Learning Invariant Representation of Tasks for Robust Surgical State Estimation

Yidan Qin\(^1\), Max Allan\(^3\), Yisong Yue\(^2\), Joel W. Burdick\(^2\), Mahdi Azizian\(^1\)

Abstract—Surgical state estimators in robot-assisted surgery (RAS) - especially those trained via learning techniques - rely heavily on datasets that capture surgeon actions in laboratory or real-world surgical tasks. Real-world RAS datasets are costly to acquire, are obtained from multiple surgeons who may use different surgical strategies, and are recorded under uncontrolled conditions in highly complex environments. The combination of high diversity and limited data calls for new learning methods that are robust and invariant to operating conditions and surgical techniques. We propose StiseNet, a Surgical Task Invariance State Estimation Network with an invariance induction framework that minimizes the effects of variations in surgical technique and operating environments inherent to RAS datasets. StiseNet’s adversarial architecture learns to separate nuisance factors from information needed for surgical state estimation. StiseNet is shown to outperform state-of-the-art state estimation methods on three datasets (including a new real-world RAS dataset: HERNIA-20).

I. INTRODUCTION

While the number of Robot-Assisted Surgeries (RAS) continues to increase, at present they are entirely based on teleoperation. Autonomy has the potential to improve surgical efficiency and to improve surgeon and patient comfort in RAS, and is increasingly investigated [1]. Autonomy can be applied to passive functionalities [2], situational awareness [3], and surgical tasks [4], [5]. A key prerequisite for surgical automation is the accurate real-time estimation of the current surgical state. Surgical states are the basic elements of a surgical task, and are defined by the surgeon’s actions and observations of environmental changes [6]. Awareness of surgical states would find applications in surgical skill assessment [7], identification of critical surgical states, shared control, and workflow optimization [8].

Short duration surgical states, with their inherently frequent state transitions, are challenging to recognize, especially in real-time. Many prior surgical state recognition efforts have employed only one type of operational data. Hidden Markov Models [7], [9], Conditional Random Fields (CRF) [10], Temporal Convolutional Networks (TCN) [11], Long-Short Term Memory (LSTM) [12], and others have been used to recognize surgical actions using robot kinematics data. Methods based on Convolutional Neural Networks (CNN), such as CNN-TCN [11] and 3D-CNN [13], have been applied to endoscopic vision data. RAS datasets consist of synchronized data streams. The incorporation of multiple types of data, including robot kinematics, endoscopic vision, and system events (e.g., camera follow: a binary variable indication of if the endoscope is moving), can improve surgical state estimation accuracy in methods such as Latent Convolutional Skip-Chain CRF [14] and Fusion-KVE [6].

Prior surgical state estimators relied heavily on RAS datasets for model fitting/training. Limitations in the dataset can be propagated (and perhaps amplified) to the estimator, possibly resulting in a lack of robustness and cross-domain generalizability [14]. Many surgical activity datasets are derived from highly uniform tasks performed using the same technique in only one setting. E.g., the JHU-ISI Gesture and Skill Assessment Working Set (JIGSAWS) [15] suturing task was obtained in a bench-top setting, with suturing performed on marked pads (Fig. 1). Valuable anatomical background visual information is not present in the training data, which may lead to errors when the estimator is applied in real-world surgeries. Moreover, state estimators that are trained on datasets devoid of endoscope motion do not generalize well to new endoscopic views. Endoscope movements are frequent and spontaneous in real-world RAS. Additionally, operators in existing surgical activity datasets typically perform the task with the same technique, or were instructed to follow a predetermined workflow, which limits variability among trials. These limitations can cause state estimators to overfit to the techniques presented during training, and make inaccurate associations between surgical states and specific placements of surgical instruments and/or visual layout, instead of truly relevant features.

In real-world RAS tasks, endoscope lighting and viewing angles, surgical backgrounds, and patient health condition vary considerably among trials, as do state transition probabilities. We consider these variations as potential nuisance factors that increase the training difficulty of a robust surgical state estimator. Moreover, surgeons may employ diverse techniques to perform the same surgical task depending on patient condition and surgeon preferences. While the effects of nuisances and technique variations on estimation accuracy can be reduced by a large and diverse real-world RAS dataset, such datasets are costly to acquire.

