BIEM: Bridging Implicit and Explicit Temporal Modeling for Surgical Phase Recognition

Yongkang Zhu[1], Xinzhe Du[1,3], Shuwei Shao[1], Shixing Ma[1], Shiyu Sha[2] Yi Liu[2], Qingfeng Yin[2], Wenguang Liu[2], and Zhe Min[1,3,4(✉)]

[1] School of Control Science and Engineering, Shandong University, Jinan, China
minzhe@sdu.edu.cn
[2] The Second Hospital, Cheeloo College of Medicine, Shandong University, Jinan, China
[3] Shenzhen Loop Area Institute (SLAI), Shenzhen, China
[4] UCL Hawkes Institute, Department of Medical Physics and Biomedical Engineering, University College London, London, U.K.

Abstract. Surgical Phase Recognition (SPR) automates surgical video analysis for intraoperative guidance and skill training. Most surgical procedures are performed following standardized workflows, which impose a well-defined structure on phase order and transition. However, existing SPR methods typically predict phases from visual features alone and rarely exploit this structured prior. Therefore, we propose Bridging Implicit and Explicit Modeling (BIEM), a hybrid framework coupling a Conditional Random Field for modeling phase transition probabilities with a Transformer-based implicit encoder. Our contributions are threefold: (i) proposing BIEM with a joint cross-entropy (CE) and negative log-likelihood (NLL) loss for stable end-to-end optimization; (ii) designing Sliding-Window Conditional Random Field (SWCRF) to incorporate surgical workflow priors; (iii) introducing HipScope39, the first dedicated hip arthroscopy dataset comprising 39 procedures. Experimental results demonstrate that BIEM achieves state-of-the-art accuracy of 89.3% on HipScope39 and competitive accuracy of 94.2% (relaxed) on Cholec80 and 86.0% on AutoLaparo, demonstrating robust cross-domain generalizability.

Keywords: Surgical phase recognition · Hip arthroscopy · Temporal modeling

1 Introduction

Surgical Phase Recognition (SPR) assigns a phase label to each frame of a surgical video, thereby partitioning the video into structured surgical phases. SPR has attracted considerable attention for its potential applications in surgical navigation [5, 6, 15], intelligent surgical robotic systems [14], and surgical quality assessment and surgical skills training [16, 17, 19].

Corresponding author: Zhe Min (minzhe@sdu.edu.cn).
Yongkang Zhu and Xinzhe Du contributed equally to this work.

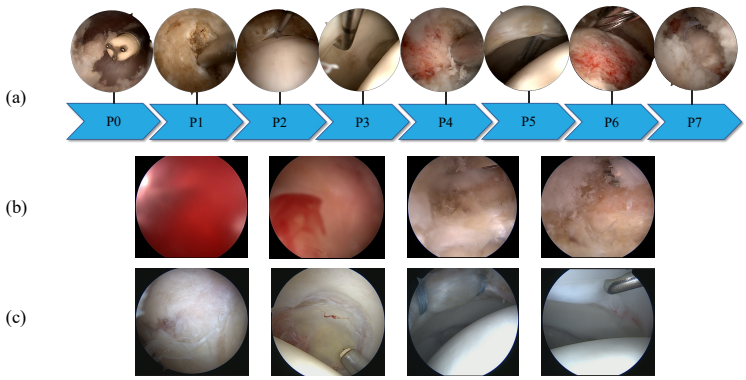

Fig. 1. (a) Standardized eight-phase surgical workflow providing strong priors. (b) Visual information loss caused by bleeding and debris. (c) Low frame contrast and high cross-phase visual similarity that challenge purely visual models.

SPR is a temporally dependent classification task that demands both accurate frame-level recognition and effective temporal context propagation. Existing approaches fall into two paradigms: implicit and explicit temporal models. Implicit methods dominate current research by learning temporal dynamics directly from feature representations. Early approaches, primarily RNNs and LSTMs [2], were constrained by their inherently sequential architecture, which prevents parallel computation and becomes infeasible for long surgical videos. Temporal Convolutional Networks (TCNs) [1,4] largely addressed this limitation by employing dilated convolutions to expand receptive fields efficiently. More recently, Vision Transformers (ViTs) [11, 12, 22] have achieved state-of-the-art performance through self-attention mechanisms that support parallelization and capture long-range dependencies. Despite their representational power, however, these implicit models often act as "black boxes" and lack explicit constraints on phase transitions.

