This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2022.3156856, IEEE Robotics and Automation Letters

SUPR-GAN: SUrgical PRediction GAN for Event Anticipation in Laparoscopic and Robotic Surgery

Yutong Ban[1,2], Guy Rosman[1,2], Jennifer A. Eckhoff[2], Thomas M. Ward[2], Daniel A. Hashimoto[2]
Taisei Kondo[3], Hidekazu Iwaki[3], Ozanan R. Meireles[2] and Daniela Rus[1]

***Abstract*—Comprehension of surgical workflow is the foundation upon which artificial intelligence (AI) and machine learning (ML) holds the potential to assist intraoperative decision making and risk mitigation. In this work, we move beyond mere identification of past surgical phases, into prediction of future surgical steps and specification of the transitions between them. We use a novel Generative Adversarial Network (GAN) formulation to sample future surgical phases trajectories conditioned on past video frames from laparoscopic cholecystectomy (LC) videos and compare it to state-of-the-art approaches for surgical video analysis and alternative prediction methods. We demonstrate the GAN formulation's effectiveness through inferring and predicting the progress of LC videos. We quantify the horizon-accuracy trade-off and explored average performance, as well as the performance on the more challenging, and clinically relevant transitions between phases. Furthermore, we conduct a survey, asking 16 surgeons of different specialties and educational levels to qualitative evaluate predicted surgery phases.**

I. Introduction

Surgical artificial intelligence (AI) ultimately aims to create machines capable of improving patient care. Whilst surgical AI and machine learning (ML) research advances, major focus has been placed on building the foundation for machines to comprehend surgical workflow [25], [31], [35], [49], [52], [53], [55], [56]. Investigations mostly concentrate on post-hoc analysis, trying to identify the current surgical phase from past video events alone. However, in order to foresee and prevent surgical complications, machines must be able to predict future events. Previous research has neglected prediction of future events or operative phases, only briefly reporting on specific predictive tasks, such as remaining surgery time [50] and tool anticipation [41]. The ability to predict multiple, non-sequential, future possible phases is especially prominent in more complex surgical cases where differences in subsequent phases could lead to drastically different outcomes. Data-driven modeling of the surgical workflow and automatic methods of risk prediction during such procedures could yield a tremendous benefit for both the patient and surgeon. With the advances in robotic surgery and synergistic integration of computer vision, AI augmented surgery may in the far future translate the potential of advanced driving assistance systems to the operating room [16]. Phase prediction and early hazard detection promises great potential for the improvement of patient outcomes, especially when paired with surgical robotic innovation ([15], ch. 12.)

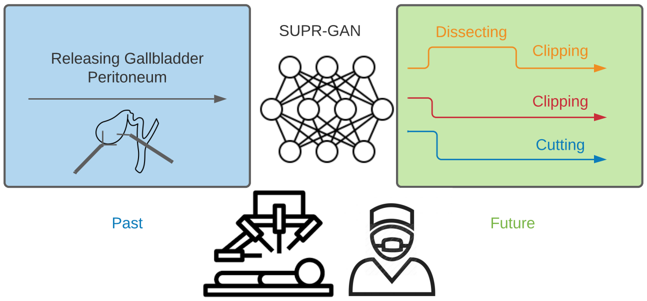


Fig. 1. Model overview: we predict a distribution of alternative future surgical phase sequences based on past surgical video frames.

In order to provide more general ability to predict future events during ongoing surgery, we need more capable models that truly understand surgical workflow. Such models should be able to handle multiple inference tasks and account for the complexity and actions throughout a surgical procedure. They should comprehend discrete characteristics of surgical procedures, such as abrupt transitions between phases, surgeons' decisions, and transitions in clinically-meaningful variables. Their predictive capability should extend beyond the mere ability to predict specific narrow tasks such as remaining surgical time, accounting for the diversity of phenomena that occurs during surgery. Finally, they should handle the uncertain nature of the sequences involved, both due to the limited field-of-view and the third party observer's lack of theory-of-mind of surgeon's intent and actions context, as well as the inherent uncertainty of human decision making – we expect predictive models to afford probabilistic reasoning that robustly handles such phenomena.

In this work, we focus on surgical phase prediction in laparoscopic cholecystectomy (LC) - removal of the gallbladder. With over 500,000 laparoscopic and robotic cases performed in the US annually, the cholecystectomy is among the most frequent surgical procedures. Although routinely performed, common bile duct injuries (with an incidence of up to 3%

Manuscript received: September 9, 2021; Revised January 6, 2022; Accepted February 3, 2022.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers comments.

This work was supported by research funding from Olympus Corporation.

