Biomedical Data Science
2026, VOL. 6, NO. 1, 205-210
DOI: 10.47297/wspbdsWSP2752-630545.20260601

Coverage-Driven Soft Labeling for Surgical Video Temporal Segmentation: Application to Radical Prostatectomy

Yihan Lang

Hangzhou Dianzi University, Hangzhou, Zhejiang, 310018, China

ABSTRACT
Radical prostatectomy, the primary curative treatment for prostate cancer, presents significant challenges due to its complex work fl ow and technical demands. Recent advances in robotic-assisted surgery and deep learning have enabled intelligent surgical assistance systems. This paper proposes a novel temporal segmentation framework combining VideoMAE feature extraction with MS-TCN++ temporal modeling, enhanced by a coverage-driven soft-label strategy for improved frame-wise classi fication in prostatectomy videos. Our method effectively addresses key challenges including background dominance, boundary ambiguity, and temporal inconsistencies. Experimental validation demonstrates superior performance in both accuracy and robustness, offering promising potential for real-time surgical monitoring and safety assistance.

KEYWORDS
Radical prostatectomy; urgical phase recognition; Temporal action segmentation; Soft labeling; Video understanding; Computer-assisted surgery

1 Introduction

Prostate cancer is one of the most prevalent malignancies in the male genitourinary system [1]. Radical prostatectomy (RP) remains the cornerstone curative option for localized and locally advanced disease [2]. The procedure—precise removal of the prostate and seminal vesicles, often with pelvic lymph node dissection—is technically demanding, lengthy, and workflow-intensive, requiring strong spatial reasoning and meticulous skills. The adoption of robotic systems (e.g., the da Vinci platform) has improved precision and reduced invasiveness [3]. In parallel, progress in deep learning for computer vision and medical video analysis has enabled intelligent surgical assistance [4]. Integrating these methods with surgical workflows to enhance intraoperative decision- making and safety has become a key research focus, supported by large repositories of surgical videos routinely archived in hospitals [5]. Despite these advances, several RP steps—such as apical dissection near the external sphincter, control of the dorsal venous complex (DVC), and preservation of neurovascular bundles (NVB)—remain challenging and are associated with complications. Together with the shortage of expert surgeons and long training cycles [6], this motivates real-time phase recognition and next-step prediction systems to trigger instrument preparation, checklist verification, and risk alerts prior to high-risk steps [7].

1.1 Technological Gaps and Limitations of Current Methods

Most approaches treat each frame (or temporal window) as a one-hot hard label and train with cross-entropy. Representative models such as ASFormer [8] (frame-wise CE with probabilistic smoothing) and MS-TCN/++ [9,10] (multi-stage re fi nement but ultimately frame-wise CE) implicitly assume a fi xed, linear phase order. Clinical reality, however, often exhibits valid variations, branches, and brief regressions (e. g., whether DVC ligation precedes or follows bladder-neck dissection) [11]. Hard labels therefore penalize legitimate alternatives, encouraging over-con fi dence and weakening generalization [12]. Two data characteristics further hinder performance: (1) Prevalent, diverse non-relevant content. Long preparations, camera motion, instrument changes, and irrigation/hemostasis yield sparse signals; single-stage models can be overwhelmed by "equivalent backgrounds," causing over/under-segmentation and label drift [13]. (2) Inherently ambiguous boundaries. Phase transitions are gradual; inter-observer variability and temporal granularity introduce boundary noise and jitter (± several frames/seconds), de- grading the effectiveness of frame-level supervision [14].

1.2 Our Contributions

We tackle temporal segmentation of complete robot-assisted RP videos. Building on MS-TCN++, we move beyond hard labels with a coverage-driven soft-label scheme:

Two-stage modeling. A surgical/non- surgical gate followed by 7-class recognition. Intra-window surgical coverage serves as a supervisory weight, improving foreground/backgroun discrimination and robustness to long background segments.

Boundary-aware soft targets. Within a symmetric window, we allocate probability to phases proportionally to their temporal overlap and normalize within the surgical portion; supervision uses soft cross-entropy. This retains uncertainty near transitions, reduces hard cuts/over-segmentation, and complements MS-TCN++ temporal smoothing for more coherent segments and fewer flickers.

