An Empirical Study on Activity Recognition in Long Surgical Videos

Zhuohong He, Ali Mottaghi, Aidean Sharghi, Muhammad Abdullah Jamal, Omid Mohareri

INTUITIVE

ML4H 2022
New Orleans, LA

Overview

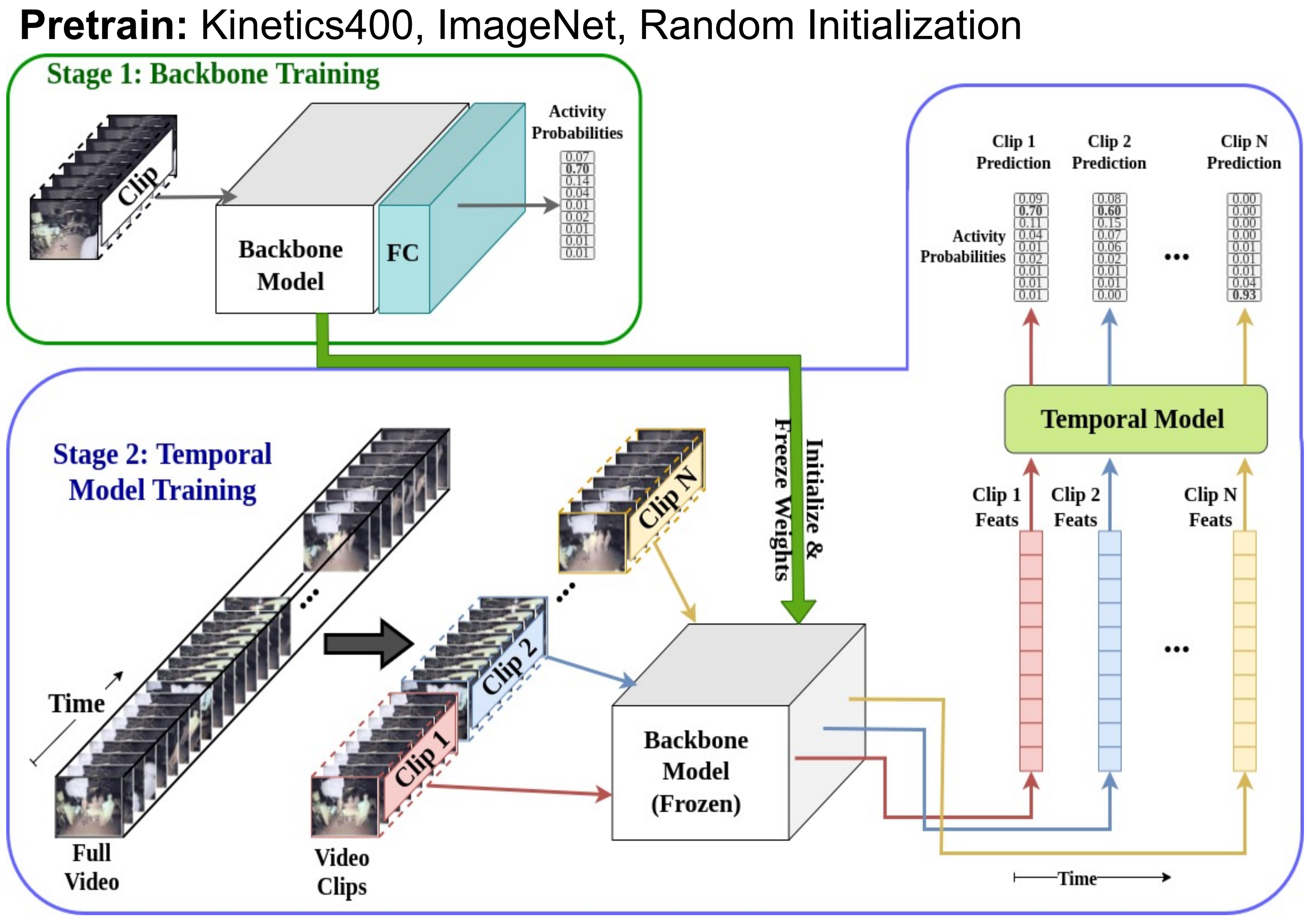
Introduction:
- Surgical videos captured by endoscopes, microscopes, and external cameras are readily available and information-dense.
- Understanding these videos provides valuable insight into operating room (OR) efficiency and safety which improve patient care.

