An Empirical Study on Activity Recognition in Long Surgical Videos

Zhuohong He[1]   ZOOEY.HE@INTUSURG.COM
Ali Mottaghi[1,2]   MOTTAGHI@STANFORD.EDU
Aidean Sharghi[3]   AIDEAN.SHARGHI@GMAIL.COM
Muhammad Abdullah Jamal[1]   ABDULLAH.JAMAL@INTUSURG.COM
Omid Mohareri[1]   OMID.MOHARERI@INTUSURG.COM

[1]*Intuitive Surgical, Sunnyvale CA, United States*
[2]*Stanford University, Stanford CA, United States*
[3]*Tubi, San Francisco CA, United States*

Abstract

Activity recognition in surgical videos is a key research area for developing next-generation devices and workflow monitoring systems. Since surgeries are long processes with highly-variable lengths, deep learning models used for surgical videos often consist of a two-stage setup using a backbone and temporal sequence model. In this paper, we investigate many state-of-the-art backbones and temporal models to find architectures that yield the strongest performance for surgical activity recognition. We first benchmark the models performance on a large-scale activity recognition dataset containing over 800 surgery videos captured in multiple clinical operating rooms. We further evaluate the models on the two smaller public datasets, the Cholec80 and Cataract-101 datasets, containing only 80 and 101 videos respectively. We empirically found that Swin-Transformer+BiGRU temporal model yielded strong performance on both datasets. Finally, we investigate the adaptability of the model to new domains by fine-tuning models to a new hospital and experimenting with a recent unsupervised domain adaptation approach.

Keywords: Surgical Activity Recognition, Transformers, 3D ConvNets,

1. Introduction

Exacerbated by nation-wide staffing shortages, hospitals are focusing more on improving efficiency across all aspects of operation. Nowhere is this efficiency more linked to patient outcomes than the operating room (OR), where time under anesthesia, staff fatigue, and case turnover directly impact the quality and quantity of care. Additionally, OR efficiency can improve hospital profits and increase staff satisfaction and retention. Thus, a growing research field is working to model and analyze OR efficiency. Within the field, activity recognition of long surgical videos is a core technology which uses computer vision and deep learning to identify the activities or phases to generate a surgery timeline. Due to the ubiquity of camera sensors, surgical videos can be quickly collected and analyzed at scale.

Two of the most information-rich types of surgical videos are the videos captured by surgical tools (endoscopes, microscopes) and the external videos capturing the OR environment. Understanding endoscopic videos of laparoscopic and robotic surgery cases give us insight into surgeon efficiency and skill. External videos of the busy OR theatre allows us to understand how the idiosyncrasies

of the hospital staff and OR might affect the surgery workflow and outcome. We denote this task as Operating Room Activity Recognition. Recently, advancements within the computer vision field have improved the state-of-the-art on long video segmentation; however, their suitability is untested on surgical videos.

Activity recognition in surgical videos presents several unique challenges over general activity recognition. First, surgical videos are long, resulting in the need to model long-reaching dependencies between activities often separated by 3 or more hours. The activities themselves range from several minutes to several hours in length. Second, surgeries are complex processes, requiring annotations from niche medical experts. Consequently, the datasets are often small, suffer from class imbalances, and contain complex patterns. Thus, it is hard to train video understanding models on these diverse datasets and the trained models don't usually generalize.

To model long-term dependencies, most existing works use a two-stage training scheme. The backbone model is first trained to extract features from sub-units of the video. A second temporal sequence model utilizes the features to generate activity predictions with long-term dependencies. Recent approaches for solving surgical activity recognition (Twinanda et al., 2017; Czempiel et al., 2020) use a frame-wise backbone model which focuses solely on creating "spatial" representations through single frames, leaving the "temporal" learning to the second stage. However, we find that the state-of-the-art activity recognition models (Bertasius et al., 2021; Feichtenhofer et al., 2019; Liu et al., 2021b) utilize short video clips (16 or 32 frames) to extract global representations.

The goal of this work is to demonstrate the capabilities of spatio-temporal models on three surgical video datasets. First, we benchmarked the models on a new large-scale dataset we call the *OR-AR* (Operating Room Activity Recognition) which was first introduced by Sharghi et al. (2020) and was later extended from 400 videos to 820 videos. Second, we show our model's state-of-the-art performance on two smaller public surgical datasets: Cholec80 (Twinanda et al., 2017) and Cataracts-101 (Schoeffmann et al., 2018)) refuting the perception that spatio-temporal models perform poorly on small datasets. We delineate the best spatio-temporal backbone/temporal model combination and analyze the models' generalizability through experiments on challenging data splits and adaptation to new domains.

Our main findings and contributions are as follows:

- To the best of our knowledge, this is the first extensive evaluation of surgical activity recognition using different deep learning architectures.

- We find that Swin-Transformer + Bidirectional GRU is the best performing two-stage model. This model forms the new state-of-the-art on Cholec80 by achieving +2.32% accuracy, +0.35% precision, and +3.44% recall when compared to the previous best. The model performs similarly well on the OR-AR dataset, and nearly matches the SOTA performance on Cataract-101 without additional processes such as contrastive learning or increased supervision.