We do not use topological priors or restricted decoding [15]; our gains mainly appear under boundary ambiguity and small temporal misalignments. The better-calibrated probabilities also provide a solid basis for future constrained

decoding (e.g., Viterbi with transition graphs [16]) or next-step prediction.

Paper Organization Section II reviews related work on surgical phase recognition and temporal segmen- tation. Section III details the methodology, including VideoMAE feature extraction, MS-TCN++ archi- tecture, and the soft-labeling strategy. Section IV presents the experimental setup, dataset, and metrics. Section V reports and discusses results. Section VI concludes and outlines future directions.

2 Background and Literature Review

2.1 Overview of Surgical Phase Recognition

Surgical Phase Recognition (SPR) assigns prede fi ned phase labels to frames or segments of surgical videos and underpins workflow analysis, real-time assistance, and skill assessment [17]. The field has evolved from traditional methods to deep learning, yet systematic reviews note persistent issues—high annotation costs, non-standardized phase de fi nitions, and the difficulty of recognizing complex procedures such as radical prostatectomy [18]. This work targets SPR for robot-assisted radical prostatectomy (RARP) using deep temporal models and reviews traditional methods (Sec. 2.2), modern temporal architectures (Sec. 2.3), loss/labeling strategies (Sec. 2.4), and RARP-specific studies (Sec. 2.5).

2.2 Traditional and Early Deep Learning Methods

Two early paths dominated. The first combined hand-crafted features with classical temporal models: visual/non-visual cues (color, texture, optical fl ow, instrument usage) feed SVMs with HMM-based consistency [19]. Limited feature expressiveness andscene variability caused poor generalization and boundary jitter, as summarized in reviews. The second path used early deep learning (CNN + backend temporal models): frame features from ResNet/Inception paired with HMM/RNN smoothing or multi-task setups (e.g., instrument detection) [20]. While effective on datasets like Cholec80, frame-independent classification con- strained long-range context, leaving temporal consistency a key bottleneck.

2.3 Modern Temporal Modeling Architectures

To capture long-range dependencies, research moved to stronger temporal architectures. TCNs leverage dilated convolutions for stable training and parallel inference, improving long-sequence modeling [21]. MS-TCN introduced multi-stage refinement and a T-MSE loss to curb over-segmentation [10]; MS-TCN++ further enhanced multi-scale pattern capture via dual dilated convolutions, becoming a mainstream backbone for action/phase segmentation [9]. Transformer-based models wer also adapted; ASFormer achieved leading segmental metrics via an asymmetric encoder–decoder and local connectivity prior [8]. Other advances, such as an explicit boundary-regression branch (e. g., ASRF), aim to improve segment-level consistency and boundary localization.

2.4 Loss Functions and Labeling Strategies

Most SPR methods use frame-wise cross- entropy with temporal smoothing (e.g., T-MSE in MS-TCN) or boundary-aware terms to curb jitter and over-segmentation [9,10]. A key limitation is the dominance of one-hot hard labels, which treat fuzzy transitions as noise and bias decision boundaries. Systematic, transition-aware soft labeling is rare in surgery; our coverage-driven soft labels encode window-level overlaps to attenuate boundary noise and strengthen supervision for temporal back- bones.

2.5 Research Specific to Robot-assisted Radical Prostatectomy

RARP has more phases, variability, and inter-patient differences than fixed-workflow procedures (e.g., LC), yet remains a key benchmark. Foundational work (Zia et al.) studied 100 RARP videos [22]; EndoVis 2022 SAR-RARP50 released a multimodal public dataset [23]; large multi-center cohorts showed feasibility for skill assessment [24]. Most studies still rely on hard labels with frame-wise CE; we instead adopt soft labels to improve robust-ness under RARP ' s complex transitions.

3 Method

3.1 Task Definition

We perform surgical workfl ow analysis on complete RARP videos, assigning a label to each sampled time point on a single timeline. Let be the ground-truth at time t. We adopt a two-stage pipeline: (i) binary detection (surgical vs. non-surgical) over the full sequence for coarse localization; (ii) 7-class recognition only inside regions predicted as surgical. The class set is $C_{surg}=\{SVVD,LN,BN,NVB,DVC,PA,BNU\}$. Both stages share an MS-TCN++ backbone and the same optimization framework, but differ in supervision construction and sampling weights. This design reduces class imbalance and label-noise effects common in long, real-world videos with large non- surgical portions and ambiguous boundaries.

