

SurgX: Neuron-Concept Association for Explainable Surgical Phase Recognition

Ka Young Kim, Hyeon Bae Kim, and Seong Tae Kim*

Kyung Hee University, Yongin, Republic of Korea
{uwrgoy7584, hyeonbae.kim, st.kim}@khu.ac.kr

Abstract. Surgical phase recognition plays a crucial role in surgical workflow analysis, enabling various applications such as surgical monitoring, skill assessment, and workflow optimization. Despite significant advancements in deep learning-based surgical phase recognition, these models remain inherently opaque, making it difficult to understand how they make decisions. This lack of interpretability hinders trust and makes it challenging to debug the model. To address this challenge, we propose SurgX, a novel concept-based explanation framework that enhances the interpretability of surgical phase recognition models by associating neurons with relevant concepts. In this paper, we introduce the process of selecting representative example sequences for neurons, constructing a concept set tailored to the surgical video dataset, associating neurons with concepts and identifying neurons crucial for predictions. Through extensive experiments on two surgical phase recognition models, we validate our method and analyze the explanation for prediction. This highlights the potential of our method in explaining surgical phase recognition. The code is available at <https://github.com/ailab-kyunghee/SurgX>.

Keywords: Surgical Phase Recognition · Endoscopic Video · Interpretability · Explainability · Neuron-Concept Association

1 Introduction

Surgical workflow analysis is crucial for computer-assisted surgery, as it enables surgical artificial intelligence systems to understand the sequential and temporal dynamics inherent in surgical procedures [14,17,21]. Surgical phase recognition stands out as one of the critical tasks in surgical workflow analysis, classifying surgical video frames into their respective phase. Reflecting its clinical importance, a growing body of work has focused on developing advanced deep learning-based surgical phase recognition models [4,5,15,16,26].

Although deep learning models have made remarkable progress in surgical phase recognition, they still encounter persistent challenges stemming from their inherent “black-box” nature, which refers to the difficulty in interpreting how these models process information and make decisions. The major challenge

* Corresponding author

is the difficulty in trusting the model’s decisions. In surgical domains, models are required to have a transparent decision-making process in compliance with regulations like the European Union’s General Data Protection Regulation (GDPR) [23,25], which emphasizes the need for sufficient explanations of model decisions. If we cannot determine which concepts the model has learned and why it makes certain decisions, effectively improving the model and trusting its decisions becomes a significant challenge. This challenge is further exacerbated in video-based models due to their inherent temporal dependency and complexity of models.

Previous studies on concept-based explainability for image-based models have shown that neurons in deep networks tend to align with human-interpretable concepts [1,3,9,12,11]. One of the earliest approaches, Network Dissect [3] associates neurons with human-understandable concepts by computing the overlap between pixel-wise segmentation masks and the feature maps of individual neurons. To address the scarcity of expensive segmentation mask datasets, Oikarinen *et al.* proposed leveraging CLIP [22], a Vision-Language Model (VLM), to perform neuron-concept association [20]. In the medical domain, where segmentation mask datasets are even more limited, [12] demonstrated that associating neurons with concepts is effective in interpreting medical diagnosis models.

While explainability in image-based models has been extensively studied, its application to surgical video models remains largely underexplored, as their inherent temporal complexities introduce unique challenges in understanding decision-making processes. To bridge this gap, we propose **SurgX**, a novel concept-based explanation framework specifically designed for interpreting surgical phase recognition models with concepts. We introduce a systematic approach for selecting representative example sequences for each neuron, constructing concept sets, and identifying the most influential neurons in each prediction. We employ SurgX to interpret two surgical phase recognition models (TeCNO [4] and Causal ASFormer [27]) and evaluate our methods by examining whether the identified concept-neuron associations are meaningful and whether they contribute to explaining the model’s predictions. Our main contributions are as follows:

- We propose SurgX, a novel concept-based explanation framework, to interpret surgical phase recognition models. To the best of our knowledge, this is the first study to provide a concept-based explanation for surgical phase recognition models.
- To ensure that concepts are appropriately associated with neurons in surgical phase recognition models, we introduce methods to construct the concept set specialized for a cholecystectomy video dataset. We then analyze how different concept sets can be used to explain model behavior more effectively, highlighting best practices for concept set selection in surgical applications.
- We validate SurgX on two surgical phase recognition models by examining whether the identified concept-neuron associations are meaningful and whether they contribute to explaining the model’s predictions.