In contrast, explicit methods, such as Hidden Markov Models (HMMs), learn transition relationships between surgical phases through a transition matrix, thereby capturing structured workflow priors and providing high interpretability. Early explicit approaches typically combined HMMs with CNNs or SVMs [10, 19], where the latter extract spatial features from individual frames. However, these methods have two key limitations: (1) Limited spatial representation: the reliance on CNNs or SVMs leads to insufficient spatial feature extraction, which constrains the performance of explicit temporal models; (2) Non-differentiable temporal components: unlike Conditional Random Fields (CRFs) [13,23], HMMs are inherently non-differentiable, making it difficult to jointly optimize spatial feature extraction and temporal modeling in an end-to-end manner. As a result, explicit methods often underperform compared to their implicit counterparts. Nonetheless, the underlying idea remains valuable: learning surgical workflow priors by modeling phase transitions.

While SPR research has been extensively studied on laparoscopic datasets (e.g., Cholec80 [19], AutoLaparo [20]), hip arthroscopy remains underexplored. Hip arthroscopy presents unique challenges: a narrow field of view, low contrast, and fluid environments that cause severe visual degradation from bleeding and debris (Fig. 1) [9]. Furthermore, high visual similarity across phases complicates recognition. However, unlike the variability observed in some laparoscopic procedures, hip arthroscopy follows a highly standardized workflow. This contrast offers a unique opportunity: leveraging strong explicit workflow priors to compensate for degraded visual information.

This raises a key question: can we integrate the powerful feature learning of implicit architectures with the interpretability and structural constraints of explicit models, within a real-time, end-to-end framework? Bridging this gap remains an open challenge.

To address this question, we propose Bridging Implicit and Explicit Modeling (BIEM), a hybrid framework coupling a Transformer backbone for implicit feature learning with a lightweight Sliding-Window Conditional Random Field (SWCRF) for explicit workflow modeling. Our contributions are threefold: (1) We introduce a joint Cross-Entropy (CE) and Negative Log-Likelihood (NLL) loss, which provides dense frame-level supervision to the implicit encoder while supervising the SWCRF's sequence-level predictions, enabling stable end-to-end optimization of BIEM. (2) We design the SWCRF to encode surgical phase transition priors into phase recognition, and its queue-based memory bank enables real-time sliding-window inference, preserving temporal coherence while retaining the interpretability of explicit models. (3) We release HipScope39, the first dataset dedicated to hip arthroscopy phase recognition, comprising 39 complete procedures.

2 Method

2.1 Overall Pipeline

As illustrated in Fig. 2, Bridging Implicit and Explicit Model- ing (BIEM) consists of two main components: a backbone network and the Sliding-Window Conditional Random Field (SWCRF) module. On one hand, we adopt Surgformer [22] as the backbone feature extractor, which encodes each input frame to capture both spatial and implicit temporal features. On the other hand, SWCRF extracts explicit temporal information and learns surgical workflow priors. It comprises three key modules: a Local Context Queue (LCQ) that stores historical frame features, a Window Memory Emission (WME) module that computes emission scores over a local temporal window, and a Surgical Phase Transition that models structured phase dependencies. Concretely, for a streaming input $\mathbf{X}_t \in \mathbb{R}^{H\times W\times C\times L}$, the backbone encodes $\mathbf{X}_t$ into an embedding $\mathbf{f}_t \in \mathbb{R}^D$ that carries both spatial and implicit temporal features, where $H, W, C, L, D \in \mathbb{Z}^+$ denote the height, width, number of channels, number of input frames per clip, and embedding dimension, respectively. This embedding is then fed into SWCRF,

