Towards Video-based Surgical Workflow Understanding in Open Orthopedic Surgery

Abdolrahim Kadkhodamohammadi[a] and Nachappa Sivanesan Uthraraj[c] and Petros Giataganas[a] and Gauthier Gras[a] and Karen Kerr[a] and Imanol Luengo[a] and Sam Oussedik[c] and Danail Stoyanov[a,b]

[a]Digital Surgery, a Medtronic Company, 230 City Road, EC1V 2QY, London, UK; [b]Wellcome/EPSRC Centre for Interventional and Surgical Sciences, University College London, UK;[a]Orthopaedic Surgery Department, University College London Hospitals, London, UK

ARTICLE HISTORY
Compiled January 7, 2021

ABSTRACT
Safe and efficient surgical training and workflow management play a critical role in clinical competency and ultimately, patient outcomes. Video data in minimally invasive surgery (MIS) has enabled opportunities for vision-based artificial intelligence (AI) systems to improve surgical skills training and assurance through post-operative video analysis and development of real-time computer assisted interventions (CAI). Despite the availability of mounted cameras for the operating room (OR), similar capabilities are much more complex to develop for recording open surgery procedures, which has resulted in a shortage of exemplar video-based training materials. In this paper, we present a potential solution to record open surgical procedures using head-mounted cameras. Recorded videos were anonymized to remove patient and staff identifiable information using a machine learning algorithm that achieves state-of-the-art results on the OR Face dataset. We then propose a CNN-LSTM-based model to automatically segment videos into different surgical phases, which has never been previously demonstrated in open procedures. The redacted videos, along with the automatically predicted phases, are then available for surgeons and their teams for post-operative review and analysis. To our knowledge, this is the first demonstration of the feasibility of deploying camera recording systems and developing machine learning-based workflow analysis solutions for open surgery, particularly in orthopedics.

KEYWORDS
Surgical workflow analysis; surgical data science; open surgery; orthopedics; Machine Learning

1. Introduction

Despite the critical nature of surgery, digital technology-assisted solutions for the modern operating room (OR) are limited (16). Research in computer-assisted interventions (CAI) has focused on the development of data-driven computational approaches to develop intelligent systems to assist the surgical team. Recognizing surgical workflows is one of the fundamental building blocks for such systems as it enables further understanding of surgical context, partitioning complex procedures into well-defined surgical steps and the identification of anomalies and best practices, hence, progress-

ing the field towards standardization of surgical practice (16; 17; 23). The majority of this effort has, however, been concentrated on minimally invasive surgery (MIS) or endovascular surgery, where a video of the surgical procedure, is inherently present. Yet some of the largest volume of surgical procedures are still performed open, for example, in specialties such as orthopedics.

Collecting data from OR equipment is often difficult or expensive as it requires either additional hardware or manual intervention (2). Most minimally invasive surgical workflow analysis approaches rely solely on intraoperative video data, as recordings can be collected directly from the video feed (8; 14; 17; 20). This has led to the availability of public datasets such as Cholec80 (21) and EndoVis (6). On the other hand, less attention has been given to understanding open surgery processes, even though it is much more complex and would therefore benefit from improved reporting and standardization. We believe this is due to two main reasons:

- contrary to MIS, capturing open surgery videos require introduction of a camera recording system to allow recording all or at least the most important part of a procedure without disrupting the surgical team;
- regulation and process requirements around data access and privacy are more complex for open surgery procedures (10).

With the rapid evolution of camera technologies, it is technologically possible for wearable devices to be used in open surgery. For example, light-weight, battery powered head-mounted cameras worn by surgeons as shown in Figure 1. The head-mounted camera makes it possible to capture the surgeon's view point of the intervention and interaction with tissues and organs, which are important for surgical workflow recognition. This is in contrast to other studies (13; 22) that rely on ceiling-mounted cameras to capture medical team activity and movement in the OR, which can have a limited view of the surgical procedure due to multiple sources of occlusion. Importantly, open surgery videos inherently contain sensitive information such as the identity of the patient or members of the surgical team by recording faces[1]. To permit utilization of such sensitive data, identifiable information must be redacted from the videos. Solutions to anonymize OR video include reducing image resolutions severely to preserve privacy (19), however, even low resolution faces can still be recognized when video sequences are available (5). In the work presented in this paper, we localized faces using a deep Convolutional Neural Network (CNN) model. The localized faces are then blurred out similarly to previous efforts in anonymization (7). We also propose a model to segment videos into surgical phases motivated by recent phase detection models, which is based on a two-stage approach. We first used *SENet154* (11) as an encoder to extract features and a Long Short-Term Memory (LSTM) to recognize surgical steps.