3.2 Coverage-driven Soft Targets

To handle weak annotations and boundary uncertainty, supervision is de fi ned on a symmetric window of length centered at $t, [t-\frac{w}{2}, t+\frac{w}{2}]$, rather than on a single frame. For annotated segments $\{[s_i,e_i],l_i\}$, the overlap inside the window is

$$\mathrm{ov}_i(t)=\max\left(0,\min\left(e_i,t+\frac{w}{2}\right)-\max\left(s_i,t-\frac{w}{2}\right)\right) \tag{1}$$

The surgical coverage ratio

$$g(t)=\frac{1}{w}\sum_{i:\mathrm{l}_i=surgical}\mathrm{ov}_i\ (t)\in[0,1] \tag{2}$$

acts as the soft target for the binary branch. For the 7-class branch, first compute the total surgical time in the window,

$$S(t)=\sum_{i:\mathrm{l}_i\in C_{surg}}\mathrm{ov}_i\ (t), \tag{3}$$

then normalize to a 7-D distribution:

$$p_k(t)=\begin{cases}\dfrac{\sum_{i:\mathrm{l}_i=k}\mathrm{ov}_i\ (t)}{S(t)}, & S(t)>0, \quad k\in C_{surg}\\ 0, & S(t)=0,\end{cases} \tag{4}$$

The scalar

$$\mathrm{cov7}(t)=\frac{S(t)}{w}\in[0,1] \tag{5}$$

is used as a sample weight. Training only uses fully annotated windows: the binary branch uses weight the 7-class branch uses This down-weights boundary-straddling or low-surgery windows, stabilizing optimization near transitions.

3.3 Model Backbone: ms-tcn++

Frame features are extracted by a pre-trained VideoMAE, transposed to $\mathrm{x}^{\top}\in\mathrm{R}^{B\times 768\times L}$, and projected via 1×1 Conv to

$$\mathrm{base}=\mathrm{Conv1d}(768\to C,\ k=1)(\mathrm{x}^{\top})\in\mathrm{R}^{B\times C\times L}. \tag{6}$$

MS-TCN++ performs multi-stage re fi nement. Stage 0 takes through dilated residual blocks (dilations $d=$ {1,2,4,8,16,32,64,128,128,128}, kernel $k=3$, non-causal padding), then a head yields For $s\geq 1$, we concatenate the previous probabilities with and project back:

$$\mathrm{X}^{(s)}=\mathrm{Conv1d}(C+K\to C,\ k=1)([\mathrm{base};\ \mathrm{P}^{(s-1)}]. \tag{7}$$

This "base + probability prior" fusion injects decision con fi dence and improves temporal consistency. The binary branch uses stages (10 blocks each); the 7-class branch uses (9 blocks each); both use width $C=64$,$k=3$, max dilation 2^7.

3.4 Loss Function and Regularization

Both branches share a unified objective: normalized classification loss with sample masking/weighting, plus second-order smoothing in the probability domain. For the binary branch, with logits $z_t^{(s)}$, probabilities $p_t^{(s)}=\sigma(z_t^{(s)})$, target g_t, and weight ,

$$\mathrm{L}_{\mathrm{bin}}^{(s)}=\frac{\sum_t w_t\ \mathrm{BCE_with_logits}\left(z_t^{(s)},g_t;pos_weight=\alpha\right)}{\sum_t w_t+\varepsilon}+\lambda_{\mathrm{tv}}\cdot\frac{1}{T-1}\sum_{t=1}^{T-1}\left(p_{t+1}^{(s)}-p_t^{(s)}\right)^2 \tag{8}$$

Total loss: . For 7-class logits $\mathrm{z}_t^{(s)}$, probabilities $\mathrm{p}_t^{(s)}$, and soft target with weights w_t,

$$\mathrm{L}_{7c}^{(s)}=\frac{\sum_t w_t\left[-\mathrm{y}_t^{\top}\log\ \mathrm{softmax}\left(\mathrm{z}_t^{(s)}\right)\right]}{\sum_t w_t+\varepsilon}+\lambda_{\mathrm{tv}}\cdot\frac{1}{T-1}\sum_{t=1}^{T-1}\left\|\mathrm{p}_{t+1}^{(s)}-\mathrm{p}_t^{(s)}\right\|_2^2 \tag{9}$$

Total: $\mathrm{L}_{7c}=\sum_s\mathrm{L}_{7c}^{(s)}$. Unlike the original MS-TCN++ (truncated MSE on log-probabilities), we smooth directly in probability space-simple, stable, and aligned with thresholded inference.

