# CataNet: Predicting remaining cataract surgery duration

Andrés Marafioti<sup>1</sup>, Michel Hayoz<sup>1</sup>, Mathias Gallardo<sup>1</sup>, Pablo Márquez Neila<sup>1</sup>, Sebastian Wolf<sup>2</sup>, Martin Zinkernagel<sup>2</sup> and Raphael Sznitman<sup>1</sup>

 AIMI, ARTORG Center, University of Bern, Switzerland
Department for Ophthalmology, Inselspital, University Hospital, University of Bern, Switzerland
andres.marafioti@artorg.unibe.ch

Abstract. Cataract surgery is a sight saving surgery that is performed over 10 million times each year around the world. With such a large demand, the ability to organize surgical wards and operating rooms efficiently is critical to delivery this therapy in routine clinical care. In this context, estimating the remaining surgical duration (RSD) during procedures is one way to help streamline patient throughput and workflows. To this end, we propose CataNet, a method for cataract surgeries that predicts in real time the RSD jointly with two influential elements: the surgeon's experience, and the current phase of the surgery. We compare CataNet to state-of-the-art RSD estimation methods, showing that it outperforms them even when phase and experience are not considered. We investigate this improvement and show that a significant contributor is the way we integrate the elapsed time into CataNet's feature extractor.

# 1 Introduction

Cataract surgery is one of the most common surgeries in the world, with over 10 million procedures conducted each year. Worldwide, 100 million people suffer from cataract-induced vision impairments and with the aging world population growing, the number of patients at risk of complete blindness is sharply increasing [1]. Yet, even though cataracts can easily be treated, the shear number of surgeries needed poses an organizational challenge of unprecedented scale.

At its core, cataract surgery involves using a surgical microscope to help replace a patient's eye lens, that has become opaque, with a synthetic clear lens. Depending on the risk of the patient [2, 3] and the experience of the operating surgeon [4, 5], the procedure can be performed in under 20 minutes, whereby the majority of delicate surgical phases last 6-15 minutes. In major outpatient cataract clinics, a single surgeon can operate over 50 patients in a given day. As such, the ability to streamline patients and prepare them for surgery plays an important role in surgical workflow and the organization around the operating room. In this context, the ability to appropriately estimate remaining surgical duration (RSD) is imperative to prepare the stream of upcoming patients and doing so as early as possible is critical.

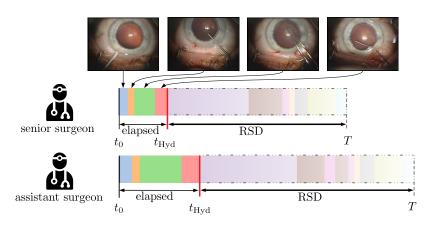


Fig. 1: Surgeon experience and surgical phases play important roles in estimating remaining surgical duration (RSD) in cataract surgery. RSD predictions at  $t_{\rm Hyd}$  allows for optimal operating room patient management.

To date, considerable efforts have been put into designing automated methods to predict RSD [6–12]. Namely, [13] presented the TimeLSTM network, which combined a CNN and an RNN to perform RSD prediction. This method, which achieves good results for cholecystectomy surgeries, pre-trained its CNN for phase recognition thus requiring phase annotations. In an attempt to avoid this requirement, [14] introduced RSDNet which only used unlabeled surgical videos to predict the RSD. Relying on the implicit progress label of the videos, the authors showed that either the surgical phase or progress labels could be effectively utilized for RSD prediction on laparoscopic surgeries. In contrast to laparoscopic procedures however, no RSD methods have focused on cataract surgery. However, there is related research such as that of Neumuth et al. [15], which proposed a surgical workflow management system potentially applicable to RSD estimation. Similarly, [16, 17] detected the current phase in cataract sequences from which RSD could be estimated. Yet these methods overlook important aspects: (1) surgeon experience plays a major factor in cataract surgery duration [4,5] and (2) assessing the risk of the patient by inspecting the initial eve anatomy plays a key role in determining the difficulty and length of the procedure [2,3].