The clinical focus of this paper is on orthopedic surgery, specifically total knee replacement (TKR) surgery. The procedure was selected due to its high volume, which is increasing yearly with an aging population (3). Given its high volume, the overall procedure is relatively standardized with little surgical variations and a low incidence of complications (13-year risk revision is about 4% for median age of 69 years old (1; 3)). The procedure is separated in a high number of instruments and surgical steps. It is therefore important for the medical team to reliably recognize the current phase of a procedure to be able to plan and act accordingly. In this paper, we relied on videos captured using head-mounted cameras to demonstrate the feasibility of using a machine learning model to both preserve privacy and recognize surgical phases: a step

[1]It is worth mentioning that as the patient is draped, the likelihood of capturing the patient's face is minimal.

Figure 1. Camera recording. Left image shows the head-mounted camera worn by two members of the clinical team. A screenshot taken from the video recorded by surgeon's camera is presented in the right image.

towards developing context-aware CAI systems.

2. Method

A deep learning based solution was used to blur any patients and clinical team faces, which may have been captured by the camera, to preserve their identities and to enable us to perform subsequent surgical workflow analysis on the recorded video.

2.1. OR Face Detection

Video face detection has advanced dramatically by the availability of large datasets, like *FDDB* (12) and *WIDER FACE* (24). We adopted the Dual Shot Face Detector (DSFD) (15) and fine-tuned the model on the OR Face dataset presented in (7). This allowed the model to adapt to be able to recognize faces in the OR, which is a non-trivial task as surgical team members are required to wear surgical masks and caps in such an environment, making standard face detection models fail. We formulated the problem of face detection as bounding box regression. To this end, we used the dual shot face detector model, which relied on two parallel streams to extract features and fed it to bounding box regressors. This model utilized a two-stream network for extracting more robust features along with auxiliary supervision at different layers. This enabled the model to construct discriminative representations and a robust predictor. VGG16 (18) was used as the backbone of the architecture.

2.2. Surgical Workflow Analysis

Surgical phase recognition is the process of segmenting a surgical procedure into different parts where the surgeon completes a task before moving to the next phase. We present a two stage model to recognize surgical phases using video data. In the first stage, we proposed to use SENet154 (11) to build a rich representation for the task of

phase recognition. This is achieved by training the network for image classification to recognize, from a single frame, to which part of the operation that frame belongs. A single image, however, does not carry enough information to accurately detect surgical objectives due to the ambiguity of surgical instruments and anatomical landmarks that could be present in multiple similar phases. We therefore proposed an LSTM network for the second stage to incorporate temporal information to the model. More specifically, we used a two layer Bidirectional LSTM network (9) to take into account information from both past and future. This model was also trained for the same objective but unlike the previous step, trained jointly with temporal windows adds long-term context, allowing the model to better understand the temporal relation between surgical phases. This also enables the network to model temporal dependencies and leverage time series data to produce temporally consistent predictions.

3. Data

Eighteen TKR procedures were recorded for which both the patients and the clinical team consented to be filmed and for the associated video data to be used for research and education purposes. Note that no patient identifiable metadata was collected during this process. Videos were uploaded to our online platform that is General Data Protection Regulation of 2018 (GDPR) compliant, and System and Organization Controls (SOC2), Health Insurance Portability and Accountability Act of 1996 (HIPAA), NHS Data Security and Protection Toolkit (DSPT) and Cyber Essentials certified.

The procedures were 75 to 130 minutes long, and performed in one OR. The procedures were recorded over the span of four months and membership of the surgical team (with the exception of the lead surgeon) varies across the procedures. The procedures were divided into 18 surgical phases, which are presented in Table 1.

A medical liaison officer specialized in orthopedic surgery annotated all 18 surgical videos with the above phases. A quality assurance check was completed by an orthopedic surgeon. In our dataset, the standard deviation of the duration of phase 16 is 10 minutes, while for all other phases excluding phases 1 and 18 it is below 3 minutes. This indicates that achieving the objective of this phase involves more complicated tasks, resulting in high variation between patients. The higher standard deviation in phases 1 and 18 is mainly because of the delay in moving patient to and off the operating table, and putting on and removing drapes.