3.5 Inference and Post-processing

We use the final-stage outputs. For the binary branch, is the surgical probability; a validation-tuned threshold is fixed for testing (an optional moving average of 5 is used only for visualization). For the 7-class branch, labels are *arg* max softmax($\mathrm{z}_t^{(S-1)}$). Evaluation follows the training masks: binary only at annotated positions; 7-class only where cov7(t)≥ 0.5.

3.6 Two-stage scheme and baselines

We compare to: (i) MS-TCN++(HARD)—single-stage, 8 classes (merging etc

with seven surgical steps), trained with one-hot CE plus the same probability-domain smoothing;

Unannotated positions are masked. (ii) Statistical baselines: Majority and Markov-1. The latter estimates transition matrix and prior and applies Viterbi to maximize

$$\hat{y}_{1:T}=arg\ \max_{y_{1:T}}\left(\log\hat{\pi}_{y_1}+\sum_{t=2}^{T}\log\hat{P}_{y_{t-1},y_t}\right) \tag{10}$$

All baselines strictly follow our evaluation masks (7- class only at annotated positions with sufficient cov- erage; binary only at annotated positions) to ensure comparability.

4 Experiments

4.1 Data and Task Setup

4.1.1 Data Source and Scale

We use a public prostatectomy video dataset with temporal segment annotations provided via three files: segments. csv, segment_metadata. csv, and video_metadata. csv. We select 203 high- quality videos for experiments. Detailed

statistics appear in Section 4.3 Table 2 and Table 3.

4.1.2 Label Taxonomy

We adopt an 8-class scheme: one non-surgical class {etc} plus seven surgical steps. Each step is de fi ned by an (Anatomy, Action) pair; {etc} aggre- gates non-operative content (e.g., setup, camera motion, irrigation) and is only used as the negative class in Stage 1.

Table 1 Surgical label set used in this study (code, anatomy, and action).

Code	Anatomy	Action
SVVD	Seminal vesicle and vas deferens	Dissect
LN	Lymph node	Dissect
BN	Bladder neck	Dissect
NVB	Neurovascular bundle	Dissect
DVC	Dorsal venous complex	Dissect
PA	Prostate apex	Dissect
BNU	Bladder neck and urethra	Suture
etc	None	None

In Stage 2, the targets are {SVVD, LN, BN, NVB, DVC, PA, BNU}; etc is handled only by the Stage 1 binary classifier.

4.1.3 Temporal Sampling and Windowing

We use a fixed-stride sliding window: sampling at 1.6 fps, window length 16 frames (≈ 10*s*), stride 8 frames (≈ 5*s*). For each center time t, we compute overlaps with annotations to obtain:

Annotation coverage covannot $\mathrm{cov}_{\mathrm{annot}}(t) \in [0,1]$;

Surgical coverage $\mathrm{cov}_{\mathrm{surg}}(t) \in [0,1]$;

7-D soft label (normalized within surgical time; zero if no surgery).

4.1.4 Task Definition

We treat SPR as frame-wise classification for long videos using a two-stage pipeline:

(1) Stage 1: binary (surgical vs. etc).

(2) Stage 2: 7-class recognition within Stage 1 surgical regions.

4.2 Data Split and Evaluation Protocol

4.2.1 Data Partitioning

We use a strict patient-level split to prevent leak- age:

Train: 138 videos; 111,116 windows; 308.7 h;

Validation: 34 videos; 23,748 windows; 66.0 h;

Test: 30 videos; 24,896 windows; 69.2 h Appendix lists video IDs for reproducibility.