The combination of limited data and high diversity calls

\(^1\)Intuitive Surgical Inc., 1020 Kifer Road, Sunnyvale, CA, 94086, USA
\(^2\)Mechanical and Civil Engineering, Caltech, Pasadena, CA, 91125, USA
Emails: Yidan.Qin@intusurg.com, Mahdi.Azizian@intusurg.com

Fig. 1: Top row: typical endoscopic images from HERNIA-20 (left), JIGSAWS (middle) and RIOUS+ (right) datasets. Bottom row: related surgical state sequence samples, where each color represents a different surgical state.
for more robust state estimation training methods, as state-of-the-art methods are not accurate enough for adoption in the safety critical field of RAS. Surgical state estimation can be made invariant to irrelevant nuisances and surgeon techniques if latent representations of the input data contain minimal information about those factors [16]. \textit{Invariant representation learning} (IRL) has been an active research topic in computer vision, where robustness is achieved through \textit{invariance induction} [16]–[20]. Zemel et al. proposed a supervised adversarial model to achieve fair classification under the two competing goals of encoding the input data correctly and obfuscating the group to which the data belongs [17]. A regularized loss function using information bottleneck also induces invariance to nuisance factors [18]. Jaiswal et al. described an adversarial invariance framework in which nuisance factors are distinguished through disentanglement [19], and bias is distinguished through the competition between goal prediction and bias obfuscation [20]. Previous work on IRL via adversarial invariance in time series data focused mostly on speech recognition [21], [22]. RAS data, arising from multiple sources, provides a new domain for IRL of high-dimensional noisy time series data.

\textbf{Contributions:} We propose \textit{StiseNet}, a surgical state estimation model that is largely invariant to nuisances and variations in surgical techniques. \textit{StiseNet}'s adversarial design pits two composite models against each other to yield an invariant latent representation of the endoscopic vision, robot kinematics, and system event data. \textit{StiseNet} learns a split representation of the input data through the competitive training of state estimation and input data reconstruction, and the disentanglement between essential information and nuisance. The influence of surgeon technique is excluded by adversarial training between state estimation and the obfuscation of a latent variable representing the technique type. \textit{StiseNet} training does not require any additional annotation apart from surgical states. Our main contributions include:

\begin{itemize}
  \item An adversarial model design that promotes invariance to nuisance and surgical technique factors in RAS data.
  \item A process to learn invariant latent representations of real-world RAS data streams, minimizing the effect of factors such as patient condition and surgeon technique.
  \item Improving frame-wise surgical state estimation accuracy for online and offline real-world RAS tasks by up to 7%, which translates to a 28% relative error reduction.
  \item Combining semantic segmentation with endoscopic vision to leverage a richer visual feature representation.
  \item Demonstrating the method on 3 RAS datasets.
\end{itemize}

\textit{StiseNet} is evaluated and demonstrated using JIGSAWS suturing [15], RIOUS+ [6], and a newly collected HERNIA-20 dataset containing real-world hernia repair surgeries. \textit{StiseNet} outperforms state-of-the-art surgical state estimation methods and improves frame-wise state estimation accuracy to 84%. This level of error reduction is crucial for state estimation to gain adoption in RAS. \textit{StiseNet} also accurately recognizes actions in a real-world RAS task even when a specific technique was not present in the training data.

\section{Methods}

\textit{StiseNet} (Figs. 2 and 3) accepts synchronized data streams of endoscopic vision, robot kinematics, and system events as inputs. To efficiently learn invariant latent representations of noisy data streams, we adopt an adversarial model design loosely following Jaiswal et al. [20] but with model architectures more suitable for time series data. Jaiswal et al.'s adversarial invariance framework for image classification separates useful information and nuisance factors, such as lighting conditions, before performing classification. \textit{StiseNet} extends this idea by separating learned features from RAS time series data into desired information for state estimation ($e_1$) and other information ($e_2$). Estimation is performed using $e_1$ to eliminate the negative effects of nuisances and variations in surgical techniques. \textit{LSTM} computational blocks are used for feature extraction and surgical state estimation. \textit{LSTMs} learn memory cell parameters that govern when to forget/read/write the cell state and memory [12]. They therefore better capture temporal correlations in time series data. \textit{StiseNet}'s components and training procedure are described next. Table I lists key concepts and notation.