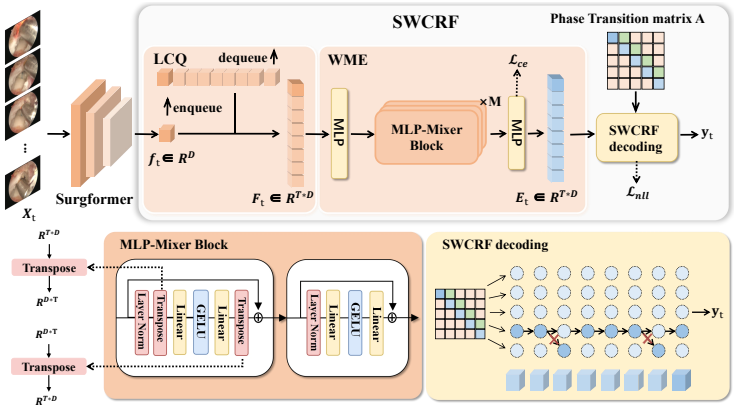

Fig. 2. Overview of BIEM framework. Given an input frame, the implicit temporal encoder (Surgformer) extracts spatiotemporal features. These are then processed by the SWCRF to capture explicit temporal information. The final prediction is obtained via SWCRF decoding. During decoding, invalid transitions are rejected, so that transient mispredictions are pulled back to the correct phase.

which leverages the learned surgical workflow to enrich $\mathbf{f}_t$ with explicit temporal context and decodes it into the predicted label sequence $\mathbf{y}_t$:

$$\mathbf{f}_t = \text{Surgformer}(\mathbf{X}_t), \quad \mathbf{y}_t = \text{SWCRF}(\mathbf{f}_{t-T+1}, \ldots, \mathbf{f}_t), \tag{1}$$

where $\mathbf{y}_t = [y_{t-T+1}, \ldots, y_t]^\top$ is the predicted phase sequence over the sliding window ending at time step t, with each $y_i \in \{0, \ldots, N-1\}$ being the phase label at the i-th time step, and $\mathbf{y}_t^{gt} = [y_{t-T+1}^{gt}, \ldots, y_t^{gt}]^\top$ the corresponding ground-truth sequence. Here T denotes the sliding window length, which we set to 32, and is independent of the clip length L.

2.2 Sliding-Window Conditional Random Field

Standard CRFs rely on global sequence information, which is unsuitable for real-time inference. SWCRF instead approximates global temporal constraints within a fixed-size local window, enabling constant-time inference per frame.

Local Context Queue. To maintain temporal context without storing the entire video, we employ a First-In-First-Out (FIFO) queue with capacity T, referred to as the LCQ. At time step t, the current feature $\mathbf{f}_t$ is enqueued, while the oldest feature is dequeued, yielding the local feature matrix $\mathbf{F}_t = [\mathbf{f}_{t-T+1}, \ldots, \mathbf{f}_t]^\top \in \mathbb{R}^{T \times D}$.

Window Memory Emission. To enrich contextual features within the local temporal window while maintaining a lightweight design, we adopt a 3-layer MLP-Mixer [18, 21] as the WME module. The MLP-Mixer first applies Token-mixing across frames within the window to aggregate temporal context,

followed by Channel-mixing to fuse information across feature channels. Given the window feature $\mathbf{F}_t$, WME computes an emission score matrix $\mathbf{E}_t \in \mathbb{R}^{T\times N}$:

$$\mathbf{E}_t = \mathrm{WME}(\mathbf{F}_t), \tag{2}$$

where N is the number of surgical phases. Each entry $\mathbf{E}_t[i,j]$ is the emission score representing the compatibility between the i-th time step and phase j before transition constraints are applied.

Surgical Phase Transition. We introduce a learnable embedding matrix $\mathbf{A} \in \mathbb{R}^{N\times N}$ as the phase transition matrix, where $\mathbf{A}[u,v]$ denotes the transition score from phase u to phase v. This matrix explicitly encodes the transition affinities between surgical phases and serves as a structured workflow prior, guiding the decoding process toward plausible phase transitions.

