ICCV)* (pp. 9992-10002). IEEE. [CrossRef]

[40] Madani, A., Namazi, B., Altieri, M. S., Hashimoto, D. A., Rivera, A. M., Pucher, P. H., ... & Alseidi, A. (2020). Artificial intelligence for intraoperative guidance: using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of surgery, 276*(2), 363. [CrossRef]

[41] Silva, B., Oliveira, B., Morais, P., Buschle, L. R., Correia-Pinto, J., Lima, E., & Vilaca, J. L. (2022). Analysis of current deep learning networks for semantic segmentation of anatomical structures in laparoscopic surgery. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 3502-3505). IEEE. [CrossRef]

[42] Khan, U., Nawaz, U., Qayyum, A., Baig, I., Alam, M., Ashraf, S., ... & Qadir, J. (2026). Surgical scene understanding in the era of foundation AI models: A comprehensive review. *IEEE Access, 14*, 137034-137089. [CrossRef]

[43] Bian, G. B., Peng, Y., Zhang, L., Li, J., & Li, Z. (2024, June). Algorithms in surgical action recognition: A survey. In *2024 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 366-371). IEEE. [CrossRef]

[44] Padoy, N. (2019). Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy & Allied Technologies, 28*(2), 82-90. [CrossRef]

[45] Guo, K., Tao, H., Zhu, Y., Li, B., Fang, C., Qian, Y., & Yang, J. (2023). Current applications of artificial intelligence-based computer vision in laparoscopic surgery. *Laparoscopic, Endoscopic and Robotic Surgery, 6*(3), 91-96. [CrossRef]

[46] Nwoye, C. I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., ... & Padoy, N. (2022). Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis, 78*, 102433. [CrossRef]

[47] Batić, D., Holm, F., Özsoy, E., Czempiel, T., & Navab, N. (2024). EndoViT: pretraining vision transformers on a large collection of endoscopic images. *International Journal of Computer Assisted Radiology and Surgery, 19*(6), 1085-1091. [CrossRef]

[48] Liu, Y., Boels, M., Garcia-Peraza-Herrera, L. C., Vercauteren, T., Dasgupta, P., Granados, A., & Ourselin, S. (2025). LoViT: Long video transformer for surgical phase recognition. *Medical Image Analysis, 99*, 103366. [CrossRef]

[49] Chadebecq, F., Lovat, L. B., & Stoyanov, D. (2023). Artificial intelligence and automation in endoscopy and surgery. *Nature Reviews Gastroenterology & Hepatology, 20*(3), 171-182. [CrossRef]

[50] Rivas-Blanco, I., Perez-Del-Pulgar, C. J., Garcia-Morales, I., & Munoz, V. F. (2021). A review on deep learning in minimally invasive surgery. *IEEE Access, 9*, 48658-48678. [CrossRef]

[51] Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., & Heng, P. A. (2021). Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging, 40*(7), 1911-1923. [CrossRef]

[52] Kadkhodamohammadi, A., Luengo, I., & Stoyanov, D. (2022). PATG: position-aware temporal graph networks for surgical phase recognition on laparoscopic videos. *International Journal of Computer Assisted Radiology and Surgery, 17*(5), 849-856. [CrossRef]

[53] Funke, I., Bodenstedt, S., Oehme, F., von Bechtolsheim, F., Weitz, J., & Speidel, S. (2019, October). Using 3D convolutional neural networks to learn spatiotemporal features for automatic surgical gesture recognition

in video. In *International conference on medical image computing and computer-assisted intervention* (pp. 467-475). Cham: Springer International Publishing. [CrossRef]

[54] Seenivasan, L., Mitheran, S., Islam, M., & Ren, H. (2022). Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robotics and Automation Letters, 7*(2), 3858-3865. [CrossRef]

[55] Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C. W., & Heng, P. A. (2017). SV-RCNet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging, 37*(5), 1114-1126. [CrossRef]