[1] Distributed Robotics Laboratory, CSAIL, MIT. 32 Vassar St, Cambridge, MA, US. {yban, rosman, rus@csail.mit.edu}

[2] SAIIL, Massachusetts General Hospital. 55 Fruit Street, Boston, MA, US. {jeckhoff, tmward, dahashimoto, ozmeireles@mgh.harvard.edu}

[3] Olympus Corporation, Tokyo, Japan. {taisei.kondo2, hidekazu.iwaki@olympus.com}

Digital Object Identifier (DOI): see top of this page.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2022.3156856, IEEE Robotics and Automation Letters

in LC) present a serious and frequent complication with major, potentially life threatening consequences for the patient [39]. Therefore, early and advanced recognition of undesirable deviations from the routine surgical course with prediction and avoidance of adverse events like bile duct injuries holds great risk mitigation potential and can improve patient outcome significantly. The prediction horizon should be sufficiently long to provide time to take action in order to prevent complications resulting from the immediate actions of the surgeon, such as accidental injury of the common bile duct or right hepatic artery. In response to consultation of a panel of board certified surgeons, a prediction horizon of 15 seconds was chosen for this work. This provides sufficient time to foresee future steps and potentially recruit further expertise and assistance in complicated cases. A longer prediction horizon may distract the surgeon from the standardized surgical workflow.

We address this prediction problem via a novel model that uses an encoder-decoder predictor based on a discrete generative adversarial network (GAN [13]). This model, illustrated in Figure 1, can predict roll-outs of the surgical process in the form of discrete label sequences for operative phases over time, by observing the recent laparoscopic video frames. Also, the model 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:

1) We define and explore a novel, yet important problem of surgical prediction and jointly estimate the phase analysis.
2) We examine models for surgical process prediction, including a new discrete GAN predictor - SUrgical PRediction GAN (SUPR-GAN).
3) We demonstrate results of the proposed model for analysis and prediction of LC. 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.

II. RELATED WORKS

Our work relates to several major topics in current AI/ML and surgical prediction research. Significant effort has been made in surgical workflow analysis. [49] was one of the first to propose deep-learning architecture to address the problem by applying a convolutional neural network (CNN). [25] first introduced CNN-LSTMs for this task, nowadays considered standard. Additional research explored the applications to more complicated procedures, such as sleeve gastrectomy in [52] and to the peroral endoscopic myotomy (POEM) in [54]. Moreover, due to the potentially lengthy duration of the procedure, [2] proposed to aggregate the long-term temporal dependencies with statistic features so that the model can capture global surgical information. Similarly, [26] proposed a temporal memory relation network to capture multi-scale temporal patterns.

Aside from phase recognition, tool and task identification [36], [49] is another active field in the surgical AI domain. A CNN was successfully applied to capsule endoscopy to detect endoluminal lesions in the bowel [45] and colour descriptors have been used to assess intraoperative bleeding [11]. [51] applied a neural network classifier to evaluate kinematic data from robotic surgery and distinguish between more and less experienced surgeons. However, little effort has been devoted to prediction of future workflow with a few notable exceptions, such as prediction of remaining surgical time as a regression task [50], or the relatively unimodal next action [37]. Retrospective automated analysis of intraoperative adverse events (bleeding, cautery injury) has been performed, emphasising the significant impact of such events in terms of increased morbidity, mortality, and hospital stay [57]. Whilst these adverse events most commonly arise from preventable errors, little research has been conducted on the prospective prediction of subsequent intraoperative phases and events/errors to enhance surgical outcome.

In context-aware systems in computer assisted interventions (CA-CAI), using both real-time visual and instrumental information has been shown to increase accuracy in surgical workflow identification [10], aiding OR situational awareness. Teaching surgical trainees technical skills, comprehension of surgical workflow, and decision making in a protected setting (without potential complications to the patient) holds great value to improve clinical education for surgeons. As shown by [18] simulation-based, video based learning and retrospective action analyses significantly improve technical as well as procedural surgical skills.

Sequence and trajectory prediction has been commonly applied to autonomous driving, where forecasting of road users' future actions allows vehicles to interact, plan and warn the driver of road risk [20]. However, the trajectories predicted are in a continuous state space. Other concerns, such as multi-agent interactions, agent goals, and environmental context, lead to different design structures [14], [23], [32], [40]. Moreover, GANs can be found in other prediction applications, such as motion prediction [3], [29], body-pose prediction [61], speech signal prediction [27]. However, they often merely predict continuous signals and hence limit the discrete aspect to either a discrete vector [23] or a mixture model [9], [21], save for a few exceptions [22], [34].

Auto-encoding and processing of discrete sequences, with different types of connectivity structures and tasks, has been prevalent in the natural language processing (NLP) literature [4], [6], [8], [44], [46]. 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. [7], [12], [28], [59], [60]. However, in these domains, often the signal of interest is directly observed, unlike surgical phases that are not directly observed in the video.

Overall GAN approach has several advantages in a surgical context: (i) GAN training w/ variety loss encourages diverse predictions [14], that are important for workflow prediction at

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2022.3156856, IEEE Robotics and Automation Letters