4. Experiments

In this section, we present the evaluation results of the model for both face detection and surgical phase recognition.

4.1. OR Face Detection

The DSFD model is among the top performing models on many challenging computer vision datasets like WIDDER Face, which includes faces at different scales, pose and levels of occlusions. The OR Face dataset (7) was used to assess the performance of DSFD on faces captured in the OR. The OR face dataset was generated from 15 surgical videos recorded in the OR using ceiling- or wall-mounted cameras. Table 2 presents the performance results following the same experimental setup as in FaceOff (7). We

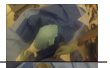
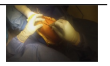
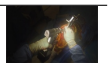
ID	Surgical Phase	Description	Duration Avg±STD	Screen shot
1	Patient Preparation	prepare the patient	280±200	
2	Mark Incision	outline incision path	732±91	
3	Stryker OrthoMap Setup & Pin Place	setup knee software	212±7	
4	Incision & Exposure	skin incision and superficial tissue dissection	290±88	
5	Medial Parapatellar Arthrotomy	access to the joint	157±61	
6	Patella Preparation and Trial Insertion	complete preparation of patella for trialing	66±34	
7	OrthoMap Sync. & Bone Registration	synchronize bone with software to create a map of the bone in software to guide bone cuts	229±125	
8	Distal Femoral Alignment & Resection	bone alignment and resection of the distal femoral surface	169±76	
9	Femoral Rotation Alignment	setting rotation for the AP chamfer block	64±35	
10	Femoral AP and Chamfer Cuts	remaining cuts of the distal femur	264±91	
11	Femoral Notch Cut	final femur cut for placement of trials and implants	288±75	
12	Proximal Tibial Alignment & Resection	bone alignment and resection of the tibial joint surface	253±73	
13	ACL/PCL Excision, Excess Soft Tissue and Bone Removal	to clear any excess soft tissue or bone to prevent trial/implant impingement/bad placement	498±130	
14	Trial Implant Insertion, ROM Assessment, & Soft Tissue Balancing	trial placement and assessment of function and gap balancing	307±138	
15	Tibial Reaming & Implantation Preparation	makes space into the tibia for the tibial implant	216±49	
16	Cementing & Implant Insertion	final and proper placement of implants	763±552	
17	Final Assessment with Tibial Trial Insert	final assessment of the knee with final implants except for tibial trial insert	463±71	
18	Tibial Insert, Wound Closure, & Pin Removal	final tibial insert and final reduction and wound closure	364±255	

Table 1. TKR surgical phases, description, and average and STD duration in seconds.

