SUrgical PRediction GAN for Events Anticipation

Yutong Ban\textsuperscript{1,2}, Guy Rosman\textsuperscript{2}, Thomas Ward\textsuperscript{1}, Daniel Hashimoto\textsuperscript{1}, Taisei Kondo\textsuperscript{3}, Hidekazu Iwaki\textsuperscript{3}, Ozanan Meireles\textsuperscript{1}, and Daniela Rus\textsuperscript{2}

\textsuperscript{1} SAIL, Massachusetts General Hospital. 55 Fruit Street, Boston, MA, US
\textsuperscript{2} Distributed Robotics Laboratory, CSAIL, MIT. 32 Vassar St, Cambridge, MA, US
\textsuperscript{3} Olympus Corporation, Tokyo, Japan

Abstract. Comprehension of surgical workflow is the foundation upon which computers build understanding of surgery. In this work, we moved beyond just identification of surgical phases to predict future surgical phases and the transitions between them. We used a novel GAN formulation that sampled the future surgical phases trajectory conditioned, on past laparoscopic video frames, and compared it to state-of-the-art approaches for surgical video analysis and alternative prediction methods. We demonstrated its effectiveness in inferring and predicting the progress of laparoscopic cholecystectomy videos. We quantified the horizon-accuracy trade off and explored average performance as well as the performance on the more difficult, and clinically important, transitions between phases. Lastly, we surveyed surgeons to evaluate the plausibility of these predicted trajectories.

1 Introduction

The past decade has seen substantial research trying to build the foundation of surgical workflow comprehension for machines \cite{13,15,24,26}. These works usually focus on post-hoc analysis, trying to identify the current surgical phase from past video events alone. Little has been done to investigate predicting future events, and the research that has been published focuses on specific predictive tasks, such as remaining surgical time \cite{25}.

The ultimate goal of this research is to create machines capable of improving intra-operative patient care. To do so, they will need to augment the surgeon’s capabilities, which requires not only understanding of past surgical events but also the ability to anticipate and predict futures ones. This prediction capability must be more than the ability to predict specific narrow tasks, like remaining surgical time. More complete models that truly understand surgical workflow are needed and should have the generic capabilities to allow them to predict multiple tasks. This should preferably be done using existing models that describe the surgical process. The use of existing models is key for rapid implementation:
completely novel models will require the generation of new annotations and cost excessive man-hours to develop.

When aiming to create such a framework, we must remember the often non-linear nature of the surgical process. This non-linearity is especially prominent in more complex surgical cases that would benefit from automatic methods of risk reduction. Another important aspect is the uncertain, yet discrete, nature of the sequences involved – the uncertainty is both due to the limited field of view of the camera, as well as, our lack of knowledge as a third person observer of the surgeon’s intent and knowledge about the case.

We address this difficult prediction problem via a new model that uses an encoder-decoder predictor based on a discrete generative adversarial network (GAN [7]). This model can predict roll-outs of the surgical process in the form of discrete label sequences for operative phases over time. This model also emits the phase estimates over the past and current time phase, allowing analysis and prediction of the surgery within a unified, multitask framework.

**Contributions:** Our contributions are as follows:

1. We present a novel formulation and model for joint surgical process analysis and prediction.
2. We examine the use of a discrete GAN within that model.
3. We demonstrate results of the proposed model for analysis and prediction of Laparoscopic Cholecystectomy surgery. We show how the model surpasses existing approaches in both sequence prediction and transition detection tasks. We further quantify the subjective plausibility of the predictions based on a survey of surgeons.

**Related works:** Our work relates to several major topics of research. In surgical workflow understanding, significant effort has been devoted to analysis of the surgical process [1,13,15,24,26,27]. However, little effort has been devoted so far to prediction of future workflow with few notable exceptions, such as prediction of remaining surgical time as a regression task [25].

Sequence prediction has been common in autonomous driving and trajectory prediction, where predicting the trajectories of road users allows vehicles to interact and plan, or warn the driver of road risk. However, the trajectories predicted are in a continuous state space, and other concerns such as multiagent interactions, agent goals, and environmental context, lead to different designs structures [8,11,16,19].

Autoencoding and processing of discrete sequences has been prevalent in the natural language processing (NLP) literature [2,3,5,22] with different types of connectivity structures and tasks. Outside of NLP, prediction and completion of discrete sequences has been more limited in its applications, with notable examples in several fields of natural sciences such as chemistry and biology [4,6,14,29,30]. However, in these domains the signal to be analyzed is directly observed, unlike surgical phases, that are not directly observed in the video.
2 Methods

In this section we introduce in detail the proposed model. The surgery can be in one of \( N_P \) possible phases during the operation.

