Title: Current Artificial Intelligence models for Laparoscopic Surgical Phase Recognition (SPR): A systematic review.

Author: Ngwaru Munodawafa Email: munodawafan@gmail.com

Introduction

After performing a surgical operation, the surgeon must write a report of what happened during the surgical procedure which includes the surgical steps they performed. This report is then used as a base document to manage the individual patient and maybe used to investigate the effect of procedures on patient outcomes in clinical research. However, studies have shown that the surgical report is not always accurate[1,2]. Sometimes steps that were not done during the surgery are reported as if they were done and vice versa[1]. Therefore, there is need to have an objective text-based record. Automatic surgical phase recognition using artificial intelligence is the first step in achieving automated surgical reports.

Laparoscopic surgery is a type of minimally invasive surgery that involves operating on organs in the abdomen through small holes, using a camera to indirectly observe the surgery. The availability of video data in this type of surgery presents an opportunity for providing automated analysis and reporting including through identification of surgical phases on laparoscopic videos [3,4] . This review studies the different models that have been applied to identify surgical phases on laparoscopic videos.

The relevance of surgical phase recognition (SPR) is critical as it serves as the prerequisite for numerous downstream applications. These applications include automated surgical skill assessment and real time feedback, providing bases for automated surgical report, generating video summaries, and providing context-aware intraoperative guidance and decision support systems. For instance, surgical phase recognition (SPR) enables systems to automatically assess safety metrics, such as the Critical View of Safety (CVS) in laparoscopic cholecystectomy. The Critical View of Safety is a specific view of the gall bladder that shows the surgeon has dissected it from the liver safely[5]. If Critical View of Safety is identified as part of surgical phase recognition, then the specific part of the surgery was performed safely. Using surgical phase recognition annotated videos can help surgeons quickly review videos. As mentioned earlier surgical phase recognition (SRP) can be used to assess the skills of surgeons; expert surgeons usually perform faster on specific phases of a surgical procedure[3].

Online SPR can be used to show progress to the rest of the staff involved in the surgery to prepare for specific equipment needed at specific stages and prepare for the next patient when the procedure is about to end and shorten turn around time in the operating theatre.

The purpose of this mini systematic review is to synthesize recent advances in deep learning-based SPR, compare contemporary methodologies across different laparoscopic contexts, identify model and data challenges, and propose a viable research niche for future exploration in data science.

The models mostly involve a model for spatial feature extraction from individual frames together with a temporal model to combine the spatial and temporal features[7]. These models perform comparably to expert surgeons at laparoscopic phase recognition in uncomplicated cases[3].

Previous studies had experimented with the use of autoencoders, reinforcement learning, graph networks and transformers [8, 9,10,11, 12, 13]. These papers were not included in this review as they were published before the inclusion period of January 2024 to September 2025.

Method

This systematic literature review is based on a structured examination of peer-reviewed scientific papers focusing on automated surgical phase recognition using deep learning models on intraoperative laparoscopic videos.

Databases and Keywords: A database search was done to identify peer reviewed papers. A Google Scholar and Scopus search for titles that include **"laparoscopic phase recognition", "phase recognition"** and **"video phase recognition"** in the title of the article. It included articles published from 01/2024 to 9/2025. A total of 19 papers were identified after applying the inclusion and exclusion criteria 8 papers were left and are part of this review these papers are 14, 15, 16,17, 18, 19, 20 and 21 in the reference.[14 15, 16, 17, 18, 19, 20, 21]. No search was done for articles that had the key phrase in other parts of the document (e.g. body, metadata) . Keywords had to be in the title.

Inclusion Criteria: Studies selected had to meet the minimum standards as follows:

1. Application of deep learning models for the automatic classification of video frames to surgical phases, segmentation of surgical phases or identification of surgical phase transition zones on laparoscopic video recording.
2. Used laparoscopic videos
3. Peer reviewed papers.
4. Published from January 2024 to September 2025,
5. Focused on laparoscopic phase recognition using video images and using a single camera view.
6. The papers that were included needed to have a description of the model that was used for the automated surgical phase recognition.