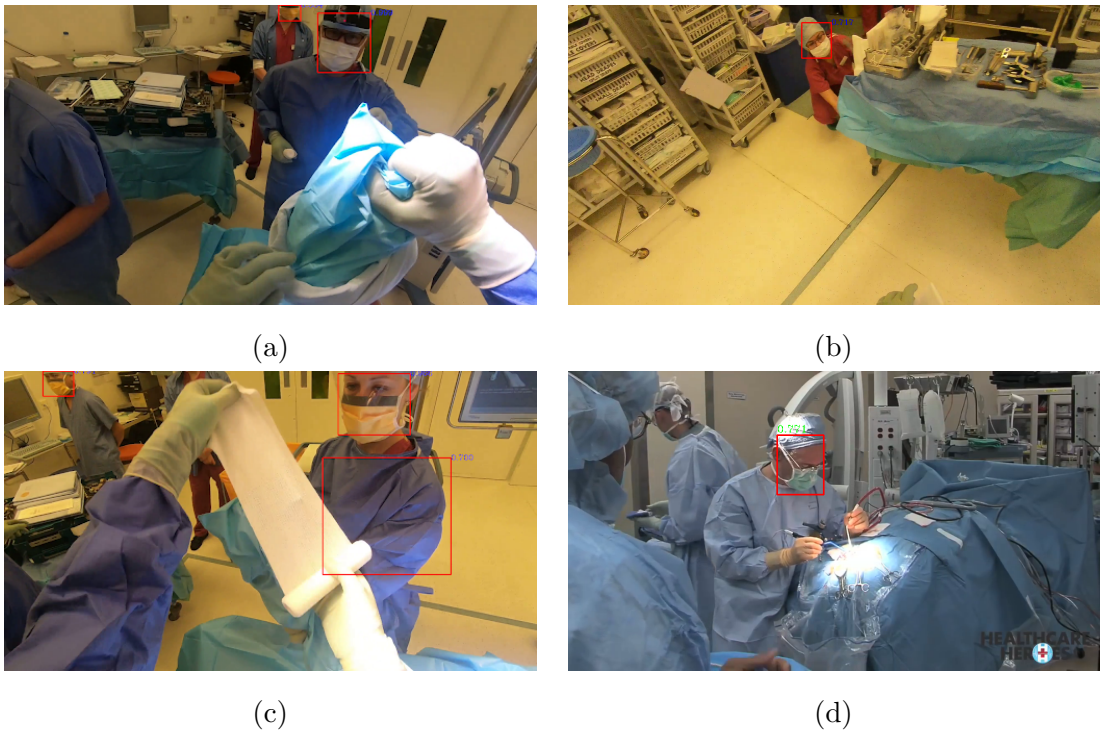


Figure 2. Sample face detection: (a-c), images from head-mounted camera recordings during total knee replacement, (d) an image from OR Face. False positive face detection is shown in (c) and failure to identify faces in (d).

evaluated the DSFD model trained on the WIDDER FACE dataset, presented in the second row of Table 2. The model outperforms FaceOff, which uses Faster R-CNN as a detector, and fine-tuning on OR Face improved the performance even more (the second row Table 2). Qualitative results are shown in Figure 2. Despite the significant viewpoint change because of the head-mounted camera, the model performed well and with only a few false positives identified, i.e. incorrect face detection on the background (see Figure 2). Missed detection was only identified in cases where the nose and eyes were not visible (see Figure 2). We should highlight that missing such faces does not jeopardize any privacy guidance.

Model	Precision	Recall	F1 Score
FaceOff (OR Face)	59.07%	93.46%	72.39%
DSFD (WIDER FACE)	88.89%	93.18%	90.98%
DSFD (finetune on OR Face)	88.04%	97.54%	92.55%

Table 2. Face detection results on OR Face. DSFD is evaluated and compared with FaceOFF on the same experimental setup.

4.2. Surgical Phase Recognition

Our SENet encoder is initialized using ImageNet pre-trained weights (4) and the number of hidden units for each layer in LSTM is set to 2048. We use our TKR phase dataset to perform 6-fold cross-validation. In each fold, we use 15 videos in the training set, one video in the validation set and two videos in the test set.

The average F1 score of predicting different phases is 91.06% with a standard deviation of 1.72%. We noticed that training the model from scratch (without pretrained weights) can lead to around 4% drop in performance due to the small sample of videos in this dataset. We also computed the average error in seconds between the ground truth phase transitions and first time a phase was predicted as another metric. The model achieved an average error of 15 ± 2.5 seconds, where the average error per fold varies between 12 to 18 seconds. In other words, if we change to a phase upon the first detection of a phase, the prediction was off by 15 seconds on average. Figure 3 shows ground truth and predicted phases using staircase plots for two test videos. For a better visibility, we shifted the staircase plots for the predicted phases upward. The vertical axis indicates the phase id and the horizontal axis the procedure progression time in seconds.

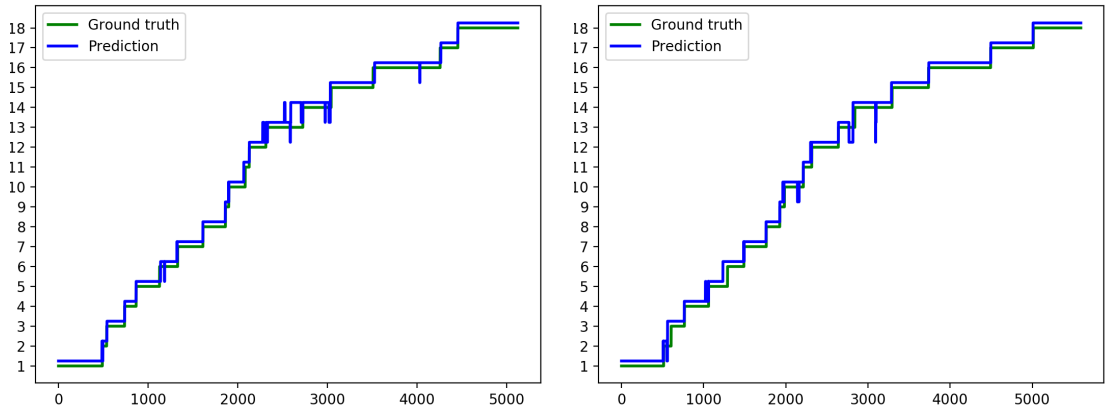


Figure 3. Surgical phase prediction for two test videos across different folds. The vertical and horizontal axes indicate phase id and the surgery progression in seconds, respectively. We shift the prediction a bit upward to allow better compression between ground truths and predictions.