2.1 Problem formulation

Our goal is to predict the surgeon’s future actions. Instead of predicting the only possible phase path, we try to predict a distribution of possible phase sequences using GAN, limiting the prediction learning to a fixed past and future horizon, as is common in prediction literature. During surgery, a surgeon makes decisions based on the recently observed \( T_p \) video frames \( I = \{I_{t_0-T_p}, \ldots, I_{t_0-1}, I_{t_0}\} \). The model prediction is represented by \( y = \{y_{t_0+1}, y_{t_0+2}, \ldots, y_{t_0+T_f}\} \), where \( T_f \) is the number of frames the model is predicting into the future. At each time step, the model predicts the surgery phases \( y_t \in \{0, 1\}^{N_P} \), where phase labels are encoded as 1-hot vectors.

2.2 Auto-encoding Sequence Models

Several auto-encoding models have been used to predict sequential data, mostly based on GANs \([7][8][10]\) and conditional variational auto-encoders \([11][20]\). Our model follows GANs and includes a generator and a discriminator model denoted by \( G \) and \( D \) respectively. As an auto-encoder GAN, the generator includes an encoder module and decoder module. The discriminator includes a past encoder module and a future encoder module and is trained to distinguish whether a provided sequence is a fake sequence created by the generator or a real sequence from the data. Usually, an additional data term is added to ensure that the generator can regenerate real data sequences, often with a variety loss \([8][23]\).
2.3 Surgical Prediction GAN (SUPR-GAN)

Unlike standard sequence prediction problems, we have additional supervisory cues in the form of annotation labels, $\hat{Y}$. We leverage the annotation labels both as the generator data term but also to train the encoder into learning a sufficient representation for the sequence. The overview of the network is given in Fig. 1.

**Generator Encoder** The generator’s encoder takes the observations then encodes all the observed information into a single vector. Since the observations are sequential, Long short-term memory (LSTM) is thus applied for the recurrence. The encoding process can be formally written as:

$$e_t = CNN(I_t), \quad h_t = LSTM_{G,\text{enc}}(h_{t-1}, e_t), \quad y_t = PhaseDecoder(h_t)$$

Where CNN uses ResNet [9] as the backbone, and MLP are multiple-layer-perceptrons.

**Generator Decoder** The decoder leverages the information from the encoder, concatenating the random noise

$$y_{t-1} = PhaseDecoder(h_{t-1}), \quad h_t = LSTM_{G,\text{dec}}(h_{t-1}, y_{t-1})$$

at each time step $t$, the estimated variables can be obtained by different heads separately:

**Discriminator** The discriminator is composed of two encoders, one for each of the past and future sequences. Both encoders feed off phase vectors,

$$h_t = LSTM_{D,}(h_{t-1}, y_{t-1}). \quad \text{(1)}$$

The last state of the future encoder is fed through a discriminator head that emits a binary label – real or fake – for the sequence. The real represents the trajectory sampled from the real data, whereas fake represents the sample generated by the prediction model.

**Discrete GAN** As the surgery workflow which we are predicting has discrete sequences, a function which can convert the generator decoder output probabilities to discrete sequences is required. Gumbel-Softmax [12] layer is used after the generator decoder output. It is a differentiable layer so that it allows us to have discrete phase prediction samples input to the discriminator, without breaking the gradient.

**Loss** The loss function used for training is a combination of GAN loss and data loss. The GAN loss penalizes the whether the predicted trajectory is reasonable or not.

$$L_{dis}(y, \hat{y}) = \min_D \max_G V(G, D), \quad \text{(2)}$$

where $V(G, D)$ is the usual GAN loss, formally written as:

$$\min_D \max_G V(G, D) = \mathbb{E}_{x \sim p_{data}(x)} \log(D(x)) + \mathbb{E}_{z \sim p_z}(\log(1 - D(z))). \quad \text{(3)}$$
z is the generator noise sample and x represent sample from the data’s label sequence, Y. The data loss in GAN-based predictors is destined to ensure the prediction is not too far from the ground truth. We use a variety loss [23], which penalizes the distance between the ground truth labels and the most similar reconstructed sequence out of a set of $N_s = 5$ samples.

$$
L_{\text{rec}}(y, \hat{y}) = \min_{j=1}^{N_s} \sum_{t=t_0+1}^{t_0+T_f} d_L(Y_t^{(j)}, \hat{y}_t).
$$

$d_L(\cdot, \cdot)$ is a distance between the labels and prediction – cross-entropy as our sequences are discrete categories.

