Towards Open, Reproducible and Collaborative Surgical Data Science

Open-Source Software for Surgical Technologies Workshop

Alejandro Granados
Lecturer in Surgical Data Science
Surgical & Interventional Engineering

alejandro.granados@kcl.ac.uk

Reproducibility

		Data: Same	Data: Different
Analysis	Same	Reproducible	Replicable
Analysis	Different	Robust	Generalisable

The Turing Way Community. *The Turing Way: A handbook for reproducible, ethical, and collaborative research*. **Zenodo** (2021)

Example One

Reproducibility of AI-related research

Article

International evaluation of an AI system for breast cancer screening

https://doi.org/10.1038/s41586-019-1799-6

Received: 27 July 2019

Accepted: 5 November 2019

Published online: 1 January 2020

Scott Mayer McKinney[1,14]*, Marcin Sieniek[1,14], Varun Godbole[1,14], Jonathan Godwin[2,14], Natasha Antropova[2], Hutan Ashrafian[3,4], Trevor Back[2], Mary Chesus[2], Greg S. Corrado[1], Ara Darzi[3,4,5], Mozziyar Etemadi[6], Florencia Garcia-Vicente[6], Fiona J. Gilbert[7], Mark Halling-Brown[8], Demis Hassabis[2], Sunny Jansen[9], Alan Karthikesalingam[10], Christopher J. Kelly[10], Dominic King[10], Joseph R. Ledsam[2], David Melnick[6], Hormuz Mostofi[1], Lily Peng[1], Joshua Jay Reicher[11], Bernardino Romera-Paredes[2], Richard Sidebottom[12,13], Mustafa Suleyman[2], Daniel Tse[1]*, Kenneth C. Young[8], Jeffrey De Fauw[2,15] & Shravya Shetty[1,15]*

Code availability

The code used for training the models has a large number of dependencies on internal tooling, infrastructure and hardware, and its release is therefore not feasible. However, all experiments and implementation details are described in sufficient detail in the Supplementary Methods section to support replication with non-proprietary libraries. Several major components of our work are available in open source repositories: Tensorflow (https://www.tensorflow.org); Tensorflow Object Detection API (https://github.com/tensorflow/models/tree/master/research/object_detection).

Screening mammography aims to identify breast cancer at earlier stages of the disease, when treatment can be more successful[1]. Despite the existence of screening programmes worldwide, the interpretation of mammograms is affected by high rates of false positives and false negatives[2]. Here we present an artificial intelligence (AI) system that is capable of surpassing human experts in breast cancer prediction. To assess its performance in the clinical setting, we curated a large representative dataset from the UK and a large enriched dataset from the USA. We show an absolute reduction of 5.7% and 1.2% (USA and UK) in false positives and 9.4% and 2.7% in false negatives. We provide evidence of the ability of the system to generalize from the UK to the USA. In an independent study of six radiologists, the AI system outperformed

Matters arising

Transparency and reproducibility in artificial intelligence

https://doi.org/10.1038/s41586-020-2766-y

Received: 1 February 2020

Accepted: 10 August 2020

Published online: 14 October 2020

Benjamin Haibe-Kains[1,2,3,4,5]✉, George Alexandru Adam[3,5], Ahmed Hosny[6,7], Farnoosh Khodakarami[1,2], Massive Analysis Quality Control (MAQC) Society Board of Directors*, Levi Waldron[8], Bo Wang[2,3,5,9,10], Chris McIntosh[2,5,9], Anna Goldenberg[3,5,11,12], Anshul Kundaje[13,14], Casey S. Greene[15,16], Tamara Broderick[17], Michael M. Hoffman[1,2,3,5], Jeffrey T. Leek[18], Keegan Korthauer[19,20], Wolfgang Huber[21], Alvis Brazma[22], Joelle Pineau[23,24], Robert Tibshirani[25,26], Trevor Hastie[25,26], John P. A. Ioannidis[25,26,27,28,29], John Quackenbush[30,31,32] & Hugo J. W. L. Aerts[6,7,33,34]

ARISING FROM S. M. McKinney et al. *Nature* https://doi.org/10.1038/s41586-019-1799-6 (2020)

