Estimated Time to Surgical Procedure Completion: An Exploration of Video Analysis Methods

Barak Ariel, Yariv Colbeci, Judith Rapoport Ferman, Dotan Asselmann, and Omri Bar(✉)

Theator Inc., Palo Alto, CA, USA
{barak,yariv,judith,dotan,omri}@theator.io

Abstract. An accurate estimation of a surgical procedure's time to completion (ETC) is a valuable capability that has significant impact on operating room efficiency, and yet remains challenging to predict due to significant variability in procedure duration. This paper studies the ETC task in depth; rather than focusing on introducing a novel method or a new application, it provides a methodical exploration of key aspects relevant to training machine learning models to automatically and accurately predict ETC. We study four major elements related to training an ETC model: evaluation metrics, data, model architectures, and loss functions. The analysis was performed on a large-scale dataset of approximately 4,000 surgical videos including three surgical procedures: Cholecystectomy, Appendectomy, and Robotic-Assisted Radical Prostatectomy (RARP). This is the first demonstration of ETC performance using video datasets for Appendectomy and RARP. Even though AI-based applications are ubiquitous in many domains of our lives, some industries are still lagging behind. Specifically, today, ETC is still done by a mere average of a surgeon's past timing data without considering the visual data captured in the surgical video in real time. We hope this work will help bridge the technological gap and provide important information and experience to promote future research in this space. The source code for models and loss functions is available at: https://github.com/theator/etc.

Keywords: Surgical Intelligence · Operating Room Efficiency · ETC · RSD · Cholecystectomy · Appendectomy · Radical Prostatectomy

1 Introduction

One of the significant logistical challenges facing hospital administrations today is operating room (OR) efficiency. This parameter is determined by many fac-

B. Ariel and Y. Colbeci—Equal contribution.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-43996-4_16.

H. Greenspan et al. (Eds.): MICCAI 2023, LNCS 14228, pp. 165–175, 2023.
https://doi.org/10.1007/978-3-031-43996-4_16

tors, one of which is surgical procedure duration that reflects intracorporeal time, which in itself poses a challenge as, even across the same procedure type, duration can vary greatly. This variability is influenced by numerous elements, including the surgeon's experience, the patient's comorbidities, unexpected events occurring during the procedure, the procedure's complexity and more. Accurate real-time estimation of procedure duration improves scheduling efficiency because it allows administrators to dynamically reschedule before a procedure has run overtime. Another important aspect is the ability to increase patient safety and decrease complications by dosing and timing anesthetics more accurately.

Currently, methods aiming for OR workflow optimization through ETC are lacking. One study showed that surgeons underestimated surgery durations by an average of 31 min, while anesthesiologists underestimated the durations by 35 min [21]. These underestimations drive inefficiencies, causing procedures to be delayed or postponed, forcing longer waiting times for patients. For example, a large variation of waiting time (47 ± 17 min) was observed in a study assessing 157 Cholecystectomy patients [15].

As AI capabilities have evolved greatly in recent years, the field of minimally invasive procedures, which is inherently video-based, has emerged as a potent platform for the harnessing of these capabilities to improve both patient care and workflow efficiency. Consequently, ETC has become a technologically achievable and clinically beneficial task.

2 Related Work

Initially, ETC studies performed preoperative estimates based on surgeon data, patient data, or a combination of these [2,10]. Later on, intraoperative estimates were performed, with some studies requiring manual annotations or the addition of external information [7,11,12,16]. Recently, a study by Twinanda et al. [23] achieved robust ETC results, even without incorporating external information, showing that video-based ETC is better than statistical analysis of past surgeons' data. However, all these studies have evaluated ETC using limited size datasets with inherent biases, as they are usually curated from a small number of surgeons and medical centers or exclude complex cases with significant unexpected events.

In this work, we study the key elements important to the development of ETC models and perform an in-depth methodical analysis of this task. First, we suggest an adequate metric for evaluation - SMAPE, and introduce two new architectures, one based on LSTM networks and one on the transformer architecture. Then, we examine how different ETC methods perform when trained with various loss functions and show that their errors are not necessarily correlated. We then test the hypothesis that an ensemble composed of several ETC model variations can significantly improve estimation compared to any single model.

