TeCNO: Surgical Phase Recognition with Multi-Stage Temporal Convolutional Networks

Tobias Czempiel[1], Magdalini Paschali[1], Matthias Keicher[1], Walter Simson[1], Hubertus Feussner[2], Seong Tae Kim[1], Nassir Navab[1,3]

[1] Computer Aided Medical Procedures, Technische Universität München, Germany
[2] MITI, Klinikum Rechts der Isar, Technische Universität München, Germany
[3] Computer Aided Medical Procedures, Johns Hopkins University, Baltimore, USA

Abstract. Automatic surgical phase recognition is a challenging and crucial task with the potential to improve patient safety and become an integral part of intra-operative decision-support systems. In this paper, we propose, for the first time in workflow analysis, a Multi-Stage Temporal Convolutional Network (MS-TCN) that performs hierarchical prediction refinement for surgical phase recognition. Causal, dilated convolutions allow for a large receptive field and online inference with smooth predictions even during ambiguous transitions. Our method is thoroughly evaluated on two datasets of laparoscopic cholecystectomy videos with and without the use of additional surgical tool information. Outperforming various state-of-the-art LSTM approaches, we verify the suitability of the proposed causal MS-TCN for surgical phase recognition.

Keywords: Surgical Workflow · Surgical Phase Recognition · Temporal Convolutional Networks · Endoscopic Videos · Cholecystectomy

1 Introduction

Surgical workflow analysis is an integral task to increase patient safety, reduce surgical errors and optimize the communication in the operating room (OR) [1]. Specifically, surgical phase recognition can provide vital input to physicians in the form of early warnings in cases of deviations and anomalies [2] as well as context-aware decision support [3]. Another use case is automatic extraction of a surgery's protocol, which is crucial for archiving, educational and post-operative patient-monitoring purposes [4].

Computer-assisted intervention (CAI) systems based on machine learning techniques have been developed for surgical workflow analysis [5], deploying not only OR signals but also intra-operative videos, which can be captured during a laparoscopic procedure, since cameras are an integral part of the workflow. However, the task of surgical phase recognition from intra-operative videos remains challenging even for advanced CAI systems [6, 7] due to the variability of patient anatomy and surgeon style [8] along with the limited availability and quality of video material [9]. Furthermore, strong similarities among phases and transition ambiguity lead to decreased performance and limited generalizability

of the existing methods. Finally, most approaches dealing with temporal information, such as Recurrent Neural Networks (RNNs) [10] leverage sliding window detectors, which have difficulties capturing long-term temporal patterns.

Towards this end, we propose a pipeline utilizing dilated Temporal Convolutional Networks (TCN) [11] for accurate and fast surgical phase recognition. Their large temporal receptive field captures the full temporal resolution with a reduced number of parameters, allowing for faster training and inference time and leveraging of long, untrimmed surgical videos.

Initial approaches for surgical phase recognition [5] exploited binary surgical signals. Hidden Markov Models (HMMs) captured the temporal information with the use of Dynamic Time Warping (DTW). However, such methods relied on whole video sequences and could not be applied in an online surgery scenario. EndoNet [12] jointly performed surgical tool and phase recognition from videos, utilizing a shared feature extractor and a hierarchical HMM to obtain temporally-smoothed phase predictions. With the rise of RNNs, EndoNet was evolved to EndoLSTM, which was trained in a two-step process including a Convolutional Neural Network (CNN) as a feature extractor and an LSTM [13] for feature refinement. Endo2N2 [14] leveraged self-supervised pre-training of the feature extractor CNN by predicting the Remaining Surgery Duration (RSD). Afterwards a CNN-LSTM model was trained end-to-end to perform surgical phase recognition. Similarly, SV-RCNet [15] trained an end-to-end ResNet [16] and LSTM model for surgical phase recognition with a prior knowledge inference scheme.