In this work, we thus present a novel approach for online RSD prediction in cataract surgery. Our approach is to explicitly incorporate information from observed surgical phases, the operating surgeon's experience and the elapsed time at any given point to infer RSD prediction. We do this by embedding the video frames with the current elapsed time of the surgery, establishing a multi-task learning problem, and jointly identifying the surgeon's experience and the surgical phase, whereby overcoming a number of important limitations from recent methods (*i.e.*, RSDNet and TimeLSTM). By doing so, our approach avoids in-

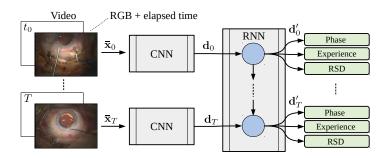


Fig. 2: The end-to-end system. The inputs are video frames concatenated with the elapsed time of the surgery. Inputs are individually fed into the CNN and aggregated by the RNN, the output of which is finally passed through three independent fully connected layers to predict surgical phase, surgeon's experience, and RSD.

troducing additional complexities and yet considerably outperforms competing methods on both average RSD measures and RSD estimates at early stages of the surgery. In addition, we present an ablation study to identify the components of our method that give rise to the performance reported<sup>3</sup>.

# 2 Approach

#### 2.1 Model

Following [5], we identify three key factors that influence the RSD: the surgeon's experience, the current surgical phase, and the elapsed time of the surgery (Fig. 1). For accurate RSD estimation, it is thus critical that the predictive model is aware of these factors when processing the input video. To that end, we incorporate the factors into the model in a number of ways. The elapsed time, readily available at both training and inference time, is appended as an additional channel to the input video frames. On the other hand, surgeon's experience and surgical phase are unknown at inference time. Instead, we train the model to estimate them from the input data.

Fig. 2 depicts our model and how these three predictive factors are incorporated into it. Formally, our model consists of a CNN  $f : [0,1]^{3+1} \to \mathcal{D}$  that maps the input tensor  $\bar{\mathbf{x}}_t$  to a frame descriptor vector  $\mathbf{d}_t \in \mathcal{D}$ , followed by a RNN  $[18]g : \mathcal{D} \to \mathcal{D}'$  that incorporates temporal information to produce a video descriptor vector  $\mathbf{d}'_t \in \mathcal{D}'$ . We pass the input tensor  $\bar{\mathbf{x}}_t = \left[\mathbf{x}_t, \mathbf{1}\frac{t}{T_{\max}}\right]$ , which contains the input frame  $\mathbf{x}_t$  at time t and the elapsed time t as an additional channel, to the CNN. The elapsed time is scaled to the range [0, 1] by dividing t

<sup>&</sup>lt;sup>3</sup> Code and instructive examples are available at github.com/aimi-lab/catanet.

by the expected maximum video length  $T_{\text{max}}$  that we set to 20 minutes. Passing the elapsed time at the image level enables the CNN to learn its embedding.

Every video descriptor vector  $\mathbf{d}'_t$  produced by the LSTM is finally processed with three independent fully connected layers  $(h^{\mathrm{exp}}, h^{\mathrm{phase}}, h^{\mathrm{rsd}})$  to estimate the surgeon's experience  $\hat{\mathbf{y}}_t^{\mathrm{exp}}$ , the surgical phase  $\hat{\mathbf{y}}_t^{\mathrm{phase}}$ , and the RSD  $\hat{y}_t^{\mathrm{rsd}}$ . A softmax non-linearity is applied to obtain the probabilities  $\hat{\mathbf{y}}_t^{\mathrm{exp}}$  and  $\hat{\mathbf{y}}_t^{\mathrm{phase}}$ .

## 2.2 Training objectives

Our training dataset is a collection of tuples  $(\{\mathbf{x}_t\}_t, \{y_t^{\text{rsd}}\}_t, \{y_t^{\text{phase}}\}_t, y^{\text{exp}})$  consisting of a video sequence  $\mathbf{x}_t$ , the corresponding remaining surgical duration  $y_t^{\text{rsd}}$  per frame, surgical phases  $y_t^{\text{phase}}$  per frame, the surgeon's experience label  $y^{\text{exp}}$  per sequence. The index t is the elapsed time of the sequence.