3 Methods

3.1 Evaluation Metrics

Mean Absolute Error (MAE). The evaluation metric used in prior work was MAE.

$$MAE(y, \hat{y}) = \frac{1}{T} \cdot \sum_{t=0}^{T-1} |y_t - \hat{y_t}| \tag{1}$$

where T is a video duration, y is the actual time left until completion, and $\hat{y}$ is the ETC predictions. A disadvantage of MAE is its reliance on the magnitude of values, consequently, short videos are likely to have small errors while long videos are likely to have large errors. In addition, MAE does not consider the actual video duration or the temporal location for which the predictions are made.

Symmetric Mean Absolute Percentage Error (SMAPE). SMAPE is invariant to the magnitude and keeps an equivalent scale for videos of different duration, thus better represents ETC performance [3,4,20].

$$SMAPE(y, \hat{y}) = \frac{1}{T} \cdot \sum_{t=0}^{T-1} (\frac{|y_t - \hat{y_t}|}{|y_t| + |\hat{y_t}|} \cdot 100) \tag{2}$$

3.2 Datasets

We focus on three different surgical video datasets (a total of 3,993 videos) that were curated from several medical centers (MC) and include procedures performed by more than 100 surgeons. The first dataset is Laparoscopic Cholecystectomy that contains 2,400 videos (14 MC and 118 surgeons). This dataset was utilized for the development and ablation study. Additionally, we explore two other datasets: Laparoscopic Appendectomy which contains 1,364 videos (5 MC and 61 surgeons), and Robot-Assisted Radical Prostatectomy (RARP) which contains 229 videos (2 MC and 14 surgeons). The first two datasets are similar, both are relatively linear and straightforward procedures, have similar duration distribution, and are abdominal procedures with similarities in anatomical views. However, RARP is almost four times longer on average. Therefore, it is interesting to explore how methods developed on a relatively short and linear procedure will perform on a much longer procedure type such as RARP. Table 3 in the appendix provides a video duration analysis for all datasets. The duration is defined as the difference between surgery start and end times, which is the time interval between scope-in and scope-out. All datasets were randomly divided into training, validation, and test sets with a ratio of 60/15/25%.

3.3 Loss Functions

Loss values are calculated by comparing ETC predictions ($\hat{y_t}$) for each timestamp to the actual time left until the procedure is complete (y_t). The final loss for each video is the result of averaging these values across all timestamps.

MAE Loss. The MAE loss is defined by:

$$L_{MAE}(y, \hat{y}) = \frac{1}{T} \cdot \sum_{t=0}^{T-1} |y_t - \hat{y}_t| \tag{3}$$

Smooth L1 Loss. The smooth L1 loss is less sensitive to outliers [9].

$$L(y, \hat{y}) = \frac{1}{T} \cdot \sum_{t=0}^{T-1} SmoothL1(y_t - \hat{y}_t) \tag{4}$$

in which

$$SmoothL1(x) = \begin{cases} 0.5x^2 & |x| < 1 \\ |x| - 0.5 & otherwise \end{cases} \tag{5}$$

SMAPE Loss. Based on the understanding that SMAPE (Sect. 3.1) is a good representation of the ETC problem, we also formulated it as a loss function:

$$L_{SMAPE}(y, \hat{y}) = \frac{1}{T} \cdot \sum_{t=0}^{T-1} \left(\frac{|y_t - \hat{y}_t|}{|y_t| + |\hat{y}_t|} \cdot 100\right) \tag{6}$$

Importantly, SMAPE produces higher loss values for the same absolute error as the procedure progresses, when the denominator is getting smaller. This property is valuable as the models should be more accurate as the surgery nears its end.

Corridor Loss. A key assumption overlooked in developing ETC methods is that significant and time-impacting events might occur during a procedure. For example, a prolonged procedure due to significant bleeding occurring after 30 min of surgery is information that is absent from the model when providing predictions at the 10 min timestamp. To tackle this problem, we apply the corridor loss [17] that considers both the actual progress of a procedure and the average duration in the dataset (see Fig. 2 in the appendix for a visual example). The corridor loss acts as a wrapper (π) for other loss functions:

$$Corridor(Loss(y, \hat{y})) = \frac{1}{T} \cdot \sum_{t=0}^{T-1} \pi(y, t) \cdot Loss(y_t, \hat{y}_t) \tag{7}$$