2 Methodology

2.1 Concept Set Construction

To ensure accurate neuron-concept annotation, it is essential to use a properly collected set of concepts that the neurons may have learned from the dataset. We introduce three concept sets focused on cholecystectomy-related concepts. First, we leverage the action triplets labels provided by the CholecT45 [19] dataset to build two concept sets, CholecT45-W and CholecT45-S. CholecT45-W comprises 30 words (e.g., grasper, clipping, gallbladder) derived from the action triplet labels in CholecT45. CholecT45-S contains 100 sentences created from the same action triplets following the prompting approach as in [28] (e.g., I use a hook to dissect the cystic plate). Lastly, the ChoLec-270 set was constructed by collecting a total of 270 concepts extracted from 11 cholecystectomy lecture videos available on the open surgical e-learning platform WebSurg, as well as from 4 cholecystectomy-related articles [7,13,18,24].

2.2 Neuron Representative Sequence Selection

In this work, each neuron’s concept association is determined by measuring the similarity between its representative sequences and candidate concepts, which underscores the importance of selecting sequences that accurately reflect each neuron’s learned patterns. To this end, we follow a two-step process, illustrated in Figure 1. First, we select the example frames from the video probing set that highly activate each neuron. Since we use activations after the ReLU function, negative values are not considered. Second, we construct sequences by including n previous frames with a dilation rate of m from each selected example frame, considering that in surgical phase recognition models, the activation of a neuron at a given feature is influenced not only by the selected frame but also by previous frames.

2.3 Neuron-Concept Annotation

Figure 1 shows an overview of the neuron concept annotation. Using the concept set constructed in Section 2.1 and the neuron representative example set selected in Section 2.2, we perform concept annotation by following [1]. First, we extract text features from the concept set and visual features from neuron representative example set. Then, we compute the cosine similarity between the text and visual features.

$$s_{n,c} = \frac{1}{E} \sum_{e=1}^E \cos(v_{n,e}, t_c), \quad (1)$$

where $s_{n,c}$ denotes the concept score between n -th neuron and c -th concept which quantifies the extent to which a neuron represents a concept. For each of the E frames in the n -th neuron’s representative example set, we compute the cosine similarity between the visual feature $v_{n,e}$ of e -th frame and the text feature t_c of

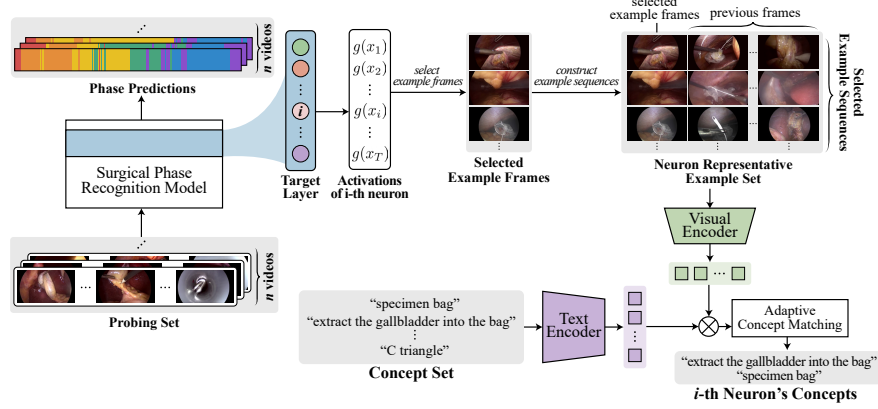


Fig. 1. Neuron concept annotation method for surgical phase recognition model. To interpret the neuron representations with concepts, neuron representative examples are selected, and a text-based concept set is prepared. The neuron representative examples are constructed based on the frames that are highly activated for the neuron in the video probing set. Both the example set and the concept set are projected into the vision-language model embedding space, where the neuron’s concept is identified through a similarity-based concept matching method.

