



This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

PhaseMamba: Two-Stage Learning from Short Clips to Full Procedures for Online Surgical Phase Recognition

Shaheer Mohamed^{1*}, Harshala Gammulle¹, Tharindu Fernando¹, Sridha Sridharan¹, Simon Lucey², and Clinton Fookes¹

¹ Signal Processing, Artificial Intelligence and Vision Technologies Group (SAIVT),
Queensland University of Technology
mohamedshaheer@hdr.qut.edu.au

{pranali.gammulle, t.warnakulasuriya, s.sridharan, c.fookes}@qut.edu.au

² Australian Institute for Machine Learning (AIML), Adelaide University
simon.lucey@adelaide.edu.au

Abstract. Online surgical phase recognition aims to identify temporal segments in long surgical videos under real-time constraints. Unlike generic action recognition, it faces challenges such as redundancy, noisy frames, abrupt transitions, and low inter-class variability. We propose *PhaseMamba*, a two-stage framework that progressively learns from short clips to full procedures. Stage 1 pretrains a local phase learner to capture fine-grained temporal patterns and transitions, while Stage 2 integrates these representations to train an end-to-end model for efficient online inference. Both stages leverage Mamba-based architectures for scalable long-sequence modeling. A phase-order refinement strategy is introduced to further improve robustness under low-confidence predictions. Experiments on Cholec80 and AutoLaparo demonstrate state-of-the-art performance.

Keywords: Surgical Phase Recognition · Endoscopic Video

1 Introduction

Online surgical phase recognition is essential for real-time, context-aware surgical workflow analysis. It enables trainee feedback, improves operating room efficiency, and supports procedural review. Unlike offline recognition, it relies solely on past frames, making real-time prediction more challenging. Transformer-based models have recently gained popularity [10, 16, 4], but their quadratic complexity limits scalability to long surgical videos and real-time deployment. Mamba-based state-space models provide linear-time sequence modeling [6, 20]. Although Mamba has been explored for surgical phase recognition [1, 19], existing approaches remain limited in handling redundancy, variable phase lengths, and ambiguous transitions in online settings.

* Corresponding author.

To address these challenges, we propose a two-stage framework: (i) Local Phase Prediction and (ii) End-to-End Phase Recognition. Stage 1 pretrains a bidirectional Mamba-based local phase learner on short clips to capture fine-grained temporal patterns and transitions, producing transition-aware features. Stage 2 employs a forward-only Mamba model for end-to-end training and causal online inference, fusing frame-level and complementary transition aware local-phase features from stage 1 for robust long-range modeling.

Surgical procedures follow a logical phase order (e.g., suturing cannot precede tissue removal), yet online predictions can be unstable due to frame-level noise. To address this, we introduce a phase-order heuristic that applies transition priors derived from training data to penalize implausible low-confidence transitions. The main contributions of *PhaseMamba* are summarized as follows:

1. We propose *PhaseMamba*, a novel two-stage framework for Online Surgical Phase Recognition that progressively learns from short local windows to full procedures.
2. In Stage 1, a bidirectional Mamba-based local learner jointly predicts phases and transitions to extract transition-aware features.
3. In Stage 2, a forward Mamba-based online model fuses frame-level and transition-aware features via gated fusion to capture local and global temporal context efficiently.
4. We further introduce a phase-order refinement strategy to correct low-confidence mispredictions during real-time inference.

2 Related Works

Early surgical phase recognition methods such as SV-RCNet [7] and TMR-Net [8] relied on ResNet-based feature extraction with LSTM or TCN temporal modeling, but struggled to capture long-range dependencies. Transformer-based models [14] later became dominant due to their strong sequence modeling ability. Trans-SVNet [4] combines ResNet-based spatial features with TCN temporal modeling through a hybrid transformer used primarily for feature fusion. LoViT [9] introduces a pretrained video transformer with cascaded attention to model local and global temporal dependencies, employing sparse attention to reduce computation. However, its inference cost still grows with input length.