[56] Yang, Z., Wen, J., Huang, D., Yang, A., Zhang, R., Ren, B., ... & Qin, N. (2025). Efficient segmentation of intraoperative anatomical landmarks in laparoscopic cholecystectomy based on deep learning. *Computer Methods and Programs in Biomedicine*, 109012. [CrossRef]

[57] Leifman, G., Golany, T., Rivlin, E., Khoury, W., Assalia, A., & Reissman, P. (2024). Real-time artificial intelligence validation of critical view of safety in laparoscopic cholecystectomy. *Intelligence-Based Medicine, 10*, 100153. [CrossRef]

[58] Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., ... & Padoy, N. (2023). The endoscapes dataset for surgical scene segmentation, object detection, and critical view of safety assessment: Official splits and benchmark. *Scientific Data, 12*. [CrossRef]

[59] Hong, W. Y., Kao, C. L., Kuo, Y. H., Wang, J. R., Chang, W. L., & Shih, C. S. (2020). Cholecseg8k: a semantic segmentation dataset for laparoscopic cholecystectomy based on cholec80. *arXiv preprint arXiv:2012.12453*. [CrossRef]

[60] Alabi, O., Toe, K. K. Z., Zhou, Z., Budd, C., Raison, N., Shi, M., & Vercauteren, T. (2025). Cholecinstanceseg: A tool instance segmentation dataset for laparoscopic surgery. *Scientific Data, 12*(1), 825. [CrossRef]

[61] Allan, M., Kondo, S., Bodenstedt, S., Leger, S., Kadkhodamohammadi, R., Luengo, I., ... & Speidel, S. (2020). 2018 robotic scene segmentation challenge. *arXiv preprint arXiv:2001.11190*. [CrossRef]

[62] Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC medical imaging, 15*(1), 29. [CrossRef]

[63] Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., & Navab, N. (2020, September). Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *International conference on medical image computing and computer-assisted intervention* (pp. 343-352). Cham: Springer International Publishing. [CrossRef]

[64] Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., ... & Padoy, N. (2023). Dissecting self-supervised learning methods for surgical computer vision. *Medical Image Analysis, 88*, 102844. [CrossRef]

[65] Özsoy, E., Örnek, E. P., Eck, U., Czempiel, T., Tombari, F., & Navab, N. (2022, September). 4d-or: Semantic scene graphs for or domain modeling. In *International conference on medical image computing and computer-assisted intervention* (pp. 475-485). Cham: Springer Nature Switzerland. [CrossRef]

[66] Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., Fiorillo, C., ... & Padoy, N. (2022). Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. *Annals of surgery, 275*(5), 955-961. [CrossRef]

[67] Van der Velden, B. H., Kuijf, H. J., Gilhuijs, K. G., & Viergever, M. A. (2022). Explainable artificial intelligence (XAI) in deep learning-based medical image analysis. *Medical image analysis, 79*, 102470. [CrossRef]

[68] Özsoy, E., Pellegrini, C., Bani-Harouni, D., Yuan, K., Keicher, M., & Navab, N. (2026). Specialized foundation models for intelligent operating rooms. *NPJ Digital Medicine*. [CrossRef]

Hafsa Gulzar is currently a Ph.D. researcher in Institute of Biomedical and Health Engineering, Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. She received her master's degree in engineering computer science and technology from Donghua University, Shanghai, China, in 2023. Her research interests include artificial intelligence, computer vision, medical image analysis, surgical video understanding, and machine learning for healthcare applications. (Email: hafsa@mail.dhu.edu.cn)

Muhammad Bilal is currently a Ph.D. researcher in Computer Science at Iowa State University, Ames, USA. He received his bachelor's degree from the Pakistan Institute of Engineering and Applied Sciences (PIEAS), Pakistan. His research interests span computer vision, machine learning, deep learning, large language models, multimodal learning, and artificial intelligence. (Email: mbilal@iastate.edu)