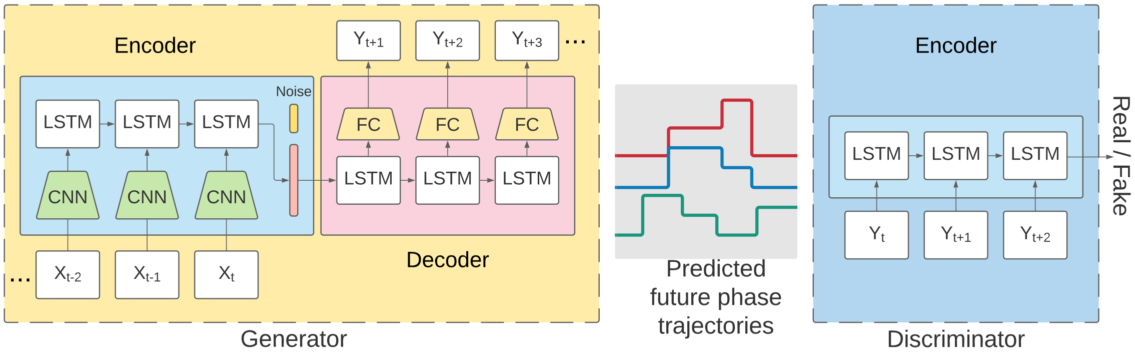


Fig. 2. Overview of the proposed SUPR-GAN model. Our model includes a CNN-LSTM-based encoder and a discrete decoder.

decision-making junction during surgical procedures. (ii) The GAN model can be easily modified for discrete sequences [30] as surgical phases and intra-operative events are often discrete labels. (iii) It extends gracefully to the prediction of multiple complex entities, as needed to explain surgical processes at adverse events. Therefore we consider GAN to be the most appropriate framework to address this surgical prediction question.

III. METHODS

In this section we elaborate the research question and the detailed design of the proposed prediction model.

A. Problem formulation

A surgeon can currently be performing one of N_P possible phases during an operation. Our goal is to predict the surgeon's future actions. Instead of predicting the only possible phase trajectory, we try to predict a distribution of possible phase trajectories 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 $\mathbf{I} = \{\mathbf{I}_{t_0-T_p}, \dots, \mathbf{I}_{t_0-1}, \mathbf{I}_{t_0}\}$. The model prediction is represented by $\mathbf{y} = \{\mathbf{y}_{t_0+1}, \mathbf{y}_{t_0+2}, \cdots \mathbf{y}_{t_0+T_f}\}$, where T_f is the number of frames the model is predicting into the future. At each time step, the model predicts the surgical phases $\mathbf{y}_t \in \{0,1\}^{N_P}$, where phase labels are encoded as 1-hot vectors.

B. Auto-encoding Sequence Models

Several auto-encoding models have been used to predict sequential data, mostly based on GANs [13], [14], [19] and conditional variational auto-encoders [23], [42]. Our model follows GANs and includes a generator and a discriminator model denoted by **G** and **D** respectively. The generator includes an encoder and a decoder. The discriminator includes a past encoder and a future encoder. It is trained to distinguish whether a presented trajectory is a fake trajectory created by the generator, or whether it is a real trajectory from the data. Usually, an additional data term is added to ensure that the generator can regenerate real data trajectories, often with a variety loss [14], [47].

C. Surgical Prediction GAN (SUPR-GAN)

Unlike standard sequence prediction problems, we have additional supervisory cues in the form of annotation labels, $\hat{\mathbf{Y}}$. The overview of the network is given in Fig. 2.

Generator Encoder The generator's encoder utilizes the observations and 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:

$$\begin{aligned} \boldsymbol{e}_I &= CNN(\mathbf{I}_t) \\ \boldsymbol{h}_t &= LSTM_{G,\mathbf{enc}}(\boldsymbol{h}_{t-1}, \boldsymbol{e}_I) \\ \mathbf{y}_t &= PhaseHead(\boldsymbol{h}_t) \end{aligned} \tag{1}$$

Where CNN uses ResNet [17] as the backbone, and $\boldsymbol{h}$ represents the hidden state of the LSTM. Both the encoder and the decoder hidden state share the same dimension $H = 32$. The $PhaseHead$ is a fully-connected layer which generates the phase probabilities.

Generator Decoder The decoder leverages the information from the encoder. While initializing the hidden state of the decoder, we simply take the last encoder hidden state then concatenate it with a random noise vector. A fully-connected layer is used to map the vector to size H.

$$\begin{aligned} \mathbf{y}_{t-1} &= PhaseHead(\boldsymbol{h}_{t-1}) \\ \boldsymbol{h}_t &= LSTM_{G,\mathbf{dec}}(\boldsymbol{h}_{t-1}, \mathbf{y}_{t-1}) \end{aligned} \tag{2}$$

at each time step t, the estimated variables can be obtained by phase emission 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,

$$\boldsymbol{h}_t = LSTM_{D,\cdot}(\boldsymbol{h}_{t-1}, \mathbf{y}_{t-1}). \tag{3}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/LRA.2022.3156856, IEEE Robotics and Automation Letters

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 [24] 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.

$$\mathcal{L}_{dis}(\mathbf{y}, \hat{\mathbf{y}}) = \min_{D} \max_{G} V(G, D), \tag{4}$$

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

$$\min_{D} \max_{G} V(G, D) = \tag{5}$$
$$\mathbf{E}_{x \sim p_{data}(x)} \log(D(x)) + \mathbf{E}_{z \sim p_z(z)}(\log(1 - D(G(z))).$$

z is the generator noise sample and x represent sample from the data's label sequence $\hat{\mathbf{Y}}$. 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 [47], which penalizes the distance between the ground truth labels and the most similar reconstructed sequence out of a set of $N_s = 10$ samples.

$$\mathcal{L}_{\text{rec}}(\mathbf{y}, \hat{\mathbf{y}}) = \min_{j=1}^{N_s} \sum_{t=t_0+1}^{t_0+T_f} d_L(\mathbf{y}_t^{(j)}, \hat{\mathbf{Y}}_t). \tag{6}$$

$d_L(\cdot, \cdot)$ 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:

$$\mathcal{L}_{\text{past}}(\mathbf{y}, \hat{\mathbf{y}}) = \sum_{t_0 - T_p}^{t_0} d_L(\mathbf{y}_t, \hat{\mathbf{y}}_t), \tag{7}$$

where we use cross-entropy loss as before.