The confusion matrix for the two test video in Figure 3 has been reported in Figure 4. The numbers on the vertical and horizontal axis indicate phase IDs. Most of the phase has been classified correctly and miss classification is happened among consecutive phases.

5. Discussion

The face detection results in Table 2 and Figure 2 indicate that the CNN-based model reliably detects and localizes faces in OR environments. Due to the critical importance of this step to comply with GDPR and preserve identities, we needed to generate a fully annotated dataset from recordings under similar viewpoint. Our recordings were filmed using head-mounted cameras, which reduced the likelihood of capturing faces. One of the major limitation is the low number of videos captured to date. Therefore,

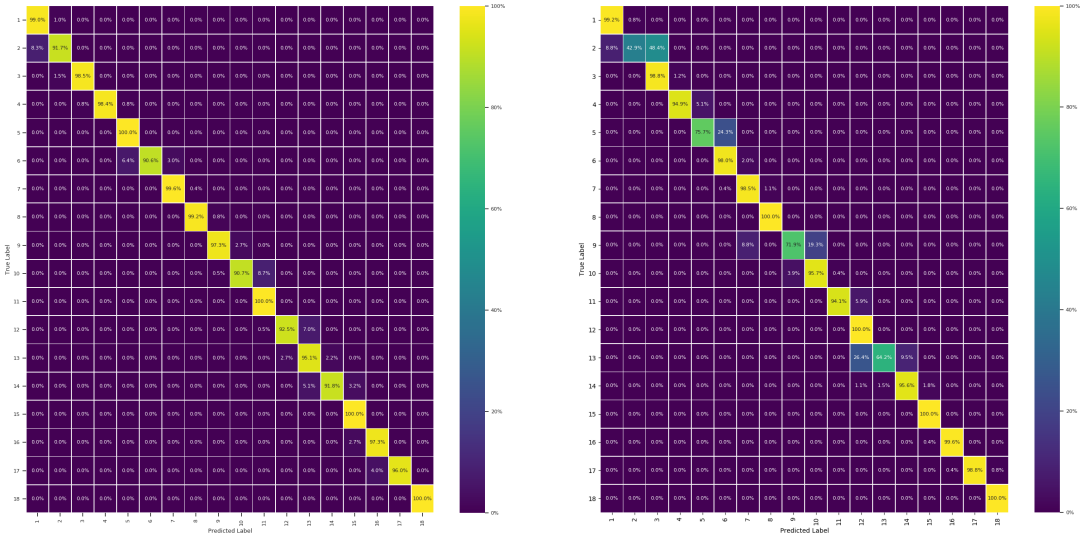


Figure 4. Confusion Matrix. The numbers along the vertical and horizontal axis indicate the phase IDs as per Table 1.

to improve the model, we are planning to collect a diverse dataset and include other objects like screens and whiteboards, which could also potentially reveal sensitive information.

The phase recognition model accurately segments videos into different surgical phases. Even though a small dataset was available to date, the model achieved an average error of 15 seconds in predicting phase transitions. We will explore collecting larger datasets and further investigating the importance of training model from scratch versus pre-trained weights. It was noticed that the model sometimes confused phases 12, 13 and 14. We believe that this is due to the fact that in all these phases the surgeon might require to remove some soft tissues, which are phase agnostic and have similar appearance in all phases. Another type of error is the early or late transition between phases (Figure 4). We should however note that the average errors of 15 second error in phase transitions are negligible for this type of procedure that could take more than one hour. The availability of more training data and more detailed annotation of key steps for each could potentially result in a more robust model.

These results demonstrate that video data alone is sufficient for developing phase recognition models for open surgery, despite the inherent challenges such as specular highlights, severe motion blurring due to head movement and dynamic background changes. It should be noted that total knee replacement has a rather complex workflow with 18 phases and a wide range of instruments. This work paves the way towards developing context-aware CAI system by detecting the progress of a procedure and providing relevant assistive information to the clinical team for better coordination, education, risk minimization and ultimately better outcomes for patients.