Interval L1 Loss. The losses described above focus on the error between predictions and labels for each timestamp independently. Influenced by the total variation loss, we suggest considering the video's sequential properties. The interval L1 loss focuses on jittering in predictions between timestamps in a pre-defined interval, aiming to force them to act more continuously. $\hat{y}_t$ are the predictions per timestamp, and S is an interval time span (jump) between two timestamps.

$$L^S_{IntervalL1}(\hat{y}_t) = \sum_{t=0}^{T-S-1} |\hat{y}_{t+S} - \hat{y}_t| \tag{8}$$

Total Variation Denoising Loss. This loss is inspired by a 1D total variation denoising loss and was modified to fit as part of the ETCouple model.

$$L_{squared\_error}(y, \hat{y}) = \frac{1}{2T} \cdot \sum_{t=1}^{T-1}((y_t - \hat{y}_t)^2 + (y_{t-S} - \hat{y}_{t-S})^2) \tag{9}$$

$$L_{total\_variation\_denoising}(y, \hat{y}) = L_{squared\_error}(y, \hat{y}) + \lambda \cdot L_{IntervalL1}^{S=120}(\hat{y}) \tag{10}$$

3.4 ETC Models

Feature Representation. All models and experiments described in this work are based on fixed visual features that were extracted from the surgical videos using a pre-trained model. This approach allows for shorter training cycles, less computing requirements, and benefits from a model that was trained on a different task [6,14]. Previous works showed that pre-training could be done with either progress labels or surgical steps labels and that similar performances are achieved, with a slight improvement when using the steps label pre-training [1,23]. In this work, we use a pre-trained Video Transformer Network (VTN) [13] model with a Vision Transformer (ViT) [8] backbone as a feature extraction module. It was originally trained using the same training set (Sect. 3.2) for the step recognition task with a similar protocol to the one described by [5].

Inferring ETC. Our ETC architectures end with a single shared fully connected (FC) layer and a Sigmoid that outputs two values: ETC and *progress*. ETC is inferred by averaging the predicted ETC value and the one calculated from the *progress*.

$$ETC = T - t_{el} = \frac{t_{el}}{progress} - t_{el} \tag{11}$$

where T is the video duration and t_{el} marks the elapsed time. Inspired by [12], we also incorporate t_{max} which is defined as the expected maximum video length. t_{max} is applied to scale the elapsed time and ensures values in a range [0, 1].

ETC-LSTM. A simple architecture that consists of an LSTM layer with a hidden size of 128. Following hyperparameters tuning on the validation set, the ETC-LSTM was trained using an SGD optimizer with a constant learning rate of 0.1 and a batch size of 32 videos.

ETCouple. ETCouple is a different approach to applying LSTM networks. In contrast to ETC-LSTM and similar methods which predict ETC for a single timestamp, here we randomly select one timestamp from the input video and

set it as an anchor sample. The anchor is then paired with a past timestamp using a fixed interval of S = 120 s. The model is given two inputs, the features from the beginning of the procedure up to the anchor and the features up to the pair location. Instead of processing the entire video in an end-to-end manner, we only process past information and are thus able to use a bi-directional LSTM (hidden dimension is 128). The rest of the architecture contains a dropout layer ($P = 0.5$), the shared FC layer, and a Sigmoid function. We explored various hyperparameters and the final model was trained with a batch size of 16, an AdamW optimizer with a learning rate of $5 \cdot 10^{-4}$, and a weight decay of $5 \cdot 10^{-3}$.

ETCformer. LSTM networks have been shown to struggle with capturing long-term dependencies [22]. Intuitively, ETC requires attending to events that occur in different temporal locations throughout the procedure. Thus, we propose a transformer-based network that uses attention modules [24]. The transformer encoder architecture has four self-attention heads with a hidden dimension of size 512. To allow this model to train in a framework where all the video's features are the input to the model but still maintain a causal system, we used a forward direction self-attention [18,19]. This is done by masking out future samples for each timestamp, thus relying only on past information. Best results on the validation set were achieved when training with a batch size of two videos, an AdamW optimizer with a learning rate of 10^{-4}, and a weight decay of 0.1.

Table 1. Comparing validation set results of the mean SMAPE and MAE performance when training ETC models with one or a combination of a few loss functions.