c -th concept. The final score $s_{n,c}$ is obtained by averaging these similarities across all selected frames. If $s_{n,c}$ exceeds a threshold $\theta_{concept}$, c -th concept is annotated to n -th neuron. The threshold is calculated as $\theta_{concept} = \beta \times \max(s_{n,c})$, where β is fixed at 0.95 in this paper.

2.4 Explaining Model’s Decision using Highly Contributed Neurons

After annotating each neuron with concepts, the model’s predictions can be explained based on the concepts of important neurons that significantly contributed to its predictions. The process of explaining a prediction is illustrated in Figure 2. When a surgical phase recognition model predicts the phase of a specific frame, various neurons in the penultimate layer, each representing different concepts, contribute to the prediction with different extents. Inspired by [10,2], we first calculate the contribution of each neuron to the final prediction to explain the model decisions as follows:

$$c_{p,i}(x) = |f_{\theta}(x) - f_{\theta}(x; a_{p,i} \leftarrow 0)| = |a_{p,i} \nabla_{a_{p,i}} f_{\theta}(x)|, \quad (2)$$

where $f_{\theta}(x)$ is the prediction of model with parameter θ for input x , and $a_{p,i}$ is the activation of the i -th neuron for class p . The term $f_{\theta}(x; a_{p,i} \leftarrow 0)$ represents the prediction without this neuron’s influence, making $|f_{\theta}(x) - f_{\theta}(x; a_{p,i} \leftarrow 0)|$ be a measure of its contribution. Then, we identify the neurons that had a significant impact on the prediction. By utilizing the concepts matched to these highly contributing neurons, interpretable explanations of why the model made a particular decision can be provided.

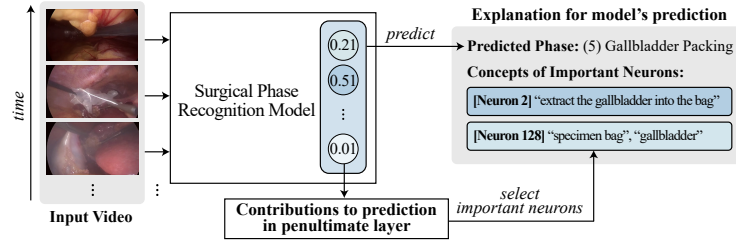


Fig. 2. Explanation of model’s decisions. Important neurons are selected based on the contribution of the neurons in the penultimate layer to the model’s prediction. The concepts annotated to these important neurons are used to explain the model decision.

3 Experiments

3.1 Experimental Setup

Model. As most surgical phase recognition models rely on either TCN or Transformer architectures, we select TeCNO [4] and Causal ASFormer [27] as target models to interpret and highlight the generality of SurgX. TeCNO is a TCN-based model commonly used for surgical phase recognition. On the other hand, ASFormer is a Transformer-based model widely used for temporal action segmentation in the computer vision domain. To reflect real-world clinical settings, we modify ASFormer to operate causally using only current and past frames, and we refer to it as Causal ASFormer in this paper. LoViT [15] features are used as the spatial features for Causal ASFormer.

Additionally, for extracting visual features from neuron representative sequences and text features from concepts, we employ SurgVLP [28], a vision-language model specifically trained for surgical domain. Using the visual encoder of SurgVLP to extract visual features from representative frames avoids the gap between the embedding spaces of the phase recognition model and SurgVLP.

Dataset. We employ the Cholec80 dataset [26], which comprises 80 laparoscopic cholecystectomy videos annotated with surgical phases. Following previous works [6,8], we split the dataset into 40 training videos and 40 test videos. The training set is used for both model training and probing set in Section 2.2, while the test set is reserved for evaluating the interpretability of SurgX.