We use the labeled data to train our model by minimizing two different loss functions. First, the CNN loss  $\ell_{cnn}$  is used to train the standalone CNN, without the RNN, to classify the phase and experience of individual frames. To this end, we append two temporary linear layers, akin to  $h^{phase}$  and  $h^{exp}$  above, acting on the output of the CNN  $\mathbf{d}_t$  to produce frame-level predictions  $\hat{\mathbf{y}}_{cnn,t}^{phase}$  and  $\hat{\mathbf{y}}_{cnn,t}^{exp}$ . The CNN loss minimizes the cross-entropies of both predictions,

$$\ell_{\rm cnn} = \mathcal{H}(\hat{\mathbf{y}}_{{\rm cnn},t}^{\rm phase}, y_t^{\rm phase}) + \mathcal{H}(\hat{\mathbf{y}}_{{\rm cnn},t}^{\rm exp}, y_t^{\rm exp}).$$
(1)

The RNN loss  $\ell_{\rm rnn}$ , on the other hand, is used with video sequences to train the RNN and to fine-tune the entire model end-to-end. It is a combination of the cross-entropies on phase and experience predictions, and the L1-norm of RSD predictions,

$$\ell_{\rm rnn} = \alpha \left| \hat{y}_t^{\rm rsd} - y_t^{\rm rsd} \right| + \mathcal{H}(\hat{\mathbf{y}}_t^{\rm phase}, y_t^{\rm phase}) + \mathcal{H}(\hat{\mathbf{y}}_t^{\rm exp}, y_t^{\rm exp}), \tag{2}$$

where the hyperparameter  $\alpha$  weights the relative contribution of the L1-norm.

### 3 Experiments

#### 3.1 Training and test data

We used the cataract-101 dataset [5] containing 101 videos (1'263'116 frames) with a resolution of  $720 \times 540$  pixels acquired at 25 fps. We did not choose a minimum video length, but used every video in the dataset. Each video is annotated with 10 surgical phases and the experience of the operating surgeon. Surgeries were performed by four different surgeons, divided in two senior surgeons (56 surgeries) and two assistant surgeons (45 surgeries). In addition, we manually labelled the start and end of each surgery, respectively, as the start of the first incision and the last tool interaction with the patient's eye.

The dataset was randomly split into 81 training and 20 test videos, so that 5 videos per surgeon remained in the test set. In the following experiments, we perform 6-fold cross-validation on the training split for model selection and hyper-parameter tuning. For inference, the output of all models is averaged.

#### **3.2** Implementation and baseline methods

Our CNN uses a DenseNet-169 [19] architecture pre-trained on ImageNet. Input images are reshaped and cropped to  $224 \times 224$ , and the network produces descriptor vectors  $\mathbf{d}_t$  of 1664 dimensions. We implement our RNN as a LSTM [18] with two layers of 128 cells, producing 128-dimensional video descriptor vectors  $\mathbf{d}'_t$ .

Training is performed in four stages: (1) First, to tackle class imbalance in surgical phases, we apply stratified sampling over the whole training dataset and sample 8000 frames per phase. We train using the Adam optimizer with early stopping in all training stages. The CNN is trained to minimize  $\ell_{\rm cnn}$  for 3 epochs with a learning rate of  $10^{-4}$ , batch size of 100 and early stopping on sub-epoch validation loss. (2) We minimize  $\ell_{\rm rnn}$  to train the RNN on full video sequences, temporally downsampled to 2.5 fps, for 50 epochs and a learning rate  $10^{-3}$ . The weights of the CNN are frozen during this stage. (3) The entire model is trained end-to-end minimizing  $\ell_{\rm rnn}$ . We apply truncated back-propagation on sub-sequences of 48 frames and setting the learning rate to  $5 \cdot 10^{-4}$  for 10 epochs. (4) Finally, we fine-tune the RNN minimizing  $\ell_{\rm rnn}$  for another 20 epochs while keeping the learning rate at  $5 \cdot 10^{-4}$ . The weights of the CNN are frozen during this stage. For the  $\ell_2$  loss, we set  $\alpha = 1$ . We implemented our method with PyTorch 1.6 and trained models using two Nvidia GeForce GTX 1080 Ti GPUs.

Given that no method for cataract RSD estimation exists, we compare our approach to two methods originally designed for laparoscopic surgery:

- **TimeLSTM** [13]: A ResNet CNN trained for phase recognition, followed by a LSTM trained for RSD prediction.
- **RSDNet** [14]: A modified version of [13], where the CNN is trained for progress prediction and the elapsed time is concatenated to the LSTM's output.