Table 1 | Essential hyperparameters for reproducing the study for each of the three models

	Lesion	Breast	Case
Learning rate	Missing	0.0001	Missing
Learning rate schedule	Missing	Stated	Missing
Optimizer	Stochastic gradient descent with momentum	Adam	Missing
Momentum	Missing	Not applicable	Not applicable
Batch size	4	Unclear	2
Epochs	Missing	120,000	Missing

…ce (AI) hold enormous potential …d go even beyond human perfor-…[1] showed the high potential of AI … the lack of details of the methods … scientific value. Here, we iden-…nt and reproducible AI research …vide solutions to these obstacles with implications for the broader field.

The work by McKinney et al.[1] demonstrates the potential of AI in medical imaging, while highlighting the challenges of making such reporting-standards). Publication of insufficiently documented research does not meet the core requirements underlying scientific discovery[2,3]. Merely textual descriptions of deep-learning models can hide their high level of complexity. Nuances in the computer code may have marked effects on the training and evaluation of results[4], potentially leading to unintended consequences[5]. Therefore, transparency in the form of the actual computer code used to train a model and arrive at its final set of parameters is essential for research reproducibility. McKinney et al.[1] stated that the code used for training the models has "a large number of dependencies on internal tooling, infrastructure

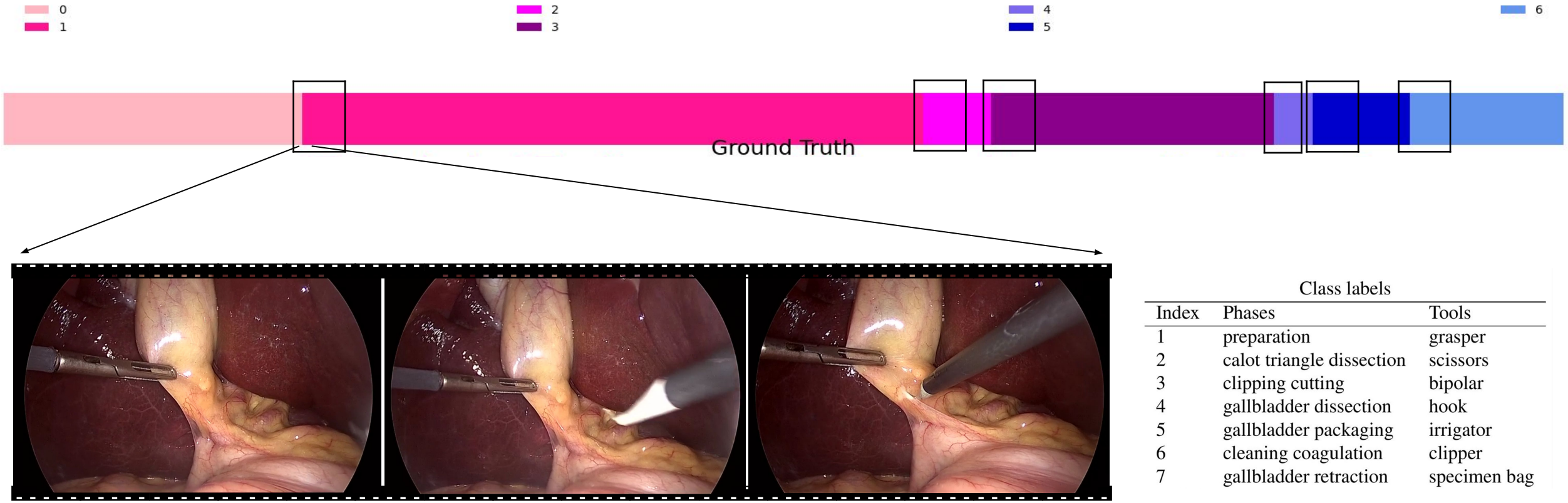
Example Two

Reproducibility of AI-related research

KING'S College LONDON

Surgical Phase Recognition

Cholec-80 dataset



Class labels

Index	Phases	Tools
1	preparation	grasper
2	calot triangle dissection	scissors
3	clipping cutting	bipolar
4	gallbladder dissection	hook
5	gallbladder packaging	irrigator
6	cleaning coagulation	clipper
7	gallbladder retraction	specimen bag