SWCRF Decoding. Following the chain-structured CRF formulation, we model the conditional probability of the ground-truth sequence $\mathbf{y}_t^{gt}$ given the window features $\mathbf{F}_t$ as:

$$P(\mathbf{y}_t^{gt} \mid \mathbf{F}_t) = \frac{1}{Z(\mathbf{F}_t)} \exp\left(\sum_{i=1}^{T} \mathbf{E}_t[i, y_i^{gt}] + \sum_{i=2}^{T} \mathbf{A}[y_{i-1}^{gt}, y_i^{gt}]\right), \tag{3}$$

where $Z(\mathbf{F}_t)$ is the partition function that normalizes over all N^T possible label sequences:

$$Z(\mathbf{F}_t) = \sum_{\mathbf{y}'} \exp\left(\sum_{i=1}^{T} \mathbf{E}_t[i, y_i'] + \sum_{i=2}^{T} \mathbf{A}[y_{i-1}', y_i']\right), \tag{4}$$

where $\mathbf{y}' = (y_1', \ldots, y_T')$ is a dummy sequence ranging over all labelings, with $y_i' \in \{0, \ldots, N-1\}$. Both $Z(\mathbf{F}_t)$ and the optimal sequence are computed in $O(T \cdot N^2)$ time via dynamic programming [23], using the forward algorithm for the partition function and the Viterbi algorithm for decoding, instead of enumerating all N^T sequences; the optimal sequence within the current window is obtained as:

$$\mathbf{y}_t = \arg\max_{\mathbf{y}'} \left(\sum_{i=1}^{T} \mathbf{E}_t[i, y_i'] + \sum_{i=2}^{T} \mathbf{A}[y_{i-1}', y_i']\right). \tag{5}$$

The predicted phase at the current time step is the last element y_t of $\mathbf{y}_t$. Since the decoding cost is independent of the video length, real-time inference is enabled.

2.3 Training loss Function

Standard NLL loss suffers from vanishing gradients, failing to effectively update the feature encoder. To resolve this, we propose a Joint Optimization Objective combining a frame-level CE loss with a sequence-level NLL loss $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda\mathcal{L}_{\text{NLL}}$, where λ is a balancing weight.

Table 1. Comparison with state-of-the-art methods on SPR datasets.

Method	Dataset	Evaluation	Acc. ↑	Prec. ↑	Rec. ↑	Jacc. ↑
SAHC(2022) [4]	HipScope39	Unrelaxed	84.8 ± 14.4	78.1	76.3	65.6
Surgformer(2024) [22]			86.1 ± 9.5	81.7	79.8	67.8
SPRMamba(2026) [24]			87.5 ± 9.1	82.8	79.5	70.3
Ours			**89.3 ± 9.3**	**83.9**	**80.9**	**73.1**
Trans-SVNet(2021) [8]	Cholec80	Unrelaxed	89.1 ± 7.0	84.7	83.6	72.5
SAHC(2022) [4]			91.8 ± 8.1	90.3	90.0	81.2
SKiT(2023) [12]			92.5 ± 5.1	84.6	88.5	76.7
Surgformer(2024) [22]			92.4 ± 6.4	87.9	89.3	79.9
LoViT(2025) [11]			91.5 ± 6.1	83.1	86.5	74.2
SPRMamba(2026) [24]			93.1 ± 4.6	89.3	90.1	81.4
Ours			**92.5 ± 5.0**	**86.5**	**87.0**	**77.7**
EndoNet(2016) [19]	Cholec80	Relaxed	81.7 ± 4.2	73.7	79.6	-
Trans-SVNet(2021) [8]			90.3 ± 7.1	90.3	89.5	79.3
SKiT(2023) [12]			93.4 ± 5.2	90.9	91.8	82.6
Surgformer(2024) [22]			93.4 ± 6.4	91.9	92.1	84.1
LoViT(2025) [11]			92.4 ± 6.3	89.9	90.6	81.2
SPRMamba(2026) [24]			93.6 ± 5.3	90.6	92.4	83.4
Ours			**94.2 ± 5.0**	**90.8**	**92.1**	**82.1**
TeCNO(2020) [3]	AutoLaparo	Unrelaxed	77.3	66.9	64.6	50.7
SKiT(2023) [12]			82.9 ± 6.8	81.8	70.1	59.9
LoViT(2025) [11]			81.4 ± 7.6	85.1	65.9	56.0
Surgformer(2024) [22]			85.7 ± 6.9	82.3	75.7	66.7
SPRMamba(2026) [24]			85.4 ± 9.8	84.7	71.9	63.7
Ours			**86.0 ± 6.1**	**81.1**	**74.8**	**65.4**