\subsection{Feature extraction}

Fig. 2 depicts the extraction of features from endoscopic vision, robot kinematics, and system events data. Visual features are extracted by a CNN-LSTM model [24], [25]. To eliminate environmental distractions in the endoscopic view, a previously trained and frozen surgical scene segmentation model based on U-Net [26] extracts a pixel-level semantic mask for each frame. We use two scene classes: tissue and surgical instrument. The semantic mask is concatenated to the unmodified endoscope image as a fourth image channel. This RGB-Mask image $I_t \in \mathbb{R}^{h \times w \times 4}$ is then input to the CNN-LSTM. We implemented a U-Net-style feature map to extract visual features, $x_t^{vis}$, since a condensed surgical scene representation can be taken advantage of by adapting U-Net weights of the semantic segmentation model trained on a large endoscopic image dataset. We implemented an LSTM encoder to better capture temporal correlations in visual CNN features. This helps the visual processing system to extract visual features that evolve in time. At time $t$, a visual latent state, $h^{vis}_t \in \mathbb{R}^{n_{vis}}$, is extracted with the LSTM model.

\begin{table}[h]
\begin{tabular}{|c|c|}
\hline
\textbf{Notation} & \textbf{Description} \\
\hline
$H$ & Concatenated vision, kinematics, and event features \\
$s$ & Surgical state \\
$T_{obs}$ & Observational window size \\
$\psi$ & All factors pertinent to the estimation of $s$ \\
$\eta$ & All other factors (nuisance factors), which are of no interest to goal variable estimation [23] \\
$l$ & Latent variable (type of surgical technique) \\
$\overline{d}$ & Mean silhouette coefficient quantifying clustering quality \\
\hline
$E$ & Encoder encodes $H$ into $e_1$ and $e_2$ \\
$M$ & Estimator infers $s$ from $e_1$ \\
$\psi$ & Dropout \\
$R$ & Reconstructor attempts to reconstruct $H$ from $\psi(e_1), e_2$ \\
$f_1$ & Disentangler infers $e_1$ from $e_1$ \\
$f_2$ & Disentangler infers $e_1$ from $e_2$ \\
$D$ & Discriminator estimates $l$ from $e_1$ \\
\hline
\end{tabular}
\caption{Key variables, concepts, and notation.}
\end{table}
Fig. 2: Features $h^{\text{vis}}$, $h^{\text{kin}}$, and $h^{\text{cut}}$ are respectively extracted from endoscopic vision, robot kinematics, and system events. A semantic mask is appended to the endoscopic vision data to form an RGB-Mask vision input.

Kinematics data are recorded from the Universal Patient-Side Manipulator (USM) of the da Vinci® Surgical System. Kinematics features are extracted using an LSTM encoder with attention mechanism [27] to identify the important kinematics data types [24]. A multiplier $\alpha_i$, whose elements weight each type of kinematics data, was learned as follows:

$$\alpha_i = \text{softmax} \left\{ u^T \tanh \left( W (h_{t-1}^{\text{kin}}, c_t^{\text{kin}}) + VX_t^{\text{kin}} \right) \right\}, \tag{1}$$

where $h_{t-1}^{\text{kin}}$ is the latent state from the previous frame, $c_t^{\text{kin}}$ is the LSTM cell state, and $X_t^{\text{kin}} = (x_t^{\text{kin}}_{t-1}, \ldots, x_t^{\text{kin}}_{t+1})$ denote the kinematic data inputs. $u$, $W$, and $V$ are learnable parameters. The weighted kinematics data feature vector $h_t^{\text{kin}} \in \mathbb{R}^{n_{\text{kin}}}$ is calculated as:

$$h_t^{\text{kin}} = \text{LSTM}(h_{t-1}^{\text{kin}}, \alpha_t \cdot x_t^{\text{kin}}) \tag{2}$$

The da Vinci® Xi Surgical System also provides system event data (details in Section III). The event features $h_t^{\text{cut}}$ are extracted via the same method as kinematics.

B. Feature encoder and Surgical state estimator

As shown in Fig. 2, Encoder $E$ extracts useful information for estimation from the latent feature data $H$. If we assume that $H$ is composed of a set of factors of variation, then $H$ is composed of mutually exclusive subsets:

- $e_1$: all the factors pertinent to the estimation of the goal variable (the current surgical state $s$);
- $e_2$: all other factors (nuisance factors), which are of no interest to goal variable estimation [23].

Encoder $E$ is a function trained to partition $H$: $[e_1, e_2] = E(H)$. A fully-connected (FC) layer maps $H$ to $e_1$, and another FC layer maps $H$ to $e_2$. Once distinguished, the surgical state $s$ at time $t$ is estimated from the history of the useful signal $\{e_1, t-T_{\text{data}}+1, \ldots, e_1, t\}$ using an LSTM decoder $M$ following [24]. By learning the parameters in $M$ using $e_1$ instead of $H$, we avoid learning inaccurate associations between nuisance factors and the goal variable.