4.2.2 Training and Evaluation Encoding

We define two masks:

$$\text{known_full}(t) = I[\mathrm{cov}_{\mathrm{annot}}(t) \geq 1 - 10^{-3}] \quad (11)$$

$$\text{known_any}(t) = I[\mathrm{cov}_{\mathrm{annot}}(t) > 0] \quad (12)$$

Training: only clean windows with known_full$(t) = 1$ (99.99% in both train/val).

Evaluation: positions with and $\mathrm{cov}_{\mathrm{surg}}(t) \geq 0.50$ (test effective ratio 99.6%).

4.2.3 Evaluation Metrics

Binary: threshold via validation grid search (F1-oriented); report Accuracy, Precision, Re- call, F1.

7-class: report frame-wise MoF; plus segment- level F1@{10,25,50}, Edit, and Boundary-F1 (±0.5s).

4.3 Dataset Statistics

The train/val sets show near-perfect valid/clean ratios (99.99%). The test set maintains a 99.6% ef- fective evaluation ratio. Total durations are 308.7 h (train), 66.0 h (val), and 69.2 h (test). The class distribution is imbalanced (e.g., BNU 31.5%, DVC 3.2%), reflecting real surgical timing.

Table 2 Detailed statistics of dataset splits

Split	Videos	Total Windows	Valid Windows	Valid Ratio (%)	Clean Windows	Clean Ratio (%)
Train	138	111,116	111,116	100.0	111,100	99.99
Validation	34	23,748	23,748	100.0	23,746	99.99
Test	30	24,896	24,789	99.6	–	–

Table 3 Surgical step duration distribution in the training set

Step	SVVD	LN	BN	NVB	DVC	PA	BNU
Ratio (%)	14.6	9.0	13.6	22.1	3.2	6.0	31.5

4.4 Ethical Statement

The dataset is public; original videos come from open educational sources under their terms. We only use timestamps and label metadata for research, with no re-identification.

4.5 Experimental Details

To enable reproducibility, we summarize key settings.

4.5.1 Feature Preprocessing and Caching

We extract VideoMAE features offline at fps, sliding windows of frames (10s) with stride 8 (5s). We save 768-D features with timestamps, soft labels, and masks in .npz files; training uses in-memory loading with padding for variable-length sequences.

4.5.2 Training Setup

We use AdamW (lr 5×10^{-4}, weight decay 1×10^{-4}), 3-epoch linear warm-up, cosine annealing, mixed precision, cuDNN benchmark, and grad clip (max-norm 5.0). Binary branch: 50 epochs; 7-class: 20 epochs. Batch size 2 (GPU) / 1 (CPU).

4.5.3 Masks and Sampling Weights

Training uses clean windows with known_full(t) = 1, and coverage-based weights. Validation/testing restrict to and $\text{cov}_{\text{surg}}(t) \geq 0.5$.

4.5.4 Evaluation Protocol

Binary: pick on validation via grid search; report Accuracy/Precision/Recall/F1. 7-class: apply temporal smoothing, then *arg max*; report MoF, F1@{10,25,50}, Edit, Boundary-F1.

4.5.5 Runtime Environment

Experiments run on Ubuntu 22.04, PyTorch 2.3, CUDA 12.1, Python 3.11, single 24GB GPU. Com- mon deps: numpy, pandas. Code will be released upon publication.

5 Results

5.1 Baseline Comparison

We compare our two-stage soft-label approach against Majority, Markov-1, and MS-TCN++ (hard). For binary detection, τ^* is tuned on validation and fixed on test. For 7-class, we score only annotated frames with using MoF, F1@{10,25,50}, Edit, and Boundary-F1.

Binary detection. Our model attains the best Re- call and F1 while keeping competitive Precision (Table 4), indicating that coverage-weighted training and the two-stage design effectively recover surgical in- tervals and reduce misses.

Seven-class recognition. Our method improves MoF and segmental F1 over MS-TCN++ (hard) and yields markedly higher Boundary-F1 (Table 5), reflecting better temporal consistency and boundary lo- calization.

