MICCAI Doctoral Consortium 2025

UNIVERSITÄT KLAGENFURT

Data-Efficient Learning for Generalizable Surgical Video Understanding

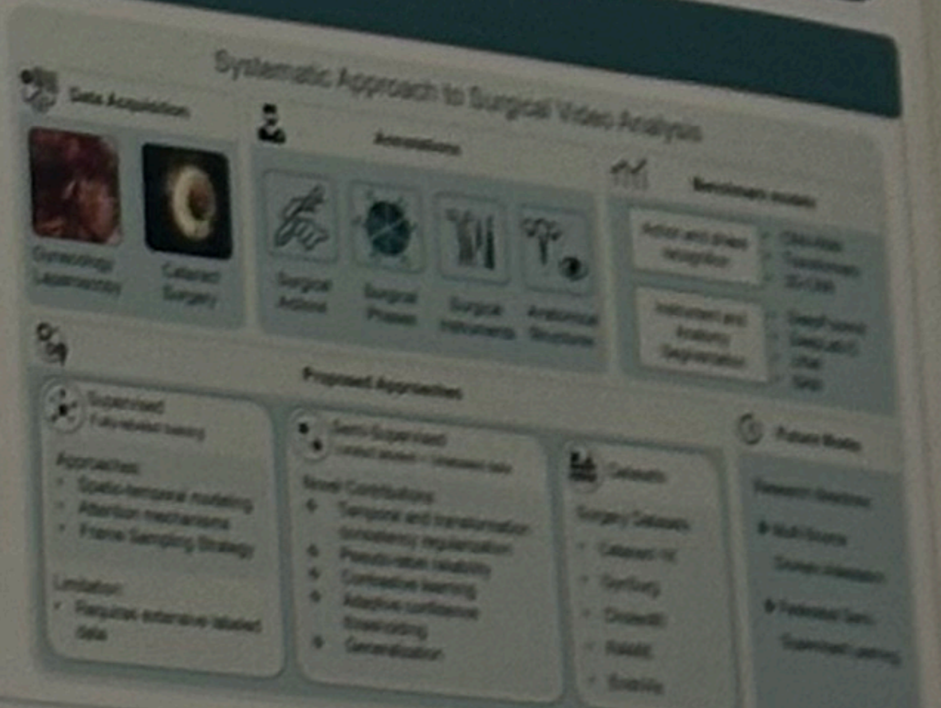
Sahar Nasirihaghighi
Department of Information Technology (ITEC), University of Klagenfurt

Research Problem and Motivation

The Challenge: Operating rooms are becoming intelligent, data-driven environments through computer-assisted systems that recognize surgical workflows and enable procedural indexing, performance evaluation, and report generation. While deep learning powers automated phase/action recognition and semantic segmentation, effectively utilizing surgical video data remains challenging.

Key Obstacles:
- Data Scarcity: Expert annotation is costly and time-consuming
- Visual Complexity: Motion blur, occlusions, smoke, blood degrade video quality
- Domain Gaps: Variations across institutions prevent reliable model generalization

Research Questions:
- How can we reduce reliance on large-scale expert annotations?
- How can we leverage abundant unlabeled surgical video data to improve intra-domain performance?
- How can we design data-efficient, temporally-aware, and generalizable models that perform reliably across diverse surgical settings?

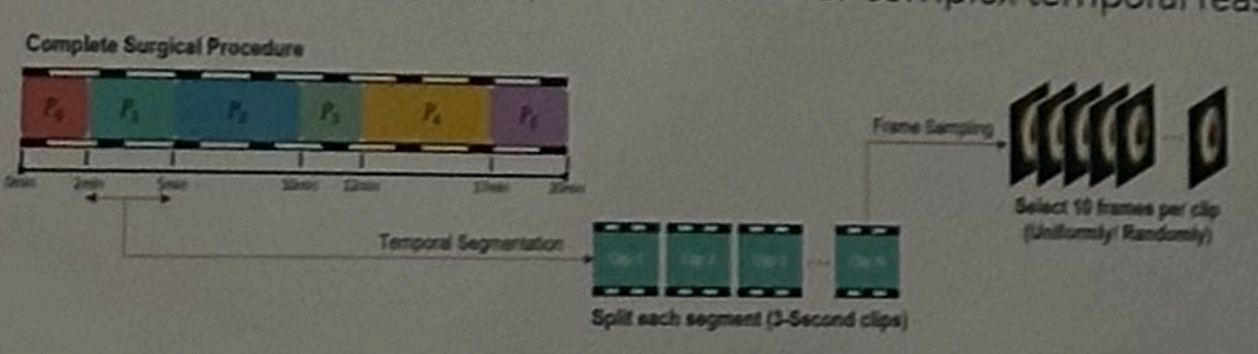


Proposed Approaches

❑ Supervised Action and Phase Recognition

CNN-RNN Architecture:
- Integrates spatial feature extractors with stacked recurrent layers
- Captures temporal dependencies for short, variable actions
- Frame sampling strategy enhances temporal diversity and robustness to surgeon-specific variations

CNN-Transformer Framework for Critical Events:
- Addresses critical event detection (bleeding, abdominal access, needle passing, coagulation)
- Transformer encoders model long-range temporal dependencies
- Multi-head self-attention processes spatial features for complex temporal reasoning