**Past Encoding Loss** Unlike domains where GAN-based predictors are decoding the raw signals (such as images and trajectory prediction), we are decoding annotated labels that are much more costly to obtain and are not the same as the encoded signal (images). This allows us to add an additional data term measuring how well the encoder recognizes the phase, even in past frames. This is similar to phase recognition costs and is expressed as:

$$
L_{\text{past}}(y, \hat{y}) = \sum_{t=t_0-T_p}^{t_0} d_L(y_t, \hat{y}_t),
$$

where we use cross-entropy loss as before.

During the experiment, the overall loss is $L = \omega_1 L_{\text{dis}} + \omega_2 L_{\text{rec}} + \omega_3 L_{\text{past}}$, where $\omega_1 = 0.6$ and $\omega_2 = 0.2$ and $\omega_3 = 0.2$.

### 3 Experiments

We now describe the experiments conducted with our model. We have tested both the objective measures of accuracy relating to the model’s predictive power, as well as the perceived plausibility of the predicted trajectories when gauged by clinical experts (surgeons).

**Datasets** The proposed model is evaluated on two large surgical video datasets. *Cholec80* Cholec80 is a public available dataset which contains 80 laparoscopic cholecystectomy. The dataset is divided into 40/40 split for training and testing. The dataset is annotated into 7 different phases.

*MGH200* MGH200 consists of 200 laparoscopic cholecystectomy videos, where 150 videos are used in training and 50 videos are employed for testing. The dataset is annotated into 12 surgical phases, which contain more variability and clinically meaningful phase transitions.

**Model Parameters and Training Strategy** In the experiment, the videos are re-sampled at 1 fps and input to the model. The number of the MoN samples is set to 10. During model training, the generator encoder is pre-trained with the surgical phase recognition task for 20 epochs. The pre-training is accomplished with the same dataset; therefore, no additional data are used. During GAN training, we use small epochs to train the generator and the discriminator in an iterative fashion, where the epoch size is 64 and the number of epoch is 2000.
Fig. 2. Examples of prediction results in MGH200 dataset. For each example, on the left, a diagram of the phases of the operation performed. The horizontal axis indicates the time, ranging from the past 15s to the future 15s. The vertical cyan bar indicates the “current” time point associated with the video image on the right side. Red horizontal segments indicate the ground truth trajectories. And the horizontal color bars (orange, blue, black) indicate the different samples predicted by the SUPR-GAN. (a) Video 20, frame 433 (b) video 41, frame 205 (c) video 20, frame 847, (d) video 47, frame 2535.

**Prediction settings** In the prediction, we use 15 seconds of the past video and have a prediction horizon of 15 seconds. The choice is made due to several reasons: (i) 15 seconds of the video segment in the past should contain the recent phase labels, and therefore rich information to predict future phase labels. (ii) The prediction horizon of 15 seconds is a reasonable length for the potential surgical applications, giving the surgeon and their team notice of potential predicted events. (iii) LSTMs are limited in their ability to numerically propagate information over large timescales [21], and to and cover a complex set of predictions as the prediction horizon increases [10,18].

**Evaluation Metrics** To evaluate the performance of the prediction models, we employ two metrics:

- **Per-transition accuracy:** Every time the ground truth transits to the new phase, if the new phase is predicted before it happened within a time window of $\delta$ seconds, we consider the transition to be well predicted. We set $\delta$ to 15.
- **Levenshtein distance:** Levenshtein distance [17] measures the minimum number of operations required to transform one sequence into the other. It is widely applied to NLP for comparing strings and to Biology for comparing DNA sequences. In our evaluation, we calculate the average Levenshtein distance between the prediction and the ground truth.
Table 1. The per-transition accuracy on Cholec80 Dataset

|                                | Constant Model | HMM | LSTM | SUPR-GAN (Proposed) |
|--------------------------------|----------------|-----|------|---------------------|
| Preparation                   |                |     |      |                     |
| Calot Triangle Dissection     | 7.5%           | 5.3%| 72.5%| 87.5%               |
| Clipping and Cutting          | 2.5%           | 42.5%| 82.5%| 57.5%               |
| Gallbladder Dissection        | 7.5%           | 35% | 70%  | 65%                 |
| Gallbladder Packaging         | 12.5%          | 10% | 50%  | 30%                 |
| Cleaning and Coagulation      | 42.8%          | 31.4%| 42.9%| 54.3%               |
| Gallbladder Retraction        | 85.0%          | 17.5%| 50%  | 30%                 |
| Overall                       | 26.0%          | 31.5%| 59.6%| 60.9%               |

