Autonomous Hierarchical Surgical State Estimation during Robot-assisted Surgery through Deep Neural Networks

Yidan Qin[1,2], Max Allan[1], Joel W. Burdick[2], Mahdi Azizian[1]

***Abstract*—Many operations in robot-assisted surgery (RAS) can be viewed in a hierarchical manner. Each surgical task is represented by a superstate, which can be decomposed into finer-grained states. The estimation of these discrete states at different levels of temporal granularity provides a temporal perception of the current surgical scene during RAS, which is a crucial step towards many automated surgeon-assisting functionalities. We propose Hierarchical Estimation of Surgical States through Deep Neural Networks (HESS-DNN), a deep learning-based system that concurrently estimates the current super- and fine-grained states. HESS-DNN incorporates endoscopic vision, robot kinematics, and system events data from the da Vinci® Xi surgical system. HESS-DNN is evaluated on a real-world robotic inguinal hernia repair surgery dataset: HERNIA-20, and achieves accurate state estimates of both surgical superstate and the corresponding fine-grained surgical state. We show that HESS-DNN improves state-of-the-art fine-grained state estimation across the entire HERNIA-20 RAS procedure through its hierarchical design. We also analyze the relative contributions of each input data type and HESS-DNN's design to surgical (super)state estimation accuracy.**

***Index Terms*—Surgical Robotics: Laparoscopy, Deep Learning Methods, AI-Based Methods, Medical Robots and Systems.**

I. INTRODUCTION

ROBOT-ASSISTED surgery (RAS) technology has been adopted for the treatment of a wide range of medical conditions [1]. Compared to open surgeries, RAS patients have reported lower complication rates, less blood loss, and quicker recoveries [2]. Compared to laparoscopy, robotic surgical systems are more intuitive to control and they eliminate hand tremors [1]. Additionally, surgical robots can provide synchronized endoscopic vision, robot kinematics, and surgical system events data. These data sources provide a rich and comprehensive representation of a surgery [3]–[5], which has a wide range of potential applications. Artificial intelligence (AI) applications in RAS are emerging research topics, such as autonomy [6], surgeon skill evaluation [7], workflow analysis [8], etc. The use of AI in robotic surgical systems helps to enable passive virtual fixtures [9], presentation of advisory information [10], and the automation of surgical tasks [11], [12]. A cardinal prerequisite for these AI applications is an accurate awareness of the current stage of the surgery. This temporal awareness should include multiple levels of temporal granularity, ranging from the current surgical task to the fine-grained action performed by the surgeon.

Since many robotic surgical procedures progress through standard operative steps, the surgical tasks that occur during a specific type of procedure, e.g. a hernia repair, are usually consistent [13]. Many surgical tasks, such as knot-tying and running suture, are also comprised of consistent fine-grained surgical states [7]. These fine-grained states include the surgeon's actions and environmental observations, and are the basic elements of a surgical task [3]. While surgical ontology [14] is an important modeling methodology for surgical procedures, in this paper we model an RAS procedure as a surgical Hierarchical FSM (HFSM) [15] (Fig. 1) consisting of discrete superstates, where each superstate is a surgical task. Each superstate in turn may be an FSM consisting of finer-grained states (surgeon actions and observations). A simple definition is found in Section II. HFSM models represent at each time step the current surgery state at multiple levels of temporal granularity, which better capture the temporal progressions of surgeries. The estimation of surgical (super)states has various applications. Autonomous recognition of the current surgical task allows for surgical skill evaluation, operating room workflow coordination, and post-operative analysis [8], [16]. Fine-grained state estimation within a surgical task is a prerequisite for automation of this task [11]. Simultaneous state estimation at multiple levels of temporal granularity provides a comprehensive understanding of the surgical progress and is a key step toward deploying AI applications in RAS.

The process of hierarchical surgical state estimation can be roughly divided into the estimation of fine-grained surgical states that last for a few seconds [17]–[20], such as pushing a needle through tissue, and the estimation of a surgical task that lasts for minutes [21], [22], such as dissecting a hernia sac. Each of these tasks presents multiple challenges. Fine-grained surgical states are difficult to classify due to their short duration and spontaneous state transitions [3], [5]. Estimating the surgical task, however, requires the capture of long-term temporal correlations in the input data, as superstates last from $<$1 minute to 10-30 minutes. Real-world RAS data is costly to acquire and has high diversity. The viewing angles, endoscopic brightness, and surgical backgrounds vary considerably across patients, even for the same surgery types. Additionally, surgeons may employ different maneuvers depending on patient

Manuscript received: February 10, 2021; Revised May 8, 2021; Accepted June 5, 2021.

This paper was recommended for publication by Editor Pietro Valdastri upon evaluation of the Associate Editor and Reviewers' comments. This work was supported by Intuitive Surgical Inc.

[1]Intuitive Surgical Inc., 1020 Kifer Road, Sunnyvale, CA,94086, USA

[2]Mechanical and Civil Engineering, Caltech, Pasadena, CA, 91125, USA

Emails: Ida.Qin@intusurg.com, Mahdi.Azizian@intusurg.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.2021.3091728, IEEE Robotics and Automation Letters

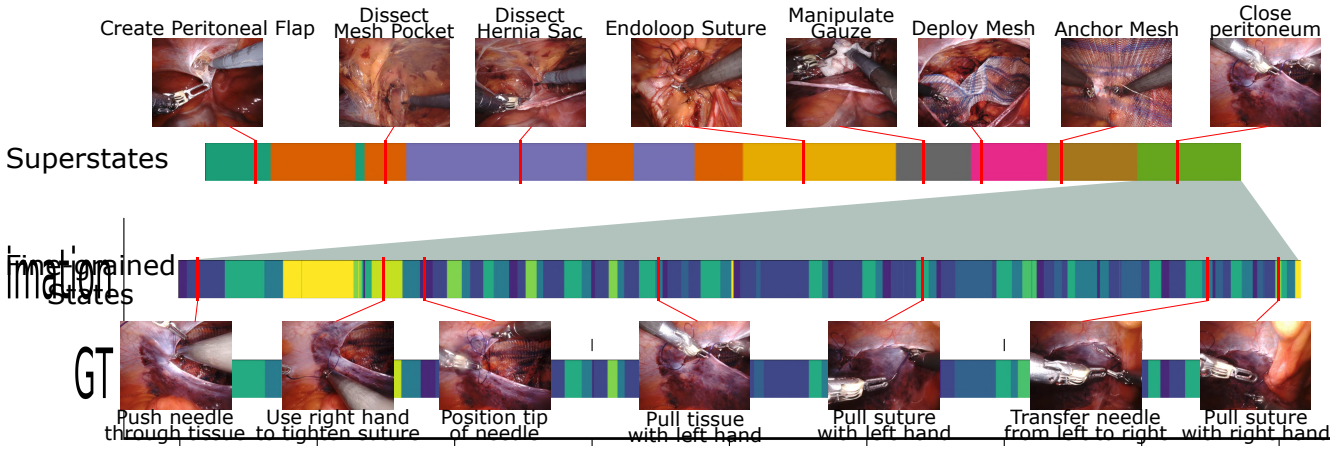

Fig. 1: An example robotic inguinal hernia repair surgery consists of multiple surgical tasks, which are superstates in an HFSM (top row). A superstate is an FSM consisting of fine-grained surgical states. An example FSM for the superstate "*close peritoneum*" is shown in the bottom row.

conditions and surgeon preferences. The limited size of real-world RAS datasets coupled with their high variability presents challenges to data-driven surgical state estimation methods [5].