Evaluation Metrics. To quantitatively evaluate SurgX, we are inspired by the evaluation metrics in [1,20,12] and evaluate the interpretability of our approach based on two key aspects. First, we define Concept Alignment Score to evaluate how well concepts are annotated to neurons by measuring the cosine similarity between the concepts assigned to neurons in the final layer and the ground-truth class as in [1,20,12]. Second, we define Prediction Interpretability Score to evaluate how well the model’s decisions can be explained by measuring the cosine similarity between the model’s prediction and the concepts annotated to highly contributing neurons in the penultimate layer. For both aspects, we compute the cosine similarity between the concept and the phase expressed in word form, as

Table 1. Analysis of concept set construction methods.

Concept Set	Concept Alignment			Prediction Interpretability			Unique Concept
	Word	Sentence	Avg	Word	Sentence	Avg	
CholecT45-W	0.4078	0.3681	<u>0.3880</u>	0.5978	0.5099	<u>0.5539</u>	7
CholecT45-S	0.3000	0.4538	0.3769	0.4679	0.5564	0.5122	<u>11</u>
ChoLec-270	0.4153	0.4796	0.4475	0.6532	0.5452	0.5992	34

Table 2. Ablation study on the neuron representative frame selection methods.

Method	Concept Alignment			Prediction Interpretability		
	Word	Sentence	Avg	Word	Sentence	Avg
Global Threshold	0.3222	0.4621	0.3922	0.5586	0.4819	0.5203
Global TopK	0.3099	0.4622	0.3861	0.5369	0.4531	0.4950
Video-wise Threshold	0.4153	0.4796	<u>0.4475</u>	0.6532	0.5452	0.5992
Video-wise TopK	0.4146	0.4824	0.4485	0.6454	0.5403	<u>0.5929</u>

well as between the concept and the phase expressed in sentence form provided in [28], and then average the two scores, since all the concepts and frames are encoded by SurgVLP [28] which can be influenced by the text format. To interpret the scores, we present the (mean, std) of the cosine similarities between all phase texts and each concept set: CholecT45-W (0.17, 0.19), CholecT45-S (0.15, 0.20), and Cholec-270 (0.19, 0.20). Additionally, when evaluating concept sets, we count the number of Unique Concepts in the penultimate layer, which indicates how finely the concepts learned by neurons can be identified when using each concept set.

3.2 Quantitative Evaluation of SurgX

This section presents the ablation studies for each part introduced in Section 2, along with the quantitative results of SurgX. The ablation studies were conducted on Causal ASFomer. The Concept Alignment Score was measured on the output of the last Conv1d layer, the final layer, where the output dimension corresponds to the number of phases (i.e., 7). Additionally, the prediction interpretability score was measured on the output of the last transformer block, the penultimate layer, where the output dimension is 256.

Concept Set Construction. To evaluate the concept sets, we conducted experiments while keeping other settings fixed as Video-wise Threshold Selection and Dilated Sequences Selection with a 5-second interval. Table 1 shows that the Concept Alignment Score and the Prediction Interpretability Score of ChoLec-270 were the highest, demonstrating that using ChoLec-270 explains the model best among the compared concept sets.

Neuron Representative Frame Selection. This study is conducted under two settings: **Global setting**, where representative frames are selected from all video frames, and **Video-wise setting**, where frames are first selected for each video and then aggregated. In each setting, two selection approaches are applied:

Table 3. Ablation study on the neuron representative sequence selection methods using cosine similarity. The number in parentheses (#) denotes the dilation size.