Qualitative Results

The result examples of the proposed model, with different clinically meaningful cases are shown in Fig. 2. In the example (a), the model predictions align well with the ground truth during the transitions to 'Checkpoint 1' and 'Clip Cystic Duct'. Similarly, in the example (b), the transitions from 'Clip Cystic Duct' to 'Dissection of Calot’s Triangle' is well captured by the black prediction sample. Moreover, the model is able to predict the different possible future phases in the bifurcation areas (e.g. samples in orange and blue shows the model is predicting it is also possible to transit to 'Clip the Cystic Duct' if the cystic duct is not fully divided). Some other results are also shown in (c) and (d). The proposed model can not only give accurate prediction about the phase transitions, it can also provide diverse samples which cover the different possible transitions.

Quantitative Results

We compare the performance of the proposed model on two datasets with several baseline methods. Constant prediction model is a simple baseline using the last frame of the past encoding head to perform the prediction; its performance suffers in the transition areas. We also employ Hidden-Markov Model (HMM) as another baseline. Once the past phase encoding likelihoods are obtained, they are used as the observation for HMM. Baum-Welch algorithm [28] is used to estimate the HMM internal parameters. We also compared to a variant of the proposed model, which is the GAN generator trained with only data loss and past encoding loss, which is referred to as LSTM in Table 2. The experimental results on Cholec80 dataset are shown in Table 1 and the results on MGH200 are shown in Table 2. The SUPR-GAN has achieved best overall per-transition accuracy on both datasets. We further show the average Levenshtein distance between the predicted sample and the ground truth, The constant model is having a good performance measured with Levenshtein distance, especially when calculating over all the samples. This is due to the following main reasons (i) when there’s no transitions, the constant model makes the reasonable predictions. (ii) For the segments with transition areas, the constant model can still give correct prediction before the actual transition happens. However, after the transition occurs the prediction accuracy can drop to very low, which can be further illustrated by the low per-transition accuracy in Table 2. Compared with other methods, the proposed model has achieved good performance by having a high per-transition accuracy and maintaining a low Levenshtein distance.
### Transition to Phase

| Block | Port placement | Fundus retraction | Release GB peritoneum | Dissection of Calot’s triangle | Checkpoint 1 | Block 2 | Divide Cystic Artery | Divide Cystic Duct | Checkpoint 2 | Block 3 | Remove GB from liver bed | Bagging |
|-------|----------------|-------------------|-----------------------|-------------------------------|--------------|--------|---------------------|-------------------|--------------|--------|--------------------------|---------|
|       | 0%             | 0%                | 91.8%                 | 12.3%                         | 0.0%         |        | 1.8%                | 0%                | 4.4%         |        | 64.6%                    | 91.892% |
|       | 0%             | 0%                | 32.1%                 | 69.5%                         | 100%         |        | 12.7%               | 0%                | 13.3%        |        | 35.4%                    | 100%    |
|       | 0%             | 33.3%             | 67.9%                 | 61.9%                         | 33.3%        |        | 7.3%                | 60%               | 66.7%        |        | 81.9%                    | 83.8%   |
|       | 0%             | 0%                | 62.3%                 | 57.1%                         | 44.4%        |        | 25.5%               | 48.9%             | 57.8%        |        | 54.2%                    | 86.8%   |

|       |      |       |                       |                               |              |        |                     |                   |              |        |                        |         |
|-------|------|-------|-----------------------|-------------------------------|--------------|--------|---------------------|-------------------|--------------|--------|--------------------------|---------|
|       |      |       |                       |                               |              |        |                     |                   |              |        |                        |         |
|       |      |       |                       |                               |              |        |                     |                   |              |        |                        |         |

|       |      |       |                       |                               |              |        |                     |                   |              |        |                        |         |
|-------|------|-------|-----------------------|-------------------------------|--------------|--------|---------------------|-------------------|--------------|--------|--------------------------|---------|
|       |      |       |                       |                               |              |        |                     |                   |              |        |                        |         |
|       |      |       |                       |                               |              |        |                     |                   |              |        |                        |         |

### Table 2. The per-transition accuracy on MGH200 Dataset

| Metrics                  | Constant Model | HMM | LSTM | SUPR-GAN(Proposed) |
|--------------------------|----------------|-----|------|--------------------|
| LD (transitions)         | 9.53           | 11.67 | 9.26 | 9.15               |
| LD (Overall)             | 3.47           | 13.15 | 4.27 | 3.36               |