The estimation of both super- and fine-grained surgical states has been previously explored in separate efforts. Existing fine-grained surgical state estimation methods have used various surgical robot data sources to infer the current surgical state. Some studies model state transitions using the surgical robot's kinematics, including hidden Markov Models [23], [24], temporal clustering [22], [25], and conditional random fields [26]. Deep learning methods that exploit temporal correlations among adjacent entries in time series data have also been proposed to infer the current fine-grained surgical states. These methods aim to capture temporal features in a surgical robot's kinematics or endoscopic vision data through Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as Long-Short Term Memory (LSTM) [18], [27], [28]. When available, the use of multiple data sources, such as endoscopic vision, robot kinematics, and system events, yields more accurate and robust fine-grained surgical state estimation [3], [29]. Multiple input sources allow data-driven state estimation methods to extract a richer representation of surgical states, which results in improved estimation accuracy, especially in real-world RAS settings [5]. Prior work in surgical phase/task segmentation has mainly relied on endoscopic video data. Twinanda et al. used annotated endoscopic videos from laparoscopic surgeries to train a CNN-LSTM [30]. Jin et al. introduced the post-processing of predictions using prior knowledge inference [31]. A multi-stage temporal convolutional network [32] and the integration of 2D and 3D CNNs [33] were proposed for richer temporal feature learning. Zia et al. collected robot kinematics and system events data to perform surgical phase recognition [22]. To the authors' knowledge, there is yet to be an attempt at hierarchical state estimation in the surgical setting. Existing methods for hierarchical video temporal segmentation and content characterization have focused on activity and surveillance videos [34], [35]. Gunsel et al. divided video characteristics such as motion vectors and color histograms into categories to segment a video sequence into shots [35]. De Menthon et al. proposed a spatio-temporal segmentation method based on a mean shift analysis that maps each frame in a video to a feature vector describing its motion characteristics [36]. Lan et al. proposed an unsupervised spatio-temporal segmentation method that proposes action-related spatial regions and clusters fine-grained temporal segments into higher-level segments [37].

Although existing fine-grained state estimation methods model a surgical task as a set of states, there is not yet an accepted definition of an RAS procedure from an estimation perspective. Additionally, environmental changes and observations are omitted in the annotations of popular surgical datasets such as JIGSAWS [38] and Cholec80 [30]. Non-action surgical states, such as bleeding and the presence/absence of a surgical instrument, are clinically crucial for some applications, such as the automation of surgical tasks [3].

Contributions: We propose a holistic definition of an RAS procedure and as a surgical Hierarchical Finite State Machine (HFSM). Each surgical task is in turn a surgical FSM, whose states model key parts of the surgical procedure, including surgical actions and observations that occur during this task. We demonstrate Hierarchical Estimation of Surgical States through Deep Neural Networks (HESS-DNN): a hierarchical surgical state estimation technique that concurrently performs surgical task estimation and fine-grained surgical state estimation. HESS-DNN uses multiple types of input data collected from the da Vinci® Xi surgical system, including endoscopic vision, robot kinematics, and system events. The goal is to achieve surgical state estimation at multiple levels of temporal granularity during RAS. Our main contributions include:

- Achieving accurate surgical state estimation at two levels of temporal granularity during RAS;
- Incorporating semantic segmentation of endoscopic vision for a more efficient visual feature representation;
- Improving fine-grained surgical state estimation accuracy comparing to state-of-the-art fine-grained estimation methods by up to 16.3% in a large set of diverse states throughout real-world RAS procedures;
- Demonstrating the necessity and advantages of a multi-input hierarchical state estimation approach with limited

data availability in a complex real-world RAS setting.

HESS-DNN is evaluated with HERNIA-20: a real-world RAS dataset containing 20 inguinal hernia repair surgeries performed on da Vinci® Xi surgical systems [5]. The performance of HESS-DNN is demonstrated at two levels of temporal granularity: surgical superstates and fine-grained states. An ablation study investigates how endoscopic vision, robot kinematics, and system events inputs contributes to HESS-DNN's estimation performance, respectively. HESS-DNN achieves a frame-wise surgical superstate estimation accuracy of 87.6% and an overall fine-grained surgical state estimation accuracy of 80.4% across 23 fine-grained states in various superstates. HESS-DNN improves state-of-the-art fine-grained state estimation methods' accuracy by up to 16.3% when training for all 23 fine-grained states in HERNIA-20 due to its hierarchical design. HESS-DNN's accurate performance in a highly diverse real-world RAS setting with a limited dataset demonstrates the advantage of a hierarchical process model and the robustness of our proposed multi-input state estimation method. Limitations on HESS-DNN's approach, such as a lack of correlational coupling between the superstate and fine-grained state estimation processes, are discussed in Section V.

II. Methods

First, we define a surgical estimation FSM that models a robotic surgical task for estimation purposes and the surgical HFSM that models an entire RAS procedure, within the scope of our consideration. A surgical estimation FSM model does not necessarily constrain any state transition probabilities due to the complex and highly dynamic real-world RAS setting.

Definition 1. A surgical estimation Finite State Model (FSM) $M(S, \Sigma, F)$ is comprised of:
S: a finite non-empty set of states;
Σ: the input symbols (or data) to the system;
P: a set of allowable transitions between states

Definition 2. A surgical Hierarchical FSM (HFSM) is a surgical FSM whose states could be other surgical FSMs. The states of a surgical HFSM are superstates.

Definition 3. A superstate models the coarsest division of a surgery into tasks, phases, or operative steps.