Table 4 Binary Classification (Surgical vs. Non-surgical) – Test Set Performance

Method	Accuracy	Precision	Recall	F1
Majority baseline	0.634	0.634	1.000	0.776
MS-TCN++ (hard, binary)	0.749	0.751	0.905	0.821
Ours (two-stage, soft labels)	0.753	0.737	0.949	0.830

Table 5 Seven-Class Overall Performance (Evaluation Mask: Annotated & Coverage cov ⩾ 0.5) – Test Set Results

Method	MoF	F1@10	F1@25	F1@50	Edit	Boundary-F1
Majority baseline	0.3425	—	—	—	0.153	0.806
Markov-1 baseline	0.3425	0.157	0.157	0.157	0.153	0.806
MS-TCN++ (hard, 7-class)	0.717	0.569	0.532	0.451	0.595	0.025
Ours (two-stage, soft labels)	0.778	0.592	0.558	0.471	0.598	0.540

Table 6 Seven-Class Per-Class Recall (Test Set; Class Order: SVVD, LN, BN, NVB, DVC, PA, BNU)

Method \ Class	SVVD	LN	BN	NVB	DVC	PA	BNU
Majority baseline	0.000	0.000	0.000	0.000	0.000	0.000	1.000
Markov-1 baseline	0.000	0.000	0.000	0.000	0.000	0.000	1.000
MS-TCN++ (hard, 7-class)	0.448	0.515	0.718	0.730	0.208	0.567	0.994
Ours (two-stage, soft labels)	0.588	0.751	0.618	0.777	0.335	0.545	0.985

Table 7 Ablation: Binary Classification (Surgical vs. Non-surgical, Test Set)

Setting	Acc	Prec	Rec	F1
MS-TCN++ (Hard labels)	0.749	0.751	0.905	0.821
Soft-Unnormalized (ablation)	0.731	0.737	0.895	0.809
Soft-Normalized + Cov-Weighted (EXP1, Cov−)	0.731	0.741	0.885	0.807
Ours (main, Cov+)	0.753	0.737	0.949	0.830

Per-class analysis. Gains concentrate on SVVD, LN, NVB, DVC—phases with ambiguous bound- aries—while overall MoF and segmental metrics also increase (Table 6).

5.2 Ablation Studies

We isolate key factors under identical back- bones/features/masks: (i) hard vs. soft supervision; (ii) within-surgery normalization; (iii) coverage weighting.

Binary detection ablations. Soft labels + coverage weighting deliver the best Recall and F1 (Table 7). Hard labels reduce

Recall (0.949→ 0.905); removing coverage weighting or normalization further degrades stability.

Seven-class ablations. Each component con- tributes (Table 8). Hard labels reduce MoF (0.778→ 0.717); removing within-surgery normalization weakens class separation; disabling coverage weighting lowers MoF and segmental F1, showing the benefit of down-weighting boundary/low-surgery windows.

Table 8 Ablation: Seven-Class Phase Recognition (Test Set; Evaluation Mask: Annotated & Coverage cov ⩾ 0.5)

Setting	MoF	F1@10	F1@25	F1@50	Edit	Boundary-F1
MS-TCN++ (Hard labels)	0.717	0.569	0.532	0.451	0.595	0.025
Soft-Unnormalized (ablation)	0.715	0.577	0.538	0.464	0.597	0.618
Soft-Normalized + Cov-Weighted (EXP1, Cov-)	0.767	0.568	0.515	0.448	0.604	0.552
Ours (main, Cov+)	0.778	0.592	0.558	0.471	0.598	0.540

6 Conclusion

Coverage-driven soft labels calibrate transitions; within-surgery normalization sharpens class separation; coverage weighting suppresses background dominance. Together, they consistently improve Recall, MoF, segmental F1, and boundary quality over strong baselines.

References

[1] FBRAY F, LAVERSANNE M, SUNG H, et al. Global cancer statistics 2022:Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries[J].CA: A Cancer Journal for Clinicians,2024,74(3):229-263.

[2] CORNFORD P, VAN DEN BERGH R C N, BRIERS E, et al. EAU-EANM-ESTRO-ESUR-ISUP-SIOG guidelines on prostate cancer—2024 update. part i: Screening, diagnosis, and local treatment with curative intent[J].European Urology,2024,86(2):148-163.

[3] BINDER J, KRAMER W. Robotically-assisted laparoscopic radical prostatectomy[J].BJU International,2001,87(4):408-410.