During the experiment, the overall loss is

$$\mathcal{L} = \omega_1 \mathcal{L}_{dis} + \omega_2 \mathcal{L}_{rec} + \omega_3 \mathcal{L}_{past} \tag{8}$$

where in practice $\omega_1 = 0.6$ and $\omega_2 = 0.2$ $\omega_3 = 0.2$.

IV. EXPERIMENTS

In the following we test both 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).

A. Datasets

We experiment on two large-scale surgical video datasets. *Cholec80* [49] is a publicly available dataset which contains 80 videos of LC. The dataset is divided into equal subsets for training and testing with 40 videos each. The dataset is annotated into 7 different phases.
MGH200 [2] consists of 200 LC videos, where 150 videos are used for training and 50 for testing. The dataset is annotated into 12 surgical phases, which is more granular than the Cholec80 datasets. MGH200 also contains more variability and clinically meaningful phase transitions.

B. Model Parameters and Training Strategy

In the experiment, the videos are re-sampled at 1 fps and fed to the model. During model training, the generator encoder is pre-trained in surgical phase recognition for 20 epochs. The pre-training is accomplished using the same dataset. Therefore, no additional data is used. During GAN training, we iteratively train the generator and discriminator, using small epochs, where the epoch size is 64 and the number of epochs is 2000. We used an Adam optimizer with a learning rate of 10^{-4}.

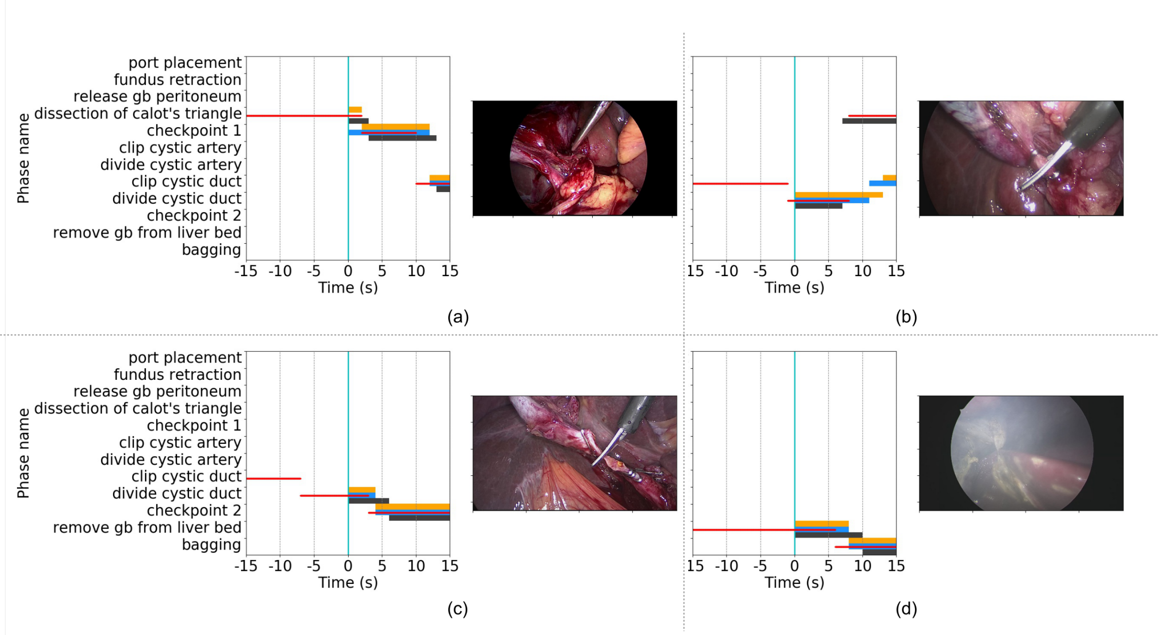
C. Prediction settings

Throughout prediction, we use the past 15 seconds to predict the upcoming future 15 seconds. This was determined for several reasons: (i) the past 15 seconds of the video segment should contain sufficient information about the operative phase to predict the future phase; (ii) a prediction horizon of 15 seconds into the future is an adequate time period for the surgeon to assess, intervene, and prevent potential adverse events; (iii) LSTMs are limited in their ability to numerically propagate information over large timescales [43] and insufficient in covering a complex set of predictions as the prediction horizon increases [19], [38]. In IV-G, we also compared the effect of using different prediction horizons.

D. Evaluation Metrics

To evaluate the prediction models, we employ two metrics:

- *Per-transition accuracy:* Every time the ground truth transits to a new phase, if said new phase is predicted correctly within δ seconds, we consider the transition to be well predicted. We set δ to 15 seconds.
- *Levenshtein distance:* Levenshtein distance (LD) [33] measures the minimum number of operations required to transform one sequence into another. It is widely applied to NLP for comparing strings and used to compare DNA sequences in biology [1]. In our evaluation, we calculate the average Levenshtein distance between the prediction and the ground truth.