Both methods were originally proposed for cholecystectomy surgeries and did not provide implementations. Therefore, we use our own implementations for both baselines, following the respective publications.

We measure the quality of RSD predictions with the mean absolute error (MAE) per video,  $MAE = \frac{1}{T} \sum_{t=0}^{T-1} |\hat{y}_t^{rsd} - y_t^{rsd}|$ . Similarly, we also provide MAE averaged over the last two (MAE-2) and five (MAE-5) minutes, as well as at the end of *Hydrodissection* phase (MAE@Hyd). The latter metric is of clinical relevance in cataract surgery, as it highlights an appropriate time to prepare the following patient for surgery. In addition, we compute frame-wise accuracy (ACC) and F1-score per video to quantify surgical phase classification.

#### 3.3 Results

Table 1 shows the RSD prediction performance for all methods grouped by surgeon experience level. We group results by surgeon's experience level as both MAE and MAE@Hyd indirectly depend on the duration of the surgery and these take 5.6 and 11.8 minutes on average for senior and assistant surgeons, respectively. CataNet outperforms both RSDNet and TimeLSTM in all but one metric.

e videos, the i	ast two and	live minutes, a	na at the end	of <i>Hydrouissee</i>
	Exp	CataNet	RSDNet	TimeLSTM
MAE@Hyd	All	$1.66 \pm 1.35$	$2.32 \pm 1.27$	$2.34 \pm 1.54$
	Senior	$1.22 \pm 0.97$	$2.86 \pm 1.31$	$3.30 \pm 1.06$
	Assistant	$2.10 \pm 1.56$	$1.78 \pm 1.02$	$1.39 \pm 1.37$
MAE-5	All	$0.64 \pm 0.56$	$1.37\pm0.83$	$1.47\pm0.78$
	Senior	$0.78 \pm 0.60$	$1.98\pm0.73$	$2.06\pm0.70$
	Assistant	$0.51 \pm 0.23$	$0.76\pm0.28$	$0.88\pm0.14$
MAE-2	All	$0.35 \pm 0.20$	$1.23\pm0.53$	$1.22\pm0.32$
	Senior	$0.37 \pm 0.22$	$1.42\pm0.45$	$1.43\pm0.32$
	Assistant	$0.34 \pm 0.18$	$1.04\pm0.56$	$1.03\pm0.13$
MAE	All	$0.99 \pm 0.65$	$1.59\pm0.69$	$1.66\pm0.79$
	Senior	$0.83 \pm 0.64$	$1.97\pm0.73$	$2.11\pm0.70$
	Assistant	$1.15 \pm 0.65$	$1.19\pm0.36$	$1.20\pm0.59$

Table 1: RSD prediction results. The MAE (mean $\pm$ std in minutes) is shown for entire videos, the last two and five minutes, and at the end of *Hydrodissection*.

At the end of the critical Hydrodissection phase, for all experiences CataNet performs 0.66 minutes better than RSDNet and 0.68 minutes better than TimeL-STM. At the end of this phase, CataNet is considerably better than the baselines for senior surgeons, but worse for assistant surgeons. Considering the prediction over the whole video, CataNet performs on average 0.6 minutes better than RS-DNet and 0.67 minutes better than TimeLSTM. This can be explained by the fact that CataNet achieves comparable results for both senior and assistant surgeons. Overall, detection of the surgeon's experience is achieved with  $0.92 \pm 0.16$ accuracy and can thus exploit the fact that senior surgeons show low variance in surgery duration, however we do not claim that this accuracy would translate to new surgeons. The competing methods, on the other hand, tend to overestimate the duration of surgeries performed by senior surgeons.

We visualize CataNet's results for individual videos shown in Fig. 3 (see Supplementary material for more examples). Here, we see that predicting the surgeon experience on every frame can be beneficial in determining the confidence in RSD predictions. That is, given that the experience of the surgeon is known by the operating staff, an incorrect classification in experience can serve as an easy and interpretable indicator when the system is performing poorly (*i.e.* overestimating the RSD for false *assistant* predictions, or underestimating it for false *senior* predictions). Additionally, considering that *experience* is not a binary label, but a multi factored and scaled concept, our approach could be used to help assistant surgeons detect which phases of the surgery they could improve on. Finally, in two test set sequences, the surgeons fails to correctly perform the *lens implantation* phase, leading to unexpected extensions of the surgeries by 2-3 minutes and consequently underestimate of RSD before the mistake. However, our approach corrects the RSD predictions shortly thereafter. Details of these two sequences can be found in the supplementary material.