Motivation:
- Previous SOTA on surgical activity recognition use lightweight frame-wise backbones due to the small dataset sizes.
- No previous study on peri-operative activity recognition on large-scale OR dataset.

Purpose: (1) To empirically study spatio-temporal clip-wise models on surgical datasets. (2) To investigate unsupervised and supervised domain adaptation techniques.

Methods

Pretrain: Kinetics400, ImageNet, Random Initialization

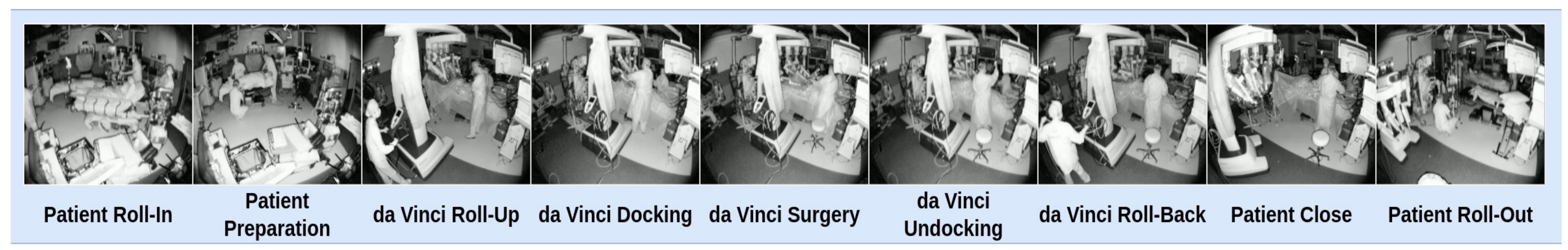


Models

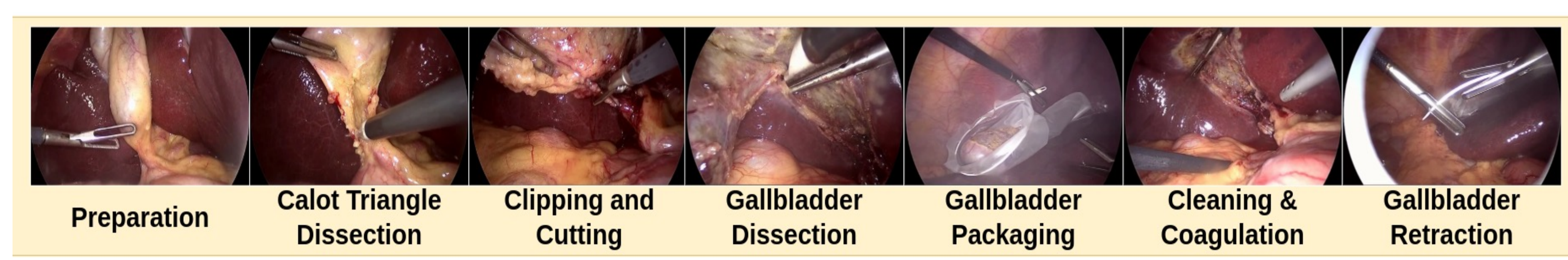
Due to the length of surgical videos, we employ a two-stage model. A backbone model extracts features for short video clips. Using the features, a temporal model predicts classes with long-term dependencies.
Backbones: We benchmarked four SOTA spatio-temporal backbones: Inflated 3DConvNet (I3D), SlowFast, TimeSformer, Video Swin Transformer
Temporal Model: We selected the Gated Recurrent Unit (GRU), Temporal Convolution Network (TCN), and Transformer models.