MTRCNet-CL [17] approached surgical phase classification as a multi-task problem. Extracted frame features were used to predict tool information while also serving as input to an LSTM model [13] for the surgical phase prediction. A correlation loss was employed to enhance the synergy between the two tasks. The common factor of the methods mentioned above is the use of LSTMs, which retain memory of a limited sequence, that cannot span minutes or hours, which is the average duration of a surgery. Thus, they process the temporal information in a slow, sequential way prohibiting inference parallelization, which would be beneficial for their integration in an online OR scenario.

Temporal convolutions [11] were introduced to hierarchically process videos for action segmentation. An encoder-decoder architecture was able to capture both high- and low-level features in contrast to RNNs. Later, TCNs adapted dilated convolutions [18] for action localization and achieved improvement in performance due to a larger receptive field for higher temporal resolution. Multi-Stage TCNs (MS-TCNs) [19] were introduced for action segmentation and consisted of stacked predictor stages. Each stage included an individual multi-layer TCN, which incrementally refined the initial prediction of the previous stages.

In this paper our contribution is two-fold: (1) We propose, for the first time in surgical workflow analysis, the introduction of causal, dilated MS-TCNs for accurate, fast and refined online surgical phase recognition. We call our method TeCNO, derived from **Te**mporal **C**onvolutional **N**etworks for the **O**perating room. (2) We extensively evaluate TeCNO on the challenging task of surgical

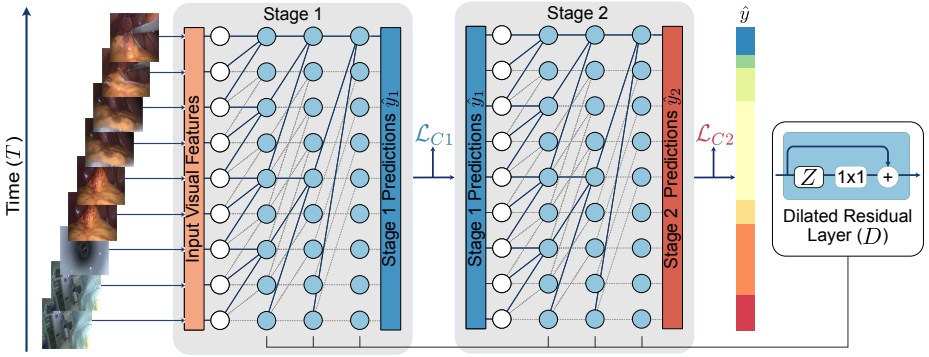

Fig. 1. Overview of the proposed TeCNO multi-stage hierarchical refinement model. The extracted frame features are forwarded to Stage 1 of our TCN, which consists of 1D dilated convolutional and dilated residual layers D. Cross-Entropy Loss is calculated after each stage and aggregated for the joint training of the model.

phase recognition on two laparoscopic video datasets, verifying the effectiveness of the proposed approach.

2 Methodology

TeCNO constitutes a surgical workflow recognition pipeline consisting of the following steps: 1) We employ a ResNet50 as a visual feature extractor. 2) We refine the extracted features with a 2-stage causal TCN model that forms a high-level reasoning of the current frame by analyzing the preceding ones. The refinement 2-stage TCN model is depicted in Fig. 1.

2.1 Feature Extraction Backbone

A ResNet50 [16] is trained frame-wise without temporal context as a feature extractor from the video frames either on a single task for phase recognition or as a multi-task network when a dataset provides additional label information, for instance tool presence per frame. In the multi-task scenario for concurrent phase recognition and tool identification, our model concludes with two separate linear layers, whose losses are combined to train the model jointly. Since phase recognition is an imbalanced multi-class problem we utilize softmax activations and weighted cross entropy loss for this task. The class weights are calculated with median frequency balancing [20]. For tool identification, multiple tools can be present at every frame, constituting a multi-label problem, which is trained with a binary-cross entropy loss after a sigmoid activation.