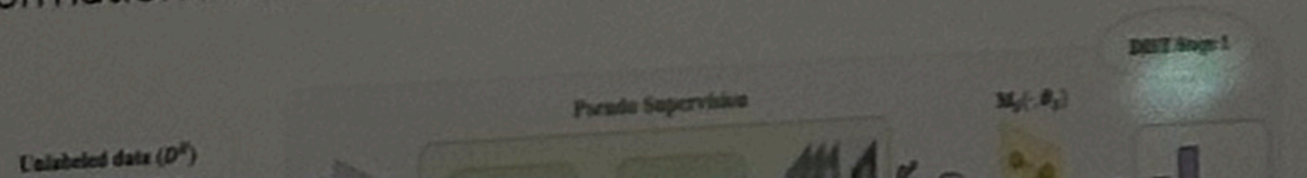


❑ Semi-Supervised Learning

❖ *Dual Invariance Self-Training for Surgical Phase Recognition.*

Two-Stage Framework:
Stage 1: Teacher model generates pseudo-labels for unlabeled clips ⟶ Student training
Stage 2: Student becomes teacher ⟶ process repeats

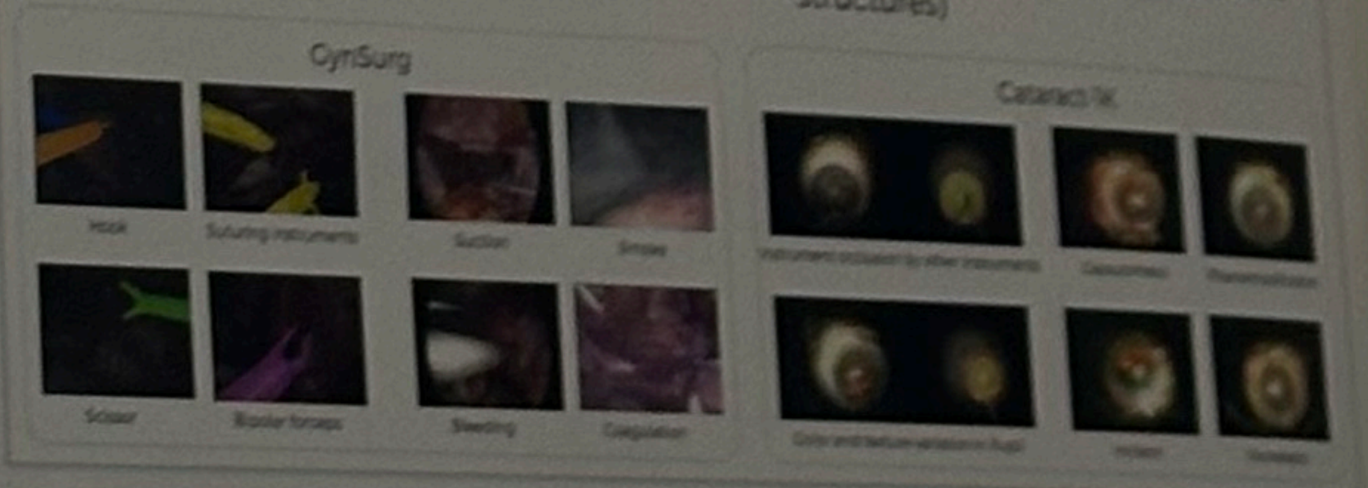
Pseudo Supervision:
I. **Reliability estimation:** Uses predictions from 3 checkpoints (T(n/3), T(2n/3), T(n)) - retains top 50% consistent labels
II. **Dual Invariance Constraints:**
- Temporal invariance across frame sampling strategies
- Transformation invariance under strong augmentations



Dataset Curation

Gynecology laparoscopy surgery (GynSurg):
- Actions: 152 videos (Actions and side-effects)
- Semantic Segmentation: 12362 frames (surgical instruments and anatomical structures)

Cataract surgery (Cataract-1k):
- Phases: 56 videos (12 phases)
- Semantic Segmentation: 2256 frames (surgical instruments and anatomical structures)



Results and Contributions

Our models were evaluated on six diverse surgical datasets (GynLap6, GynSurg, Cataract-1K, RAMIE, Cholec80, EndoVis) spanning action/phase recognition and instrument and anatomy segmentation tasks.

Dataset	Model	1/16 Acc.	1/16 F1	1/32 Acc.	1/32 F1
Cataract-1K	CMPL	83.26	79.87	72.47	[illegible]
	SVFormer	84.06	80.28	73.55	67.14
	DIST Stage1	91.22	90.26	85.93	84.02
	DIST Stage2	92.78	91.60	85.97	82.62

Dataset	Model	1/16 Acc.	1/16 F1	1/32 Acc.	1/32 F1
Cholec80	SVFormer	75.84	70.21	64.26	60.51
	DIST Stage1	79.52	79.43	74.63	73.44
	DIST Stage2	79.95	79.87	75.46	74.81

Dataset	Model	[illegible]	[illegible]
Cataract-1K	AD-MT	[illegible]	[illegible]
	AD-MT+ENCORE	[illegible]	[illegible]

Dataset	Model	[illegible]	[illegible]
Cataract-1K	Switch	[illegible]	[illegible]
	Switch+ENCORE	[illegible]	[illegible]

Dataset	Model	[illegible]	[illegible]
EndoVis	UniMatch	[illegible]	[illegible]
	UniMatch+ENCORE	[illegible]	[illegible]

As shown in this table, DIST achieves strong results on challenging low-label splits (e.g., 85.97% accuracy with only one labeled video in the 1/32 split) on both Cataract-1K and Cholec80.

This table summarizes ENCORE performance improvements [illegible] the-art baselines on the [illegible] low-label splits.