KING'S College LONDON

Surgical Phase Recognition SOTA (2021)

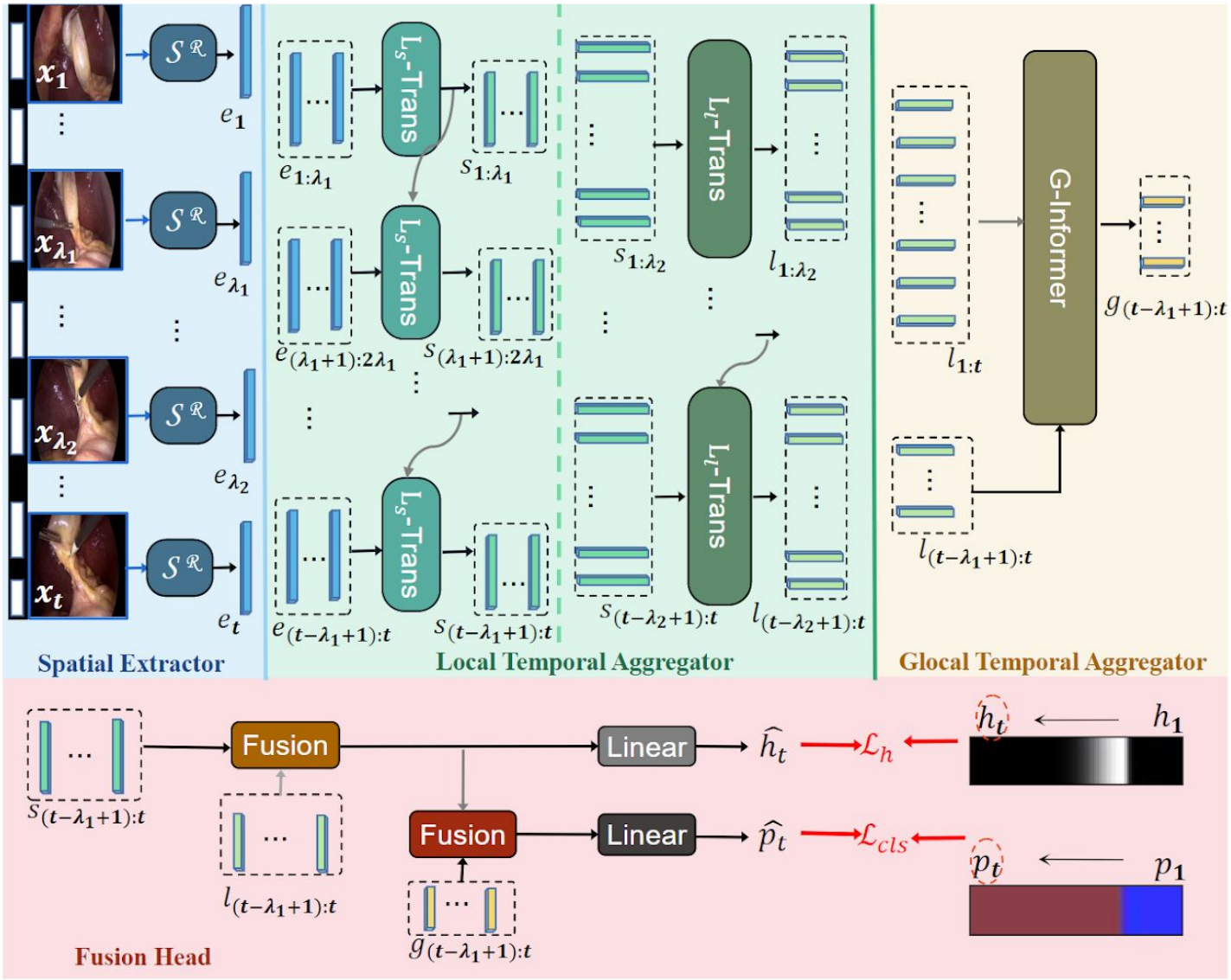
Method	Cholec80				M2CAI16				#param
	Accuracy	Precision	Recall	Jaccard	Accuracy	Precision	Recall	Jaccard	
EndoNet* [26]	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9	—	—	—	—	—	58.3M
EndoNet+LSTM* [27]	88.6 ± 9.6	84.4 ± 7.9	84.7 ± 7.9	—	—	—	—	—	68.8M
MTRCNet-CL* [14]	89.2 ± 7.6	86.9 ± 4.3	88.0 ± 6.9	—	—	—	—	—	29.0M
PhaseNet [24,26]	78.8 ± 4.7	71.3 ± 15.6	76.6 ± 16.6	—	79.5 ± 12.1	—	—	64.1 ± 10.3	58.3M
SV-RCNet [13]	85.3 ± 7.3	80.7 ± 7.0	83.5 ± 7.5	—	81.7 ± 8.1	81.0 ± 8.3	81.6 ± 7.2	65.4 ± 8.9	28.8M
OHFM [30]	87.3 ± 5.7	—	—	67.0 ± 13.3	85.2 ± 7.5	—	—	68.8 ± 10.5	47.1M
TeCNO [4]	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9	86.1 ± 10.0	85.7 ± 7.7	**88.9 ± 4.5**	74.4 ± 7.2	24.7M
Trans-SVNet (ours)	**90.3 ± 7.1**	**90.7 ± 5.0**	**88.8 ± 7.4**	**79.3 ± 6.6**	**87.2 ± 9.3**	**88.0 ± 6.7**	87.5 ± 5.5	**74.7 ± 7.7**	24.7M