We adopt a two-stage approach so that our temporal refinement pipeline is independent of the feature extractor and the available ground truth provided in the dataset. As we will discuss in Section 4, TCNs are able to refine the

predictions of various features extractors regardless of their architecture and label information.

2.2 Temporal Convolutional Networks

For the temporal phase prediction task, we propose TeCNO, a multi-stage temporal convolutional network that is visualized in Fig. 1. Given an input video consisting of $x_{1:t},\ t \in [1, T]$ frames, where T is the total number of frames, the goal is to predict $y_{1:t}$ where y_t is the class label for the current time step t. Our temporal model follows the design of MS-TCN and contains neither pooling layers, that would decrease the temporal resolution, nor fully connected layers, which would increase the number of parameters and require a fixed input dimension. Instead, our model is constructed solely with temporal convolutional layers.

The first layer of Stage 1 is a 1x1 convolutional layer that matches the input feature dimension to the chosen feature length forwarded to the next layer within the TCN. Afterwards, dilated residual (D) layers perform dilated convolutions as described in Eq. 1 and Eq. 2. The major component of each D layer is the dilated convolutional layer (Z).

$$Z_l = ReLU(W_{1,l} * D_{l-1} + b_{1,l}) \tag{1}$$

$$D_l = D_{l-1} + W_{2,l} * Z_l + b_{2,l} \tag{2}$$

D_l is the output of D (Eq. 2), while Z_l is the result of the dilated convolution of kernel $W_{1,l}$ with the output of the previous layer D_{l-1} activated by a $ReLU$ (Eq. 1). $W_{2,l}$ is the kernel for the 1x1 convolutional layer, $*$ denotes a convolutional operator and $b_{1,l}, b_{2,l}$ are bias vectors.

Instead of the acausal convolutions in MS-TCN [19] with predictions $\hat{y_t}(x_{t-n}, ..., x_{t+n})$ which depend on both n past and n future frames, we use causal convolutions within our D layer. Our causal convolutions can be easily described as 1D convolutions with kernel size 3 with a dilation factor. The term causal refers to the fact that the output of each convolution is shifted and the prediction $\hat{y}$ for time step t does not rely on any n future frames but only relies on the current and previous frames i.e. $\hat{y_t}(x_{t-n}, ..., x_t)$. This allows for intra-operative online deployment of TeCNO, unlike biLSTMs that require knowledge of future time steps [21–23].

Increasing the dilation factor of the causal convolutions by 2 within the D layer for each consecutive layer we effectively increase the temporal receptive field RF of the network without a pooling operation (Eq.3). We visualize the progression of the receptive field of the causal convolutions in Fig. 1. A single D layer with a dilation factor of 1 and a kernel size of 3 can process three time steps at a time. Stacking 3 consecutive D layers within a stage, as seen in Fig. 1, increases the temporal resolution of the kernels to 8 time steps. The size of the temporal receptive field depends on the number of D layers $l \in [1, N]$ and is given by:

$$RF(l) = (2)^{l+1} - 1 \tag{3}$$

This results in a exponential increase of the receptive field, which significantly reduces the computational cost in comparison to models that achieve higher receptive field by increasing the kernel size or the amount of total layers [18].

Multi-Stage TCN The main idea of the multi-stage approach is to refine the output of the first stage S_1 by adding M additional stages to the network $S_{1...M}$ [24]. The extracted visual feature vectors for each frame of a surgical video $x_{1:T}$ are the input of S_1, as explained above. The output of S_1 is directly fed into the second stage S_2. As seen in Fig. 1, the outputs of S_1 and S_2 have independent loss functions and the reported predictions are calculated after S_2, where the final refinement is achieved.