These definitions are a subset of a typical (hierarchical) finite state machine model: our current study does not incorporate transition dynamics between states, which is a subject of future work. Hence, an HFSM is akin to a graph that models the relationships between states and hierarchy. Because our study data set does not have any forbidden transitions between surgical states, in this paper we do not incorporate the potential impact of forbidden transitions on estimator performance.

The hierarchical surgical state estimator, HESS-DNN, estimates, from the input data stream, surgical states at two levels of temporal granularity: superstates (tasks) that last for minutes and fine-grained states that last for seconds. While many surgeries will have a deeper hierarchical structure, we focus on a simple hierarchy as a first step. HESS-DNN's components and training process are described next.

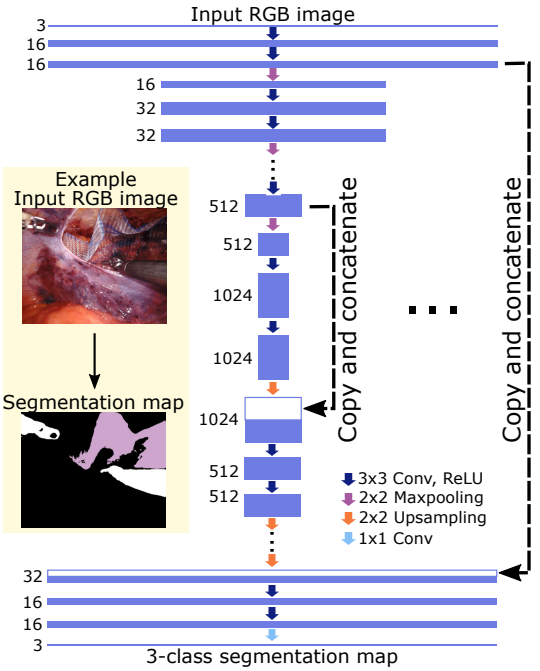

Fig. 2: U-Net architecture for surgical scene semantic segmentation. Blue boxes represent feature maps with their number of channels. White boxes denote copied feature maps. Different operations are denoted by colored arrows. U-Net input and output examples are shown on the left.

A. Feature extraction

The RAS input data (endoscopic vision, robot kinematics, and system events) is processed as follows. To leverage a more efficient visual feature representation, a semantic mask is extracted from the endoscopic view. The semantic mask is generated using a trained and frozen surgical scene segmentation model based on U-Net [39] and eliminates the background variability in the endoscopic view. The semantic segmentation model (Fig. 2) uses three scene classes: surgical instruments, tissue, and others. HESS-DNN's feature extraction component is diagrammed in Fig. 4. At time t, a CNN extracts information $\boldsymbol{x}_t^{\text{vis}}$ from the semantic segmentation map of the endoscopic vision. Since HESS-DNN accepts time series data as input, we implemented an LSTM encoder to embed the temporal correlations among $\boldsymbol{x}^{\text{vis}}$ from adjacent frames into a latent representation $\boldsymbol{h}_t^{\text{vis}} = \text{LSTM}(\boldsymbol{h}_{t-1}^{\text{vis}}, \boldsymbol{x}_t^{\text{vis}})$ [40]. The latent representation at time t therefore includes information from prior time steps. An LSTM unit contains a memory cell with cell state c_t at time t. Three sigmoid gates governs the access and modification to the memory cell: forget gate f_t, input gate i_t, and output gate o_t. An LSTM unit is then updated from the previous time step $t-1$ as:

$$f_t = \phi(\mathbf{W}_f(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) + \mathbf{b}_f) \quad (1)$$
$$i_t = \phi(\mathbf{W}_i(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) + \mathbf{b}_i) \quad (2)$$
$$o_t = \phi(\mathbf{W}_o(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) + \mathbf{b}_o) \quad (3)$$
$$c_t = f_t \cdot c_{t-1} + i_t \cdot \tanh(\boldsymbol{W}_c(\boldsymbol{h}_{t-1}, \boldsymbol{x}_t) + \mathbf{b}_c) \quad (4)$$
$$\boldsymbol{h}_t = o_t \cdot \tanh(c_t) \quad (5)$$

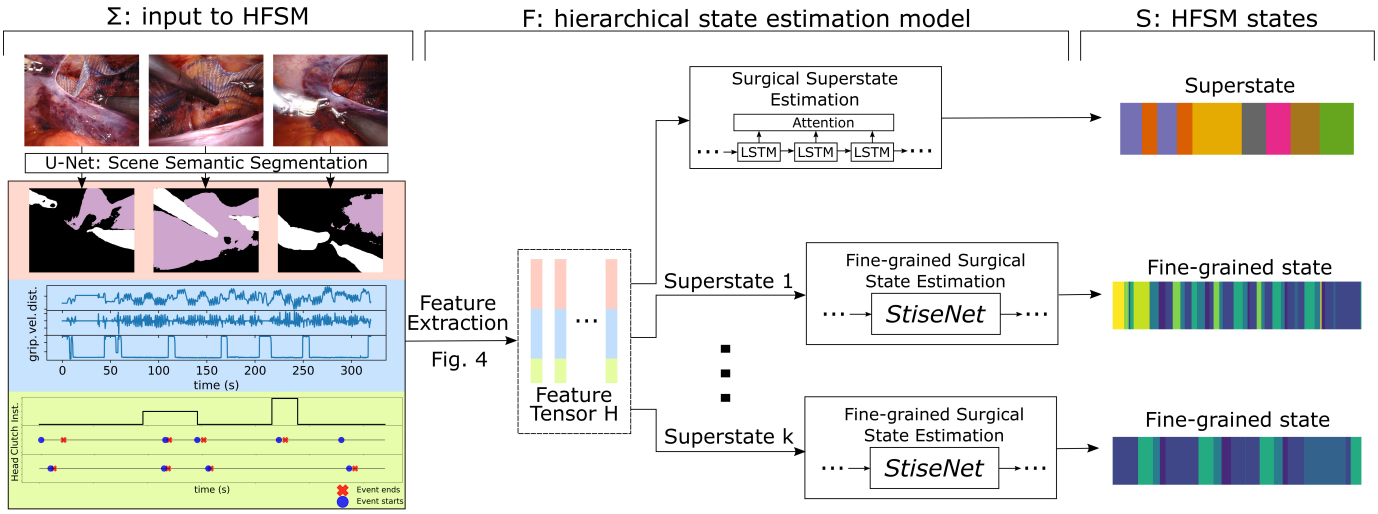

Fig. 3: Parameters of an RAS procedure HFSM and HESS-DNN's architecture. The inputs to the HFSM include endoscopic vision, robot kinematics, and system events. A feature extraction component embeds information in input data for hierarchical surgical (super)state estimation. HESS-DNN plays the role of state transition function F that determines the current system state given the previous state and input data. Our previous work *StiseNet* [5] is implemented for fine-grained state estimation.