(a)

(b)

(c)

(d)

Fig. 3. Examples of prediction results in MGH200 dataset. For each example, on the left, a diagram of the phases of the operation was shown. The horizontal axis indicates the time, ranging from the past 15s to future 15 seconds. 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, which are randomly chosen from the predicted samples. (a) Video 20, frame 433 (b) video 41, frame 205 (c) video 20, frame 847, (d) video 47, frame 2535.

Transition to Phase	Constant Model	HMM	Ours w/o Dis.	SUPR-GAN (Ours)
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%	17.5%	50%	30%
Overall	26.0%	31.5%	59.6%	**60.9%**

TABLE I
THE PER-TRANSITION ACCURACY ON CHOLEC80 DATASET. OUR PREDICTOR SIGNIFICANTLY OUTPERFORMS THE BASELINES.

E. Qualitative Results

The exemplary results of the proposed model are shown in Fig 3. In example (a), the model predictions align with the ground truth during the transitions between 'Checkpoint 1' and 'Clip Cystic Duct'. Similarly, in the example (b), the transition from 'Clip Cystic Duct' to 'Dissection of Calot's Triangle' is well captured by the black prediction trajectory. Moreover, the model is able to detect the different possible future phases in more challenging and questionable phase transitions (e.g.in Figure 3 (b): the black trajectory shows the model is able to transit back from 'Clip the Cystic Duct' to further 'Dissection of the Calot's Triangle' as a reasonable surgical action). Other examples, confirming model and ground truth accordance, are shown in (c) and (d). The proposed model can not only give accurate prediction about the phase transitions, it can also predict alternative trajectories and therefore cover various possible transitions.

F. 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 [58] 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 Ours w/o Dis. in Table II. The experimental results on Cholec80 dataset are shown in Table I and the results on MGH200 are shown in Table II. The SUPR-GAN has achieved best overall per-transition accuracy on both datasets, as measured by the same variety loss used for training. In addition to the prediction module, the proposed SUPR-GAN also contains a past phase recognition module, the past phase recognition accuracy is 82.3%. For future predictions, 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 mainly due to the following reasons: (i) when there's no transitions, the constant model makes the reasonable predictions; (ii) For the

	Transition to Phase	Constant Model	HMM	Ours w/o Dis.	SUPR-GAN(Ours)
Block 1	Port placement	0%	0%	0%	100%
	Fundus retraction	0%	0%	83.3%	83.3%
	Release GB peritoneum	91.8%	32.1%	67.9%	62.3%
	Dissection of Calot's triangle	12.3%	69.5%	61.9%	57.1%
	Checkpoint 1	0.0%	100%	33.3%	44.4%
Block 2	Clip Cystic Artery	1.8%	12.7%	7.3%	25.5%
	Clip Cystic Duct	2.1%	8.5%	29.8%	53.2%
	Divide Cystic Artery	0%	39.6%	11.3%	47.2%
	Divide Cystic Duct	0%	0%	60%	48.9%
	Checkpoint 2	4.4%	13.3%	66.7%	57.8%
Block 3	Remove GB from liver bed	64.6%	35.4%	81.9%	54.2%
	Bagging	91.8%	100%	83.8%	86.8%
	Overall	15%	42%	50.4%	**53.5%**

TABLE II

THE PER-TRANSITION ACCURACY ON MGH200 DATASET. THE SUPR-GAN OUT-PERFORMS THE OTHER METHODS IN OVERALL ACCURACY.

Metrics	Constant Model	HMM	Ours w/o Dis.	SUPR-GAN (Ours)
LD (transitions)	9.53	11.67	9.26	**9.15**
LD (Overall)	3.47	13.15	4.27	**3.36**

TABLE III

THE LEVENSHTEIN DISTANCE (LD) ON MGH200 DATASET (THE LOWER THE BETTER). SUPR-GAN ACHIEVED THE LOWEST LD IN BOTH OVERALL AND TRANSITION SETTINGS.

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 significantly, which can be further illustrated by the low per-transition accuracy in Table II. Compared to other methods, the proposed model has achieved good performance by having a high per-transition accuracy and maintaining a low Levenshtein distance.

Visualizing the predictions, the discriminator encourages a diversity of predictions, while keeping the predictions reasonable. In certain operative phases the surgeon may have freedom of choice on their next phase (e.g. clipping of cystic artery first or duct first or further dissection of the hepatocystic triangle). The discriminator encourages such alternate approaches, resulting in a higher accuracy. However, sequences, where steps are more deterministic and unique ((e.g. remove GB from liverbed), the discriminator may introduce some noise. We note the relative small average performance increase when adding the discriminator compared variety-loss only training. This phenomena has been observed in other prediction tasks [14], [47]. Yet the increase is still significant – the change in LD gave a p-value of 0.029 using a pairwise t-test on the per-video scores.