Method	Concept Alignment			Prediction Interpretability		
	Word	Sentence	Avg	Word	Sentence	Avg
Single-Frame	0.3576	0.4569	0.4073	0.6426	0.5513	0.5970
Contiguous-Sequence	0.3641	0.4711	0.4176	0.6458	0.5487	0.5973
Dilated-Sequence (3)	0.3534	0.4920	<u>0.4227</u>	0.6545	0.5499	<u>0.6022</u>
Dilated-Sequence (5)	0.4153	0.4796	0.4475	0.6532	0.5452	0.5992
Dilated-Sequence (10)	0.3684	0.4004	0.3844	0.6856	0.5548	0.6202

(1) selecting the top-K frames based on activation values (K=40 for the global setting and K=1 for the video-wise setting) and (2) applying an adaptive thresholding method ($\alpha=0.95$) from Adaptive Neuron Representative Image Selection in [12]. Table 2 shows that the Video-wise Threshold method, which first selects frames for each video using an adaptive threshold before aggregation, achieves the best performance. In the global setting, frames are often concentrated in a few specific videos and include many adjacent frames, whereas the video-wise setting selects more diverse frames across videos, enabling the model to capture common concepts across different videos more effectively.

Representative-Frame-Based Sequence Construction. To consider the temporal attributes of the Surgical Phase Recognition model, we conducted an ablation study to construct a neuron representative frame sequence. Table 3 presents a comparison of performance between using a single frame and a sequence of 10 frames. The sequence of 10 frames was configured to include frames selected at 1-second intervals (Contiguous-Sequence) as well as frames selected at 3-second, 5-second, and 10-second intervals (Dilated-Sequence (3), Dilated-Sequence (5), and Dilated-Sequence (10), respectively). The results demonstrated that using a sequence of frames outperform annotating concepts with a single frame. Additionally, the application of the dilation technique, which enables the sequence to incorporate broader contextual information from preceding frames based on the representative frame, is found to improve performance.

3.3 Discussion

Figure 3 shows examples of explanations for predictions made by Causal ASFormer and TeCNO, visualizing the best-contributing neuron to the predictions and the corresponding concepts annotated to the neuron. We analyze the reason for the model’s predictions based on the explanation examples.

Analysis of Correct Prediction. Figure 3 (a) illustrates the concepts of the neurons that contribute most to correct predictions. In the left example of (a), although the current frame does not contain a port, both Causal ASFormer and TeCNO incorporate information from previous frames, allowing neurons associated with “Insert a port” and “Pushed into the port” to play a crucial role in accurately predicting the “(1) Preparation” phase. In the right example of (a), despite the darkness of the current frame, neurons associated with “hepatocystic