Exclusion Criteria: Excluded papers that were:

1. Not peer reviewed
2. Published before January 2024 or after September 2025,
3. Used other forms of surgical videos other than laparoscopy e. g. intramural endoscopic surgery, arthroscopy, thoracoscopy
4. Used other modalities e.g. speech recognition
5. Had multi camera input.
6. Did not detail the type of deep learning model used

Screening Steps: Identified papers that mat the inclusion criteria and removed papers according to the exclusion criteria. Eight papers were identified. A lot of the papers that were published in surgical journal lacked detail on the model and the training procedure[1]. Whilst other venues had more detailed description of the model and the training procedure.

Findings:Themes

A. Improvements on Training Approaches (Pretraining, Joint Spatial-Temporal training, Changes to loss function)

Different studies used different datasets for pretraining e.g. Kinetics Dataset[14] and ImageNet 2012 [16, 17, 18, 20]. However, none of the models used a dataset of surgical images for pretraining. The other studies did not disclose what the datasets that were used for the pretraining. For Training, validation and testing the studies used publicly available laparoscopic datasets that have surgical phases annotated already. The cholec80 dataset was used by 7/8 of the studies either alone or together with other datasets e.g. AutoLaparo

Traditional methods had focused on extracting the spatial features from a single frame through training to classify each frame to specific phase then use the trained spatial feature extractor (Usually a CNN with the final layer removed) to give input to temporal model e.g. LSTM or Temporal Convolutional Network[7]. This method has challenges as some images of the surgical field maybe identical but belong to different phases of the surgery. This is further worsened by the nature of surgical videography which might be shaky at times, be obscured by blood or smoke from the electrocautery. Five of the studies out of the eight applied strategies that trained the models on both temporal and spatial features at the same time. Li Y et al 2024[21] and Liu Y et al 2025[17] used method sequence of short video clips to train spatial-temporal feature extractor. Two other studies[14, 16] used 3D CNN or 2D CNN.

Most of the papers applied Categorical Cross-Entropy loss or binary cross entropy[14, 15,]. Some papers applied innovative loss functions. The HecVL model which mapped short clip level videos to narrations, surgical phase level video clips to a summary description of the phase and the whole video a summary of the surgery used different loss functions at different levels[18]. At clip level they used different loss functions for at different levels. Liu Y *et al* 2025[17] LoViT paper used a custom loss function that applied a dynamic penalty to misclassification of frames close transition zones using a heat map.

$$\mathcal{L}^* = \mathcal{L}_1(\hat{h}, h) + \mathcal{L}_{CE}(\hat{p}, p),$$

The first term, $\mathcal{L}_1(\hat{h}, h)$, is the L1 loss that refines the model's sensitivity to phase transitions by comparing predictions to ground truth. The second term, $\mathcal{L}_{CE}(\hat{p}, p)$, is the cross-entropy loss that evaluates phase classification accuracy. Together, they align transition detection with phase prediction to improve performance on surgical video data[17]. Xia y et al 2025[20] improved on this loss function by adding the term $\mathcal{L}_{IoU}$.

$$\mathcal{L} = \mathcal{L}_1(\hat{h}, h) + \mathcal{L}_{CE}(\hat{p}, p) + \mathcal{L}_{IoU} \qquad \mathcal{L}_{IoU} = 1 - \frac{\min(\hat{s}_{start}, s_{start}) + \min(\hat{s}_{end}, s_{end})}{\max(\hat{s}_{start}, s_{start}) + \max(\hat{s}_{end}, s_{end})}$$

Where S_{end}, S_{start} are the actual values and $\hat{}s_{end}$, $\hat{}s_{start}$ the predicted distance between the centre point and the start or end point of a phase. Given that this model had the best accuracy changes to the loss function may give better performance[20].

Post processing: Most of the papers applied post processing step after the individual video frames were classified this might lead to inaccurate reflection of the model's performance.

B. Central role of Convolutional Neural Networks (CNN)

In all the studies CNNs were used as part of the model. CNN were used as stand-alone feature extraction models[15,19], as visual feature embedding models for transformers[17,18, 20, 21] or as part of the model architecture in 3D CNN (34-layer 3D-ResNet) or 2D CNN [14, 16]. ResNet 50 was the most used model for these tasks in most of the models.