G. Influence of the Sequence Horizons

The topic of past and future horizon length is inherent to problem formulations in GAN-based temporal analysis [5] We evaluate the performance of the proposed method with different prediction lengths, as displayed in TABLE IV. Since LD distance is proportional to the prediction length, we normalize it to the length of 15 seconds. The table shows, that with increased prediction horizon the prediction performance drops, from 2.99 with prediction horizon of 10 seconds to 7.8 with prediction horizon of 45 seconds. However, this does not decrease significantly at the clinically chosen prediction horizon of 15 seconds (see Sec I). We also note the degradation of performance with fewer past frames supplied to the model, as the model lacks context information.

Past steps	Future steps	CM	HMM	Ours w/o Dis.	Ours
5	15	11.6	14.81	5.48	4.28
10	15	7.89	9.37	4.61	3.82
15	10	7.15	13.07	3.9	2.99
15	15	3.47	13.15	4.27	3.36
15	30	9.24	13.31	5.31	3.95
15	45	16.53	13.9	5.51	7.8

TABLE IV

THE NORMALIZED LD OF THE DIFFERENT FUTURE/PAST HORIZONS ON MGH200 DATASET (LOWER LD IS BETTER). AS THE PREDICTION HORIZON INCREASES OR THE PAST HORIZON IS REDUCED, THE PERFORMANCE OF THE MODEL DROPS.

H. Surgeon Survey on Phase Identification and Prediction

To augment the objective measures used and verify the realism of the model's phase prediction, we conducted a survey among a total of 16 surgeons of different educational stages comparing the surgeons' and model prediction. The participants were each presented with 5 out of 20 randomly selected video segments of a LC, 15 seconds each, where the video was stopped at a certain point. The participants were asked to identify the currently performed surgical step. In addition, they were asked to estimate the time remaining for that surgical task until the surgeon proceeds to the next upcoming step. Furthermore participants were asked to evaluate three different possible future trajectories from said point and rank them in terms of likeliness and plausibility to predict the next upcoming surgical step. Two of the presented future trajectories were chosen uniformly from trajectories generated by the model whilst one presented the ground truth – in a broad sense, forming a variant of an *imitation game* [48], as participants distinguish between a real and a computationally-created trajectory.

Firstly to note that the surgeons' future trajectory classification accuracy varied among participating surgeons and was associated with professional and educational level. Overall, $36,67\%$ of surgical faculty, 33.33% of fellows and $53,33\%$ of residents incorrectly predicted the future surgical step. Overall only 53.3% of questionnaire answers correctly predicted the future operative step according to the previously determined ”ground truth” (see Fig. 4). Furthermore, the surgeons' estimation of time until the next operative step displayed a high variance, not linked to experience or type of step, indicating that the estimation of remaining time for future tasks is obviously not a trivial exercise.

Upon a manual inspection of the 20 presented videos and predicted trajectories by a surgeon, some divergence between the model trajectories and ground truth was observed; but only one model trajectory appeared to be entirely unreasonable, displaying impossible surgical steps. This can be seen in Fig. 5 (a) , where yellow (removal of the gallbladder from the liver bed) represents the ground truth, the model proposed trajectories display surgical steps that chronologically would be performed before this (clipping and dissection). We note that in this scenario our diversity-enhancing loss term resulted in heterogenous samples even though the posterior distribution should not have been diverse, and this can be mitigated by better accuracy-coverage trade-offs.

In most videos, the model proposed trajectories differing from the ground truth represented feasible alternative surgical approaches. An example can be seen in Fig. 5 (b): the ground truth is displayed by the blue trajectory (dissection of calot's triangle), although the yellow (clip cystic duct) and black (clip cystic artery) differ entirely from the ground truth, they still represent plausible courses of surgical action. Here only 17% of surgeons selected the correct answer in blue, which is consistent with the heterogenity of surgeons selections in other cases of multiple plausible future steps of the procedure. This shows that the model is not only able to rule out unrealistic future trajectories, but also display multiple alternative, plausible approaches. Overall the model predictions and surgeons' answers aligned well with each other with an accuracy of 53.5% in SUPR-GAN prediction and 53.3% surgeons' choice of suspected ground truth.

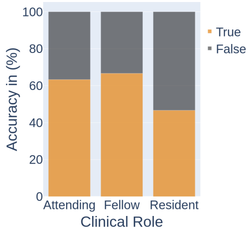


Fig. 4. Surgeon survey: percentage of correct and incorrect phase prediction according to clinical role and experience.

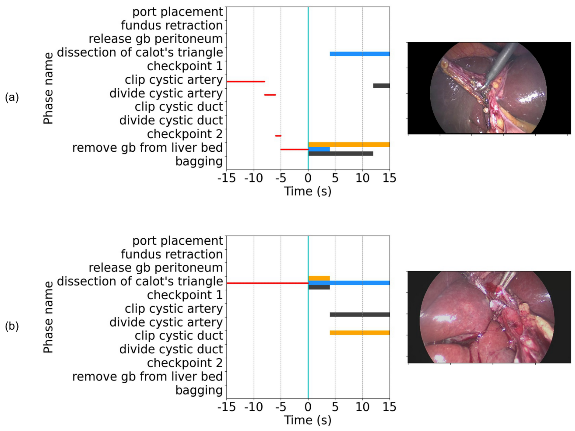


Fig. 5. Examples in the survey. (a) A scenario with unrealistic predicted trajectories – Yellow represents the ground truth, blue and black show the model predictions. The predictions are unrealistic in the sense that the temporal order of the phases would not occur in an actual surgery. (b) Model proposed trajectories in surgeon survey illustrating alternative future surgical paths. – Blue represents the ground truth, black and yellow show the model predictions.