After each stage $S_{1...M}$ we use a weighted cross-entropy loss to train our model, as described in Eq. 4. Here, y_t is the ground truth phase label and $\hat{y}_{mt}$ is the output prediction of each stage $m \in [1, M]$. The class weights w_c are calculated using median frequency balancing [20] to mitigate the imbalance between phases. Our TeCNO model is trained utilizing exclusively phase recognition labels without requiring any additional tool information.

$$\mathcal{L}_C = \frac{1}{M}\sum_{m}^{M}\mathcal{L}_{Cm} = -\frac{1}{M}\frac{1}{T}\sum_{m}^{M}\sum_{t}^{T} w_c y_{mt} \cdot log(\hat{y}_{mt}) \tag{4}$$

3 Experimental Setup

Datasets We evaluated our method on two challenging surgical workflow intra-operative video datasets of laparoscopic cholecystectomy procedures for the resection of the gallbladder. The publicly available Cholec80 [25] includes 80 videos with resolutions 1920×1080 or 854×480 pixels recorded at 25 frames-per-second (fps). Each frame is manually assigned to one of seven classes corresponding to each surgical phase. Additionally, seven different tool annotation labels sampled at 1fps are provided. The dataset was subsampled to 5fps, amounting to ∼92000 frames. We followed the split of [12, 17] separating the dataset to 40 videos for training, 8 for validation, and 32 for testing.

Cholec51 is an in-house dataset of 51 laparoscopic cholecystectomy videos with resolution 1920×1080 pixels and sampling rate of 1fps. Cholec51 includes seven surgical phases that slightly differ from Cholec80 and have been annotated by expert physicians. There is no additional tool information provided. 25 videos were utilized for training, 8 for validation and 18 for test. Our experiments for both datasets were repeated 5 times with random initialization to ensure reproducability of the results.

Model Training TeCNO was trained for the task of surgical phase recognition using the Adam optimizer with an initial learning rate of 5e-4 for 25 epochs. We report the test results extracted by the model that performed best on the validation set. The batch size is identical to the length of each video. Our method was

Table 1. Ablative testing results for different feature extraction CNNs and increasing number of stages for Cholec80. Average metrics over multiple runs are reported (%) along with their respective standard deviation.

	AlexNet			**ResNet50**		
	Acc	Prec	Rec	Acc	Prec	Rec
No TCN	74.40 ± 4.30	63.06 ± 0.32	70.75 ± 0.05	82.22 ± 0.60	70.65 ± 0.08	75.88 ± 1.35
Stage I	84.04 ± 0.98	79.82 ± 0.31	79.03 ± 0.99	88.35 ± 0.30	$\mathbf{82.44 \pm 0.46}$	84.71 ± 0.71
Stage II	85.31 ± 1.02	81.54 ± 0.49	79.92 ± 1.16	$\mathbf{88.56 \pm 0.27}$	81.64 ± 0.41	$\mathbf{85.24 \pm 1.06}$
Stage III	84.41 ± 0.85	77.68 ± 0.90	79.64 ± 1.60	86.49 ± 1.66	78.87 ± 1.52	83.69 ± 1.03

implemented in PyTorch and our models were trained on an NVIDIA Titan V 12GB GPU using Polyaxon[4]. The source code for TeCNO is publicly available[5].

Evaluation Metrics To comprehensively measure the results of the phase prediction we deploy three different evaluation metrics suitable for surgical phase recognition [5], namely Accuracy (Acc), Precision (Prec) and Recall (Rec). Accuracy quantitatively evaluates the amount of correctly classified phases in the whole video, while Precision, or positive predictive value, and Recall, or true positive rate, evaluate the results for each individual phase [22].

Ablative Testing To identify a suitable feature extractor for our MS-TCN model we performed experiments with two different CNN architectures, namely AlexNet [26] and ResNet50 [16]. Additionally we performed experiments with different number of TCN stages to identify which architecture is best able to capture the long temporal associations in our surgical videos.