	MAE loss	SMAPE loss	Interval L1 loss (S = 1)	Corridor loss	Mean SMAPE	MAE
ETC-LSTM	✓				**20.68**	**7.67**
	✓	✓			21	7.9
	✓	✓	✓		20.81	7.96
	✓	✓		✓	20.85	7.79
	✓	✓	✓	✓	20.92	7.97
	MAE loss	SMAPE loss	Total variation denoising loss		Mean SMAPE	MAE
ETCouple	✓				21.84	**7.81**
	✓	✓			21.7	8.82
	✓	✓	✓		**21.6**	8.1
	MAE loss	SMAPE loss	Interval L1 loss (S = 1)	Corridor loss	Mean SMAPE	MAE
ETCformer	✓				21.13	7.79
	✓	✓			21.06	7.65
	✓	✓	✓		20.78	**7.38**
	✓	✓		✓	20.68	7.72
	✓	✓	✓	✓	**20.54**	7.67

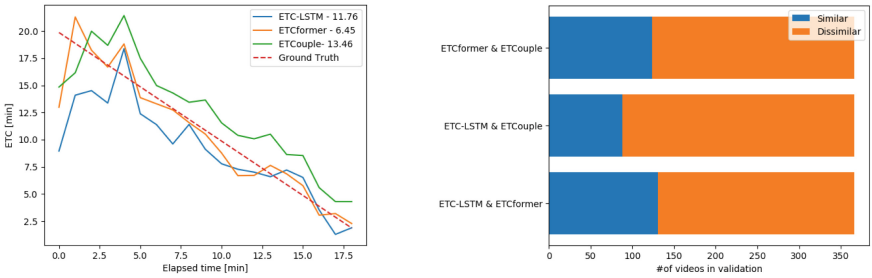
4 Results

4.1 Ablation Experiments

This section provides ablation studies on the Cholecystectomy dataset.

Loss Functions. Table 1 provides a comparison on the same validation set when using one or the sum of a few loss functions. The classic approach of using LSTM produces the best results when using only MAE loss. However, ETCouple and ETCformer benefit from the combination of several loss functions.

Error Analysis. To test whether the errors of the various models are correlated, we compared the predictions made by the different models on a per-video basis. We use SMAPE and analyze the discrepancy by comparing the difference of every two model variations independently. Then, we divided the videos into a *similar* and a *dissimilar* group, by using a fixed threshold, i.e., if the SMAPE difference is smaller than the threshold the models are considered as providing *similar* results. The threshold was empirically set to 2, deduced from the ETC curves, which are almost identical when the SMAPE difference is smaller than 2 (Fig. 1(a) and appendix Fig. 4. We demonstrate these results visually in Fig. 1(b). Interestingly, there are significant differences in SMAPE between different models (disagreement in more than 50%). ETC-LSTM and ETCouple show the highest disagreement.



(a) ETC Curve (b) Error analysis comparison

Fig. 1. (a) ETC per minute on a Cholecystectomy video. Each color represents a different ETC model, the numbers are the SMAPE scores. The dashed lines represents the ground truth progress. (b) An error analysis comparison was performed by measuring the difference in SMAPE per video between two independent models (each row is a pair of two models). The blue color represents the number of *similar* videos, and the orange color the number of *dissimilar* videos in the validation set (total of 366). (Color figure online)

Baseline Comparison. We reproduce RSDNet [23] on top of our extracted features and use it as a baseline for comparison. We followed all methodical details described in the original paper, only changing the learning rate reduction policy to match the same epoch proportion in our dataset. Table 2 shows that ETC-LSTM and RSDNet have similar results, ETC-LSTM achieves better SMAPE scores while RSDNet is more accurate in MAE. These differences can be the product of scaling the elapsed time using t_{max} *vs.* s_{norm} and shared *vs.* independent FC layer. The ETCformer model reaches the best SMAPE results but is still short on MAE.

Ensemble Analysis. There are many tasks in machine learning in which data can be divided into easy or hard samples. We argue that the ETC task is different in these regards. Based on the error analysis, we explored how an ensemble of models performs and if it produces better results (Table 2). In contrast to a classic use case of models ensemble, in which the same model is trained with bootstrapping on different folds of data, here we suggest an ensemble that uses different models, which essentially learn to perform differently on the same input video. Figure 3 in the appendix illustrates the MAE error graph for the ensemble, presenting the mean and SD of the MAE. All model variations' performance is also provided in the appendix in Table 4. When using more than one model, the ETC predictions for each model are averaged into a single end result.