V. CONCLUSIONS

In this paper, we proposed a prediction framework for surgical workflow based on a discrete encoder-decoder GAN to foresee surgical phases in laparoscopic cholecystectomy. Our evaluation on objective metrics, alongside the results of the performed survey of perceived plausibility demonstrate the effectiveness and reliability of the approach. Furthermore, our research suggests several possible opportunities to extend this approach to additional predictive tasks and more exhaustive 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. Due to broad extendability to diverse and more complex surgical procedures, this model holds the potential to significantly enhance surgical decision making and augment risk mitigation for ultimately improved patient outcome.

Acknowledgement This work uses de-identified surgical videos for model learning. Mass General Brigham IRB (protocol number 2018P002340) approved secondary use of surgical data for this project.

REFERENCES

[1] M. M. Al Aziz, D. Alhadidi, and N. Mohammed. Secure approximation of edit distance on genomic data. *BMC medical genomics*, 10(2):55–67, 2017.
[2] Y. Ban, G. Rosman, T. Ward, D. Hashimoto, T. Kondo, H. Iwaki, O. Meireles, and D. Rus. Aggregating long-term context for learning laparoscopic and robot-assisted surgical workflows. In *ICRA*, 2021.
[3] E. Barsoum, J. Kender, and Z. Liu. HP-GAN: Probabilistic 3D human motion prediction via GAN. In *CVPR workshops*, 2018.
[4] S. Bowman, L. Vilnis, O. Vinyals, A. Dai, R. Jozefowicz, and S. Bengio. Generating sentences from a continuous space. In *SIGNLL*, pages 10–21, 2016.
[5] E. Brophy, Z. Wang, Q. She, and T. Ward. Generative adversarial networks in time series: A survey and taxonomy. *arXiv:2107.11098.*
[6] 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*, 2014.
[7] Z. Costello and H. G. Martin. How to hallucinate functional proteins. *arXiv preprint arXiv:1903.00458*, 2019.
[8] A. M. Dai and Q. V. Le. Semi-supervised sequence learning. In *NIPS*, pages 3079–3087, 2015.
[9] N. Deo and M. M. Trivedi. Multi-modal trajectory prediction of surrounding vehicles with maneuver based lstms. In *IVS*, 2018.
[10] O. Dergachyova, D. Bouget, A. Huaulmé, X. Morandi, and P. Jannin. Automatic data-driven real-time segmentation and recognition of surgical workflow. *IJCARS*, 11(6):1081–1089, 2016.

[11] A. Garcia-Martinez, J. M. Vicente-Samper, and J. M. Sabater-Navarro. Automatic detection of surgical haemorrhage using computer vision. *Artificial intelligence in medicine*, 78:55–60, 2017.

[12] R. Gómez-Bombarelli and et al. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[13] 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*, page 2672, 2014.

[14] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi. Social GAN: Socially acceptable trajectories with generative adversarial networks. In *CVPR*, pages 2255–2264, 2018.

[15] D. A. Hashimoto, G. Rosman, and O. R. Meireles. *Artificial Intelligence in Surgery: Understanding the Role of AI in Surgical Practice*. McGraw-Hill Education / Medical, 1 edition, July 2021.

[16] D. A. Hashimoto, G. Rosman, D. Rus, and O. R. Meireles. Artificial intelligence in surgery: promises and perils. *Annals of surgery*, 2018.

[17] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[18] M. Higgins and et al. Development and decay of procedural skills in surgery: A systematic review of the effectiveness of simulation-based medical education interventions. *The Surgeon*, 2021.

[19] 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. *RA-L*, 2020.

[20] X. Huang, S. G. McGill, J. A. DeCastro, L. Fletcher, J. J. Leonard, B. C. Williams, and G. Rosman. Carpal: Confidence-aware intent recognition for parallel autonomy. *RA-L*, 6(3):4433–4440, 2021.

[21] X. Huang, S. G. McGill, B. C. Williams, L. Fletcher, and G. Rosman. Uncertainty-aware driver trajectory prediction at urban intersections. In *ICRA*, pages 9718–9724, 2019.

[22] X. Huang, G. Rosman, I. Gilitschenski, A. Jasour, S. G. McGill, J. J. Leonard, and B. C. Williams. HYPER: Learned hybrid trajectory prediction via factored inference and adaptive sampling. In *ICRA*, 2022. accepted.

[23] B. Ivanovic, K. Leung, E. Schmerling, and M. Pavone. Multimodal deep generative models for trajectory prediction: A conditional variational autoencoder approach. *RA-L*, 6(2):295–302, 2020.

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

[25] 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-TMI*, 37(5):1114–1126, 2017.

[26] Y. Jin, Y. Long, C. Chen, Z. Zhao, Q. Dou, and P.-A. Heng. Temporal memory relation network for workflow recognition from surgical video. *IEEE-TMI*, 2021.