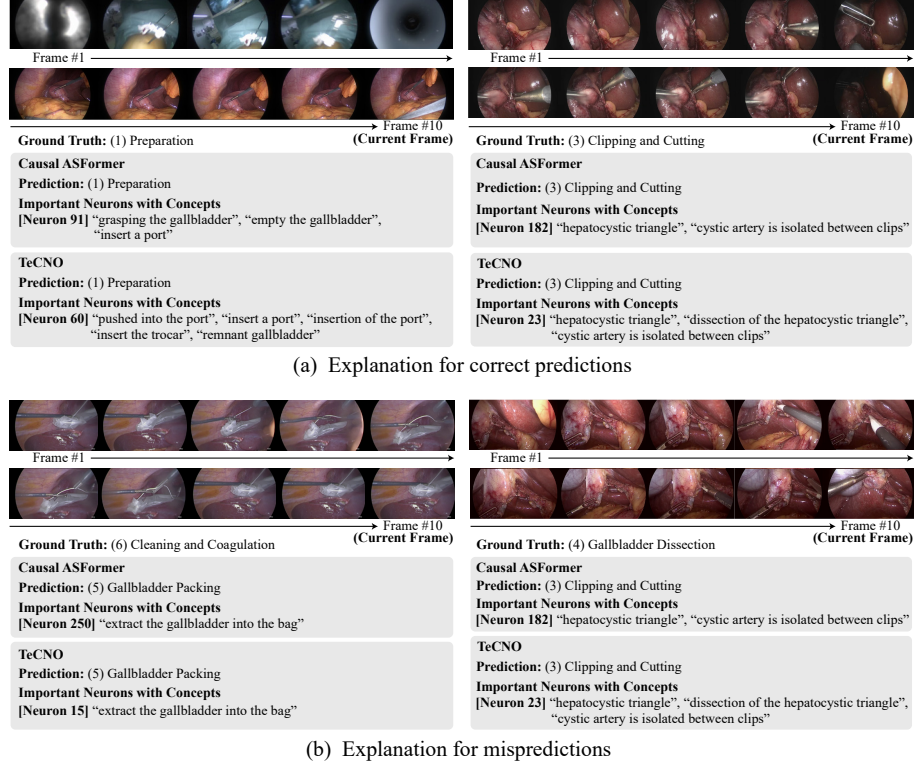


Fig. 3. Qualitative results of SurgX. Predictions of Causal ASFormer and TeCNO for the current frame, along with the concepts of the neurons that strongly contribute to those predictions. Only the best-contributing neuron is shown for each explanation.

triangle" and "cystic artery is isolated between clips" in both models contribute the most to accurately predicting the "(3) Clipping and Cutting". These examples demonstrate that Causal ASFormer and TeCNO do not rely solely on the current frame but instead leverage information from previous frames to learn and utilize relevant concepts for accurate predictions.

Analysis of Incorrect Prediction. Figure 3 (b) shows that SurgX also enables interpreting incorrect predictions. In the left example of (b), a neuron annotated with "extract the gallbladder into the bag" misleads the model to predict "(5) Gallbladder Packing", while the ground truth phase is "(6) Cleaning and Coagulation". In the right example of (b), the presence of visible clips leads the model to detect the concept "cystic artery is isolated between clips," resulting in the misprediction "(3) Clipping and Cutting". When the GT phase is "(4) Gallbladder Dissection," mispredictions as "(3) Clipping and Cutting" frequently occur in the models. In these cases, 88.22% involves neurons associated with "cystic artery is isolated between clips" In contrast, in correctly predicted cases, 92.88% do not involve these neurons.

4 Conclusion

In this study, we propose SurgX, a concept-based explanation framework for surgical phase recognition. SurgX includes concept construction, neuron representative sequence selection, and concept annotation methods designed specifically for surgical phase recognition models. Experiments on two surgical phase recognition models using the Cholec80 dataset demonstrate that SurgX provides meaningful concept-neuron associations, improving model explainability. By offering insights into model predictions with surgical terminology-based concepts, SurgX enhances the transparency and reliability of surgical AI models.

Acknowledgements. This work was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00334321), by the Institute of Information & Communications Technology Planning and Evaluation(IITP) grant funded by the Korea government (MSIT) (No. RS-2022-II220078, No. RS-2024-00509257, No. RS-2022-00155911) and ITRC(Information Technology Research Center) grant (IITP-2025-RS-2023-00258649).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahn, Y.H., Kim, H.B., Kim, S.T.: Wwv: A unified framework for explaining what where and why of neural networks by interpretation of neuron concepts. In: CVPR. pp. 10968–10977 (2024)
2. Ahn, Y.H., Park, G.M., Kim, S.T.: Line: Out-of-distribution detection by leveraging important neurons (2023), <https://arxiv.org/abs/2303.13995>
3. Bau, D., Zhou, B., Khosla, A., Oliva, A., Torralba, A.: Network dissection: Quantifying interpretability of deep visual representations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6541–6549 (2017)
4. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 343–352 (2020)
5. Czempiel, T., Paschali, M., Ostler, D., Kim, S.T., Busam, B., Navab, N.: Opera: Attention-regularized transformers for surgical phase recognition. In: International Conference on Medical Image Computing and Computer Assisted Intervention. pp. 604–614 (2021)
6. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: 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 (2021)
7. Hassler, K., Collins, J., Philip, K., Jones, M.: Laparoscopic Cholecystectomy. StatPearls Publishing (2025)