Sequence-level NLL Loss. $\mathcal{L}_{\text{NLL}}$ maximizes the probability of the ground-truth label sequence, enforcing temporal consistency:

$$\mathcal{L}_{\text{NLL}} = -\log P(\mathbf{y}_t^{gt} \mid \mathbf{F}_t) = \log Z(\mathbf{F}_t) - \left(\sum_{i=1}^{T} \mathbf{E}_t[i, y_i^{gt}] + \sum_{i=2}^{T} \mathbf{A}[y_{i-1}^{gt}, y_i^{gt}] \right). \quad (6)$$

Frame-level CE Loss. $\mathcal{L}_{\text{CE}}$ provides a direct gradient pathway for stable optimization of the backbone and WME parameters:

$$\mathcal{L}_{\text{CE}} = -\log \frac{\exp\left(\mathbf{E}_t[T, y_t^{\text{gt}}]\right)}{\sum_{j=1}^{N} \exp(\mathbf{E}_t[T, j])}. \quad (7)$$

CE is applied only to the last time step of each window ($i = T$). As the window slides with a stride of one, every frame eventually occupies the last position and receives direct supervision.

3 Experiments

Datasets. We evaluate our method on three surgical video datasets: Cholec80 [19] (used under CC BY-NC 4.0), AutoLaparo [20] (CC-BY-NC-SA 4.0), and a

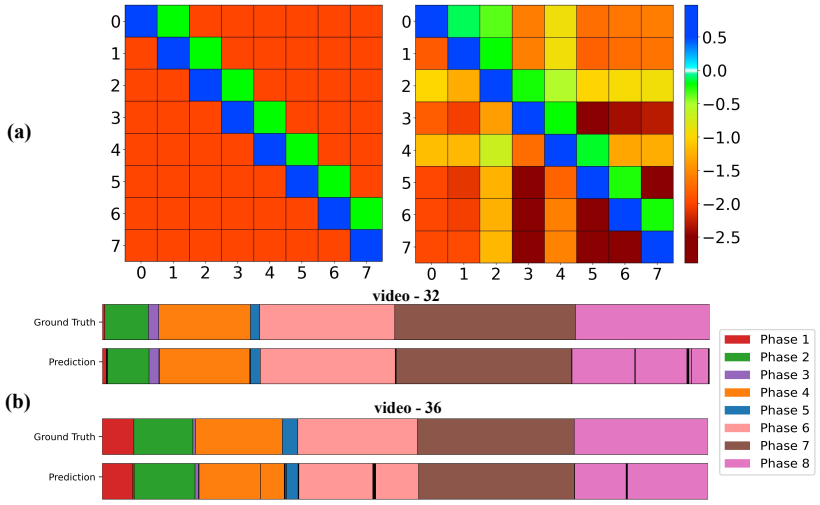

Fig. 3. Qualitative results on the HipScope39 dataset. (a) The predefined surgical workflow prior and the phase-transition matrix learned by SWCRF. (b) Ground-truth and predicted phase sequences for two representative videos.

hip arthroscopy dataset, HipScope39 (ethical approval number KYLL2024915). Cholec80 contains 80 laparoscopic cholecystectomy videos (40/40 train/validation split). AutoLaparo comprises 21 laparoscopic hysterectomy videos (10/4/7 train /validation/test split). HipScope39 consists of 39 full-length hip arthroscopy procedures (25 fps). Each video is annotated by senior orthopedic surgeons into eight clinically meaningful phases: *Preparation*, *Capsulotomy*, *Suction and Release*, *Acetabular Rim Resection*, *Anchor Insertion*, *Labral Repair*, *Femoral Osteoplasty*, and *Capsular Closure*, corresponding to P0–P7 in Figure 1. The phase definitions were determined through discussion among three senior orthopedic surgeons, and all 39 videos were strictly annotated following the same standardized protocol. The dataset is split into 20 training, 10 validation, and 9 test cases. Notably, all 39 procedures in HipScope39 strictly follow the standard eight-phase workflow, whereas only 81.3% (65/80) of Cholec80 and 33.3% (7/21) of AutoLaparo videos adhere to their respective standard workflows.

Evaluation metrics. We adopt the most commonly used metrics in the field of SPR: Accuracy, Precision, Recall, and Jaccard score. For Accuracy, we adopt two evaluation protocols: unrelaxed and relaxed [7]. Unrelaxed Accuracy counts a prediction as correct only if it exactly matches the ground-truth label. Relaxed Accuracy sets a tolerance window around phase boundaries.