[27] L. Juvela, B. Bollepalli, J. Yamagishi, and P. Alku. Gelp: Gan-excited linear prediction for speech synthesis from mel-spectrogram. *arXiv:1904.03976*, 2019.

[28] A. Kadurin and et al. The cornucopia of meaningful leads: Applying deep adversarial autoencoders for new molecule development in oncology. *Oncotarget*, 8(7):10883, 2017.

[29] J. N. Kundu, M. Gor, and R. V. Babu. Bihmp-gan: Bidirectional 3d human motion prediction gan. In *AAAI*, 2019.

[30] M. J. Kusner and J. M. Hernández-Lobato. GANS for sequences of discrete elements with the gumbel-softmax distribution. *arXiv:1611.04051*, 2016.

[31] F. Lalys and P. Jannin. Surgical process modelling: a review. *IJCARS*, 9(3):495–511, 2014.

[32] 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 *CVPR*, pages 336–345, 2017.

[33] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union, 1966.

[34] X. Li, G. Rosman, I. Gilitschenski, B. Araki, C.-I. Vasile, S. Karaman, and D. Rus. Learning an explainable trajectory generator using the automaton generative network (AGN). *RA-L*, 7(2):984–991, Apr. 2022.

[35] O. R. Meireles, G. Rosman, M. S. Altieri, L. Carin, G. Hager, A. Madani, N. Padoy, C. M. Pugh, P. Sylla, T. M. Ward, et al. Sages consensus recommendations on an annotation framework for surgical video. *Surgical endoscopy*, 35(9):4918–4929, 2021.

[36] C. I. Nwoye, C. Gonzalez, T. Yu, P. Mascagni, D. Mutter, J. Marescaux, and N. Padoy. Recognition of instrument-tissue interactions in endoscopic videos via action triplets. In *MICCAI*, pages 364–374. Springer, 2020.

[37] J. Park and C. H. Park. Recognition and prediction of surgical actions based on online robotic tool detection. *IEEE Robotics and Automation Letters*, 6(2):2365–2372, Apr. 2021.

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

[39] B. W. Renz, F. Bösch, and M. K. Angele. Bile duct injury after cholecystectomy: surgical therapy. *Visceral medicine*, 2017.

[40] N. Rhinehart, R. McAllister, K. Kitani, and S. Levine. Precog: Prediction conditioned on goals in visual multi-agent settings. In *ICCV*, pages 2821–2830, 2019.

[41] D. Rivoir, S. Bodenstedt, I. Funke, F. von Bechtolsheim, M. Distler, J. Weitz, and S. Speidel. Rethinking anticipation tasks: Uncertainty-aware anticipation of sparse surgical instrument usage for context-aware assistance. In *MICCAI*, pages 752–762. Springer, 2020.

[42] K. Sohn, H. Lee, and X. Yan. Learning structured output representation using deep conditional generative models. *NIPS*, 28:3483–3491, 2015.

[43] I. Sutskever. *Training recurrent neural networks*. PhD thesis, 2013.

[44] I. Sutskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[45] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR,2016*, 2016.

[46] K. S. Tai, R. Socher, and C. D. Manning. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*, pages 1556–1566, 2015.

[47] L. A. Thiede and P. P. Brahma. Analyzing the variety loss in the context of probabilistic trajectory prediction. In *ICCV*, 2019.

[48] A. M. Turing. Computing machinery and intelligence. In *Parsing the turing test*, pages 23–65. Springer, 2009.

[49] 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-TMI*, 36(1):86–97, 2016.

[50] 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-TMI*, 2018.

[51] M. Uemura, M. Tomikawa, T. Miao, R. Souzaki, S. Ieiri, T. Akahoshi, A. K. Lefor, and M. Hashizume. Feasibility of an ai-based measure of the hand motions of expert and novice surgeons. *Computational and mathematical methods in medicine*, 2018.

[52] 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 *ICRA*, pages 754–759, 2017.

[53] T. M. Ward, D. M. Fer, Y. Ban, G. Rosman, O. R. Meireles, and D. A. Hashimoto. Challenges in surgical video annotation. *Computer Assisted Surgery*, 26(1):58–68, 2021.

[54] 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.

[55] T. M. Ward, D. A. Hashimoto, Y. Ban, G. Rosman, and O. R. Meireles. Artificial intelligence prediction of cholecystectomy operative course from automated identification of gallbladder inflammation. *Surgical Endoscopy*, pages 1–9, 2022.

[56] T. M. Ward, P. Mascagni, Y. Ban, G. Rosman, N. Padoy, O. Meireles, and D. A. Hashimoto. Computer vision in surgery. *Surgery*, 169(5):1253–1256, 2021.

[57] H. Wei, F. Rudzicz, D. Fleet, T. Grantcharov, and B. Taati. Intraoperative adverse event detection in laparoscopic surgery: Stabilized multi-stage temporal convolutional network with focal loss. 2021.

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

[59] K. K. Yang, Z. Wu, and F. H. Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 2019.

[60] L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: Sequence generative adversarial nets with policy gradient. In *AAAI*, volume 31, 2017.

[61] C. Zhao, G. G. Yen, Q. Sun, C. Zhang, and Y. Tang. Masked gan for unsupervised depth and pose prediction with scale consistency. *TNNLS*, 2020.