8. Jin, Y., Dou, Q., Chen, H., Yu, L., Qin, J., Fu, C.W., Heng, P.A.: Sv-renet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging* **37**(5), 1114–1126 (2017)
9. Kalibhat, N., Bhardwaj, S., Bruss, C.B., Firooz, H., Sanjabi, M., Feizi, S.: Identifying interpretable subspaces in image representations. In: *International Conference on Machine Learning*. pp. 15623–15638 (2023)
10. Khakzar, A., Baselizadeh, S., Khanduja, S., Rupprecht, C., Kim, S.T., Navab, N.: Neural response interpretation through the lens of critical pathways. In: *CVPR*. pp. 13528–13538 (2021)
11. Khakzar, A., Musatian, S., Buchberger, J., Valeriano Quiroz, I., Pinger, N., Baselizadeh, S., Kim, S.T., Navab, N.: Towards semantic interpretation of thoracic disease and covid-19 diagnosis models. In: *International Conference on Medical Image Computing and Computer Assisted Intervention*. pp. 499–508 (2021)
12. Kim, H.B., Ahn, Y.H., Kim, S.T.: Mask-free neuron concept annotation for interpreting neural networks in medical domain. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 524–533 (2024)
13. Kim, S., Donahue, T.: Laparoscopic cholecystectomy. *JAMA* **319**(17), 1834 (2018)
14. Kirtac, K., Aydin, N., Lavanchy, J.L., Beldi, G., Smit, M., Woods, M.S., Aspart, F.: Surgical phase recognition: From public datasets to real-world data. *Applied Sciences* **12**(17), 8746 (2022)
15. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: Lovit: Long video transformer for surgical phase recognition. *Medical Image Analysis* **99**, 103366 (2025)
16. Liu, Y., Huo, J., Peng, J., Sparks, R., Dasgupta, P., Granados, A., Ourselin, S.: Skit: a fast key information video transformer for online surgical phase recognition. *IEEE/CVF International Conference on Computer Vision* pp. 21074–21084 (2023)
17. Maier-Hein, L., Eisenmann, M., Feldmann, C., Feussner, H., Forestier, G., Gianarou, S., Gibaud, B., Hager, G.D., Hashizume, M., Katic, D., et al.: Surgical data science: A consensus perspective. *arXiv preprint arXiv:1806.03184* (2018)
18. Majumder, A., Altieri, M.S., Brunt, L.M.: How do i do it: laparoscopic cholecystectomy. *Annals of Laparoscopic and Endoscopic Surgery* **5** (2020)
19. Nwoye, C.I., Yu, T., Gonzalez, C., Seeliger, B., Mascagni, P., Mutter, D., Marescaux, J., Padoy, N.: Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos. *Medical Image Analysis* **78**, 102433 (2022)
20. Oikarinen, T., Weng, T.W.: Clip-dissect: Automatic description of neuron representations in deep vision networks. *ICLR* (2023)
21. Padoy, N., Blum, T., Ahmadi, S.A., Feussner, H., Berger, M.O., Navab, N.: Statistical modeling and recognition of surgical workflow. *Medical image analysis* **16**(3), 632–641 (2012)
22. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763 (2021)
23. Salahuddin, Z., Woodruff, H.C., Chatterjee, A., Lambin, P.: Transparency of deep neural networks for medical image analysis: A review of interpretability methods. *Computers in biology and medicine* **140**, 105111 (2022)
24. Strasberg, S.M.: A three-step conceptual roadmap for avoiding bile duct injury in laparoscopic cholecystectomy: an invited perspective review. *J Hepatobiliary Pancreat Sci* **26**(4), 123–127 (2019)

25. Temme, M.: Algorithms and transparency in view of the new general data protection regulation. *Eur. Data Prot. L. Rev.* **3**, 473 (2017)
26. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging* **36**(1), 86–97 (2016)
27. Yi, F., Wen, H., Jiang, T.: Asformer: Transformer for action segmentation. *British Machine Vision Conference (BMVC)* (2021)
28. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. *arXiv preprint arXiv:2307.15220* (2023)