The captured data was also used post-operatively for review of surgical phases, performance monitoring and teaching purposes by the operating surgical team. Figure 5 illustrates our web-based video platform that allows secure access of the surgical process data from any device. Compliant with GDPR and HIPAA certified, obtaining the necessary consent and redacting of the data has played a major role on allowing this data to be available for the surgeon and their team. Automatically detected phase transitions are also shown in the web interface in order to facilitate browsing through

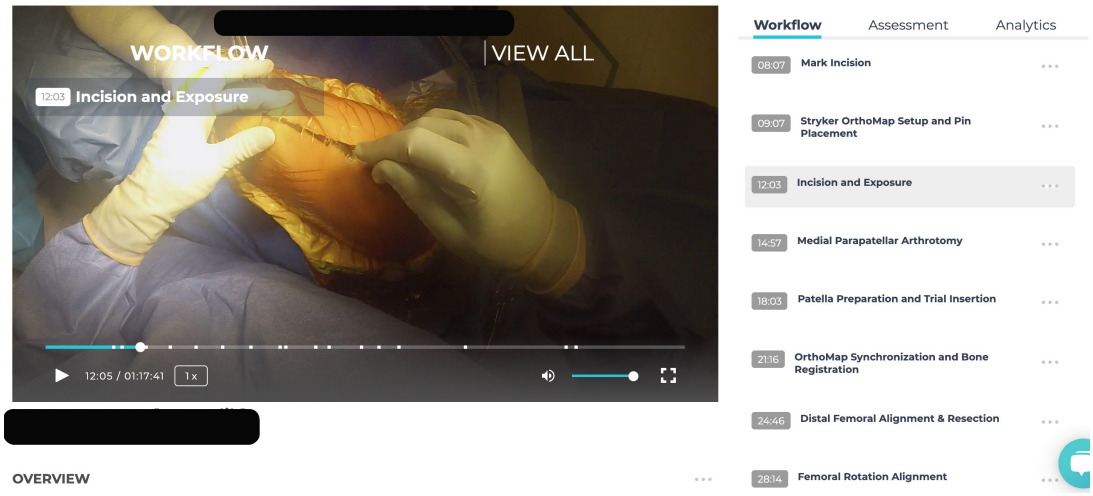


Figure 5. Video platform screenshots. Our web platform shows surgical phases and allows browsing through the videos.

the videos by clicking on them.

6. Conclusion

The ability to video capture open surgery facilitates possibilities for developing machine learning-based approaches to assist surgical teams and improve efficiency and safety within the OR. In this paper, we presented a first demonstration of a solution to record procedures using head-mounted cameras and process this information on a secure manner in a machine learning pipeline. We demonstrated that machine learning models can be used to remove any personal identifiable data in compliance with GDPR and HIPAA. The anonymized surgical video from TKR was then used to perform surgical workflow analysis through automated surgical phase prediction. A CNN-LSTM model was introduced to recognize surgical phases in the open surgery domain, which has previously not been reported. The proposed model predicted phase transitions within an average of 15 seconds from the manually annotated phase transitions. Our results indicate that head-mounted video recordings can be used to perform surgical phase recognition. Our novel web platform allows the surgeon to access the videos and use the phase information for post-operative insights. The platform can be used for postoperative review and training purposes with other team members to better understand surgical process. Future work will extend our model to work in real-time and provide context-aware intraoperative assistive information. We will also explore applying our approach to additional open surgical specialties where currently there is no solution for automated activity recognition and analysis.

References

[1] L.E. Bayliss, D. Culliford, A.P. Monk, S. Glyn-Jones, D. Prieto-Alhambra, A. Judge, C. Cooper, A.J. Carr, N.K. Arden, D.J. Beard, and A.J. Price, *The effect of patient age at intervention on risk of implant revision after total replacement*

of the hip or knee: a population-based cohort study, The Lancet 389 (2017), pp. 1424 – 1430.

[2] S. Bodenstedt, M. Wagner, D. Katić, P. Mietkowski, B. Mayer, H. Kenngott, B. Müller-Stich, R. Dillmann, and S. Speidel, *Unsupervised temporal context learning using convolutional neural networks for laparoscopic workflow analysis*, arXiv preprint arXiv:1702.03684 (2017).

[3] R. Cook, P. Davidson, and R. Martin, *More than 80% of total knee replacements can last for 25 years*, Bmj 367 (2019).