To improve efficiency, SKiT [10] proposes an input-length-independent architecture using a key pooling mechanism to incrementally record temporal information, while Surgformer [16] adopts divided spatial-temporal attention with hierarchical temporal modeling over sparsely sampled frames. Although computationally efficient, these approaches rely on sampling or heuristic mechanisms that may miss subtle transitions in visually ambiguous phases.

As discussed, the quadratic complexity of transformers limits scalability to long surgical videos, often requiring sampling or pooling strategies for efficiency. To address this, we explore Mamba [6], a state space model with linear-time

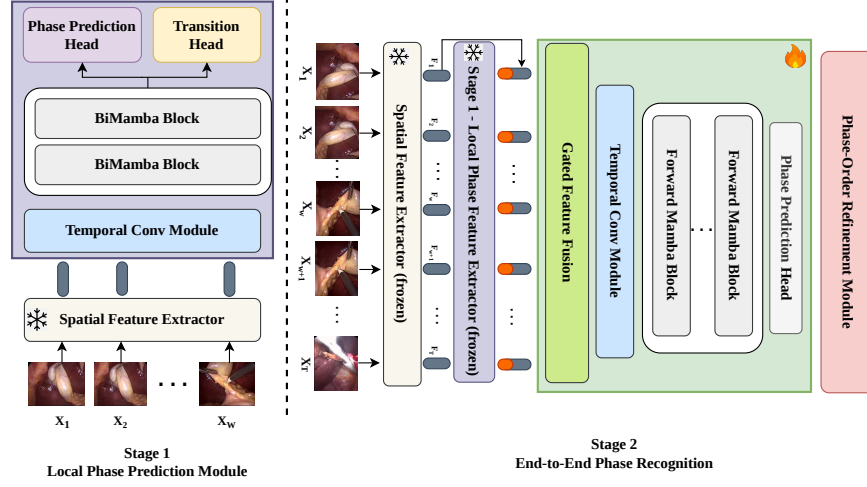


Fig. 1. Overview of the proposed *PhaseMamba* framework. **Stage 1:** A bidirectional Mamba-based local module learns short-term phase and transition patterns from video segments. **Stage 2:** A forward-only Mamba model performs online recognition by fusing frame-level and local features, followed by phase-order refinement for improved temporal coherence.

complexity that efficiently captures long-range dependencies. SR-Mamba [1] applies a bidirectional Mamba decoder for temporal modeling but operates in an offline setting. SPR-Mamba [19] combines Mamba with Transformers through a Scaled Residual TranMamba module and adopts hierarchical temporal sampling for online recognition. However, it still relies on sampling heuristics and does not fully exploit Mamba’s standalone modeling capability.

We propose a Mamba-based two-stage framework for efficient online surgical phase recognition. Stage 1 learns local temporal patterns with transition awareness, and Stage 2 performs end-to-end frame-wise recognition without input sampling. A phase-refinement strategy further improves robustness under noisy and ambiguous conditions.

3 Methodology

We propose a two-stage framework for surgical phase recognition. Stage 1 learns transition-aware local temporal representations, and Stage 2 performs end-to-end causal online recognition by integrating these features for global modeling. Both stages use Mamba-based architectures for efficient long-sequence processing, with an additional phase-order refinement module for robustness.

3.1 Stage 1: Local Phase Prediction Module

The first stage learns fine-grained temporal patterns and phase transitions from short video segments using a local phase prediction module. Operating on fixed-length windows (e.g., 1-minute clips), it predicts the dominant phase and detects transitions, suppressing noisy frames and improving boundary sensitivity via a dedicated transition head. We pretrain this module on diverse short clips sampled from long videos to balance underrepresented phases and emphasise transitions. The model takes frame-level features from a pretrained backbone (e.g., ViT [3], CLIP [11]) as input.