Baseline Comparison TeCNO was extensively evaluated against surgical phase recognition networks, namely, PhaseLSTM [12], EndoLSTM [12] and MTRCNet [17], which employ LSTMs to encompass the temporal information in their models. We selected LSTMs over HMMs, since their superiority has been extensively showcased in the literature [14]. Moreover, MTRCNet is trained in an end-to-end fashion, while the remaining LSTM approaches and TeCNO focus on temporally refining already extracted features. Since Cholec51 does not include tool labels, EndoLSTM and MTRCNet are not applicable due to their multi-task requirement. All feature extractors for Cholec80 were trained for a combination of phase and tool identification, except for the feature extractor of PhaseLSTM [25], which requires only phase labels. The CNNs we used to extract the features for Cholec51 were only trained on phase recognition since no tool annotations were available.

[4] https://polyaxon.com/
[5] https://github.com/tobiascz/TeCNO/

Table 2. Baseline Comparison for Cholec80 and Cholec51 Datasets. EndoLSTM and MTRCNet require tool labels, therefore cannot be applied for Cholec51. The average metrics over multiple runs are reported (%) along with their respective standard deviation.

	Cholec80			**Cholec51**		
	Acc	Prec	Rec	Acc	Prec	Rec
0.12	69.22 ± 0.11	69.26 ± 0.05				
PhaseLSTM [12]	79.68 ± 0.07	72.85 ± 0.10	73.45 ± 0.12	81.94 ± 0.20	68.84 ± 0.11	68.05 ± 0.79
EndoLSTM [22]	80.85 ± 0.17	76.81 ± 2.62	72.07 ± 0.64	—	—	—
MTRCNet [17]	82.76 ± 0.01	76.08 ± 0.01	78.02 ± 0.13	—	—	—
ResNetLSTM [15]	86.58 ± 1.01	80.53 ± 1.59	79.94 ± 1.79	86.15 ± 0.60	70.45 ± 2.85	67.42 ± 1.43
TeCNO	$\mathbf{88.56 \pm 0.27}$	$\mathbf{81.64 \pm 0.41}$	$\mathbf{85.24 \pm 1.06}$	$\mathbf{87.34 \pm 0.66}$	$\mathbf{75.87 \pm 0.58}$	$\mathbf{77.17 \pm 0.73}$

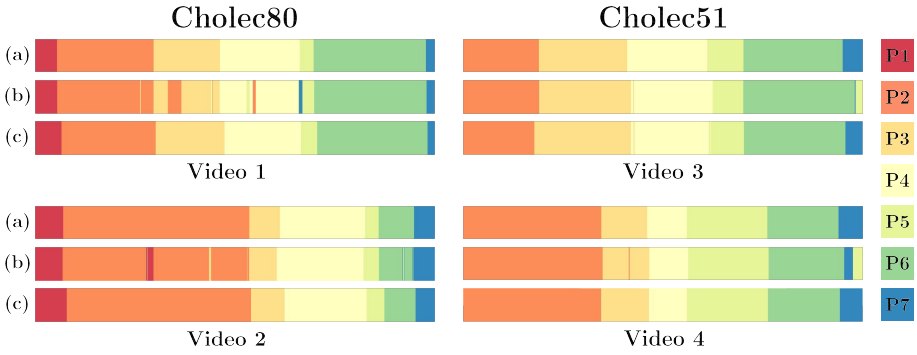

Fig. 2. Qualitative Results regarding quality of phase recognition for Cholec80 and Cholec51. (a) Ground Truth (b) ResNetLSTM Predictions (c) TeCNO Predictions. P1 to P7 indicate the phase label.

4 Results

Effect of Feature Extractor Architecture As can be seen in Table 1, ResNet50 outperforms AlexNet across the board with improvements ranging from 2% to 8% in accuracy. Regarding precision and recall, the margin increases even further. For all stages ResNet50 achieves improvement over AlexNet of up to 7% in precision and 6% in recall. This increase can be attributed to the improved training dynamics and architecture of ResNet50 [16]. Thus, the feature extractor selected for TeCNO is ResNet50.