Implementation details. The backbone follows the original Surgformer configuration with a learning rate of 1×10^{-4}. The sliding window length is set to $T = 32$. The weight λ for the NLL loss is set to 0.25.

Table 2. Ablation study of SWCRF components on HipScope39.

WME	Transition	CE Loss	NLL loss	Acc. ↑	Prec. ↑	Rec. ↑	Jacc. ↑
		✓		84.1 ± 8.6	78.9	78.3	64.8
✓		✓		84.6 ± 7.5	79.5	77.4	64.7
✓	✓		✓	81.2 ± 4.9	77.6	75.8	62.2
✓	✓	✓	✓	**89.3 ± 9.3**	**83.9**	**80.9**	**73.1**

3.1 Comparison with State-of-the-Art Methods

We compare our method against state-of-the-art approaches on three surgical phase recognition datasets. As shown in Table 1, our method achieves the best performance across all metrics on the challenging HipScope39 dataset, surpassing SPRMamba (89.3% vs. 87.5% accuracy). On Cholec80, our method achieves 92.5% under unrelaxed evaluation and 94.2% under relaxed evaluation, outperforming most competitors. On AutoLaparo, it attains 86.0% accuracy, surpassing Surgformer (85.7%) and SPRMamba (85.4%), with higher Recall and Jaccard scores than SPRMamba. Overall, our method achieves competitive performance across all three datasets, with the largest improvement on the highly standardized HipScope39 dataset (Table 1), demonstrating the benefit of integrating explicit workflow priors when the surgical workflow is highly standardized.

3.2 Ablation Studies

We conduct ablation studies on HipScope39 to evaluate the contribution of each SWCRF component. All variants are trained under the same streaming data-loading protocol. Our baseline (Surgformer + linear head) achieves 84.1% accuracy, which is slightly lower than the 86.1% reported in Table 1 due to the streaming protocol adopted for fair comparison. As shown in Table 2, adding WME marginally improves accuracy to 84.6%, while the NLL-only variant drops sharply to 81.2%, confirming that CE-based frame-level supervision is essential for learning discriminative emission scores. The full SWCRF with joint CE and NLL losses achieves the best result (89.3%). Furthermore, our model maintains a competitive inference speed of 31 FPS (vs. 33 FPS for the baseline), exceeding the 25 FPS real-time requirement.

3.3 Qualitative Analysis

Fig. 3(a) visualizes the learned transition matrix alongside the predefined surgical workflow prior. The color gradient from blue to red indicates the suppression degree for each transition from the row phase to the column phase. The learned matrix generally conforms to the prior, confirming that SWCRF effectively encodes structured phase dependencies. Fig. 3(b) presents prediction results on two video sequences from the HipScope39 dataset. Our model accurately captures the overall surgical workflow. However, local predictions occasionally exhibit phase flickering, visible as the black bars highlighted in the figure.

4 Conclusions

We propose BIEM, a hybrid framework that bridges implicit and explicit temporal modeling for surgical phase recognition. By integrating an implicit temporal backbone with the Sliding-Window Conditional Random Field (SWCRF), our method achieves strong performance on two public benchmarks and our HipScope39 dataset. As shown in Table 1, BIEM achieves its largest improvement on HipScope39, the dataset with the highest degree of surgical workflow standardization, and its smallest improvement on AutoLaparo, which exhibits the lowest standardization. Effectively learning structured workflow priors in surgeries with low standardization remains an open question for future work.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant Nos. 62303275 and U25A20145. This work was also supported by the National Natural Science Fund for Excellent Young Scientists Fund Program (Overseas) under Grant 22IAA01849, Jinan Science and Technology Bureau under Grant 202333011.