Given a temporal window $X \in \mathbb{R}^{T_w \times D_{in}}$, we first apply normalization, linear projection, and temporal smoothing. As shown in Eq. 1, the input is projected to hidden dimension D_h and processed by a Temporal Convolutional Module (TCM):

$$X' = \text{Linear}(\text{LN}(\text{TCM}(\text{Linear}(\text{LN}(X))))), \quad (1)$$

The sequence X' is then passed through bidirectional Mamba blocks to model temporal dependencies within the window. The forward and backward representations are computed as:

$$h_{\text{fwd}} = \text{Mamba}_{\rightarrow}(X'), \quad h_{\text{bwd}} = \text{flip}(\text{Mamba}_{\leftarrow}(\text{flip}(X'))), \quad (2)$$

Finally, the bidirectional features are fused via a gated residual interaction as:

$$X_{\text{out}} = X' + \text{GELU}((h_{\text{fwd}} + h_{\text{bwd}}) \odot X'), \quad (3)$$

The local module employs two output heads for phase classification and binary transition detection, enabling joint supervision within each temporal window. The transition head predicts whether a phase change occurs, improving boundary sensitivity. The training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{phase}} + \lambda \mathcal{L}_{\text{transition}}, \quad (4)$$

where $\mathcal{L}_{\text{phase}}$ is weighted cross-entropy loss to address class imbalance, $\mathcal{L}_{\text{transition}}$ is binary cross-entropy loss, and we empirically set $\lambda = 0.3$. This joint learning produces transition-aware features X_{out} , which are passed to Stage 2 and fused with frame-level features for online recognition.

3.2 Stage 2: End-to-End Phase Recognition

In Stage 2, we perform end-to-end online phase recognition. Input frames are processed by a spatial backbone to obtain features $X \in \mathbb{R}^{T \times D_{in}}$. The pretrained Stage 1 model (frozen) extracts transition-aware features $F_{\text{local}}^{(t)}$ using a sliding temporal window of size T_w , which are fused with frame-level features for online recognition. During inference, only past frames (from $\max(0, t - T_w + 1)$ to t) are used. Although Stage 1 is pretrained bidirectionally within each window, no future frames are accessed during Stage 2 inference.

To fuse frame-level features $X^{(t)} \in \mathbb{R}^{D_{\text{in}}}$ and local phase-aware features $F_{\text{local}}^{(t)} \in \mathbb{R}^{D_{\text{local}}}$, both are projected to a shared hidden space D_h , producing $f^{(t)}, p^{(t)} \in \mathbb{R}^{D_h}$. The soft gate is computed as in Eq. 5, and the fused representation is then obtained using the learned gate as in Eq. 6:

$$g^{(t)} = \sigma\left(\text{Linear}_{\text{gate}}([f^{(t)} \parallel p^{(t)}])\right), \quad (5)$$

$$F_{\text{fused}}^{(t)} = \text{ReLU}\left(g^{(t)} \odot f^{(t)} + (1 - g^{(t)}) \odot p^{(t)}\right), \quad (6)$$

The fused sequence is then temporally smoothed and processed by causal Mamba blocks for online modeling:

$$Z_{\text{out}} = \text{Mamba}_{\rightarrow}(\text{TCM}(F_{\text{fused}})) \in \mathbb{R}^{T \times D_{\text{global}}}, \quad (7)$$

3.3 Phase-Order Refinement Module

Although surgical workflows follow a logical phase order, online inference may produce implausible transitions due to noisy or ambiguous frames. To improve temporal consistency while preserving the online setting, we introduce a conditional phase-order refinement module based on transition priors. We estimate a phase transition matrix P_{trans} from the training data by counting consecutive phase pairs and applying Laplace smoothing. The row-normalised transition probabilities are computed as:

$$P_{\text{trans}}(i, j) = \frac{\text{count}(i \rightarrow j) + \lambda}{\sum_k \text{count}(i \rightarrow k) + \lambda N}, \quad (8)$$

where λ is the smoothing factor and N is the number of phases.