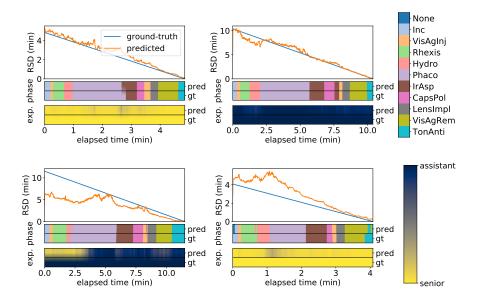


Fig. 3: Four examples of our method's outputs. For each plot, we show (**top**) the concordance between ground-truth and predicted RSD, (**middle**) the ground-truth and predicted surgical phases over time, and (**bottom**) the ground-truth and predicted probability of the surgeon's experience level.

**Ablation study:** CataNet is trained to classify the experience of the surgeon, the surgical phase, and the RSD, while its input is the video frames concatenated with the elapsed surgical time. To characterize the effects on performance of these different components, we show the performance of the following different approaches in Table 2: (i) train the CNN to only predict surgical phases and the RNN to predict both phases and RSD; (ii) train the CNN to only predict the surgeons experience and the RNN to predict both experience and RSD; (iii) train the CNN to only predict the surgeons experience and the RNN to predict both experience and RSD; (iii) train the CNN and the RNN to estimate only the RSD; (iv) same as (iii) but concatenate the elapsed time to the output of the LSTM (*i.e.*, as in RSDNet) instead of to the video frames.

From these experiments, we can see that (i) generally performs as well as CataNet, even outperforming it for some metrics. However, CataNet generally achieves a better performance for senior surgeons, who conduct the bulk of actual cataract surgeries [20]. In addition, we notice that (iii) performs better than (iv), showing that using the elapsed time as an input for the model considerably outperforms having it after the LSTM layer. Last, even when training without any labels, our approach (iii) performs better than that of RSDNet.

**Results on surgical phase classification:** Table 3 shows CataNet's performance for phase classification. Compared to the state-of-the-art by Qui et al. [21], CataNet achieves an increase of 12% in accuracy from 0.84 to 0.95. Fur-

Table 2. Ablation evaluation for h5D prediction.								
	Exp	Ca	taNet	(i)	phase	(ii) exp	(iii) RSD	(iv) elapsed
	All	1.66	$\pm 1.35$	1.43	$\pm$ 1.19	$1.82 \pm 1.63$	$1.99 \pm 1.38$	$2.28 \pm 1.34$
J		1.22	$\pm$ 0.97	1.42	$\pm 1.38$	$1.46 \pm 1.3$	$1.71 \pm 1.24$	$1.45\pm0.59$
		2.10	$\pm 1.56$	1.43	$\pm$ 1.05	$2.18 \pm 1.91$	$2.26 \pm 1.52$	$3.12 \pm 1.37$
	All	0.64	$\pm 0.46$	0.74	$\pm 0.56$	$0.87 \pm 0.62$	$0.76\pm0.41$	$0.75\pm0.34$
MAE-5	Senior	0.78	$\pm 0.60$	0.88	$\pm 0.73$	$0.99 \pm 0.77$	$0.98\pm0.44$	$0.85\pm0.31$
	Assistant	0.51	$\pm  0.23$	0.59	$\pm 0.27$	$0.76\pm0.43$	$0.55\pm0.24$	$0.64\pm0.34$
	All	0.35	$\pm 0.20$	0.35	$\pm$ 0.23	$0.51\pm0.27$	$0.39\pm0.28$	$0.44\pm0.20$
MAE-2	Senior	0.37	$\pm 0.22$	0.35	$\pm 0.26$	$0.52\pm0.36$	$0.45\pm0.38$	$0.51\pm0.22$
	Assistant	0.34	$\pm 0.18$	0.36	$\pm 0.21$	$0.50\pm0.18$	$0.33 \pm 0.10$	$0.36\pm0.15$
	All	0.99	$\pm \ 0.65$	0.98	$\pm \ 0.58$	$1.22\pm0.92$	$1.11\pm0.62$	$1.34\pm0.73$
MAE	Senior	0.83	$\pm 0.64$	0.91	$\pm 0.77$	$1.03\pm0.80$	$1.03\pm0.46$	$0.83 \pm 0.30$
	Assistant	1.15	$\pm \ 0.65$	1.04	$\pm 0.31$	$1.41 \pm 1.03$	$1.20\pm0.76$	$1.85\pm0.67$