Evaluation Metrics. We employ four frequently-used metrics in surgical phase recognition for comprehensive comparisons. These measurements are accuracy (AC), precision (PR), recall (RE), and Jaccard index (JA), which are also utilized in [13,30]. The AC is calculated at the video level, defined as the percentage of frames correctly recognized in the entire video. Since the video classes are imbalanced, the PR, RE, and JA are first computed towards each phase and then averaged over all the phases. We also count the number of parameters to indicate the training and inference speed to a certain degree.

Gao. *Trans-SVNet: Accurate Phase Recognition from Surgical Videos via Hybrid Embedding Aggregation Transformer*. **MICCAI** (2021)

Surgical Phase Recognition

LoViT: Long Video Transformer for Surgical Phase Recognition



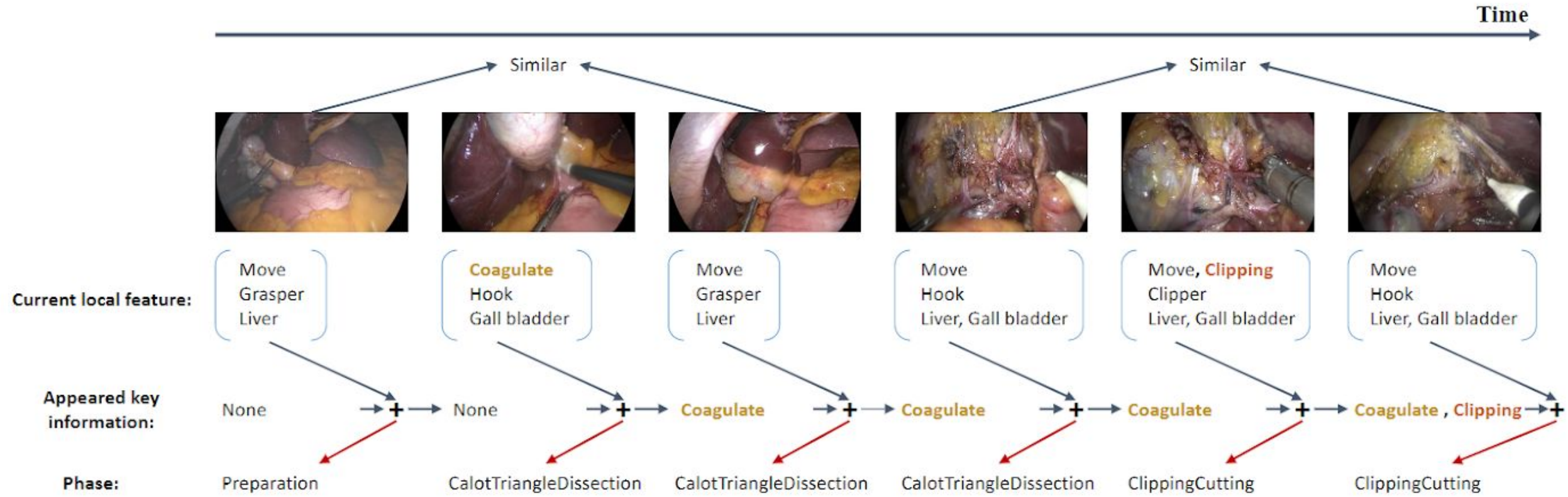
Contributions

1. temporally-rich spatial feature extractor
2. Multiscale temporal feature aggregation
3. Phase transition-aware supervision

Liu. *LoViT: Long Video Transformer for Surgical Phase Recognition.* (under review)

Surgical Phase Recognition

SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition

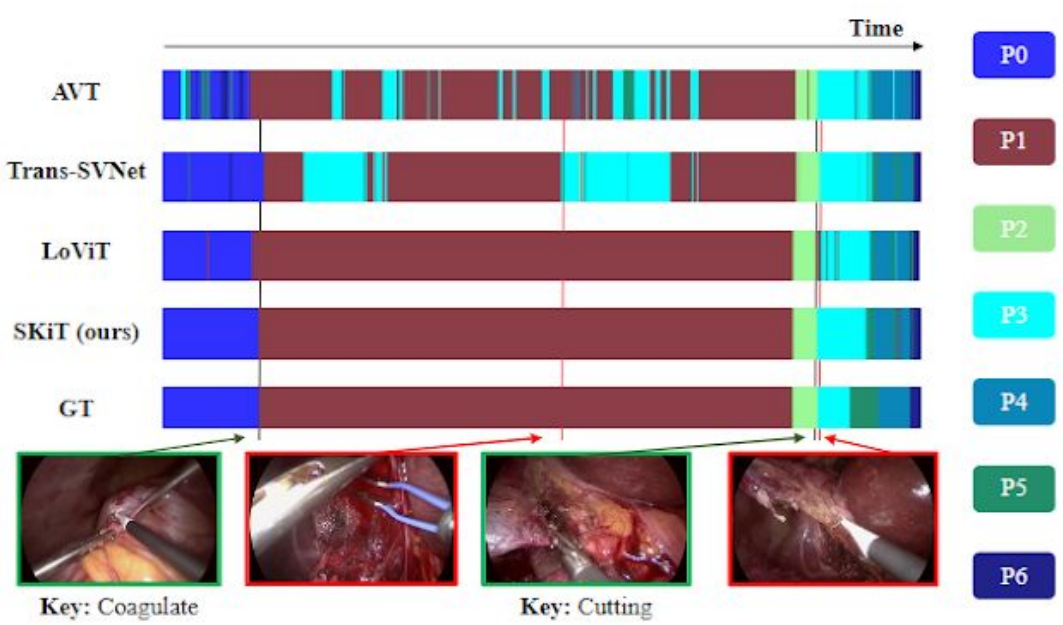


Liu. *SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition*. (under review)