C. Model Architecture and Temporal Modelling: Move towards attention-based models

Successful SPR requires modelling both spatial visual features and long-range temporal dependencies. Early models used CNN-RNN combinations (ResNet-LSTM) or CNN-Temporal Convolutional Network (TCN). Recently, investigators have used attention as part of the architecture in 6 out of the 8 selected papers [15, 17, 18, 19, 20, 21]. The other papers used 3D CNN [14] or 2D CNN (Temporal Shift Module (TSM))[16].

The models that used attention as part of their architecture can be split into the ones that used the transformer architecture (4/8)[17,18, 20, 21] and the ones that applied attention as part LSTM (2/8) [15, 19]. The main goal was to try and apply attention to both long distance and short distance relationships between video frames.

1.For the transformer architecture-based models they were mostly built upon the Trans SVNet architecture. Li. Y *et al* 2024 employed a binary classifier for each surgical phase that would be used if the multi class classifier had low confidence score. The confidence score was adjusted using temperature regularize scaling this improved performance on the original Trans SVNet on the cholec80 dataset[21]. The binary classifier only considered the current phase or the next phase. Another example is the Transformers for Long Videos (LoViT): To overcome the quadratic computational complexity of traditional self-attention in long surgical videos, the Long Video Transformer (LoViT) was developed[17]. LoViT employs a multi-scale temporal aggregator consisting of local L-Trans modules and a Global Informer (G-Informer) module utilizing ProbSparse self-attention to reduce complexity[17]. This did improve on the accuracy of the surgical phase recognition to 92.4%[17]

2. Attention was also applied to LSTM based models. For example TUNeS (Temporal U-Net with Self-Attention) combines the local inductive bias of a convolutional U-Net structure with efficient global self-attention integrated only at the bottleneck of the U-Net[19]. This approach models global dependencies on temporally down sampled, semantically rich features, balancing computational cost and comprehensive analysis[19]. Furthermore, achieving high performance often requires training the visual feature extractor (CNN) not just on individual frames, but using extended temporal context (up to 64 frames) jointly with an LSTM, which significantly improves the feature extractor's capability

D. Evaluation

Though most of the papers reported on accuracy, recall, precision and F1 score, some often left out some of the metrics. Accuracy was the most reported metric in 7/8 of the papers[14, 15, 17, 18,19, 20, 21.] The accuracy ranged from 88% - 94.3% for the papers 14, 15, 17, 19, 20 and 21 this is close to expert surgeon performance on simple cases[3]. The model HecVL had poor accuracy of 41.7%

with zero short prompting[17.] However, accuracy obscures poor performance on surgical phases that had fewer frames in the videos.

The study You, J *et al* 2024 reported on mean average precision and did report on accuracy, recall or F1 scores[16]. Li, Y. et al 2024 only reported overall accuracy[21]. In the papers 17, 18,19 and 20 reported accuracy and F1 of the entire video. The papers 14 and15 had a more detailed approach to reporting they highlighted accuracy, precision, recall and F1 score for each phase and the entire video

Gaps

CNN usage: Though all the paper utilised CNN (mostly ResNet50) no paper explored the possibility of improving ResNet 50 by fine tuning it on open-source laparoscopic datasets for example GynSurg, M2CAI16 or JIGSAWS [7]. This type of pretraining can also be applied to ViT models and evaluate their performance against models that were pretrained on general datasets like ImageNet. Though ResNet50 was used for vector embedding there was no exploration of the benefit of using other image embedding methods like patch embedding with positional embedding for visual transformers. This needs to be investigated in relation to surgical phase recognition to see if it provides better vector representations.

Loss function changes had a positive impact on performance especially using transition zones to map dynamic penalty to the loss [17, 20]. More experimentation might yield even better models for SPR.

Only one paper (1/8) tried to balance the underrepresented phases by using sampling more from them[15]. The authors showed that this improved F1 scores on the surgical phases with fewer frames in the video[15]. Applying class balancing may in other architecture may improve performance.

The reporting of model performance was haphazard, and standardization is needed with need to publish model code and datasets.

For the transformer-based models inference takes a longer time for longer videos and using alternative architecture e.g. the Mamba model. This has been tried but the paper was not peer reviewed and had limited reporting on performance[26].