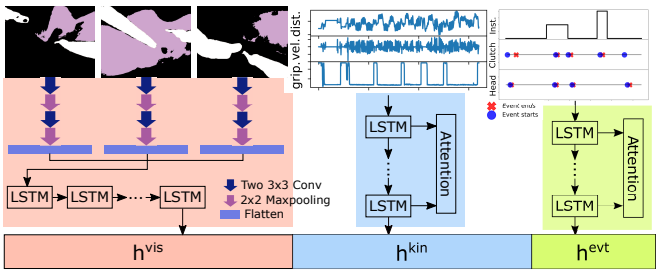

Fig. 4: Details of HESS-DNN's feature extraction component. $\boldsymbol{h}^{\text{vis}}$, $\boldsymbol{h}^{\text{kin}}$, and $\boldsymbol{h}^{\text{evt}}$ are extracted from endoscopic vision, robot kinematics, and system events, respectively.

where $\mathbf{W}$ and $\mathbf{b}$ are learnable parameters and ϕ is a sigmoid function [40]. The visual features are denoted $\boldsymbol{h}_t^{\text{vis}} \in \mathbb{R}^{n_{\text{vis}}}$.

Kinematic features are extracted from Universal Patient-side Manipulators (USMs). Since the input kinematics data include multiple robotic arms' translations, rotations, and joint angles, we implemented an LSTM encoder with an input attention mechanism [4] to identify the driving data types. The embedded latent representation of robot kinematics takes the form $\boldsymbol{h}_t^{\text{kin}} = \text{LSTM}(\boldsymbol{h}_{t-1}^{\text{kin}}, \boldsymbol{\alpha}_t \cdot \boldsymbol{x}_t^{\text{kin}})$, where multiplier $\boldsymbol{\alpha}_t$ is a vector whose elements represent the weights of all input kinematics data types. The weight vector $\boldsymbol{\alpha}_t$ is derived as:

$$\boldsymbol{\alpha}_t = \text{softmax}\left\{\boldsymbol{u}_e^T \tanh\left(\boldsymbol{W}_e(\boldsymbol{h}_{t-1}, \boldsymbol{c}_{t-1}) + \boldsymbol{V}_e \boldsymbol{x}_t\right)\right\} \quad (6)$$

where $\boldsymbol{u}_e$, $\mathbf{W}_e$, and $\mathbf{V}_e$ are learnable parameters.

System events data are also logged from the da Vinci® Xi Surgical System (details in Section III). The embedded latent representation of events $\boldsymbol{h}_t^{\text{evt}}$ is extracted following the same method as the robot kinematics.

B. Hierarchical surgical (super)state estimation

The latent representations, or "features", of endoscopic vision, robot kinematics, and system events data, are concatenated at time t to form a feature vector $\boldsymbol{H}_t \in \mathbb{R}^{n_{\text{vis}}+n_{\text{kin}}+n_{\text{evt}}}$. Data are processed over an observation window, $[\boldsymbol{H}_{t-T_{obs}}, \boldsymbol{H}_{t-T_{obs}+\delta}, ..., \boldsymbol{H}_t]$, where T_{obs} is the window duration and δ denotes the sampling interval. Since the durations of fine-grained states and superstates are significantly different, the values of T_{obs} and δ are different for surgical superstate estimation and fine-grained state estimation. For simplicity, T_{obs} and δ are chosen such that the length of input tensor $l = T_{obs}/\delta + 1$ is an integer. HESS-DNN's architecture is shown in Fig. 3.

Surgical superstates are estimated via an LSTM-based decoder network that operates on the input data tensor [4]. Cho et al. showed that the performance of an encoder-decoder network could rapidly deteriorate as the length of the input data increases [41]. We therefore implemented an addition attention mechanism [42] to select important latent representations across all time steps in an adaptive manner. At time t, the attention weights $\boldsymbol{\beta} \in \mathbb{R}^l$ are determined from the previous decoder hidden state $\boldsymbol{d}_{t-1}$ and cell state as:

$$\boldsymbol{\beta}_t^j = \text{softmax}(\boldsymbol{u}_d^T \tanh(\boldsymbol{W}_d(\boldsymbol{d}_{t-1}, \boldsymbol{c}_{t-1}) + \boldsymbol{V}_d \boldsymbol{H}_t^j)) \quad (7)$$

for $j \in [1, l]$ where $\boldsymbol{u}_d$, $\boldsymbol{W}_d$, and $\boldsymbol{V}_d$ are learnable parameters. The LSTM decoder is updated using the weighted feature $\tilde{\boldsymbol{H}}_t = \sum_{j=1}^{l} \beta_t^j \boldsymbol{H}_t^j$ and the previous system state s_{t-1}:

$$\boldsymbol{d}_t = \text{LSTM}(\boldsymbol{d}_{t-1}, [s_{t-1}, \tilde{\boldsymbol{H}}_t]) \quad (8)$$

where s_t is determined through a fully-connected layer.

The use of multiple input data types for fine-grained state estimation has been more extensively researched. Here, we deploy our previously introduced *StiseNet* [5] to estimate the fine-grained surgical states within a superstate. *StiseNet* employs an adversarial model design, which pits two model components against each other during training to produce a latent data representation from $\boldsymbol{H}_t$ that is invariant to nuisances (e.g., anatomical background, brightness, etc.) and variations in surgical style. Details of *StiseNet*'s architecture

are discussed in [5]. StiseNet's adversarial model architecture was not applied to superstate estimation: the computational complexity of learning temporal correlations across the long superstate durations (up to 30 minutes) is prohibitive.

C. Training and inference

The U-Net model for surgical scene semantic segmentation was separately trained on a large surgical image dataset, following [43]. The endoscopic vision input was resized to a 640×512 RGB image. We determined the dimensions of extracted features using grid search: $n_{\text{vis}} = 40$, $n_{\text{kin}} = 40$, and $n_{\text{evt}} = 4$. All data inputs are synchronized at 10Hz. For superstate estimation, $T_{obs} = 60\text{sec}$ and $\delta = 5$. The fine-grained state estimation parameters, $T_{obs} = 2\text{sec}$ and $\delta = 1$, were also determined via grid search. HESS-DNN's end-to-end training is guided by the sum of a superstate estimation loss and *StiseNet*'s fine-grained state estimation loss:

$$L = L_{\text{super}} + L_{\text{StiseNet}} \tag{9}$$

where L_{super} is categorical cross-entropy loss. As discussed in [5], *StiseNet*'s training is a minimax game [44]. HESS-DNN therefore inherits a scheduled adversarial optimizer [44], in which the generative or the discriminative component trains on a data batch while the other component's weights are frozen.