During inference at time t , the Stage 2 prediction head outputs the current phase probability distribution p_t . Let $\tilde{y}_t = \arg \max_j p_t(j)$ be the original current prediction, and let $\hat{y}_{t-1}$ be the previously emitted phase. The refinement module is activated only when the current prediction is low-confidence and suggests a phase change, i.e., when $\max_j p_t(j) < \delta$ and $\tilde{y}_t \neq \hat{y}_{t-1}$. When activated, each candidate phase j is scored using the current model confidence and the transition prior from the previously emitted phase:

$$s_t(j) = \log(p_t(j)) + \alpha \log(P_{\text{trans}}(\hat{y}_{t-1}, j)), \quad (9)$$

where α controls the influence of the transition prior. The current output is then refined as $\hat{y}_t = \arg \max_j s_t(j)$. Otherwise, the model directly emits the original prediction, i.e., $\hat{y}_t = \tilde{y}_t$. Importantly, this module only produces the current output $\hat{y}_t$. It does not access future frames, does not decode the full sequence, and does not revise previously emitted predictions $\hat{y}_{1:t-1}$. Therefore, the refinement remains causal and compatible with online surgical phase recognition.

Table 1. Comparison between PhaseMamba and state-of-the-art methods on the Cholec80 dataset using video-level and phase-level metrics. Where available, results are reported as mean $\pm$ standard deviation.

Method	Accuracy $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	Jaccard $\uparrow$
PhaseNet [12]	78.8 $\pm$ 4.7	71.3	76.6	–
SV-RCNet [7]	85.3 $\pm$ 7.3	80.7	83.5	–
OHFM [17]	87.3 $\pm$ 5.7	–	–	67.0
TeCNO [2]	88.6 $\pm$ 7.8	86.5	87.6	75.1
TMRNet [8]	90.1 $\pm$ 7.6	90.3	89.5	79.1
Trans-SVNet [4]	90.3 $\pm$ 7.1	90.7	88.8	79.3
Not End-to-End [18]	91.5 $\pm$ 7.1	–	88.6	77.2
LoViT [9]	92.4 $\pm$ 6.3	89.9	90.6	81.2
SKiT [10]	93.4 $\pm$ 5.2	90.9	91.8	82.6
SurgFormer [16]	93.4 $\pm$ 6.4	91.9	92.1	84.1
SR-Mamba [1]	92.6 $\pm$ 8.6	90.3	90.6	81.5
SPR-Mamba [19]	93.1 $\pm$ 4.6	89.3	90.1	81.4
PhaseMamba (Ours)	93.8 $\pm$ 6.7	93.0	91.2	85.2

4 Experiments

In this section, we evaluate *PhaseMamba* against state-of-the-art methods, describing the experimental setup, results and ablation studies.

4.1 Experimental Setup

We evaluate *PhaseMamba* on two public benchmarks, Cholec80 [13] and AutoLaparo [15]. Following prior work [10, 9, 2, 4], we use the standard 40/40 split for Cholec80 (downsampled to 1 fps) and the official 10/4/7 train/val/test split for AutoLaparo. Experiments are conducted on a single NVIDIA A6000 GPU using LoViT [9] frame-level features (768-dim) as in [10]. Stage 1 employs two Mamba blocks (hidden size 256, state size 16, window length 64) trained for 50 epochs with Adam (1×10^{-4}), while Stage 2 uses three forward Mamba blocks trained for 20 epochs at 1×10^{-5} . Following [10, 9], we report accuracy, precision, recall, and Jaccard score, and adopt the relaxed online evaluation protocol where predictions within a 10-second window around transitions are considered correct.

4.2 Comparison with State-of-the-art Methods

Quantitative comparisons between our *PhaseMamba* and recent state-of-the-art models, on Cholec80 and AutoLaparo datasets, are provided in Tables 1 and 2, respectively. On the Cholec80 dataset, *PhaseMamba* achieves the highest accuracy of 93.8%, precision of 93.0%, and Jaccard index of 85.2%, outperforming all previous methods. Compared to SR-Mamba, *PhaseMamba* achieves a notable improvement of +1.2% in accuracy and +3.7% in Jaccard, while outperforming

Table 2. Comparison between PhaseMamba and state-of-the-art methods on the AutoLaparo dataset using video-level and phase-level metrics. Where available, results are reported as mean $\pm$ standard deviation.