Effect of TCN and Number of Stages Table 1 also highlights the substantial improvement in the performance achieved by the TCN refinement stages. Both AlexNet and ResNet50 obtain higher accuracy by 10% and 6% respectively with the addition of just 1 TCN Stage. Those results signify not only the need for temporal refinement for surgical phase recognition but also the ability of TCNs to improve the performance of any CNN employed as feature extractor, regardless

of its previous capacity. We can also observe that the second stage of refinement improves the prediction of both architectures across our metrics. However, Stage 2 outperforms Stage 3 by 1% in accuracy for AlexNet and 2% for ResNet50. This could indicate that 3 stages of refinement lead to overfitting on the training set for our limited amount of data.

Comparative Methods In Table 2 we present the comparison of TeCNO with different surgical phase recognition approaches that utilize LSTMs to encompass the temporal information in their predictions. PhaseLSTM [27] and EndoLSTM [27] are substantially outperformed by ResNetLSTM and TeCNO by 6% and 8% in terms of accuracy for both datasets respectively. This can be justified by the fact that they employ AlexNet for feature extraction, which as we showed above has limited capacity. Even though MTRCNet is trained in an end-to-end fashion, it is also outperformed by 4% by ResNetLSTM and 6% by TeCNO, which are trained in a two-step process. Comparing our proposed approach with ResNetLSTM we notice an improvement of 1-2% in accuracy. However, the precision and recall values of both datasets are substantially higher by 6%-10%. The higher temporal resolution and large receptive field of our proposed model allow for increased performance even for under-represented phases.

Phase Recognition Consistency In Fig. 2 we visualize the predictions for four laparoscopic videos, two for each dataset. The results clearly highlight the ability of TeCNO to obtain consistent and smooth predictions not only within one phase, but also for the often ambiguous phase transitions. Compared against ResNetLSTM, TeCNO can perform accurate phase recognition, even for the phases with shorter duration, such as P5 and P7. Finally, TeCNO showcases robustness, since Video 3 and 4 are both missing P1. However, the performance of our model does not deteriorate.

5 Conclusion

In this paper we proposed TeCNO, a multi-stage Temporal Convolutional Neural Network, which was successfully deployed on the task of surgical phase recognition. Its full temporal resolution and large receptive field allowed for increased performance against a variety of LSTM-based approaches across two datasets. Online and fast inference on whole video-sequences was additionally achieved due to causal, dilated convolutions. TeCNO increased the prediction consistency, not only within phases, but also in the ambiguous inter-phase transitions. Future work includes evaluation of our method on a larger number of videos from a variety of laparoscopic procedures.

Acknowledgements

Our research is partly funded by the DFG research unit 1321 PLAFOKON and ARTEKMED in collaboration with the Minimal-invasive Interdisciplinary

Intervention Group (MITI). We would also like to thank NVIDIA for the GPU donation.