- We find that I3D backbone is the most efficient model achieving 98.9% of the Swin-Transformer's performance on OR-AR using 66% less parameters and 80% less FLOPs.

- We used an unsupervised technique to adapt our model to a new domain, a

new hospital. Using an unseen hospital's dataset, which is 8% the size of OR-AR, we achieved 93% of the performance (on OR-AR) with no labels. Using labels, we achieved 96% of the performance.

2. Related Work

In this section, we mostly review the current trends and approaches for surgical activity recognition.

Surgical activity recognition on *laparoscopic videos* has been extensively studied in the literature thanks to well-made public datasets such as the Cholec80, first introduced by Twinanda et al. (2017). Twinanda et al. (2017) also developed a multi-task learning approach to extract features from cholecystectomy video frames that capture both low-level visual information as well as tool-presence signals. Yengera et al. (2018) replaced the traditional HMMs with recurrent neural networks (RNN). Jin et al. (2020) proposed a correlation loss to capture the relevance between tool detection and activity recognition tasks to enrich the representation learning and utilized an RNN to learn the temporal dependencies between frames. TeCNO (Czempiel et al., 2020) used temporal convolution networks (Lea et al., 2017) on top of frame-wise features generated using ResNet to achieve state-of-the-art performance on Cholec80 and Cholec51 datasets.

On the other hand, surgical activity recognition in *robotic OR videos* was first introduced by Sharghi et al. (2020), which collected the large-scale OR activity recognition dataset (OR-AR) and developed a supervised model which can recognize the start and end of significant steps. They followed a two-stage training where I3D (Carreira and Zisserman, 2017), a 3D convolution-based model, was first pretrained and used for clip-wise feature extraction. Then, they

trained a LSTM (Hochreiter and Schmidhuber, 1997) using extracted features to model the global dependencies between different activities. Schmidt et al. (2021) introduced a multi-view model using the same building blocks. Jamal and Mohareri (2022) proposed a multi-modal unsupervised approach for low-data regime.

Finally, surgical activity recognition in *cataract surgery videos* is another burgeoning research area benefiting from the Cataract-101 (Schoeffmann et al., 2018). Using Cataract-101, Qi et al. (2019) explored a two-path ResNet architecture combining the RGB frame and edge detection frames to produce frame-wise, independent activity predictions. Xia and Jia (2021) developed an end-to-end model using a triplet loss consisting of activity recognition, phase recognition, and contrastive learning resulting in the state-of-the-art on the dataset.

3. Preliminaries

In studying and bench-marking model performance, we selected combinations between 4 state-of-the-art backbone models and 4 temporal models resulting in 16 model pairs. Our model selection explore both convolution (Carreira and Zisserman, 2017; Feichtenhofer et al., 2019) and Transformer-based (Bertasius et al., 2021; Liu et al., 2021b) backbones, as well as CNN (Lea et al., 2017), RNN (Cho et al., 2014), and Transformer (Vaswani et al., 2017) temporal models.

3.1. Backbone Architectures

During training of long videos, the stage-one backbone model usually learns to generate discriminative features for smaller units (either a single frame or a clip consisting of few frames) within the video. We focus on generating features for clip units due to their added ability to encapsulate the temporal motion of the objects in the scene when compared to only the spatial layout in a single

frame. We explore two prominent groups of "spatio-temporal" backbone models: 3D CNN and Transformer-based models.

3.1.1. 3D CNN-based Backbones.

3D CNNs are an intuitive method to extend single frame learning to video clip learning. First introduced in (Ji et al., 2013), 3D CNNs form the basis for a family of model architectures. Two landmark architectures – the Inflated 3D ConvNet (I3D)(Carreira and Zisserman, 2017) and the SlowFast (Feichtenhofer et al., 2019) – are explored in our work.

Inflated 3D ConvNets (I3D) I3D uses a Inception-V1 2D convolution model pretrained on ImageNet cloned over the temporal dimension to train on 3D video clips.

SlowFast This model proposed by Feichtenhofer et al. (2019) is a two stream path for video recognition. The "slow" path subsamples the input clip at a low frame rate and uses large spatial convolutions to extract visual information. The "fast" path subsamples the input clip at a fast frame rate and uses spatially small, temporally deep convolutions to capture rapid motions. The two pathways are connected through lateral connections at each layer.

3.1.2. Transformer-based Backbones.

The Transformer (Vaswani et al., 2017) which has revolutionized the Natural Language Processing (NLP) field has two distinct capabilities when processing sequences: (1) the ability to capture long-range dependencies, and (2) incredible compute parallelization, leading to scalability. Transformers are the state-of-the-art on many tasks including machine translation and question-answering (Chen et al., 2018; Dai et al., 2019). Recently, Transformers have been successfully applied to visual tasks. Vision Transformers (ViT) (Dosovitskiy et al.,

2020) treats images patches as input tokens to the Transformer's self-attention units.