III. Experimental Evaluations

We demonstrate HESS-DNN's hierarchical (super)state estimation performance using a real-world RAS dataset, HERNIA-20. The surgeries in HERNIA-20 are annotated at two levels of temporal granularity: surgical tasks and fine-grained surgical states [5] (Table I). The performance of HESS-DNN is evaluated in two experimental settings. An ablation study was performed to understand the necessity and contributions of different input data combinations to an accurate and robust hierarchical (super)state estimation.

A. Dataset

The HERNIA-20 dataset contains 20 real-world robotic inguinal hernia repair surgeries performed on da Vinci Xi ® Surgical Systems using the transabdominal preperitoneal repair method [13]. HERNIA-20 contains 12 unilateral and 8 bilateral inguinal hernia repair procedures. The procedure contains 8 tasks, which correspond to superstates in an HFSM. Five superstates (SS0-SS4) were further divided into fine-grained surgical states. These superstates occurred frequently in HERNIA-20 and have high clinical importance. Overlaps of fine-grained states exist among different superstates, and the repeated states were treated as the same class during the training and evaluation of the flat fine-grained state estimation methods that we compare against. The determination of (super)state labels were obtained from surgical textbooks and practicing surgeons. HESS-DNN does not pose any constraint on state transition probabilities in its state estimation process.

HERNIA-20 surgeries last 15-70 minutes. Three types of da Vinci® surgical instruments (ProGrasp™ forceps, large needle driver, and monopolar curved scissors) and two types of laparoscopic instruments (tack fixation device and suction irrigator) are involved. Endoscopic vision data is collected with a da Vinci endoscope. The robot kinematics data (end-effector positions, velocities, gripper angles, and endoscope positions) is collected from four Universal Patient-Side Manipulators (USMs). 10 system events are collected, including six binary events (surgeon head in/out of the console, endoscope follow, instrument follow, two master clutches on surgeon-side console, and energy pedal) and four categorical events (the installed da Vinci® surgical instrument on USMs). HERNIA-20 data is fully anonymized so that the identity of each procedure's surgeon(s) is unknown.

B. Metrics

The quality of our hierarchical surgical state estimation model is quantified by the percentage of time steps with accurate state estimates in the test set, as judged by comparison with the ground truth annotation reviewed by experts. Since some autonomy applications require the real-time knowledge

Superstate ID	**Inguinal Hernia Repair**	Percentage of instances (%)	Mean duration (s)
SS0	Create peritoneal flap	16.5	74.6
SS1	Dissect mesh pocket	19.7	88.1
SS2	Dissect hernia sac	16.9	136.1
SS3	Deploy mesh	18.0	138.7
SS4	Close peritoneum	16.8	202.3
SS5	Endoloop suture	3.1	131.3
SS6	Anchor mesh	5.3	117.2
SS7	Manipulate gauze	3.7	81.7

Fine-grained Surgical States

Create Peritoneal Flap	Mean duration (s)
Cut peritoneum with monopolar scissors	5.7
Stretch peritoneum with left hand	3.1
Adjust endoscope	2.9
Dissect Mesh Pocket	**Mean duration (s)**
Pull tissue with left hand	4.2
Local energized cut	2.8
Push and cut tissue	3.9
Adjust endoscope	3.1
Dissect Hernia Sac	**Mean duration (s)**
Push and cut tissue	4.9
Pull and cut tissue	4.1
Pull tissue with left hand	4.3
Push tissue with left hand	2.6
Adjust endoscope	3.7
Deploy Mesh	**Mean duration (s)**
Unfold mesh	6.6
Push mesh to tissue with left hand	2.9
Push mesh to tissue with right hand	2.6
Pull tissue with left hand	3.8
Pull tissue with right hand	3.4
Manipulate mesh	7.0
Adjust endoscope	2.8
Close Peritoneum	**Mean duration (s)**
Reach for the needle	3.9
Position the tip of the needle	3.3
Push needle through the tissue	4.2
Pull tissue with left hand	3.6
Transfer needle from left to right	3.7
Orient needle	6.6
Pull suture with left hand	5.8
Pull suture with right hand	4.8
Transfer needle from right to left	4.6
Use right hand to tighten suture	4.3
Adjust endoscope	3.8

Table I: HERNIA-20 superstates and fine-grained states descriptions and mean durations.

of surgical (super)states, we evaluate HESS-DNN performance in non-causal and causal settings. In a non-causal setting, the estimator has access to information from preceding and future time steps. Accordingly, we implemented a bi-directional LSTM units [40]. In the causal setting, HESS-DNN only has access to data from the current and preceding time steps; therefore, forward LSTM units were implemented. A non-causal setting is suitable to evaluate post-operative hierarchical state estimation performance, while state estimation performance in a causal setting is crucial for real-time applications. The dataset was divided into training and test sets following an 80:20 split (as the surgeon identity is not available) and a five-fold cross validation was performed.

C. Ablation study

We evaluated the contributions of various input data to surgical (super)state estimation through an "ablation study" by removing subsets of input data types and comparing the ablated methods' performances to the complete estimator (This shall not be mistaken with surgical ablation). Specifically, HESS-DNN's performance with only one type of input data was compared against its performance with the full input data. Additionally, we examined the effectiveness of semantic segmentation on visual feature extraction by comparing HESS-DNN against an ablated version, HESS-DNN-NU, which omits the semantic segmentation of endoscopic vision and instead extracts features directly from raw endoscopic videos. The ablation study demonstrates the necessity of each input data type and validates HESS-DNN's design.

IV. Results and Discussions

Tables II and III quantify HESS-DNN's estimation performance in non-causal and causal settings. State estimation accuracy is shown for both surgical superstates (tasks) and all fine-grained states in Table I. Table III also compares HESS-DNN's overall fine-grained state estimation accuracy against state-of-the-art fine-grained state estimation methods. Fig. 5 presents a state sequence from one HERNIA-20 surgery in order to visualize the surgical (super)state estimation quality of HESS-DNN and its ablated versions. The processing time of HESS-DNN is 8.9 frames per second on a workstation equipped with an NVIDIA GTX 1080 Ti graphics card, an Intel Core i7-6700 CPU, and 16GB of RAM.