Disclosure of Interests

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

References

1. Bai, L., Ma, B., Wang, R., Wang, G., Cui, B., Jiang, Z., Islam, M., Min, Z., Lai, J., Navab, N., et al.: Multimodal graph representation learning for robust surgical workflow recognition with adversarial feature disentanglement. Information Fusion **123**, 103290 (2025)
2. Cheng, K., You, J., Wu, S., et al.: Artificial intelligence-based automated laparoscopic cholecystectomy surgical phase recognition and analysis. Surgical Endoscopy **36**(5), 3160–3168 (May 2022). https://doi.org/10.1007/s00464-021-08619-3
3. 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. Springer (2020). https://doi.org/10.1007/978-3-030-59716-0_33
4. Ding, X., Li, X.: Exploring segment-level semantics for online phase recognition from surgical videos. IEEE Transactions on Medical Imaging **41**(11), 3309–3319 (2022). https://doi.org/10.1109/TMI.2022.3182995
5. Du, X., Ma, S., Zhang, Z., Song, R., Li, Y., Meng, M.Q.H., Min, Z.: Registration after completion: Towards sparse and partial point set registration for computer-assisted orthopedic surgery. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7710–7717. IEEE (2025)

6. Du, X., Zhang, Z., Zhang, A., Song, R., Li, Y., Meng, M.Q.H., Min, Z.: Robust generalized partial-to-full point set registration with overlap-guided bidirectional hybrid mixture models for computer-assisted orthopedic surgery. IEEE Transactions on Automation Science and Engineering **23**, 7189–7208 (2026)
7. Funke, I., Rivoir, D., Speidel, S.: Metrics matter in surgical phase recognition. arXiv preprint arXiv:2305.13961 (2023)
8. 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. Springer (2021). https://doi.org/10.1007/978-3-030-87193-2_56
9. Hassan, M.M., Farooqi, A.S., Feroe, A.G., Lee, A., Cusano, A., Novais, E., Wuerz, T.H., Kim, Y.J., Parisien, R.L.: Open and arthroscopic management of femoroacetabular impingement: a review of current concepts. Journal of hip preservation surgery **9**(4), 265–275 (2022)
10. Lalys, F., Riffaud, L., Morandi, X., Jannin, P.: Surgical phases detection from microscope videos by combining svm and hmm. In: International MICCAI Workshop on Medical Computer Vision. pp. 54–62. Springer (2010). https://doi.org/10.1007/978-3-642-18421-5_6
11. 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)
12. 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). https://doi.org/10.1109/ICCV60755.2023.02065
13. Loebens, N., de Andrade Porto, J.V., de Andrade Porto, K.R., Souza, J.S., Pistori, H.: Types of markov random fields in deep learning: A review. Neurocomputing p. 132353 (2025)
14. Min, Z., Lai, J., Ren, H.: Innovating robot-assisted surgery through large vision models. Nature Reviews Electrical Engineering **2**(5), 350–363 (2025)
15. Min, Z., Zhang, A., Zhang, Z., Wang, J., Song, S., Ren, H., Meng, M.Q.H.: 3-d rigid point set registration for computer-assisted orthopedic surgery (caos): A review from the algorithmic perspective. IEEE Transactions on Medical Robotics and Bionics **5**(2), 156–169 (2023)
16. Padoy, N.: Machine and deep learning for workflow recognition during surgery. Minimally Invasive Therapy & Allied Technologies **28**(2), 82–90 (2019). https://doi.org/10.1080/13645706.2019.1584116
17. Shao, S., Zhu, K., Ma, S., Du, X., Zhang, B., Min, Z.: Depth any endoscopy: Towards self-supervised generalizable depth estimation in monocular endoscopy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 34126–34137 (2026)
18. Tolstikhin, I., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., et al.: Mlp-mixer: An all-mlp architecture for vision. arXiv preprint arXiv:2105.01601 **1**(2), 3 (2021)
19. 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 (2017). https://doi.org/10.1109/TMI.2016.2593957

20. 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). https://doi.org/10.1007/978-3-031-16449-1_46
21. Xia, J., Zhuge, M., Geng, T., Fan, S., Wei, Y., He, Z., Zheng, F.: Skating-mixer: Long-term sport audio-visual modeling with mlps. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 2901–2909 (2023). https://doi.org/10.1609/aaai.v37i3.25392
22. 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). https://doi.org/10.1007/978-3-031-72089-5_55
23. Yu, B., Fan, Z.: A comprehensive review of conditional random fields: variants, hybrids and applications. Artificial Intelligence Review **53**(8), 4289–4333 (2020). https://doi.org/10.1007/s10462-019-09793-6
24. 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. IEEE Sensors Journal (2026). https://doi.org/10.1109/JSEN.2025.3534567