TimeSformer Bertasius et al. (2021) introduced the TimeSformer which extends the ViT into video understanding by performing self-attention across compact embeddings of image patches from each frame in a clip. They further investigated several self-attention designs over the space-time volume and assessed each on benchmark datasets. However, TimeSformer's computational complexity grows quadratically with image dimension and clip duration.

Video Swin Transformer Liu et al. (2021b) proposed the Video Swin Transformer which extended Swin Transformers (Liu et al., 2021a) to introduce a TimeSformer alternative that yielded linear computational complexity growth with respect to image and time dimensions. The model performs self-attention across patches within separate image windows; then "shifts" the window in the following layers to merge the features from the previous layer (*swin*=*s*hifted *win*dows). In this paper, we denote the Video Swin Transformer as "Swin".

3.2. Temporal Sequence Models

A powerful backbone serves as a great foundation in many video understanding tasks; however, when dealing with videos comprising of multiple long activities, in order to obtain reliable (global) predictions, we still need to model the temporal dependencies between the extracted features. In this work, we focus on CNN, RNN, and Transformer-based temporal sequence models.

Temporal Convolution Networks. Lea et al. (2017) introduced a new class of time-series models consisting of hierarchical temporal convolutions to capture long-range dependencies in videos and used it to perform
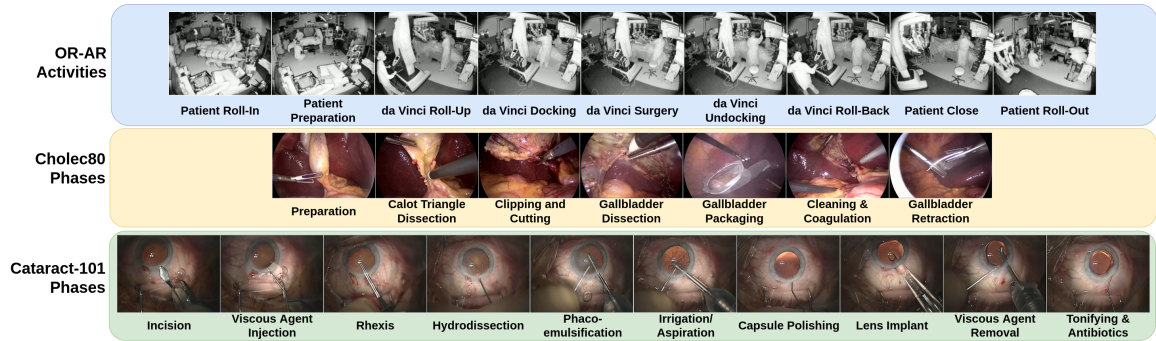
Figure 1: The predicted activity classes for each dataset with corresponding illustrations.

action segmentation on videos. Assuming a low number of layers, these models are faster to train than traditional RNNs.

Recurrent Neural Networks. RNNs have been used in a wide range of applications and require little introduction. They follow a sequential nature where the output at any time depends on the "memory" of previous inputs as well as the current input. Since the memory compactly summarizes all the previous inputs, RNNs can suffer when learning long-range dependencies. Long-Short Term Memory Networks (LSTM) (Hochreiter and Schmidhuber, 1997) alleviate this issue by introducing several gates responsible in determining what information is worth storing in the memory. Gated Recurrent Network (GRU) (Cho et al., 2014) (a simplified LSTM) have shown incredible performance in sequence modeling tasks and hence we include it in our experiments. A bi-directional GRU architecture contains a second GRU which operates on the sequence in reverse order, giving the model knowledge of future data.

Transformers. Transformers are very efficient for training when the input sequence length is significantly smaller than the feature dimensions. However, this is not always the case when dealing with videos of variable length. Nevertheless, we include it (Vaswani et al., 2017) in our experiments.

4. Datasets

OR-AR We benchmarked the backbone & temporal model combinations on three different datasets. Our first dataset, the base Operating Room Activity Recognition (OR-AR) dataset was first introduced by Sharghi et al. (2020). It contains 820 videos taken using time-of-flight sensors placed around the OR during robotic surgery. Each video frame is a weighted sum of intensity and depth frames. The data was collected on 27 surgeons and includes 30 types of surgeries across two ORs over a period of two years. This dataset contains an additional 66 videos taken from a second hospital which will only be used for model adaptation experiments. Video annotations assign regions of the video into 9 class shown in Figure 1.

We created 5 data splits that measure the generalizability of the models. The random split separates the dataset into an 80-20 train-test split. The OR, surgeon, procedure, and chronological splits are described in detail in the Appendix B.1.