Surgical Superstate Estimation Accuracy (%)

Model/Input data	Non-causal	Causal
Trivial	12.5	12.5
System events	40.6 ± 17.87	39.4 ± 19.04
Raw endoscopic vision	47.1 ± 10.10	41.4 ± 8.91
Robot kinematics	65.9 ± 8.41	52.3 ± 8.75
Endoscopic vision semantic mask	70.1 ± 5.17	64.5 ± 9.67
HESS-DNN-NU	76.6 ± 4.04	70.6 ± 4.77
HESS-DNN	**87.6 ± 3.91**	**84.1 ± 4.11**

Table II: HERNIA-20's Surgical superstate estimation performance of HESS-DNN and its ablated versions.

Overall Fine-grained State Estimation Accuracy (%)

Model/Input data	Non-causal	Causal
Trivial	4.3	4.3
TCN (vision) [45]	45.7 ± 10.04	41.9 ± 12.66
TCN (kinematics) [45]	48.9 ± 16.40	43.4 ± 17.40
Forward LSTM [18]	50.1 ± 8.11	49.7 ± 11.54
Bidir. LSTM [18]	54.7 ± 7.97	-
Fusion-KVE [3]	62.0 ± 6.05	59.9 ± 7.17
StiseNet [5]	64.1 ± 8.66	61.4 ± 8.08
HESS-DNN-NU	70.1 ± 6.71	66.9 ± 6.94
HESS-DNN	**80.4** ± 5.60	**75.7** ± 5.31

Table III: Overall fine-grained surgical state estimation performance in HERNIA-20.

Each type of input data contributes to HESS-DNN's hierarchical surgical state estimation to different degrees. Table II compares the surgical superstate estimation performance of HESS-DNN and its ablated versions when applied to HERNIA-20. When endoscopic vision, robot kinematics, and system events input data are included, HESS-DNN achieves a superior frame-wise superstate estimation accuracy in both non-causal and causal settings. The significant improvement in estimation accuracy comparing to HESS-DNN with a single type of input (first four rows) confirms the advantage of including multiple sources of input data. As suggested in our previous work [3], different surgical states contain different representative features, which may be recognizable through certain types of input data but not others. For instance, the SS4 (*Close peritoneum*) and SS5 (*Endoloop suture*) superstates both involve suturing; therefore, the installed surgical instrument for both superstates is a large needle driver. The system events data therefore cannot distinguish between these two superstates; however, the visual and kinematics features of SS4 and SS5 are significantly different and distinguishable. On the contrary, the dissection surgical superstates (SS0-SS2) require an energy instrument for cautery. The system event associated with pressing the energy pedals is therefore useful in distinguishing SS0-SS2 from other superstates. In comparison to fine-grained surgical state estimation, there has been less prior work on surgical superstate estimation, with little open-source code to facilitate comparisons; however, HESS-DNN achieves accurate superstate estimation in a complex real-world RAS environment with limited data.

Semantic segmentation of endoscopic vision yields a more effective visual feature extraction. We analyzed how the extraction of semantic segmentation maps from endoscopic vision data affects surgical (super)state estimation. HESS-DNN yields up to a 13.5% improvement in superstate estimation accuracy as compared to HESS-DNN-NU, which omits the U-Net model and extracts visual features directly from raw endoscopic videos. Two factors contribute to this improvement. The extraction of a surgical scene semantic segmentation map significantly reduces environmental distractions and variability. For instance, the surgical mesh implants and gauze deployed during inguinal hernia repair have various colors and forms, depending upon their manufacturer. The variation in mesh appearances is irrelevant to surgical superstate estimation. Through U-Net semantic segmentation, such distractions are eliminated. The U-Net model was separately trained and frozen on an extensive surgical scene dataset containing 25,000 semantically annotated images from real-world RAS. HESS-DNN can take advantage of a condensed representation of the endoscopic view provided by extensive training data for semantic segmentation purposes. This is especially valuable in real-world RAS datasets, like HERNIA-

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.2021.3091728, IEEE Robotics and Automation Letters

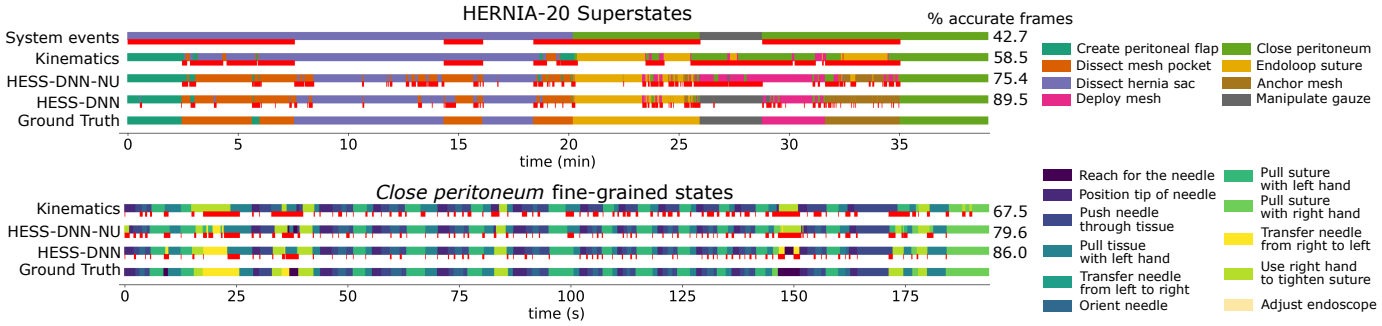

Fig. 5: An example HERNIA-20 superstate sequence (top) and the fine-grained state sequence of the *close peritoneum* superstate (bottom). The causal estimation results of HESS-DNN and its highest-performance ablated versions are compared against the ground truth. The discrepancies between (super)state estimation results and ground truth are marked in red.

20, where data is limited and costly to obtain.

HESS-DNN improves fine-grained surgical state estimation accuracy through the prior determination of surgical superstates. In Table III, we compare the overall fine-grained surgical state estimation of HESS-DNN against state-of-the-art fine-grained state estimation methods and our prior work. The fine-grained state estimation methods were trained to estimate all fine-grained states listed in Table I, with the repeated states treated as the same class. The high complexity of RAS and limited training data hinders flat (non-hierarchical) fine-grained state estimation performances, as there were 23 states in total in a highly dynamic RAS setting. HESS-DNN improves the fine-grained state estimation accuracy by up to 16.3% comparing to flat estimation techniques, as the superstate estimation is carried out prior to fine-grained state estimation and therefore narrows down the pool of candidate fine-grained surgical states.