[4] MAIER-HEIN L, EISENMANN M, SARIKAYA D, et al. Surgical data science – from concepts toward clinical translation[J].Medical Image Analysis, 2022,76:102306.

[5] YOUSSEF C S, HARAM K, NOËL J, et al. Evolution of the digital operating room: the place of video technology in surgery[J]. Langenbeck's Archives of Surgery,2023,408(1):95.

[6] MEARA J G, LEATHER A J M, HAGANDER L, et al. Global surgery 2030: evidence and solutions for achieving health, welfare, and economic development[J].The Lancet,2015,386(9993):569-624.

[7] GARROW C R, KOWALEWSKI K F, LI L, et al. Machine learning for surgical phase recognition: A systematic review[J].Annals of Surgery,2021,273 (4):684-693.

[8] YI F C, WEN Y Z, LIU X, et al. Asformer: Transformer for action segmentation[EB/OL].(2021) [2026-01-24]. https://arxiv.org/abs/2110.08568.

[9] LI S J, ABU FARHA Y, LIU Y, et al. MS-TCN++: Multi-stage temporal convolutional network for action segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2020.

[10] ABU FARHA Y, GALL J.MS-TCN: Multi-stage temporal convolutional network for action segmentation[EB/OL].(2019) https://openaccess.thecvf.com/content_CVPR_2019/html/Abu_Farha_MS-TCN_Multi-Stage_Temporal_Convolutional_Network_for_Action_Segmentation_CVPR_2019_paper.html.

[11] LALYS F, JANNIN P. Surgical process modelling: a review[J]. International Journal of Computer Assisted Radiology and Surgery, 2014, 9(3): 495-511.

[12] Sunil Thulasidasan, Gopinath Chennupati, Jeff Bilmes, et al. On mixup training: Improved calibration and predictive uncertainty for deep neural networks.
In Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[13] Yazan Abu Farha and Juergen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3575–3584, June 2019.

[14] WARD T M, FER D M, BAN Y Q, et al. Challenges in surgical video annotation[J].Computer Assisted Surgery,2021,26(1):58-68

[15] PADOY N, BLUM T, AHMADI S A, et al. Statistical modeling and recognition of surgical workflow[J].Medical Image Analysis,2012,16(3):632-641.

[16] FORNEY G D. The viterbi algorithm[J].Proceedings of the IEEE,1973,61(3):268-278.

[17] MAIER-HEIN L, EISENMANN M, SARIKAYA D, et al. Surgical data science – from concepts toward clinical translation[J].Medical Image Analysis, 2022,76:102306.

[18] GARROW C R, KOWALEWSKI K F, LI L, et al. Machine learning for surgical phase recognition: A systematic review[J].Annals of Surgery,2021,273 (4):684-693.

[19] PADOY N, BLUM T, AHMADI S A, et al. Statistical modeling and recognition of surgical workflow[J].Medical Image Analysis,2012,16(3):632-641.

[20] TWINANDA A P, SHEHATA S, MUTTER D, et al. EndoNet: A deep architecture for recognition tasks on laparoscopic videos[J].IEEE Transactions on Medical Imaging,2017,36(1):86-97.

[21] Colin Lea, Michael D. Flynn, René Vidal, et al. Temporal convolutional networks for action segmentation and detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pages 156– 165, 2017.

[22] ZIA A, HUNG A J, YUH K, et al. Automated surgical activity recognition in robot-assisted radical prostatectomy using deep learning[J]. IEEE Transactions on Biomedical Engineering,2018,65(11):2381-2392.

[23] PSYCHOGYIOS D, COLLEONI E, VAN AMSTERDAM B, et al.SAR-RARP50: Segmentation of surgical instrumentation and action recognition on robot-assisted radical prostatectomy challenge[EB/OL].(2024) [2026-01-24]. https://arxiv.org/abs/2401.00496..

[24] ZHAO X, TAKENAKA S, IUCHI S, et al. Automatic recognition of surgical phase of robot-assisted radical prostatectomy based on artificial intelligence deep-learning model and its application in surgical skill evaluation: a joint study of 18 medical education centers[J]. Surgical Endoscopy,2025,39(9):5623-5635.