Cholec80 Our second dataset, Cholec80 (Twinanda et al., 2017), contains 80 endoscopic videos of cholecystectomy procedures
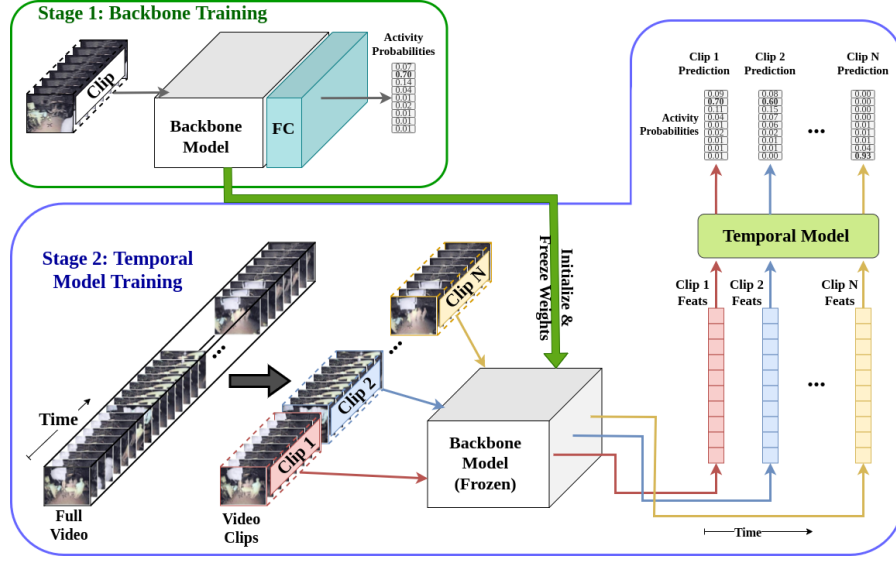
Figure 2: Two-stage model training process. In stage 1, the backbone model learns to classify clips as different surgical activities. In stage 2, a video is cut into N clips and passed through the backbone to extract N clip features. Then, the sequence of features trains a temporal model to classify each clip as one of the activities.

performed by 13 surgeons and is labeled with 7 surgical phases by a senior surgeon. The surgical phases are defined in Figure 1. The dataset also includes presence annotations for 7 surgical tools which several other papers use to aid phase predictions. The videos are recorded at 25fps using a 1920x1080 resolution. Following previous works, the first 40 videos are used for training and the remaining 40 for testing.

Cataract-101 The Cataract-101 dataset contains 101 videos from cataract surgeries captured from a microscope and labelled by a senior ophthalmic surgeon in Klinikum Klagenfurt, Austria (Schoeffmann et al., 2018). The start & end times for 10 phases (defined in Figure 1) are given. Videos are recorded at 25fps using 720x540 resolution with an average 8.3 min. length. These phases are less orderly than the other datasets. Following

Qi et al. (2019), we randomly split the data into 73 train & 28 test videos.

5. Experiments

Metrics. For the Cholec80 and Cataract-101 datasets, we report the average test accuracy, precision, and recall on the epoch with the lowest training loss. On the OR-AR dataset, we report the average test-time mean average precision (mAP) score at the last training epoch.

Backbone Training Every experiment involves two stages – backbone training and temporal model training – which are illustrated in detail in Figure 2.

Once the models were trained, we extracted features from the videos for each model by saving the output of the penultimate layer for each clip. For OR-AR, we noticed that dropping the clip size from 32 to 16
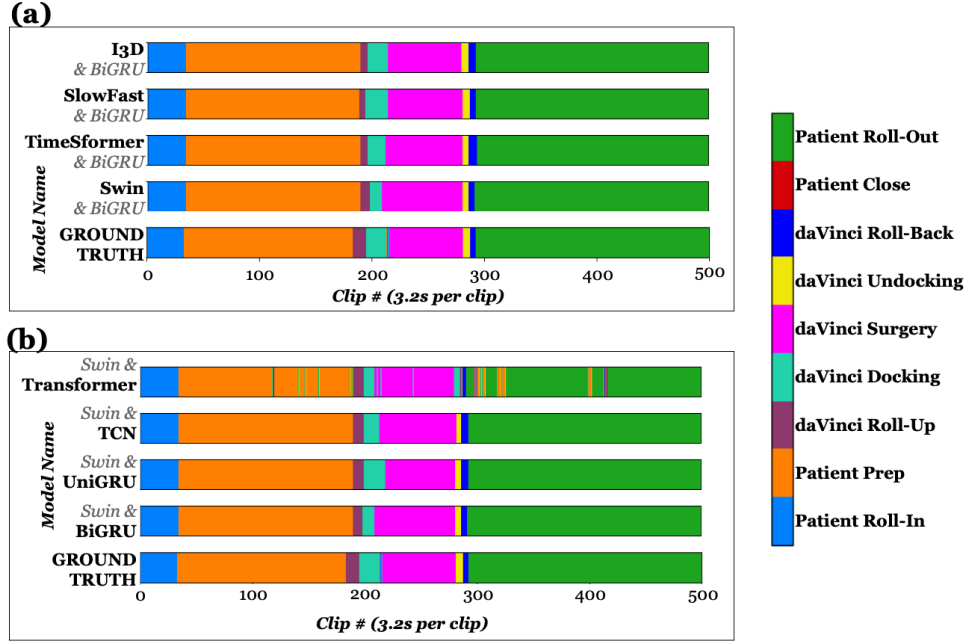
Figure 3: Prediction comparison of different models for one video in the OR-AR dataset. (a) The performance of various backbones with BiGRU. (b) The performance of various temporal models with the Swin backbone.

frames during extraction improved the performance of the subsequent temporal model. We conjecture this change increases the information density in the features resulting stronger performance.