Table 2: Ablation evaluation for RSD prediction

thermore, CataNet reliably detects the *Hydrodissection* phase, which is critical in the clinical context. Indeed, knowing the RSD at the end of this phase will improve the OR management since it corresponds to the moment where the next patient could be prepared for surgery.

Inference speed: RSD estimation is intended to be performed on realtime. We measured the execution time using a GeForce MX250 and avoided any overhead produced by other components of the system. We first run 100 frames through the GPU after which we measured the inference time on the next 1000 frames. The average time per frame was  $34.3 \pm 1.9$  ms, which corresponds to 29.09 fps. Considering that we sample the videos at 2.5 fps, we conclude that CataNet can easily be applied at 10 times real-time speed.

## 4 Conclusion

We have proposed a novel real-time method for estimating RSD for cataract surgeries from video feeds. Our approach jointly predicts the RSD, the surgeon's experience and the surgical phase, as these three elements are interconnected. Even when training our method without any labels, it outperforms the previous state-of-the-art RSD estimation models. We investigated the sources of this improvement and attribute these to (1) concatenating the video frames with the elapsed time and (2) including the phase and experience labels. Predicting the experience on every frame additionally increases the clinical applicability of our method by identifying low method confidence by observing predicted and real experience levels. Moving forward, a major challenge is in establishing large datasets to evaluate generalization capabilities and major clinical impact [22], for which assuring data consistency will be critical [23]. In the future, we plan to investigate this and how pre-operative data can be used to further improve RSD predictions for cataract surgery.

0			
	CataNet	Qui et al. [21]	TimeLSTM-CNN [13]
F1	$0.93 \pm 0.06$	-	$0.80 \pm 0.07$
F1-Hyd	$0.94 \pm 0.08$	-	$0.84 \pm 0.17$
ACC	$0.95 \pm 0.05$	$0.84\pm0.06$	$0.84\pm0.07$

Table 3: Macro F1-score and micro accuracy averaged over the 6-fold models.

Acknowlegements: This work was partially supported by the Haag-Streit Foundation and the University of Bern.

# References

- W. Wang, W. Yan, A. Müller, and M. He, "A global view on output and outcomes of cataract surgery with national indices of socioeconomic development," *Investigative Ophthalmology and Visual Science*, vol. 58, pp. 3669–3676, 07 2017.
- A. Achiron, F. Haddad, M. Gerra, E. Bartov, and Z. Burgansky-Eliash, "Predicting cataract surgery time based on preoperative risk assessment," *European Journal* of Ophthalmology, vol. 26, no. 3, 2016.
- M. Lanza, R. Koprowski, R. Boccia, K. Krysik, S. Sbordone, A. Tartaglione, A. Ruggiero, and F. Simonelli, "Application of artificial intelligence in the analysis of features affecting cataract surgery complications in a teaching hospital," *Frontiers in Medicine*, vol. 7, 2020.
- S. P. Devi, K. S. Rao, and S. S. Sangeetha, "Prediction of surgery times and scheduling of operation theaters in ophthalmology department," *Journal of Medical* Systems, vol. 36, no. 2, pp. 415–430, 2012.
- K. Schoeffmann, M. Taschwer, S. Sarny, B. Münzer, M. J. Primus, and D. Putzgruber, "Cataract-101 - Video dataset of 101 cataract surgeries," *Proceedings of* the 9th ACM Multimedia Systems Conference, MMSys 2018, pp. 421–425, 2018.
- N. Padoy, T. Blum, H. Feussner, M. O. Berger, and N. Navab, "On-line recognition of surgical activity for monitoring in the operating room," in *Proceedings of the National Conference on Artificial Intelligence*, vol. 3, 2008.
- S. Franke, J. Meixensberger, and T. Neumuth, "Intervention time prediction from surgical low-level tasks," *Journal of Biomedical Informatics*, vol. 46, no. 1, 2013.
- 8. A. C. Guédon, M. Paalvast, F. C. Meeuwsen, D. M. Tax, A. P. van Dijke, L. S. Wauben, M. van der Elst, J. Dankelman, and J. J. van den Dobbelsteen, "'It is Time to Prepare the Next patient' real-time prediction of procedure duration in laparoscopic cholecystectomies," *Journal of Medical Systems*, vol. 40, no. 12, 2016.
- N. Spangenberg, M. Wilke, and B. Franczyk, "A big data architecture for intrasurgical remaining time predictions," in *Proceedia Computer Science*, vol. 113, 2017.
- M. Maktabi and T. Neumuth, "Online time and resource management based on surgical workflow time series analysis," *International Journal of Computer Assisted Radiology and Surgery*, vol. 12, no. 2, 2017.
- S. Bodenstedt, M. Wagner, L. Mündermann, H. Kenngott, B. Müller-Stich, M. Breucha, S. T. Mees, J. Weitz, and S. Speidel, "Prediction of laparoscopic procedure duration using unlabeled, multimodal sensor data," *International Journal* of Computer Assisted Radiology and Surgery, vol. 14, no. 6, 2019.
- 12. D. Rivoir, S. Bodenstedt, F. von Bechtolsheim, M. Distler, J. Weitz, and S. Speidel, "Unsupervised Temporal Video Segmentation as an Auxiliary Task for Predicting