Table 2. ETC models comparison on the test set, using mean SMAPE and standard deviation (SD), median SMAPE, 90th percentile (90p) SMAPE, and MAE. The ensemble achieves the best results in most metrics.

		Mean SMAPE [SD]	Median SMAPE	90p SMAPE	MAE (min)
Cholecystectomy	RSDNet	20.97 [8.2]	19.06	32.6	7.48
	ETC-LSTM	20.75 [8.1]	18.82	31.33	7.88
	ETCouple	20.99 [8.3]	18.75	33.18	8.15
	ETCformer	20.06 [7.8]	18.22	31.34	7.56
	Ensemble	**19.57** [7.9]	**17.56**	**30.87**	**7.33**
Appendectomy	RSDNet	20.6 [8.5]	19.18	32.55	10.98
	ETC-LSTM	21.92 [9.2]	19.73	33.98	11.88
	ETCouple	21.49 [7.6]	20.08	31.76	8.49
	ETCformer	22.74 [8.3]	20.58	33.88	12.24
	Ensemble	**19.87** [7.8]	**17.9**	**30.75**	**7.47**
RARP	RSDNet	17.03 [5.9]	15.81	25.07	**21.01**
	ETC-LSTM	16.85 [4.7]	16.07	24.26	23.8
	ETCouple	19.05 [9.4]	17.58	30.57	30.88
	ETCformer	17.55 [6.8]	15.94	29.28	26.18
	Ensemble	**15** [5.9]	**13.94**	**23.43**	21.35

4.2 Appendectomy and RARP Results

We examine the results on two additional datasets to showcase the key elements explored in this work (Table 2). In Appendectomy, the ensemble also achieves the best results on the test set with a significant drop in SMAPE and MAE scores. The ETCformer performs the worst compared to other model variations, this might be because transformers require more data for training, therefore additional data could show its potential as seen in Cholecystectomy. The RARP dataset contains fewer videos, but they are of longer duration. Here too, the ensemble achieves better SMAPE scores.

5 Discussion

In this work, we examine different architectures trained with several loss functions and show how SMAPE can be utilized as a better metric to compare ETC models. In the error analysis, we conclude that each model learns to operate differently on the same videos. This led us to explore an ensemble of models which eventually achieves the best results. Yet, this conclusion can facilitate future work, focusing on understanding the differences and commonalities of the models' predictions and developing a unified or enhanced model, potentially reducing the complexity of training several ETC models and achieving better generalizability. Future work should also incorporate information regarding the surgeon's experience, which may improve the model's performance.

This work has several limitations. First, the proposed models need to be evaluated across other procedures and specialties for their potential to be validated further. Second, the ensemble's main disadvantage is its requirement for more computing resources. In addition, there may be data biases due to variability in the time it takes surgeons to perform certain actions at different stages of their training. Finally, although our model relies on video footage only, and no annotations are required for ETC predictions, manual annotations of surgical steps are still needed for pre-training of the feature extraction model.

Real-time ETC holds great potential for surgical management. First, in optimizing OR scheduling, and second as a proxy to complications that cause unusual deviations in anticipated surgery duration. However, optimizing OR efficiency with accurate procedural ETC, based on surgical videos, has yet to be realized. We hope the information in this study will assist researchers in developing new methods and achieve robust performance across multiple surgical specialties, ultimately leading to better OR management and improved patient care.

Acknowledgements. We thank Ross Girshick for providing valuable feedback on this manuscript and for helpful suggestions on several experiments.