Temporal Model Training. Once the features are extracted, each video in the training set serves as a single sample for training the temporal models. The goal of the temporal model is to learn the global logic of phase predictions across clips. Since several videos in the same batch can have vastly different lengths, we padded the shorter sequences with zeros; however, we make sure to prevent the padding from influencing the loss using a masking scheme.

5.1. Pretraining

To converge on the smaller Cholec80 and Cataract-101 datasets, all backbone mod-

els required pretraining on the Kinetics-400 dataset (K400) (Kay et al., 2017), a dataset consisting of 240K training clips of human actions. For the larger OR-AR dataset, we used random initialization (He et al., 2015).

6. Results and Discussion

OR-AR. The Video Swin Transformer backbone shows a clear performance edge over other backbones as illustrated in the random split results in Table 1 and Figure 3. Interestingly, the 3D CNN-based backbones (SlowFast, I3D) outperforms the attention-based TimeSformer backbone. We believe this is due to differences between the datasets. OR-AR requires a deeper understanding of motion cues because camera position is not fixed between videos and the OR scene is more complex. The TimeSformer's "divided" spatio-temporal attention strug-

362

Table 1: Average Validation mAP scores on the last epoch over 2 trials on the OR-AR dataset using the Random split (refer to Appendix B.1 for split details).

		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
Backbone	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**

Table 2: Performance comparison of top bench-marked models on Cholec80. Average metrics (%) over multiple runs along with their standard deviations.

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**

Table 3: Cataract-101 performance comparison between our approach and other baselines.

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

Table 4: Performance of the model adaptation techniques from hospital A to hospital B. All techniques use the Swin+BiGRU model.

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

gles against the joint spatio-temporal modelling of Swin and 3D CNN-based backbones. From our experiments, we also conclude that Swin is better at recognizing patterns across different environments. When testing on an unseen OR (refer to Appendix Table 7), the performance gap between Swin and other backbones is even more pronounced. Other models may rely on OR-specific visual cues resulting in lower generalizability.

Cholec80. The clip-wise models that we analyzed achieved the **new state-of-the-art on the Cholec80 dataset**. Table 2 demonstrates that all clip-wise models performed on-par or better than baselines from existing literature, with the Swin+UniGRU yielding the best performance in accuracy (+2.32%) and recall (+0.35%) while yielding strong performance in precision (+3.43%). This is especially impressive considering several baselines utilize additional information (tool presence) to perform activity recognition. TimeSformer+UniGRU and Slow-Fast+UniGRU still produced strong results. Results for of all backbone & temporal combinations are shown in Appendix Table 11.

Cataract-101. Our benchmarked models achieved strong performance on the Cataract-101 dataset (see Table 3). Our best model *Swin+UniGRU* model achieved numbers slightly below SOTA (accuracy: -

1.84%, precision: -1.84%, recall: -3.08%). The state-of-the-art CB-RCNeSt uses additional supervision by grouping the 7 phases into 4 broader phases and training the model to predict both. Additionally, it uses a contrastive branch to differentiate visually similar frames. Since cataract surgery phases are often non-sequential, we reason that our two-stage model's global understanding may be unhelpful for this task.

Backbone & Temporal Combinations. The results from each backbone & temporal model combination are shown in Table 1 for the OR-AR random split and Appendix Table 11 for the Cholec80 dataset. Overall, the Swin+BiGRU combination performed the best across all datasets. Pertaining to temporal model performance, it is expected that the future-seeing models (Bi-GRU, Transformer) should outperform the "online" models (UniGRU, TCN). It appears that Transformers do not meet this expectation. We believe this is due to the Transformer's complexity. The Transformer can access the global sequence from each time-step. While usually an advantage, this makes the model incredibly difficult to optimize when there is a natural order of activities in a very long sequence especially when the dataset is small and video lengths are irreg-

ular. In contrast, the BiGRU is inherently sequential and delivers strong performance.

6.1. Model Adaptation

We investigated the transfer learning techniques for the OR-AR task on an additional 66 recordings from another hospital (56 train, 10 test) which is 8% the size of the OR-AR dataset. Our results are presented in Table 4. Using the unsupervised technique from Mottaghi et al. (2022) to fine-tune Swin, we achieved an impressive 88.99% test mAP (93% of the OR-AR supervised performance). Fine-tuning using labelled data, we achieve 91.64% mAP. Please see Appendix A.3 for implementation details.

7. Conclusion

In this paper, we focus on exploring spatio-temporal backbone architectures for surgical activity recognition and activity detection in the OR. We explore both 3D CNN-based and transformer-based architectures and show their superior performance on benchmark datasets. Furthermore, we argue that while no pretraining is required in the presence of a large labeled dataset, by benefiting from existing action recognition datasets we can overcome the performance and optimization issues that often serve as an obstacle to adoption of these models on small medical datasets. Finally, we combine the backbone models with the three predominant categories of temporal modeling and show that the Video Swin Transformer + Bi-Directional GRU produces the best performance on both Cholec80 and OR-AR datasets.