### Table 3. The Levenshtein distance (LD) on MGH200 Dataset (the lower the better)

**Surgeon evaluation** In addition to objective measures, we have also compared our predictions to surgeon’s perception of future events. We have curated 20 surgical video segments of 15 seconds, sampled around transitions in the surgery. We have divided the segments into 4 groups, and asked 15 surgeons to view them. For each example, we provide the surgeon annotators 3 future trajectories, where two of them are generated by the model and one is the ground truth. We asked the surgeons to pick the most plausible one. From the results collected from the surgeons, 47% of the examples the surgeons didn’t find the true trajectories, which further states that our model is able to make reasonable future surgical workflow predictions.

### 4 Conclusions

In this paper, we proposed a prediction framework for surgical workflow based on a discrete encoder–decoder GAN. Our evaluation on objective metrics as well as study of perceived plausibility demonstrate the effectiveness of the approach, and suggest several directions to extend this approach to additional predictive tasks and more complete exploration of video-based surgical process prediction and its relation to the surgical mindset as an statistical inference problem of both analytical and practical value.
References

1. Y. Ban, G. Rosman, T. Ward, D. Hashimoto, T. Kondo, O. Meireles, and D. Rus. Aggregating long-term context for learning surgical workflows. *arXiv preprint arXiv:2009.00681*, 2020.

2. S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

3. K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax, Semantics and Structure in Statistical Translation*, page 103, 2014.

4. Z. Costello and H. G. Martin. How to hallucinate functional proteins. *arXiv preprint arXiv:1903.00458*, 2019.

5. A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In *Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2*, pages 3079–3087, 2015.

6. R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

7. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

8. A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2255–2264, 2018.

9. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

10. X. Huang, S. G. McGill, J. A. DeCastro, L. Fletcher, J. J. Leonard, B. C. Williams, and G. Rosman. Diversitygan: Diversity-aware vehicle motion prediction via latent semantic sampling. *IEEE Robotics and Automation Letters*, 5(4):5089–5096, 2020.

11. B. Ivanovic, K. Leung, E. Schmerling, and M. Pavone. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *IEEE Robotics and Automation Letters*, 6(2):295–302, 2020.

12. E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

13. 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*, 37(5):1114–1126, 2017.

14. A. Kadurin, A. Aliper, A. Kazennov, P. Mamoshina, Q. Vunhaelen, K. Khrabrov, and A. Zhavoronkov. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 8(7):10883, 2017.

15. F. Lalys and P. Jannin. Surgical process modelling: a review. *International journal of computer assisted radiology and surgery*, 9(3):495–511, 2014.

16. N. Lee, W. Choi, P. Vernaza, C. B. Choy, P. H. Torr, and M. Chandraker. Desire: Distant future prediction in dynamic scenes with interacting agents. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 336–345, 2017.
17. V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

18. T. Phan-Minh, E. C. Grigore, F. A. Boulton, O. Beijbom, and E. M. Wolff. Covernet: Multimodal behavior prediction using trajectory sets. pages 14074–14083, 2020.

19. N. Rhinehart, R. McAllister, K. Kitani, and S. Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2821–2830, 2019.

20. K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems*, 28:3483–3491, 2015.

21. I. Sutskever. *Training recurrent neural networks*. PhD thesis, 2013.

22. K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1556–1566, 2015.

23. L. A. Thiede and P. P. Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9954–9963, 2019.

24. 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(1):86–97, 2016.

25. 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*, 38(4):1069–1078, 2018.

26. M. Volkov, D. A. Hashimoto, G. Rosman, O. R. Meireles, and D. Rus. Machine learning and coresets for automated real-time video segmentation of laparoscopic and robot-assisted surgery. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 754–759. IEEE, 2017.

27. T. M. Ward, D. A. Hashimoto, Y. Ban, D. W. Rattner, H. Inoue, K. D. Lillemoe, D. L. Rus, G. Rosman, and O. R. Meireles. Automated operative phase identification in peroral endoscopic myotomy. *Surgical Endoscopy*, pages 1–8, 2020.

28. L. R. Welch. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter*, 53(4):10–13, 2003.

29. K. K. Yang, Z. Wu, and F. H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.

30. L. Yu, W. Zhang, J. Wang, and Y. Yu. Segan: Sequence generative adversarial nets with policy gradient. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.