A closer inspection of the surgical (super)state estimation results in Fig. 5 further supports these observations. In an example HERNIA-20 superstate sequence, HESS-DNN can recognize SS0-SS2 from other superstates when it only has access to system events; however, it is unable to further distinguish among superstates, resulting in poor estimation results. When using robot kinematics data, HESS-DNN is unable to distinguish among superstates *close peritoneum*, *endoloop suture*, and *anchor mesh* (25min - 34min). Since these three superstates all involve suturing maneuvers, distinguishable kinematics features were difficult to extract. The incorporation of endoscopic vision data overcame this challenge. Many intermittent errors are observed in HESS-DNN-NU for both surgical superstate and fine-grained state estimations, which suggests ineffective visual features are extracted from limited amounts of data. The frequent fluctuations in state identification, which causes the intermittent pattern of errors, further indicates the infirmity of HESS-DNN-NU. Clearly, a full HESS-DNN model with multiple types of input data achieves the most accurate estimation given a limited dataset in a complex real-world RAS setting.

V. Conclusions and Future Work

This paper proposed a simple definition of an RAS procedure as a surgical HFSM with states at multiple levels of temporal granularity. Each superstate in a surgical HFSM is an FSM consisting of fine-grained surgical actions and observations. We demonstrated *HESS-DNN*: a hierarchical surgical state estimation method that simultaneously estimates the current super and fine grained state of an RAS using multiple types of input data. The performance of HESS-DNN was illustrated through its application to HERNIA-20: a real-world RAS dataset containing surgical states at two levels of temporal granularity. HESS-DNN accepts synchronized endoscopic vision, robot kinematics, and system events data streams. HESS-DNN's surgical superstate estimation narrows down the pool of candidate fine-grained states and improves the fine-grained state estimation performance of state-of-the-art methods by up to 16.3% when evaluated on a large set of diverse surgical states in a real-world RAS setting. Additionally, we showcased the necessity and contributions of each type of input data to surgical super(state) estimations through an ablation study. HESS-DNN's architecture and performance have limitations, and there are many opportunities for future improvements. Currently, the surgical superstate estimation and fine-grained state estimation are conducted through two uncoupled network architectures with little shared knowledge. While we aimed to illustrate the process of hierarchical surgical state estimation, HESS-DNN's architecture can be further optimized for more efficient training and inference, and the incorporation of correlation across states in the hierarchy. Additionally, we are working to extend the modeling hierarchy to more levels of temporal granularity. We also plan to apply HESS-DNN to other types of RAS procedures such as prostatectomy, and expand HERNIA-20 to include more surgeries as well as to publish the dataset. Despite its limitations, our exploratory effort with HESS-DNN demonstrates the benefits of hierarchical surgical state estimation, and in the future it could support surgical autonomy and shared control applications in RAS.

Acknowledgment

This work was funded by Intuitive Surgical, Inc. We would like to thank Dr. Seyedshams Feyzabadi, Dr. Sandra Park, Dr. Humphrey Chow, and Dr. Wenqing Sun for their support.

References

[1] G. I. Barbash, "New technology and health care costs–the case of robot-assisted surgery," *New Eng. J. Medicine*, vol. 363, no. 8, p. 701, 2010.

[2] S. B. Stringfield, L. Parry, S. Eisenstein, S. Horgan, C. J. Kane, and S. L. Ramamoorthy, "Ten-year review of robotic surgery at an academic medical center," *J. Amer. Coll. Surg.s*, vol. 225, no. 4, p. S79, 2017.

[3] Y. Qin, S. A. Pedram, S. Feyzabadi, M. Allan, A. J. McLeod, J. W. Burdick, and M. Azizian, "Temporal segmentation of surgical sub-tasks through deep learning with multiple data sources," in *IEEE Int. Conf. Robotics and Automation*, 2020, pp. 371–377.

[4] Y. Qin, S. Feyzabadi, M. Allan, J. W. Burdick, and M. Azizian, "davincinet: Joint prediction of motion and surgical state in robot-assisted surgery," *arXiv preprint arXiv:2009.11937*, 2020.

[5] Y. Qin, M. Allan, Y. Yue, J. W. Burdick, and M. Azizian, "Learning invariant representation of tasks for robust surgical state estimation," *IEEE Robotics Automation Lett.s*, vol. 6, no. 2, pp. 3208–3215, 2021.

[6] G.-Z. Yang, J. Cambias, K. Cleary, E. Daimler, J. Drake, P. E. Dupont, N. Hata, P. Kazanzides, S. Martel, R. V. Patel, *et al.*, "Medical robotics—regulatory, ethical, and legal considerations for increasing levels of autonomy," *Science Robotics*, vol. 2, no. 4, p. 8638, 2017.

[7] Y. Gao, S. S. Vedula, C. E. Reiley, N. Ahmidi, B. Varadarajan, H. C. Lin, L. Tao, L. Zappella, B. Béjar, D. D. Yuh, *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *MICCAI Workshop: M2CAI*, vol. 3, 2014, p. 3.

[8] N. Padoy, "Machine and deep learning for workflow recognition during surgery," *Minimally Invasive Therapy & Allied Technologies*, vol. 28, no. 2, pp. 82–90, 2019.

[9] M. Selvaggio, G. A. Fontanelli, F. Ficuciello, L. Villani, and B. Siciliano, "Passive virtual fixtures adaptation in minimally invasive robotic surgery," *IEEE Robot. Auto. Lett.s*, vol. 3, no. 4, pp. 3129–3136, 2018.

[10] P. Chalasani, A. Deguet, P. Kazanzides, and R. H. Taylor, "A computational framework for complementary situational awareness in surgical assistant robots," in *IEEE Int. Conf. Robotic Comp.*, 2018, pp. 9–16.

[11] A. Shademan, R. S. Decker, J. D. Opfermann, S. Leonard, A. Krieger, and P. C. Kim, "Supervised autonomous robotic soft tissue surgery," *Sci. Trans. Med.*, vol. 8, no. 337, pp. 337ra64–337ra64, 2016.

[12] A. Attanasio, B. Scaglioni, M. Leonetti, A. F. Frangi, W. Cross, C. S. Biyani, and P. Valdastri, "Autonomous tissue retraction in robotic assisted minimally invasive surgery–a feasibility study," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6528–6535, 2020.

[13] F. Ito, D. Jarrard, and J. C. Gould, "Transabdominal preperitoneal robotic inguinal hernia repair," *J. Laparoendoscopic & Adv. Surg. Techn.s*, vol. 18, no. 3, pp. 397–399, 2008.

[14] B. Gibaud, G. Forestier, C. Feldmann, G. Ferrigno, P. Gonçalves, T. Haidegger, C. Julliard, D. Katić, H. Kenngott, L. Maier-Hein, *et al.*, "Toward a standard ontology of surgical process models," *Int. J. Comp. Assist. radiology and surgery*, vol. 13, no. 9, pp. 1397–1408, 2018.