Acknowledgments

References

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *CoRR*, abs/2102.05095, 2021. URL https://arxiv.org/abs/2102.05095.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. *CoRR*, abs/1705.07750, 2017. URL http://arxiv.org/abs/1705.07750.

Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George F. Foster, Llion Jones, Niki Parmar, Mike Schuster, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. The best of both worlds: Combining recent advances in neural machine translation. *CoRR*, abs/1804.09849, 2018. URL http://arxiv.org/abs/1804.09849.

KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014. URL http://arxiv.org/abs/1409.1259.

Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8,*

2020, Proceedings, Part III, volume 12263 of *Lecture Notes in Computer Science*, pages 343–352. Springer, 2020. doi: 10.1007/978-3-030-59716-0_33. URL https://doi.org/10.1007/978-3-030-59716-0_33.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019. URL http://arxiv.org/abs/1901.02860.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL https://arxiv.org/abs/2010.11929.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6201–6210. IEEE, 2019. doi: 10.1109/ICCV.2019.00630. URL https://doi.org/10.1109/ICCV.2019.00630.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1026–1034. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.123. URL https://doi.org/10.1109/ICCV.2015.123.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL https://doi.org/10.1162/neco.1997.9.8.1735.

Muhammad Abdullah Jamal and Omid Mohareri. Multi-modal unsupervised pre-training for surgical operating room workflow analysis. *arXiv preprint arXiv:2207.07894*, 2022.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):221–231, 2013. doi: 10.1109/TPAMI.2012.59. URL https://doi.org/10.1109/TPAMI.2012.59.

Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: Workflow recognition from surgical videos using recurrent convolutional network. *IEEE Trans. Medical Imaging*, 37(5):1114–1126, 2018. doi: 10.1109/TMI.2017.2787657. URL https://doi.org/10.1109/TMI.2017.2787657.

Yueming Jin, Huaxia Li, Qi Dou, Hao Chen, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Anal.*, 59, 2020. doi: 10.1016/j.media.2019.101572. URL https://doi.org/10.1016/j.media.2019.101572.

Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. URL http://arxiv.org/abs/1705.06950.

Colin Lea, Michael D. Flynn, René Vidal, Austin Reiter, and Gregory D. Hager. Temporal convolutional networks for action segmentation and detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1003–1012. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.113. URL https://doi.org/10.1109/CVPR.2017.113.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 9992–10002. IEEE, 2021a. doi: 10.1109/ICCV48922.2021.00986. URL https://doi.org/10.1109/ICCV48922.2021.00986.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. *CoRR*, abs/2106.13230, 2021b. URL https://arxiv.org/abs/2106.13230.

Ali Mottaghi, Aidean Sharghi, Serena Yeung, and Omid Mohareri. Adaptation of surgical activity recognition models across operating rooms, 2022. URL https://arxiv.org/abs/2207.03083.

Baolian Qi, Xiaolin Qin, Jia Liu, Yang Xu, and Yuwen Chen. A deep architecture for surgical workflow recognition with edge information. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1358–1364, 2019. doi: 10.1109/BIBM47256.2019.8983269.

Adam Schmidt, Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Multi-view surgical video action detection via mixed global view attention. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2021 - 24th International Conference, Strasbourg, France, September 27 - October 1, 2021, Proceedings, Part IV*, volume 12904 of *Lecture Notes in Computer Science*, pages 626–635. Springer, 2021. doi: 10.1007/978-3-030-87202-1_60. URL https://doi.org/10.1007/978-3-030-87202-1_60.

Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In Pablo César, Michael Zink, and Niall Murray, editors, *Proceedings of the 9th ACM Multimedia Systems Conference, MMSys 2018, Amsterdam, The Netherlands, June 12-15, 2018*, pages 421–425. ACM, 2018. doi: 10.1145/3204949.3208137. URL https://doi.org/10.1145/3204949.3208137.

Aidean Sharghi, Helene Haugerud, Daniel Oh, and Omid Mohareri. Automatic operating room surgical activity recognition for robot-assisted surgery. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention - MICCAI 2020 - 23rd International Conference, Lima, Peru, October 4-8, 2020, Proceedings, Part III*, volume 12263 of *Lecture Notes in Computer Science*, pages 385–395. Springer, 2020. doi: 10.1007/978-3-030-59716-0_37. URL https://doi.org/10.1007/978-3-030-59716-0_37.

367

Andru Putra Twinanda. *Vision-based approaches for surgical activity recognition using laparoscopic and RBGD videos. (Approches basées vision pour la reconnaissance d'activités chirurgicales à partir de vidéos laparoscopiques et multi-vues RGBD)*. PhD thesis, University of Strasbourg, France, 2017. URL https://tel.archives-ouvertes.fr/tel-01557522.

Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. Endonet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Trans. Medical Imaging*, 36(1):86–97, 2017. doi: 10.1109/TMI.2016.2593957. URL https://doi.org/10.1109/TMI.2016.2593957.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Tong Xia and Fucang Jia. Against spatial–temporal discrepancy: contrastive learning-based network for surgical workflow recognition. *International Journal of Computer Assisted Radiology and Surgery*, 16(5):839–848, May 2021. ISSN 1861-6429. doi: 10.1007/s11548-021-02382-5. URL https://doi.org/10.1007/s11548-021-02382-5.

Gaurav Yengera, Didier Mutter, Jacques Marescaux, and Nicolas Padoy. Less is more: Surgical phase recognition with less annotations through self-supervised pretraining of CNN-LSTM networks. *CoRR*, abs/1805.08569, 2018. URL http://arxiv.org/abs/1805.08569.

Appendices

8. Implementation Details

8.1. Backbone Training

To train the backbone, we selected short clips of activities from the long surgical videos; then we sub-sampled the frames within each clip according to Appendix Table 5 to form the input frames to the backbone. For Cholec80 and Cataract-101, we kept the scope of the backbones to 1-second long clips to match the prediction frequency of the existing works which form our baselines for comparison. We used categorical cross entropy loss on the output probabilities and used stochastic gradient descent (SGD) with momentum to update the backbone weights. We used a cosine learning rate scheduler (with Swin,Timesformer LR=0.01; SlowFast,I3D LR=0.1). We trained Swin, TimeSformer, SlowFast, and I3D for 50, 60, 196, and 100 epochs respectively for all datasets. SlowFast and I3D models also used a linear warm-up for 34 and 10 epochs respectively. The only applied augmentation technique was random flipping.

Table 5: Clip forming strategies. N(M/S) represents N-frame model inputs formed by sampling M-frame video clips by S.

Dataset	Training	Extraction
OR-AR	16(32/2)	16(16/1)
Cholec80	8(25/3.125)	8(25/3.125)
Cataract101	8(25/3.125)	8(25/3.125)

Backbone params, FLOPs, and performance on the OR-AR dataset can be seen in Appendix Table 6. The attention-based models have higher number of parameters and FLOPs compared to their 3D Con-vNet counterparts. TimeSformer has the highest number of parameters (and FLOPs) due to global connection between the image patches. Swin transformer reduces the complexity by only connecting the neighboring patches instead.

8.2. Temporal Model Training

For the GRU models, we used 0.1 drop-out on input, a 64 dimension hidden state across 4 GRU layers, and a linear last layer. We report numbers for both the unidirectional and bidirectional GRU. For the TCN model, we used 7 layers with the number of channels between layers equivalent to the input feature dimension. For the Transformer model, we used an FC layer to decrease the input feature dimension to 128, followed by an 1 layer, 2 headed encoder. Lastly, we used an linear layer to obtain the final output. The Transformer was trained with 0.1 drop-out on the encoder parameters.

8.3. Model Adaptation

For the OR-AR dataset, we explored domain adaptation techniques to understand the Swin backbone's ability to converge on an additional 66 recordings from another hospital which we randomly split into 56 training and 10 testing videos. We applied the unsupervised adaptation method proposed by Mottaghi et al. (2022) which trains the model on strongly augmented data from the new domain using auto-generated pseudo-labels. Additionally, we fine-tuned the Swin backbone using fully-labeled data for 25 epochs starting from a model trained on the original domain. We applied a tiered learning rate with cosine decay which mostly froze lower layers (by-layer lr: $[0, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}]$). For the temporal model in both techniques, we retrained the BiGRU from random under supervision for 20 epochs with a learning rate of 0.001. We found that train-

Table 6: Further information on backbone models including details about model sizes and backbone Top-1 error during testing.

Model	# of Params	FLOPs	Top-1 Acc. on test sets	
			Cholec80	OR-AR (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

ing BiGRU from random and from a model trained on the original domain both yielded similar performance.

8.4. Hardware.

We performed most benchmarks on four NVIDIA A100 (40GB) GPUs using the AMD EPYC 7742 64-Core CPU with 1TB RAM. Remaining experiments were conducted using four NVIDIA RTX A4000 (16GB) GPUs using the AMD Ryzen Threadripper PRO 3975WX 32-Cores CPU with 500GB RAM.

9. Splitting Strategies

9.1. OR-AR Splitting Strategies

The Operating Room Activity Recognition (OR-AR) dataset contains roughly 820 robotic surgery case-videos (.mp4) from ≈205 cases, on 27 surgeons and 30 types of surgeries, across two ORs, and over a period of two years. We placed four sensors around the OR to capture multiple viewpoints of each case. In order to comprehensively assess our models, we defined several train-test data splits. Each split is designed to evaluate models' generalizability across different domains.

I. Random. We randomly split the cases into a train-test split at a ratio of 80:20. Multiple views from the same case were grouped into the same set, either train or test. This splitting strategy creates homogeneous sets, and thus expresses the models' upper bound of performance.