Conclusion

In conclusion, current AI models for laparoscopic surgical phase recognition (SPR) demonstrate strong performance, with accuracies approaching expert surgeon levels in simple cases with clear steps[3]. Most models combine CNN-based spatial feature extraction with attention-based temporal modelling, particularly transformers. However, challenges remain, including reliance on non-surgical pretraining datasets, inconsistent evaluation metrics, prolonged inference time and limited robustness to intraoperative variability. Future work should focus on standardized evaluation reporting, domain-specific pretraining, loss function experimentation and model generalization across surgical phases and contexts to enable reliable, real-world clinical deployment of SPR systems and exploration of other attention architectures.

References

1. Sharma, V., Gettman, M. T., Boorjian, S. A., Asselmann, D., & Tollefson, M. K. Enhancing Accuracy of Operative Reports with Automated Artificial Intelligence Analysis of Surgical Video. *Journal of the American College of Surgeons*, *240*(5), 739–746. (2025). https://doi.org/10.1097/XCS.0000000000001352

2. Wauben, L. S. G. L., Van Grevenstein, W. M. U., Goossens, R. H. M., Van Der Meulen, F. H., & Lange, J. F. Operative notes do not reflect reality in laparoscopic cholecystectomy. *Journal of British Surgery*, *98*(10), 1431-1436. (2011) https://doi.org/10.1002/bjs.7576

3. Golany, T., Aides, A., Freedman, D. *et al.* Artificial intelligence for phase recognition in complex laparoscopic cholecystectomy. *Surg Endosc* **36**, 9215–9223 (2022). https://doi.org/10.1007/s00464-022-09405-5

4. Takeuchi, M., Collins, T., Ndagijimana, A. *et al.* Automatic surgical phase recognition in laparoscopic inguinal hernia repair with artificial intelligence. *Hernia* **26**, 1669–1678 (2022). https://doi.org/10.1007/s10029-022-02621-x

5. Kaya, B., Fersahoglu, M. M., Kilic, F., Onur, E., & Memisoglu, K. Importance of critical view of safety in laparoscopic cholecystectomy: a survey of 120 serial patients, with no incidence of complications. *Annals of hepato-biliary-pancreatic surgery*, *21*(1), 17–20. (2017). https://doi.org/10.14701/ahbps.2017.21.1.17

6. Shinozuka, K., Turuda, S., Fujinaga, A. *et al.* Artificial intelligence software available for medical devices: surgical phase recognition in laparoscopic cholecystectomy. *Surg Endosc* **36**, 7444–7452 (2022). https://doi.org/10.1007/s00464-022-09160-7

7. Liao, W., Zhu, Y., Zhang, H., Wang, D., Zhang, L., Chen, T., Zhou, R., & Ye, Z. Artificial intelligence-assisted phase recognition and skill assessment in laparoscopic surgery: a systematic review. *Frontiers in surgery* **12**, 1551838. (2025). https://doi.org/10.3389/fsurg.2025.1551838

8. Zhai, Y., Chen, Z., Zheng, Z. *et al.* Artificial intelligence for automatic surgical phase recognition of laparoscopic gastrectomy in gastric cancer. *Int J CARS* **19**, 345–353 (2024). https://doi.org/10.1007/s11548-023-03027-5

9. Kadkhodamohammadi, A., Luengo, I. & Stoyanov, D. PATG: position-aware temporal graph networks for surgical phase recognition on laparoscopic videos. *Int J CARS* **17**, 849–856 (2022). https://doi.org/10.1007/s11548-022-02600-8

10. Zou, X., Liu, W., Wang, J., Tao, R., & Zheng, G. ARST: auto-regressive surgical transformer for phase recognition from laparoscopic videos. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization* ***11***(4), 1012–1018, (2022). https://doi.org/10.1080/21681163.2022.2145238

11. Eckhoff, J.A., Ban, Y., Rosman, G. *et al.* TEsoNet: knowledge transfer in surgical phase recognition from laparoscopic sleeve gastrectomy to the laparoscopic part of Ivor–Lewis esophagectomy. *Surg Endosc* **37**, 4040–4053 (2023). https://doi.org/10.1007/s00464-023-09971-2

12. Konduri, P. S., & Rao, G. S. N. Surgical phase recognition in laparoscopic videos using gated capsule autoencoder model. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, *11*(5), 1973–1995, (2023). https://doi.org/10.1080/21681163.2023.2203280