Method	Accuracy $\uparrow$	Precision $\uparrow$	Recall $\uparrow$	Jaccard $\uparrow$
SV-RCNet [7]	75.6	64.0	59.7	47.2
TMRNet [8]	78.2	66.0	61.5	49.6
TeCNO [2]	77.3	66.9	64.6	50.7
Trans-SVNet [4]	78.3	64.2	62.1	50.7
AVT [5]	77.8	68.0	62.2	50.7
LoViT [9]	81.4 $\pm$ 7.6	85.1	65.9	56.0
SKiT [10]	82.9 $\pm$ 6.8	81.8	70.1	59.9
PhaseMamba (Ours)	83.7 $\pm$ 6.6	88.2	75.0	67.4

SPR-Mamba by +0.7% in accuracy and +3.8% in Jaccard. These consistent gains across both metrics demonstrate the effectiveness of our enhancements over prior transformer and mamba-based architectures, particularly in achieving better temporal consistency and phase boundary precision. While SurgFormer achieves a slightly higher recall of 92.1%, PhaseMamba offers a better overall balance across all metrics, indicating superior generalisation and robustness under ambiguous scenarios. Further, beyond recognition performance, we conduct runtime analysis under the temporal-module setting reported in SKiT [10]. PhaseMamba requires 3.7 ms at 3000 frames, compared with 7.0 ms for SKiT and 54.5 ms for LoViT, demonstrating the efficiency of Mamba-based temporal modeling for online inference.

On the more challenging AutoLaparo dataset, *PhaseMamba* achieves state-of-the-art results with an accuracy of 83.7%, precision of 88.2%, and Jaccard of 67.4%, surpassing all prior methods. Compared to the previous best-performing method, SKiT (82.9% accuracy, 59.9% Jaccard), PhaseMamba delivers a +0.8% improvement in accuracy and a substantial +7.5% gain in Jaccard, indicating superior temporal consistency and segmentation quality. These results confirm the robustness and generalisability of our two-stage Mamba framework and phase-order refinement across diverse surgical procedures.

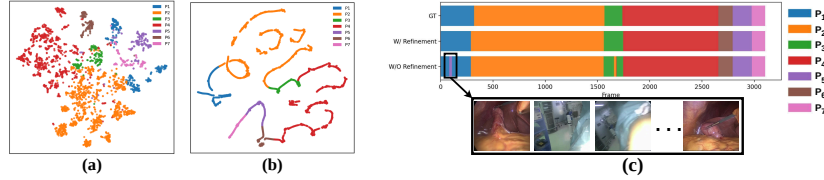
4.3 Ablation Study

We conduct ablation studies on the Cholec80 dataset to systematically evaluate the impact of our proposed components.

Effect of the Two-Stage Framework: To assess the impact of our two-stage framework and dual-head supervision, we perform a progressive ablation study. We first train the end-to-end model (Sec. 3.2) directly on frame-level features without Stage 1 pretraining. Next, we introduce a phase-only Stage 1 predictor and fuse its features with the end-to-end model using gated fusion. Finally, we incorporate transition prediction for joint supervision during Stage 1 pretraining to obtain richer transition-aware features. As shown in Table 3, each component

Table 3. Ablation study of the two-stage framework, demonstrating the impact of Stage 1 pretraining and dual-head supervision in the local phase prediction module.

Method	Accuracy	Precision	Recall	Jaccard
Frame-only baseline	89.5 ± 6.7	86.8	84.8	74.5
+ Phase-only Pretraining	93.0 ± 6.8	91.5	92.5	85.0
+ Phase + Transition (Ours)	93.8 ± 6.7	93.0	91.2	85.2

**Fig. 2.** Qualitative results. (a) t-SNE of raw LoViT features. (b) t-SNE after local feature fusion, showing improved phase separation. (c) Phase-order refinement correcting low-confidence mispredictions.

yields consistent performance improvements. t-SNE visualisation for Video 41 (Fig. 2 (a)–(b)) shows that fused features yield smoother trajectories and clearer phase separation than raw LoViT embeddings [9].