II. Operating Room. Because the data is captured from two ORs, we can study a model's generalizability to an unseen environment by creating a train-test split on the ORs. We used data from one OR for training and use the other OR for testing. This resulted in a 63:37 train-test split.

III. Surgeon. Surgeons lead the procedure and significantly affect the surgical workflow. In this experiment, we train our model on videos from 13 surgeons and assess the model's performance on videos from the remaining 8 surgeons. The surgeons are split such that we obtain an 80%-20% train-test split on case videos.

IV. Procedure. With this split, we assess if the model can generalize to novel procedures by training on 21 specific procedures (Lobectomy, Colectomy, etc.) and testing on 7 other unseen procedures (Umbilical Hernia Repair, etc.). The resulting split has a train-test split of 80:20 on case videos.

V. Chronological. We would like to see our models keep up with the evolution of OR workflow over time. We thus introduced a split where the first 80% of chronologically-ordered cases were placed in the train set with the following 20% in the test set.

Table 1 in the main text shows how the models perform on Random data split. We assess the models on the remaining data splits in Tables 7, 8, 9 and 10 in the supplementary material (see the next page).

Table 7: Average Validation mAP on the last epoch scores over 2 Trials on the OR-AR Dataset using the Operating Room split

		Temporal Model			
		Transformer	Bi-GRU	Uni-GRU	TCN
	I3D	49.15±2.41	71.39±2.62	65.77±1.68	67.66±0.25
	SlowFast	49.48±1.33	72.18±0.67	68.29±0.02	67.40±0.35
Backbone	TimeSformer	44.08±0.60	66.80±2.28	62.44±1.05	63.88±2.57
	Swin	**55.41±1.40**	**75.92±2.06**	**73.18±4.12**	**76.39±1.85**

Table 8: Average Validation mAP scores on the last epoch over 2 Trials on the OR-AR Dataset using the Surgeon split

		Temporal Model			
		Transformer	Bi-GRU	Uni-GRU	TCN
	I3D	67.61±1.25	88.89±1.24	84.00±1.32	83.79±0.06
	SlowFast	68.66±0.83	88.81±0.91	83.78±1.56	83.68±1.18
Backbone	TimeSformer	64.66±1.12	86.51±0.77	82.24±0.67	81.65±0.02
	Swin	**73.16±1.22**	**90.70±0.43**	**86.27±1.16**	**86.08±0.21**

Table 9: Average Validation mAP scores on the last epoch over 2 Trials on the OR-AR Dataset using the Procedure split

		Temporal Model			
		Transformer	Bi-GRU	Uni-GRU	TCN
	I3D	72.20±0.89	91.29±0.40	86.80±0.43	86.61±1.46
	SlowFast	73.27±0.27	89.96±0.63	87.60±1.42	86.82±0.39
Backbone	TimeSformer	71.21±0.06	**91.38±0.04**	85.75±0.43	86.85±0.37
	Swin	**77.92±2.52**	91.23±0.86	**88.98±0.37**	**89.53±0.36**

Table 10: Average Validation mAP scores on the last epoch over 2 Trials on the OR-AR Dataset using the Chronological split

		Temporal Model			
		Transformer	Bi-GRU	Uni-GRU	TCN
	I3D	78.99±0.16	93.65±0.63	92.17±0.09	89.64±0.07
	SlowFast	79.28±0.13	92.22±1.00	90.02±1.36	90.97±1.32
Backbone	TimeSformer	76.35±1.24	92.93±0.40	89.72±1.02	90.20±0.92
	Swin	**83.35±1.70**	**93.91±0.61**	**92.64±0.77**	**93.01±0.16**

Table 11: Performance comparison of different backbones and temporal models on the Cholec80 Dataset.

Model	Accuracy	Precision	Recall
I3D+Transformer	73.46±0.99	68.42±1.20	66.02±1.53
I3D+TCN	81.21±2.49	81.75±0.46	75.64±0.45
I3D+UniGRU	88.27±1.04	80.18±0.20	80.58±1.97
I3D+BiGRU	91.81±0.06	85.20±1.44	85.10±4.18
SlowFast+Transformer	74.12±0.03	69.73±0.32	65.99±1.15
SlowFast+TCN	85.35±1.39	80.53±0.69	79.86±0.50
SlowFast+UniGRU	90.47±0.46	83.12±2.09	82.33±1.22
SlowFast+BiGRU	92.74±0.23	87.71±1.27	84.69±1.00
TimeSformer+Transformer	73.46±0.99	73.93±0.32	68.27±2.28
TimeSformer+TCN	87.22±0.24	91.75±0.46	81.95±0.69
TimeSformer+UniGRU	90.42±0.47	86.05±1.13	83.20±1.80
TimeSformer+BiGRU	92.82±1.91	89.70±1.34	86.18±2.67
Swin+Transformer	80.10±0.72	74.35±0.98	74.37±0.69
Swin+TCN	87.23±0.34	82.57±0.75	82.51±0.31
Swin+UniGRU	90.88±0.01	85.07±1.74	85.59±0.53
Swin+BiGRU	93.87±0.04	89.96±0.79	89.65±0.58