References

1. 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*, vol. 1, no. 9, pp. 691–696, 2017.
2. A. Huaulmé, P. Jannin, F. Reche, J. L. Faucheron, A. Moreau-Gaudry, and S. Voros, "Offline identification of surgical deviations in laparoscopic rectopexy," *Artificial Intelligence in Medicine*, vol. 104, no. May 2019, 2020.
3. N. Padoy, "Machine and deep learning for workflow recognition during surgery," *Minimally Invasive Therapy and Allied Technologies*, vol. 28, no. 2, pp. 82–90, 2019.
4. 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.
5. N. Padoy, T. Blum, S. A. Ahmadi, H. Feussner, M. O. Berger, and N. Navab, "Statistical modeling and recognition of surgical workflow," *Medical Image Analysis*, vol. 16, no. 3, pp. 632–641, 2012.
6. G. Lecuyer, M. Ragot, N. Martin, L. Launay, and P. Jannin, "Assisted phase and step annotation for surgical videos," *International Journal of Computer Assisted Radiology and Surgery*, 2020.
7. 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, pp. 1089–1095, 2019.
8. I. Funke, S. T. Mees, J. Weitz, and S. Speidel, "Video-based surgical skill assessment using 3D convolutional neural networks," *International Journal of Computer Assisted Radiology and Surgery*, vol. 14, no. 7, pp. 1217–1225, 2019.
9. U. Klank, N. Padoy, H. Feussner, and N. Navab, "Automatic feature generation in endoscopic images," *International Journal of Computer Assisted Radiology and Surgery*, vol. 3, no. 3, pp. 331–339, 2008.
10. H. Al Hajj, M. Lamard, P. H. Conze, S. Roychowdhury, X. Hu, G. Maršalkaitė, O. Zisimopoulos, M. A. Dedmari, F. Zhao, J. Prellberg, M. Sahu, A. Galdran, T. Araújo, D. M. Vo, C. Panda, N. Dahiya, S. Kondo, Z. Bian, A. Vahdat, J. Bialopetravičius, E. Flouty, C. Qiu, S. Dill, A. Mukhopadhyay, P. Costa, G. Aresta, S. Ramamurthy, S. W. Lee, A. Campilho, S. Zachow, S. Xia, S. Conjeti, D. Stoyanov, J. Armaitis, P. A. Heng, W. G. Macready, B. Cochener, and G. Quellec, "CATARACTS: Challenge on automatic tool annotation for cataRACT surgery," *Medical Image Analysis*, vol. 52, pp. 24–41, 2019.
11. C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 9915 LNCS, pp. 47–54, 2016.
12. 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*, vol. 36, no. 1, pp. 86–97, 2017.
13. S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
14. G. Yengera, D. Mutter, J. Marescaux, and N. Padoy, "Less is More: Surgical Phase Recognition with Less Annotations through Self-Supervised Pre-training of CNN-LSTM Networks," 2018.
15. Y. Jin, Q. Dou, H. Chen, L. Yu, J. Qin, C. W. Fu, and P. A. Heng, "SV-RCNet: Workflow recognition from surgical videos using recurrent convolutional network," *IEEE Transactions on Medical Imaging*, vol. 37, no. 5, pp. 1114–1126, 2018.
16. K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," tech. rep.
17. Y. Jin, H. Li, Q. Dou, H. Chen, J. Qin, C. W. Fu, and P. A. Heng, "Multi-task recurrent convolutional network with correlation loss for surgical video analysis," *Medical Image Analysis*, vol. 59, 2020.
18. A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WAVENET: A GENERATIVE MODEL FOR RAW AUDIO," tech. rep.
19. Y. A. Farha and J. Gall, "MS-TCN: Multi-Stage Temporal Convolutional Network for Action Segmentation," tech. rep.
20. D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2650–2658, 2015.
21. T. Yu, D. Mutter, J. Marescaux, and N. Padoy, "Learning from a tiny dataset of manual annotations: a teacher/student approach for surgical phase recognition," 2018.
22. A. P. Twinanda, N. Padoy, M. J. Troccaz, and G. Hager, "Vision-based Approaches for Surgical Activity Recognition Using Laparoscopic and RBGD Videos," *Thesis*, no. Umr 7357, 2017.
23. A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional LSTM networks for improved phoneme classification and recognition," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 3697 LNCS, pp. 799–804, 2005.
24. A. Newell, K. Yang, and J. Deng, "Stacked Hourglass Networks for Human Pose Estimation," tech. rep.
25. A. P. Twinanda, D. Mutter, J. Marescaux, M. de Mathelin, and N. Padoy, "Single- and Multi-Task Architectures for Surgical Workflow Challenge at M2CAI 2016," pp. 1–7, 2016.
26. A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems*, vol. 25, 2012.
27. A. P. Twinanda, J. Marescaux, M. De Mathelin, and N. Padoy, "Single-and Multi-Task Architectures for Surgical Workflow Challenge at M2CAI 2016," tech. rep.