[15] D. Harel, "Statecharts: A visual formalism for complex systems," *Science of computer programming*, vol. 8, no. 3, pp. 231–274, 1987.

[16] A. Zia, A. Hung, I. Essa, and A. Jarc, "Surgical activity recognition in robot-assisted radical prostatectomy using deep learning," in *Med. Image Comp. and Comp. Assist. Inter.*, A. F. et. al, Ed. Springer, 2018, pp. 273–280.

[17] C. Lea, R. Vidal, and G. D. Hager, "Learning convolutional action primitives for fine-grained action recognition," in *IEEE Int. Conf, Robotics and Automation*, 2016, pp. 1642–1649.

[18] R. DiPietro, C. Lea, A. Malpani, N. Ahmidi, S. S. Vedula, G. I. Lee, M. R. Lee, and G. D. Hager, "Recognizing surgical activities with recurrent neural networks," in *Int. Conf. Med. image Comp. and Comp.-Assist. Intervention*, 2016, pp. 551–558.

[19] G. Menegozzo, D. Dall'Alba, C. Zandonà, and P. Fiorini, "Surgical gesture recognition with time delay neural network based on kinematic data," in *Int. Symp. Medical Robotics*, 2019, pp. 1–7.

[20] E. Mavroudi, D. Bhaskara, S. Sefati, H. Ali, and R. Vidal, "End-to-end fine-grained action segmentation and recognition using conditional random field models and discriminative sparse coding," in *IEEE Winter Conf. App.s Computer Vision*, 2018, pp. 1558–1567.

[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," *arXiv preprint arXiv:1812.00033*, 2018.

[22] A. Zia, C. Zhang, X. Xiong, and A. M. Jarc, "Temporal clustering of surgical activities in robot-assisted surgery," *IJCARS*, vol. 12, no. 7, pp. 1171–1178, 2017.

[23] L. Tao, E. Elhamifar, S. Khudanpur, G. D. Hager, and R. Vidal, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *Int. Conf. Info. Process. in Comp.-Assist. Interv.* Springer, 2012, pp. 167–177.

[24] J. Rosen, J. D. Brown, L. Chang, M. N. Sinanan, and B. Hannaford, "Generalized approach for modeling minimally invasive surgery as a stochastic process using a discrete markov model," *IEEE Transactions on Biomedical engineering*, vol. 53, no. 3, pp. 399–413, 2006.

[25] S. Krishnan, A. Garg, S. Patil, C. Lea, G. Hager, P. Abbeel, and K. Goldberg, "Transition state clustering: Unsupervised surgical trajectory segmentation for robot learning," in *Robotics Research*. Springer, 2018, pp. 91–110.

[26] L. Tao, L. Zappella, G. D. Hager, and R. Vidal, "Surgical gesture segmentation and recognition," in *Int. Conf. Med. Image Comp. and Comp.-Assist. Intervention*. Springer, 2013, pp. 339–346.

[27] D. Zhang, B. Xiao, B. Huang, L. Zhang, J. Liu, and G.-Z. Yang, "A self-adaptive motion scaling framework for surgical robot remote control," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 359–366, 2018.

[28] B. Xiao, W. Xu, J. Guo, H.-K. Lam, G. Jia, W. Hong, and H. Ren, "Depth estimation of hard inclusions in soft tissue by autonomous robotic palpation using deep recurrent neural network," *IEEE Trans. Auto. Science Engineering*, vol. 17, no. 4, pp. 1791–1799, 2020.

[29] Y.-H. Long, J.-Y. Wu, B. Lu, Y.-M. Jin, M. Unberath, Y.-H. Liu, P.-A. Heng, and Q. Dou, "Relational graph learning on visual and kinematics embeddings for accurate gesture recognition in robotic surgery," *arXiv preprint arXiv:2011.01619*, 2020.

[30] 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, 2016.

[31] 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, 2017.

[32] T. Czempiel, M. Paschali, M. Keicher, W. Simson, H. Feussner, S. T. Kim, and N. Navab, "Tecno: Surgical phase recognition with multi-stage temporal convolutional networks," in *Int. Conf. Med. Image Comp. and Comp.-Assist. Intervention*. Springer, 2020, pp. 343–352.

[33] Y. Ding, J. Fan, K. Pang, H. Li, T. Fu, H. Song, L. Chen, and J. Yang, "Surgical workflow recognition using two-stream mixed convolution network," in *Int. Conf. Adv. Elec. Materials, Comp.s and Soft. Eng.*, 2020, pp. 264–269.

[34] S. Liu, Y. Zheng, Z. Xiaosong, H.-j. Wu, K. Zhang, A. E. Keinath, and C. Y. Chung, "Hierarchical segmentation and quality measurement for video editing," Dec. 8 2016, uS Patent App. 15/173,465.

[35] B. Gunsel, Y. Fu, and A. M. Tekalp, "Hierarchical temporal video segmentation and content characterization," in *Multimedia Storage and Arch. Syst.s II*, vol. 3229. Int. Soc. Opt.s and Phot.s, 1997, pp. 46–56.

[36] D. DeMenthon and R. Megret, *Spatio-temporal segmentation of video by hierarchical mean shift analysis*. Citeseer, 2002.

[37] T. Lan, Y. Zhu, A. Roshan Zamir, and S. Savarese, "Action recognition by hierarchical mid-level action elements," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 4552–4560.

[38] N. Ahmidi, L. Tao, S. Sefati, Y. Gao, C. Lea, B. B. Haro, L. Zappella, S. Khudanpur, R. Vidal, and G. D. Hager, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2025–2041, 2017.

[39] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. image Comp. and Comp.-Assist. intervention*. Springer, 2015, pp. 234–241.

[40] F. A. Gers and J. Schmidhuber, "Recurrent nets that time and count," in *Proc. IEEE Int. Joint Conf. Neural Net.s*, vol. 3, 2000, pp. 189–194.

[41] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint arXiv:1409.1259*, 2014.

[42] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," *arXiv preprint arXiv:1704.02971*, 2017.

[43] M. Allan, S. Ourselin, D. J. Hawkes, J. D. Kelly, and D. Stoyanov, "3-d pose estimation of articulated instruments in robotic minimally invasive surgery," *IEEE TMI*, vol. 37, no. 5, pp. 1204–1213, 2018.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Adv. Neural Info. Process. Systems*, 2014, pp. 2672–2680.

[45] C. Lea, R. Vidal, A. Reiter, and G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *European Conference on Computer Vision*. Springer, 2016, pp. 47–54.