Datasets

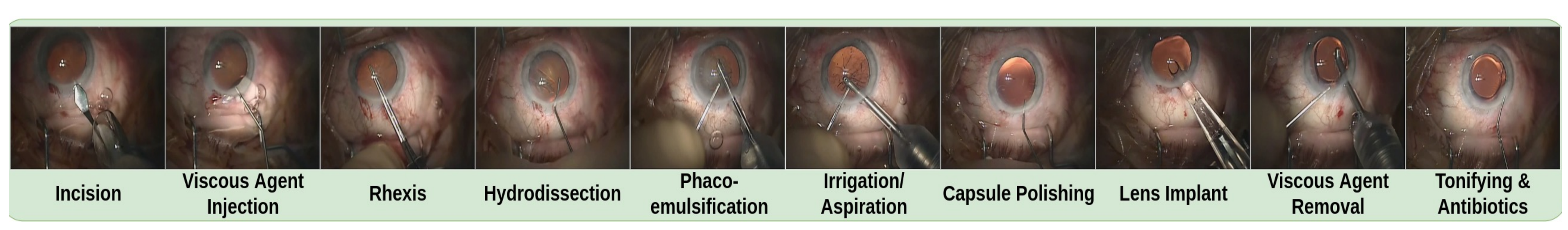
(1) Operating Room Activity Recognition (OR-AR): A large private dataset of 820 videos captured by time-of-flight sensors placed around the operating room. Each video ranges from 2-8 hours. Videos includes 30 procedure types across two ORs from 27 surgeons.



(2) Cholec80: A small public dataset containing 80 endoscopic videos of cholecystectomy procedures by 13 surgeons captured at 25 fps.



(3) Cataract-101: A small public dataset of 101 cataract surgeries captured from a microscope. The average video length is 8.3 mins.



Results

Backbone

Model	# of Params	FLOPs	Top-1 Acc. on test sets	
			Cholec80	SAR (Random)
I3D	27.2M	28.7G	71.36±2.19	85.33±0.67
SlowFast	33.6M	12.7G	71.38±3.59	86.26±1.50
TimeSformer	121.3M	196.1G	69.02±0.26	84.39±0.20
Swin Transformer	88.1M	141.0G	79.09±1.29	86.25±2.50

OR-AR

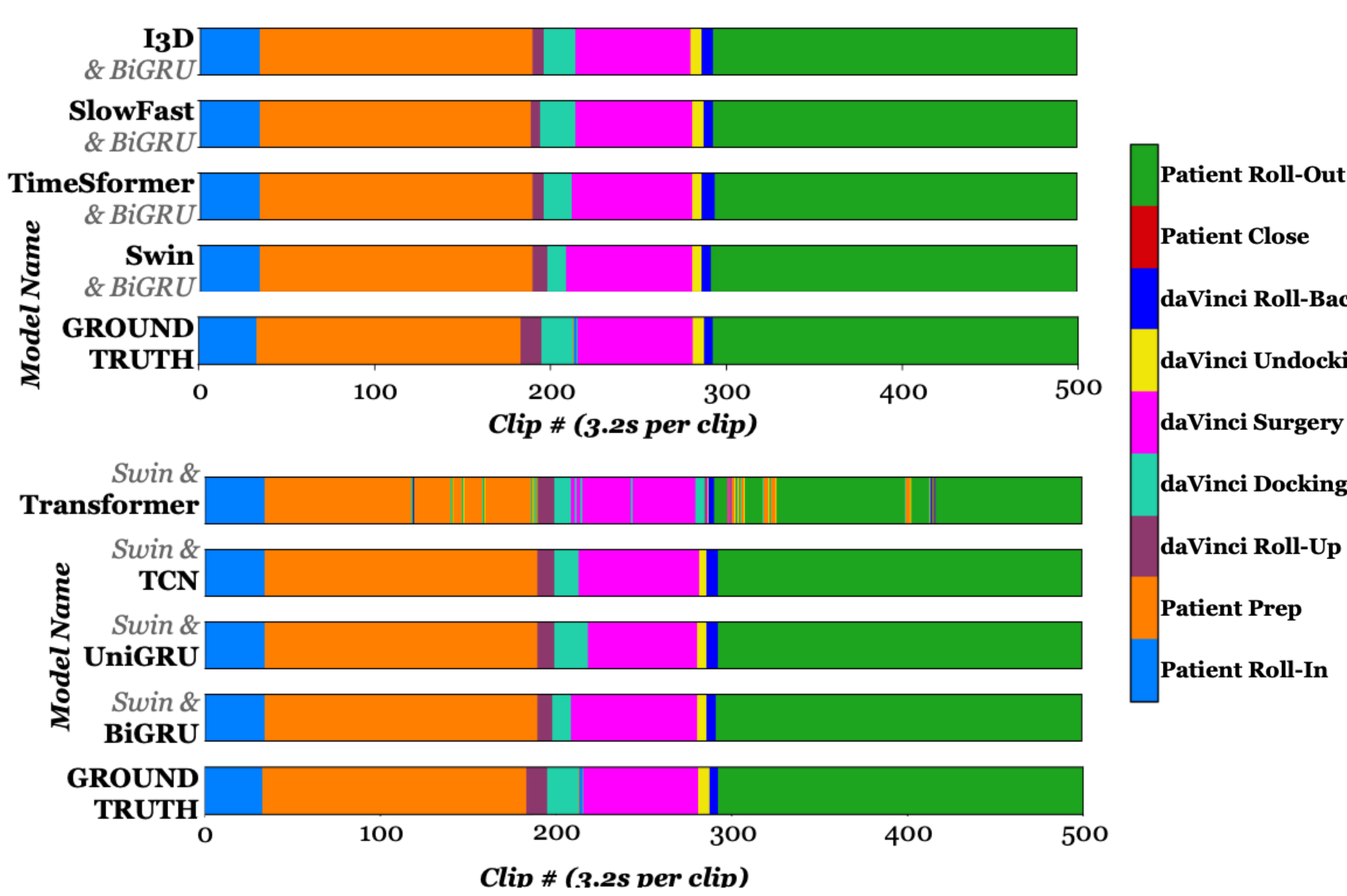
Backbone / Last Epoch (val-mAP)	*Temporal Model*			
	Transformer	Bi-GRU	Uni-GRU	TCN
I3D	79.30±0.06	94.04±0.66	90.95±0.74	91.33±0.23
SlowFast	79.42±1.71	94.33±0.19	90.70±0.04	89.79±1.08
TimeSformer	76.23±0.33	93.20±0.04	88.89±0.66	89.59±0.07
Swin	**82.50±2.35**	**95.13±0.35**	**92.02±0.69**	**91.54±0.03**