Effectiveness of Phase-Order Refinement Module: We evaluate the phase-order refinement module by comparing performance with and without its application during inference. As shown in Table 4, refinement consistently improves performance, confirming its effectiveness for robust online inference. Notably, even without refinement, PhaseMamba achieves a state-of-the-art Jaccard score of 84.2 compared with prior methods in Table 1, indicating strong temporal segmentation quality. The refinement further improves Jaccard to 85.2. A qualitative example is shown in Video 41 (Fig. 2 (c)), where the refinement corrects an early misprediction caused by brief camera noise, improving temporal coherence.

Feature Fusion Analysis: To evaluate the gated fusion mechanism, we compare it with two alternatives: (i) Concat + MLP fusion and (ii) attention-based fusion. The Concat + MLP variant concatenates frame-level and local features followed by a two-layer MLP, while the attention-based approach applies self-attention for feature integration. As shown in Table 5, our gated fusion consistently outperforms both baselines, demonstrating the benefit of adaptively weighting frame-level and local features for robust temporal modeling.

5 Conclusion

This paper presented *PhaseMamba*, a two-stage framework for online surgical phase recognition. Stage 1 learns transition-aware local temporal representations from short clips, while Stage 2 fuses these features with frame-level representations for efficient and robust online inference. A phase-order refinement module

Table 4. Effect of the phase-order refinement strategy on prediction consistency and overall performance.

Method	Accuracy	Precision	Recall	Jaccard
Without Refinement	92.2 ± 7.3	91.1	91.8	84.2
With Refinement (Ours)	93.8 ± 6.7	93.0	91.2	85.2

Table 5. Comparison of different fusion strategies for combining frame-level features with local phase and transition-aware representations.

Fusion Method	Accuracy	Precision	Recall	Jaccard
Concat + MLP Fusion	92.5 ± 7.4	91.5	92.1	84.4
Attention-Based Fusion	92.9 ± 7.0	91.9	92.2	85.0
Gated Fusion (Ours)	93.8 ± 6.7	93.0	91.2	85.2

further improves temporal coherence under ambiguous conditions. Experiments on Cholec80 and AutoLaparo demonstrate state-of-the-art performance. Future work will explore incorporating additional cues, such as tool-level annotations and multimodal extensions, to further enhance phase recognition.

Acknowledgments. This work was supported in part by the Australian Research Council (ARC) Discovery research grant DP250103634

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

References

1. Cao, R., Wang, J., Liu, Y.H.: Sr-mamba: Effective surgical phase recognition with state space model. arXiv preprint arXiv:2407.08333 (2024)
2. Czempel, 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 343–352. Springer (2020)
3. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
4. 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: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part IV 24. pp. 593–603. Springer (2021)
5. Girdhar, R., Grauman, K.: Anticipative video transformer. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13505–13515 (2021)
6. Gu, A., Dao, T.: Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 (2023)

7. 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)
8. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging* **40**(7), 1911–1923 (2021)
9. 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)
10. 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. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 21074–21084 (2023)
11. 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. PmLR (2021)
12. Twinanda, A.P., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Single- and multi-task architectures for surgical workflow challenge at m2cai 2016. *arXiv preprint arXiv:1610.08844* (2016)
13. 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)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
15. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 486–496. Springer (2022)
16. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 606–616. Springer (2024)
17. Yi, F., Jiang, T.: Hard frame detection and online mapping for surgical phase recognition. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. pp. 449–457. Springer (2019)
18. Yi, F., Yang, Y., Jiang, T.: Not end-to-end: Explore multi-stage architecture for online surgical phase recognition. In: *Proceedings of the Asian Conference on Computer Vision*. pp. 2613–2628 (2022)
19. Zhang, X., Zhang, Q., Chen, J., Zhou, C., Wang, Y., Zhang, Z., Li, X., Qian, D.: Sprmamba: Surgical phase recognition for endoscopic submucosal dissection with mamba. *arXiv preprint arXiv:2409.12108* (2024)
20. Zhu, L., Liao, B., Zhang, Q., Wang, X., Liu, W., Wang, X.: Vision mamba: Efficient visual representation learning with bidirectional state space model. In: *Forty-first International Conference on Machine Learning* (2024), <https://openreview.net/forum?id=YbHCqn4qF4>