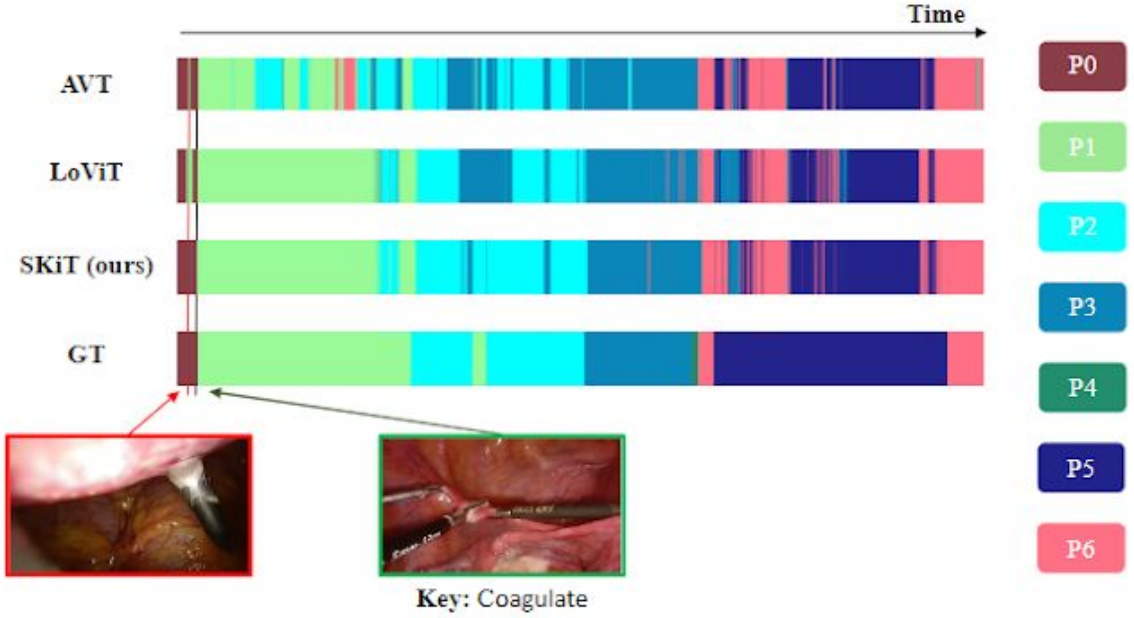
KING'S College LONDON

Surgical Phase Recognition

Cholec-80 and AutoLaparo



(a) Phase Recognition predictions on Cholec80.



(b) Phase Recognition predictions on AutoLaparo.

Liu. *LoViT: Long Video Transformer for Surgical Phase Recognition*. (under review)

Liu. *SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition*. (under review)

KING'S College LONDON

Surgical Phase Recognition

Cholec-80 and AutoLaparo

Dataset	Method	Video-level Metric	Phase-level Metric		
		Accuracy	Precision	Recall	Jaccard
Cholec80	EndoNet [36]*	81.7 ± 4.2	73.7	79.6	-
	MTRCNet-CL [21]*	89.2 ± 7.6	**86.9**	88.0	-
	PhaseNet [35]	78.8 ± 4.7	71.3	76.6	-
	SV-RCNet [20]	85.3 ± 7.3	80.7	83.5	-
	OHFM [40]	87.3 ± 5.7	-	-	67.0
	TeCNO [5]	88.56	81.64	85.24	-
	Trans-SVNet [14]	89.11 ± 7.03	84.72	83.63	72.50
	AVT [16]	86.73 ± 7.62	77.34	82.13	66.42
	LoViT [2]	91.50 ± 6.10	83.07	86.5	74.15
	SKiT (ours)	$\mathbf{92.46 \pm 5.10}$	84.59	**88.52**	**76.74**
AutoLaparo	SV-RCNet	75.62	64.02	59.70	47.15
	TMRNet [22]	78.20	66.02	61.47	49.59
	TeCNO	77.27	66.92	64.60	50.67
	Trans-SVNet	78.29	64.21	62.11	50.65
	AVT	77.81 ± 9.38	68.04	62.23	50.66
	LoViT	81.43 ± 7.35	**85.07**	65.85	55.90
	SKiT (ours)	$\mathbf{82.93 \pm 6.75}$	81.78	**70.12**	**59.86**

Liu. *LoViT: Long Video Transformer for Surgical Phase Recognition*. (under review)

Liu. *SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition*. (under review)