Cholec80

Model	Accuracy	Precision	Recall
PhaseLSTM (Twinanda et al., 2017)	79.68±0.07	72.85±0.10	73.45±0.12
EndoLSTM (Twinanda, 2017)	80.85±0.17	76.81±2.62	72.07±0.64
MTRCNet (Jin et al., 2020)	82.76±0.01	76.08±0.01	78.02±0.13
ResNetLSTM (Jin et al., 2018)	86.58±1.01	80.53±1.59	79.94±1.79
TeCNO (Czempiel et al., 2020)	88.56±0.27	81.64±0.41	85.24±1.06
I3D+UniGRU	88.27±1.04	80.18±0.20	80.58±1.97
SlowFast+UniGRU	90.47±0.46	83.12±2.09	82.33±1.22
TimeSformer+UniGRU	90.42±0.47	**86.05±1.13**	83.20±1.80
Swin+UniGRU	**90.88±0.01**	85.07±1.74	**85.59±0.53**

Cataract-101

Model	Accuracy	Precision	Recall
(Qi et al., 2019)	87.10	–	–
CB-RCNeSt (Xia and Jia, 2021)	96.37	94.89	94.69
I3D+UniGRU	93.69±0.21	91.27±0.02	91.05±0.41
SlowFast+UniGRU	92.08±0.32	89.30±1.22	88.63±0.32
TimeSformer+UniGRU	94.44±0.01	92.43±0.25	91.89±0.23
Swin+UniGRU	94.53±0.09	93.05±0.09	91.61±0.16



Backbone Method	Temp. Model Method	mAP	Accuracy	Precision	Recall
Freeze (init: hospA)	Freeze (init: hospA)	72.92	96.34	53.26	53.66
Train (init: K400)	Train (init: random)	83.01	93.47	68.84	75.99
Freeze (init: hospA)	Train (init: random)	90.85	97.15	81.65	89.01
Train (init: hospA)	Train (init: random)	91.64	97.64	87.61	84.08
Mottaghi et al. (2022)	Train (init: random)	88.99	97.26	86.72	86.02

Key Conclusions

- We establish a precedence of using clip-wise models for activity recognition in surgical video datasets. We show that pretraining is essential for convergence on small datasets.
- Video Swin Transformer & BiGRU is the strongest performing model.
- We achieved the new state-of-the-art performance on Cholec80 and OR-AR activity recognition set. Strong performance on Cataract-101.
- We demonstrate the adaptability of our model to domain shifts with minimal supervision.