Phase Recognition in Surgical Videos using Multi-Video Finetuning

Shobhit Agarwal,[1] Vedant Srinivas[1]

[1]Stanford University

Stanford Computer Science

Abstract

Surgical phase recognition is a critical task in computer-assisted surgery, requiring models that can capture both fine-grained spatial features and long-range temporal dependencies in video. In this work, we present a lightweight, two-stage architecture that combines a frozen ResNet-18 visual backbone with a causal Transformer encoder for real-time phase prediction in laparoscopic surgeries. Using the Cholec80 dataset, we explore multi-video finetuning as a strategy to improve temporal generalization and robustness, sampling frames across multiple surgeries to model higher-level procedural structure. Our ResNet+Transformer approach achieves 63.8% frame-wise accuracy with only 23.8M parameters, outperforming RNN-based baselines and approaching the performance of state-of-the-art methods like EndoNet and MSN at a fraction of the computational cost. We show that multi-video finetuning improves model robustness to visual drift and phase ambiguity, and provide qualitative analyses to support the effectiveness of our temporal reasoning framework.

Stanford University

Introduction

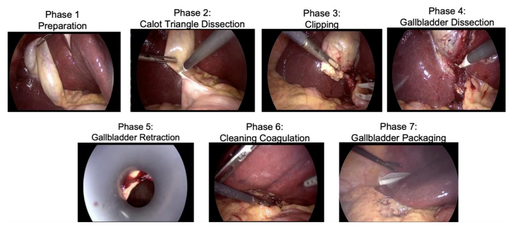
- **Problem:** Surgical phase recognition is a long-standing challenging in computer vision, that can enable **real-time decision support**, **skill assessment**, and **workflow automation** in operating rooms

Credit: Da Vinci surgical system/Getty Images

- **Limitations of Existing Methods:** Existing deep learning frameworks (e.g. LoViT, EndoNet) are **unable to capture** long-term dependencies, rely significantly on visual similarity between adjacent phases, or lack temporal generalization.

Dataset

- **Cholec80:** 80 laparoscopic surgery videos annotated with 7 surgical phases. Real-world, but limited in size.

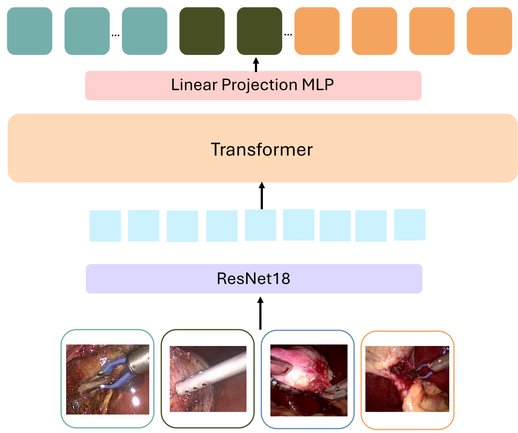


Goals

1. Lightweight architecture and parameter count
2. Architecturally-agnostic mechanism to boost performance
3. Achieve comparable accuracy to existing state-of-the-art models/baselines

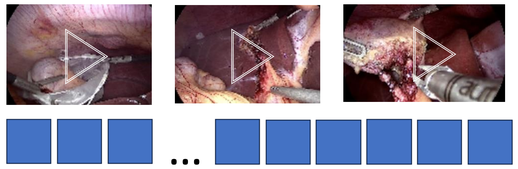
Methodology

1: ResNet + Transformer



Spatial Encoder (Frozen ResNet-18): Extracts key visual features from each frame (tools, anatomy) efficiently using a pre-trained backbone.
Temporal Encoder (Causal Transformer): Models the procedural flow by analyzing the sequence of frame features. A causal mask ensures the model only uses past information, making it suitable for live prediction.

2: Training Strategy



Multi-Video Finetuning: Our proposed method to boost test-time performance with no changes to the model architecture.

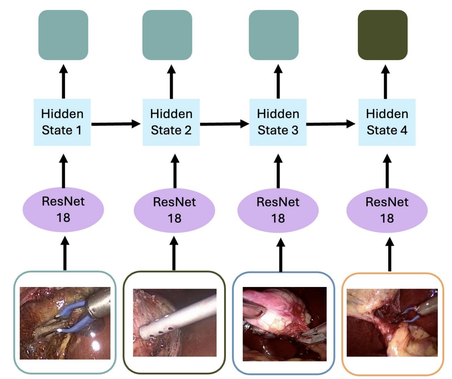
1. Divide individual videos into T/n frames, where n is the number of videos and T is the number of frames in that video.
2. Uniformly sample these frames at 1fps, and coalesce each video frame-by-frame

Theoretical advantages of proposed approach:

- "Forces" model to learn temporal relationships, not just copy previous frame label
- Accounts for inter-video variability (as well as phase variability)
- Lightweight, does not add any additional computational overhead to train using this method.

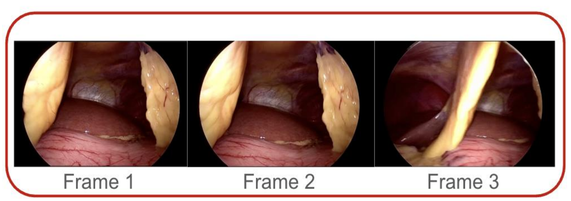
3: ResNet + GRU

Temporal Modeling: Used ResNet for frozen feature extraction. GRU is causal by nature, computationally efficient, and maintains temporal dependencies with hidden state.



4: 3D CNN

Joint Spatiotemporal Learning:
Stack consecutive video frames into 3D volume input of shape [C, T, H, W] where T is number of frames.
3D CNN backbone learns spatiotemporal features simultaneously
Final prediction made through MLP
Downside: May be less computationally efficient than other architectures since it needs to learn more parameters



Results

We use two metrics to evaluate model performance:

- Frame-wise phase classification accuracy
- Parameter count for computational efficiency

Method	Accuracy	Parameters
AlexNet	59.6	62,378,344
EndoNet (No HMM) [23]	65.6	62,435,702
MSN [3]	76.3	88,117,923
LoViT [17]	**91.5**	123,229,677
3D CNN + Transformer	38.4	52,183,337
Resnet + GRU	40.6	11,769,671
ResNet + Transformer	60.6	**23,789,639**

Table 1: Comparison of our method with baselines

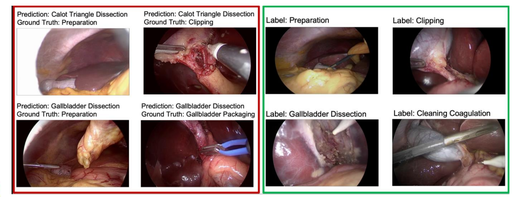
Ablation Study

Method	w/o MVT	w/ MVT
3D CNN + Transformer	38.4	39.1
Resnet + GRU	40.6	47.8
ResNet + Transformer	60.6	**63.8**

Table 2: Multi-video finetuning effects on custom-trained models.

Using Multi-Video Finetuning had a measurable impact on all the architectures we tried, suggesting a model-agnostic approach to improving test-time performance for video models.

Conclusion & Examples



- MVT increases accuracy without compromising computational efficiency
- Efficient and accurate architecture means potential for real-time surgical assistance
- Our method outperforms RNN baseline and EndoNet, but there is room for improvement