[4] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, *Arcface: Additive angular margin loss for deep face recognition*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[6] EndoVis Challenge, `https://endovis.grand-challenge.org/` (2020), accessed: March 2020.

[7] E. Flouty, O. Zisimopoulos, and D. Stoyanov, *FaceOff: Anonymizing Videos in the Operating Rooms*, in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, Springer, 2018, pp. 30–38.

[8] I. Funke, A. Jenke, S.T. Mees, J. Weitz, S. Speidel, and S. Bodenstedt, *Temporal coherence-based self-supervised learning for laparoscopic workflow analysis*, in *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, Springer, 2018, pp. 85–93.

[9] A. Graves, S. Fernández, and J. Schmidhuber, *Bidirectional LSTM networks for improved phoneme classification and recognition*, in *International Conference on Artificial Neural Networks*, 2005, pp. 799–804.

[10] C.E. Houghton, D. Casey, D. Shaw, and K. Murphy, *Ethical challenges in qualitative research: examples from practice*, Nurse researcher 18 (2010).

[11] J. Hu, L. Shen, and G. Sun, *Squeeze-and-excitation networks*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[12] V. Jain and E. Learned-Miller, *FDDB: A benchmark for face detection in unconstrained settings. university of massachusetts*, Amherst, Tech. Rep. UM-CS-2010-009 2 (2010), p. 8.

[13] A. Kadkhodamohammadi, A. Gangi, M. de Mathelin, and N. Padoy, *Articulated clinician detection using 3d pictorial structures on RGB-D data*, Medical Image Analysis 35 (2017), pp. 215 – 224.

[14] A. Kadkhodamohammadi, I. Luengo, S. Barbarisi, H. Taleb, E. Flouty, and D. Stoyanov, *Feature Aggregation Decoder for Segmenting Laparoscopic Scenes*, in *OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging*, Springer International Publishing, Cham, 2019, pp. 3–11.

[15] J. Li, Y. Wang, C. Wang, Y. Tai, J. Qian, J. Yang, C. Wang, J. Li, and F. Huang, *DSFD: Dual Shot Face Detector*, in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2019, pp. 5060–5069.

[16] L. Maier-Hein, S.S. Vedula, S. Speidel, N. Navab, R. Kikinis, A. Park, M. Eisenmann, H. Feussner, G. Forestier, S. Giannarou, M. Hashizume, D. Katic, H. Kenngott, M. Kranzfelder, A. Malpani, K. März, T. Neumuth, N. Padoy, C. Pugh, N. Schoch, D. Stoyanov, R. Taylor, M. Wagner, G.D. Hager, and P. Jannin, *Surgical*

data science for next-generation interventions, Nature Biomedical Engineering 1 (2017), pp. 691–696.

[17] N. Padoy, *Machine and deep learning for workflow recognition during surgery*, Minimally Invasive Therapy & Allied Technologies 28 (2019), pp. 82–90.

[18] K. Simonyan and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition*, in *International Conference on Learning Representations*, 2015.

[19] V. Srivastav, A. Gangi, and N. Padoy, *Human Pose Estimation on Privacy-Preserving Low-Resolution Depth Images*, in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Springer, 2019, pp. 583–591.

[20] A.P. Twinanda, E.O. Alkan, A. Gangi, M. de Mathelin, and N. Padoy, *Data-driven spatio-temporal rgbd feature encoding for action recognition in operating rooms*, International journal of computer assisted radiology and surgery 10 (2015), pp. 737–747.

[21] A.P. Twinanda, S. Shehata, D. Mutter, J. Marescaux, M. De Mathelin, and N. Padoy, *Endonet: A deep architecture for recognition tasks on laparoscopic videos*, IEEE transactions on medical imaging 36 (2016), pp. 86–97.

[22] A.P. Twinanda, P. Winata, A. Gangi, M. Mathelin, and N. Padoy, *Multi-stream deep architecture for surgical phase recognition on multi-view RGBD videos*, in *Proc. M2CAI Workshop MICCAI*, 2016, pp. 1–8.

[23] T. Vercauteren, M. Unberath, N. Padoy, and N. Navab, *Cai4cai: The rise of contextual artificial intelligence in computer assisted interventions*, Proceedings of the IEEE 108 (2020), pp. 198–214.

[24] S. Yang, P. Luo, C.C. Loy, and X. Tang, *WIDER FACE: A face detection benchmark*, in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 5525–5533.