13. Zhang, Y., Bano, S., Page, AS., Deprest, J., Stoyanov, D., Vasconcelos, F. Retrieval of Surgical Phase Transitions Using Reinforcement Learning. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. MICCAI 2022. Lecture Notes in Computer Science, vol 13437. (2022). https://doi.org/10.1007/978-3-031-16449-1_47

14. Yang, H. Y., Hong, S. S., Yoon, J. *et al.* Deep learning-based surgical phase recognition in laparoscopic cholecystectomy. Korean Association of Hepato-Biliary-Pancreatic Surgery 28 (4) 466-473 (2024). https://dx.doi.org/10.14701/ahbps.24-091

15. Liu, M., Duan, F., Ling, L. et al. Dynamic data balancing strategy-based Xception-dual-channel LSTM model for laparoscopic cholecystectomy phase recognition. Int J CARS (2025). https://doi-org.ezp.sub.su.se/10.1007/s11548-025-03509-8

16. You, J., Cai, H., Wang, Y. *et al.* Artificial intelligence automated surgical phases recognition in intraoperative videos of laparoscopic pancreatoduodenectomy. *Surg Endosc* **38**, 4894–4905 (2024). https://doi-org.ezp.sub.su.se/10.1007/s00464-024-10916-6

17. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., *et al.* LoViT: Long Video Transformer for surgical phase recognition, *Medical Image Analysis*,**99**, 103366 (2025), https://doi.org/10.1016/j.media.2024.103366

18. Yuan, K., Srivastav, V., Navab, N., Padoy, N. (2024). HecVL: Hierarchical Video-Language Pretraining for Zero-Shot Surgical Phase Recognition. In: Linguraru, M.G., *et al.* Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. MICCAI 2024.

Lecture Notes in Computer Science, vol 15006. Springer, Cham. https://doi.org/10.1007/978-3-031-72089-5_29

19. I. Funke, D. Rivoir, S. Krell and S. Speidel, "TUNeS: A Temporal U-Net With Self-Attention for Video-Based Surgical Phase Recognition," in *IEEE Transactions on Biomedical Engineering*, vol. 72, no. 7, pp. 2105-2119, July 2025, https://doi.org/10.1109/TBME.2025.3535228

20. Y. Xia, S. Song, Y. Yang, T. Lu and R. Xiao, "Surgical Video Phase Recognition Model Integrating Phase Localization Module and Temporal Context Method," *2025 5th International Conference on Sensors and Information Technology*, Nanjing, China, 2025, pp. 620-623, https://doi.org/10.1109/ICSI64877.2025.11009385.

21. Li, Y., Gupta, H., Prasanna, P., Ling I. H., Surgical Phase Recognition in Laparoscopic Cholecystectomy, Procedia Computer Science, **239**, 2006-2012, 2024, https://doi.org/10.1016/j.procs.2024.06.386

22. Yanagida, Y., Takenaka, S., Kitaguchi, D. *et al.* Surgical skill assessment using an AI-based surgical phase recognition model for laparoscopic cholecystectomy. *Surg Endosc* **39**, 5018–5026 (2025). https://doi.org/10.1007/s00464-025-11903-1

23. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, PA. Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer. In: de Bruijne, M., *et al.* Medical Image Computing and Computer Assisted Intervention – MICCAI 2021. MICCAI 2021. Lecture Notes in Computer Science(), vol 12904. Springer, Cham. (2021) https://doi.org/10.1007/978-3-030-87202-1_57

24. Cheng, K., You, J., Wu, S. *et al.* Artificial intelligence-based automated laparoscopic cholecystectomy surgical phase recognition and analysis. *Surg Endosc* **36**, 3160–3168 (2022). https://doi.org/10.1007/s00464-021-08619-3

25. Komatsu, M., Kitaguchi, D., Yura, M. *et al.* Automatic surgical phase recognition-based skill assessment in laparoscopic distal gastrectomy using multicenter videos. *Gastric Cancer* **27**, 187–196 (2024). https://doi.org/10.1007/s10120-023-01450-w

26. Cao, R., Wang, J., & Liu, Y. H. (2024). Sr-mamba: Effective surgical phase recognition with state space model. arXiv preprint arXiv:2407.08333.