KING'S College LONDON

Surgical Phase Recognition

Cholec-80 and AutoLaparo

Dataset	Method	Video-level Metric	Phase-level Metric		
		Accuracy ↑	Precision ↑	Recall ↑	Jaccard ↑
Cholec80	EndoNet [14]*+	81.7 ± 4.2	73.7 ± 16.1	79.6 ± 7.9	-
	MTRCNet-CL [18]*+	89.2 ± 7.6	86.9 ± 4.3	88.0 ± 6.9	-
	PhaseNet [39]+	78.8 ± 4.7	71.3 ± 15.6	76.6 ± 16.6	-
	SV-RCNet [12]+	85.3 ± 7.3	80.7 ± 7.0	83.5 ± 7.5	-
	OHFM [20]+	87.3 ± 5.7	-	-	67.0 ± 13.3
	TeCNO [22]+	88.6 ± 7.8	86.5 ± 7.0	87.6 ± 6.7	75.1 ± 6.9
	TMRNet [23]+	90.1 ± 7.6	90.3 ± 3.3	89.5 ± 5.0	79.1 ± 5.7
	Trans-SVNet [13]+	90.3 ± 7.1	**90.7 ± 5.0**	88.8 ± 7.4	79.3 ± 6.6
	LoViT (ours)+	**92.40 ± 6.3**	89.9 ± 6.1	**90.6 ± 4.4**	**81.2 ± 9.1**
	Trans-SVNet	89.1 ± 7.0	**84.7**	83.6	72.5
	AVT [31]	86.7 ± 7.6	77.3	82.1	66.4
	LoViT (ours)	**91.5 ± 6.1**	83.1	**86.5**	**74.2**
AutoLaparo	SV-RCNet	75.6	64.0	59.7	47.2
	TMRNet	78.2	66.0	61.5	49.6
	TeCNO	77.3	66.9	64.6	50.7
	Trans-SVNet	78.3	64.2	62.1	50.7
	AVT	77.8 ± 9.4	68.0	62.2	50.7
	LoViT (ours)	**81.4 ± 7.6**	**85.1**	**65.9**	**55.9**

Liu. *LoViT: Long Video Transformer for Surgical Phase Recognition*. (under review)

Liu. *SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition*. (under review)

Big players also facing these problems

Reproducibility of AI-related research

KING'S College LONDON

Big players also facing these problems

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Sutskever[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

The game of Go has long been viewed as the most challenging of classic games for artificial intelligence owing to its enormous search space and the difficulty of evaluating board positions and moves. Here we introduce a new approach to computer Go that uses 'value networks' to evaluate board positions and 'policy networks' to select moves. These deep neural networks are trained by a n learning from games of self-play of-the-art Monte Carlo tree searc new search algorithm that combin our program AlphaGo achieved a champion by 5 games to 0. This is full-sized game of Go, a feat previ

ELF OpenGo: An Analysis and Open Reimplementation of AlphaZero

Yuandong Tian [1] Jerry Ma [*1] Qucheng Gong [*1] Shubho Sengupta [*1] Zhuoyuan Chen [1] James Pinkerton [1] C. Lawrence Zitnick [1]

However, these advances in playing ability come at significant computational expense. A single training run requires millions of selfplay games and days of training on thousands of TPUs, which is an unattainable level of compute for the majority of the research community. When combined with the unavailability of code and models, the result is that the approach is very difficult, if not impossible, to reproduce, study, improve upon, and extend.

Towards Open, Reproducible and Collaborative Surgical Data Science

Reproducibility Crisis

Challenges to Reproducibility in Science

- confounding, multiple hypothesis testing, randomness, incomplete documentation, restricted access to data and code [Beam2020]
- data access is especially germane for medicine due to privacy barriers
- human clinical trials seem to be less at risk [Collins2014]
- openness with data supported but not mandated by publishers
- requires the interplay of skills that are necessary but not currently valued
- research success is still measured in terms of impact factor, research outputs, and grant income [UKHouseOfCommons2023]
- uncertainty in the academic job market [UKHouseOfCommons2023]
- non-replicable publications tend to be cited more often than replicable ones [Sierra-Garcia2021]

KING'S College LONDON

Challenges to Reproducibility posed by AI/ML Models

- the quality of published reports is lacking transparency, clear reporting to facilitate replicability, exploration for potential ethical concerns, and clear demonstrations of effectiveness [Vollmer2020]
- releasing open-source software often involves a mountain of unforeseen work for developers
- models typically consist of enormous number of parameters with default values that may differ between libraries and versions
- experiments are often dependent on random initialisation for iterative algorithms and data loaders [TheTuringWay2021]
- the cost to reproduce a state-of-the-art deep learning model from the beginning can be immense [Beam2020]
- neural architecture search: $1-3.2 million with 626,155 lbs of CO_2 emissions [Strubell2019]