References

1. Aksamentov, I., Twinanda, A.P., Mutter, D., Marescaux, J., Padoy, N.: Deep neural networks predict remaining surgery duration from cholecystectomy videos. In: Descoteaux, M., Maier-Hein, L., Franz, A., Jannin, P., Collins, D.L., Duchesne, S. (eds.) MICCAI 2017. LNCS, vol. 10434, pp. 586–593. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66185-8_66
2. Ammori, B., Larvin, M., McMahon, M.: Elective laparoscopic cholecystectomy. Surg. Endosc. **15**(3), 297–300 (2001)
3. Armstrong, J.S., Collopy, F.: Error measures for generalizing about forecasting methods: empirical comparisons. Int. J. Forecast. **8**(1), 69–80 (1992)
4. Armstrong, J.S., Forecasting, L.R.: From Crystal Ball to Computer, p. 348. New York (1985)
5. Bar, O., et al.: Impact of data on generalization of AI for surgical intelligence applications. Sci. Rep. **10**(1), 1–12 (2020)
6. Colbeci, Y., Zohar, M., Neimark, D., Asselmann, D., Bar, O.: A multi instance learning approach for critical view of safety detection in laparoscopic cholecystectomy. In: Proceedings of Machine Learning Research, vol. 182, pp. 1–14 (2022)
7. Dexter, F., Epstein, R.H., Lee, J.D., Ledolter, J.: Automatic updating of times remaining in surgical cases using Bayesian analysis of historical case duration data and "instant messaging" updates from anesthesia providers. Anesth. Analg. **108**(3), 929–940 (2009)
8. Dosovitskiy, A., et al.: An image is worth 16×16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Girshick, R.: Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448 (2015)
10. Macario, A., Dexter, F.: Estimating the duration of a case when the surgeon has not recently scheduled the procedure at the surgical suite. Anesth. Analg. **89**(5), 1241–1245 (1999)
11. Maktabi, M., Neumuth, T.: Online time and resource management based on surgical workflow time series analysis. Int. J. Comput. Assist. Radiol. Surg. **12**(2), 325–338 (2017)
12. Marafioti, A., et al.: CataNet: predicting remaining cataract surgery duration. In: de Bruijne, M., et al. (eds.) MICCAI 2021. LNCS, vol. 12904, pp. 426–435. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-87202-1_41
13. Neimark, D., Bar, O., Zohar, M., Asselmann, D.: Video transformer network. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3163–3172 (2021)
14. Neimark, D., Bar, O., Zohar, M., Hager, G.D., Asselmann, D.: "Train one, classify one, teach one"-cross-surgery transfer learning for surgical step recognition. In: Medical Imaging with Deep Learning, pp. 532–544. PMLR (2021)
15. Paalvast, M., et al.: Real-time estimation of surgical procedure duration. In: 2015 17th International Conference on E-Health Networking, Application & Services (HealthCom), pp. 6–10. IEEE (2015)
16. Padoy, N., Blum, T., Feussner, H., Berger, M.O., Navab, N.: On-line recognition of surgical activity for monitoring in the operating room. In: AAAI, pp. 1718–1724 (2008)
17. Rivoir, D., Bodenstedt, S., von Bechtolsheim, F., Distler, M., Weitz, J., Speidel, S.: Unsupervised temporal video segmentation as an auxiliary task for predicting the remaining surgery duration. In: Zhou, L., et al. (eds.) OR 2.0/MLCN -2019. LNCS,

vol. 11796, pp. 29–37. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32695-1_4
18. Shen, T., Zhou, T., Long, G., Jiang, J., Pan, S., Zhang, C.: DiSAN: directional self-attention network for RNN/CNN-free language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)
19. Shen, T., Zhou, T., Long, G., Jiang, J., Zhang, C.: Fast directional self-attention mechanism. arXiv preprint arXiv:1805.00912 (2018)
20. Tofallis, C.: A better measure of relative prediction accuracy for model selection and model estimation. J. Oper. Res. Soc. **66**(8), 1352–1362 (2015)
21. Travis, E., Woodhouse, S., Tan, R., Patel, S., Donovan, J., Brogan, K.: Operating theatre time, where does it all go? A prospective observational study. BMJ **349** (2014). https://doi.org/10.1136/bmj.g7182, https://www.bmj.com/content/349/bmj.g7182
22. Trinh, T., Dai, A., Luong, T., Le, Q.: Learning longer-term dependencies in RNNs with auxiliary losses. In: International Conference on Machine Learning, pp. 4965–4974. PMLR (2018)
23. Twinanda, A.P., Yengera, G., Mutter, D., Marescaux, J., Padoy, N.: RSDNet: learning to predict remaining surgery duration from laparoscopic videos without manual annotations. IEEE Trans. Med. Imaging **38**(4), 1069–1078 (2018)
24. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)