the Remaining Surgery Duration," in OR 2.0 Context-Aware Operating Theaters and Machine Learning in Clinical Neuroimaging. Springer, 2019, pp. 29–37.

- I. Aksamentov, A. P. Twinanda, D. Mutter, J. Marescaux, and N. Padoy, "Deep neural networks predict remaining surgery duration from cholecystectomy videos," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2017*, M. Descoteaux, L. Maier-Hein, A. Franz, P. Jannin, D. L. Collins, and S. Duchesne, Eds. Cham: Springer International Publishing, 2017, pp. 586–593.
- 14. A. P. Twinanda, G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "RSDNet: Learning to Predict Remaining Surgery Duration from Laparoscopic Videos Without Manual Annotations," *IEEE Transactions on Medical Imaging*, vol. 38, no. 4, 2019.
- T. Neumuth, P. Liebmann, P. Wiedemann, and J. Meixensberger, "Surgical workflow management schemata for cataract procedures process model-based design and validation of workflow schemata," *Methods of Information in Medicine*, vol. 51, no. 5, 2012.
- O. Zisimopoulos, E. Flouty, I. Luengo, P. Giataganas, J. Nehme, A. Chow, and D. Stoyanov, "DeepPhase: Surgical Phase Recognition in CATARACTS Videos," Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11073 LNCS, pp. 265–272, 2018.
- M. J. Primus, D. Putzgruber-Adamitsch, M. Taschwer, B. Münzer, Y. El-Shabrawi, L. Böszörmenyi, and K. Schoeffmann, "Frame-Based Classification of Operation Phases in Cataract Surgery Videos," *Lecture Notes in Computer Science (including* subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 10704 LNCS, pp. 241–253, 2018.
- S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.
- G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision* and pattern recognition, 2017, pp. 4700–4708.
- R. J. Campbell, S. R. El-Defrawy, S. S. Gill, M. Whitehead, E. d. L. Campbell, P. L. Hooper, C. M. Bell, and M. W. ten Hove, "Association of Cataract Surgical Outcomes With Late Surgeon Career Stages: A Population-Based Cohort Study," *JAMA Ophthalmology*, vol. 137, no. 1, pp. 58–64, 01 2019.
- B. Qi, X. Qin, J. Liu, Y. Xu, and Y. Chen, "A Deep Architecture for Surgical Workflow Recognition with Edge Information," *Proceedings - 2019 IEEE Interna*tional Conference on Bioinformatics and Biomedicine, BIBM 2019, pp. 1358–1364, 2019.
- O. Bar, D. Neimark, M. Zohar, G. D. Hager, R. Girshick, G. M. Fried, T. Wolf, and D. Asselmann, "Impact of data on generalization of AI for surgical intelligence applications," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.
- N. Ghamsarian, M. Taschwer, and K. Schoeffmann, "Deblurring cataract surgery videos using a multi-scale deconvolutional neural network," in 2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI), 2020, pp. 872–876.