Open, Reproducible and Collaborative

- community-driven guide to reproducible, ethical, inclusive and collaborative data science
- resources for reproducible research, project design, communication, collaboration, and ethical research

The Turing Way Community. *The Turing Way: A handbook for reproducible, ethical, and collaborative research*. **Zenodo** (2021)

Prospects for Increasing Reproducibility of AI/ML Studies

- large models as a commodity for productive industry-academic partnerships [Beam2020]
- a "wall-garden" approach to cope with privacy concerns
- reporting guidelines for clinical trials adapted to AI: TRIPOD-AI [Collins2021], CONSORT-AI [Liu2020]
- guidelines for scientific data management: FAIR [Wilkinson2016] [UKHouseOfCommons2023]
- interdisciplinary groups applying AI/ML to health would benefit from questions related to transparency, reproducibility, ethics, and effectiveness (TREE) [Vollmer2020]
- avoid methodological pitfalls [Kapoor2022]
- Community Data License Agreement [CDLA-Permissive-2.0]
- reproducible analytical pipelines (RAP) [Gov.uk]

KING'S College LONDON

Cross-disciplinarity challenges in healthcare

- knowledge deficit leads to misconceptions and missed opportunities for valuable and important collaborations
- there remains a substantial gap in the interdisciplinary literature around how cross-disciplinary collaborations are initiated [Priaulx2018]
- despite cross-disciplinary collaborations are (only) taking place in high-income countries, there is little evidence on how to design and implement them
- global healthcare challenges are conceptualised in varied ways with poor coordination due to problem definition and positioning [Ding2020]

KING'S College LONDON

Surgical Data Study Groups

Inspired by the Data Study Groups at the Alan Turing Institute and the monthly presentations at KHP Surgical Challenges to **make great leaps in data science and artificial intelligence** to **change surgery for better** and to **provide opportunities to future leaders** in the field to **experiment with real-world data** to **solve** the **surgical challenges** we see today and better be prepared to face those we will see tomorrow.

Towards Open, Reproducible and Collaborative Surgical Data Science

Conclusions

KING'S College LONDON

Conclusions

- surgery is a global challenge
- reproducibility crisis
- steps towards reproducibility can be the result of our own practice
- collaboratively propose automated benchmarks for validation
- The Turing Way: community-driven resources
- reporting guidelines for clinical trials adapted to AI (TRIPOD-AI, CONSORT-AI)
- FAIR principles for data management
- surgical data study groups for openly and collaboratively solve real-world problems

KING'S College LONDON

Surgical Data Science

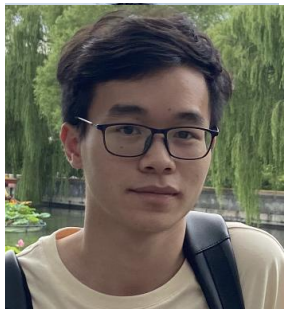
Team

Maxence Boels, *PhD*
Surgical Phase Recognition

Jingjing Peng, *PhD*
Brain Shift Prediction

Harry Robertshaw, *PhD*
Autonomous Navigation

Hassna Irzan, *Post-doc*
Multimodal AI



Yang Liu, *PhD*
Surgical Phase Recognition

Julien Quarez, *PhD*
Causal AI

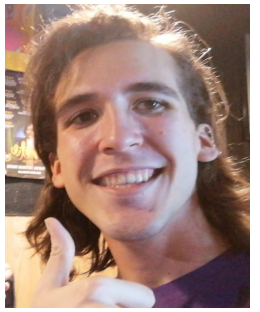
Basma Alabdullah, *PhD*
Soft Robotics

Mahrukh Saeed, *PGT*
Generative Models

Ayesha Harsh, *UG*
Sentiment Analysis

Kenaan Sarhan, *UG*
Surgical Complications

Alexis Leclercq, *UG*
Image Segmentation



Carlos Ramirez, *PGT*
Physics AI

Towards Open, Reproducible and Collaborative Surgical Data Science

Open-Source Software for Surgical Technologies Workshop

Alejandro Granados
Lecturer in Surgical Data Science
Surgical & Interventional Engineering

alejandro.granados@kcl.ac.uk