C. Learning an invariant representation

The invariance induction to nuisance and technique factors is learned via competition and adverseness between model components [28] (yellow and pink shaded components in Fig. 3). While $M$ encourages the pooling of factors relevant to surgical state estimation in signal $e_1$, a reconstructor $R$ (a function implemented as an FC layer) attempts to reconstruct from the separated signals. Dropout $\psi$ is added to $e_1$ to make it an unreliable source to reconstruct $H$ [19]. This configuration of signals prevents a convergence to the trivial solution where $e_1$ monopolizes all information, while $e_2$ contains none. The mutual exclusivity between $e_1$ and $e_2$ is achieved through adversarial training. Two FC layers $f_1$ and $f_2$ are implemented as disentanglers. $f_1$ attempts to infer $e_2$ from $e_1$, while $f_2$ infers $e_1$ from $e_2$. To achieve mutual exclusivity, we should not be able to infer $e_1$ from $e_2$ or vice versa. Hence, the losses of $f_1$ and $f_2$ must be maximized. This leads to an adversarial training objective [29]. The loss function with invariance to nuisance factors is:

$$L_{\text{nuis}} = \alpha L_M (s; M(e_1)) + \beta L_R (H; R(e_2, \psi(e_1))) + \gamma (L_f (e_1; f_1(e_2)) + L_f (e_2; f_2(e_1))) \tag{3}$$

where $\alpha$, $\beta$, and $\gamma$ respectively weight the adversarial loss terms [29] associated with architectural components $M$, $R$, and disentanglers $f_1$ and $f_2$. The training objective with invariance to nuisance factors is a minimax game [28], [30]:

$$\min_{P_1} \max_{P_2} L_{\text{nuis}} \tag{4}$$

where the loss of component $P_1 = \{E, M, R\}$ is minimized while the loss of $P_2_{\text{nuis}} = \{f_1, f_2\}$ maximized.

Besides the presence of nuisance factors, variability in $H$ could also arise from variability in surgical techniques. Variations in technique may not be entirely separable by an invariance to nuisance factors, as they may be correlated to the surgical state. StiseNet therefore adopts an adversarial debiasing design [31] that deploys a discriminator $D : e_1 \rightarrow l$ for surgical technique invariance. The latent variable $l$ represents the type of technique employed to perform a surgical task. $l$ is a trial-level categorical attribute that is inferred by k-means clustering of kinematics time series training data based on a dynamic time warping distance metric (function $\phi$) [32]. The clusters represent different surgical techniques used in the training trials. The optimal number of clusters $k$ is dataset-specific. To determine it, we implemented the elbow method using inertia [33] and the silhouette method [34]. The inertia is defined as the sum of squared distances between each cluster member and its cluster center [33] for all clusters. The inertia decreases as $k$ increases, and the elbow point is a relatively optimal $k$ value [33]. The silhouette coefficient $d_i$ for time series $i$ is:

$$d_i = \frac{\text{mean}(j \left( \min_{m \in C_i} \sum_{j \in C_m} \phi(i,j) - \sum_{j \in C_i \cap \neq k} \phi(i,j) \right))}{\max(a_i, b_i)} \tag{5}$$

where $C_i$ is the cluster of time series $i$. The operation $\min$ represents the closest time series to $i$ that does not belong to $C_i$. $\max(a_i, b_i)$ is a measure of how close each data point in one cluster is to data points in the nearest neighboring clusters. The $k$ with the highest $d_i$ is the optimal
number of clusters. The loss function with invariance to both
nuisance and surgical techniques is then:
\[
L = L_{\text{nuis}} + \delta L_D(l, D(e_1))
\]
where $\delta$ is the weight associated with the discriminator loss.
The term $P2$ contains an additional term: $P2 = \{f_1, f_2, D\}$:
\[
\min_{P1} \max_{P2} L.
\]

D. Training and inference

StiseNet’s feature extraction components were trained
following [6]. Specifically, the first three channels of the top
layer in U-Net visual feature map were initialized with the
weights from the surgical scene segmentation model. The
visual input was resized to $h = 256$ and $w = 256$. The
extracted features have dimensions $n_{vis} = 40$, $n_{kin} = 40$,
and $n_{ext} = 4$, which were determined using grid search.
All data sources are synchronized at 10Hz with $T_{obs} =
20$ samples $= 2$sec. The optimal cluster number, $k$, for
JIGSAWS, RIOUS+, and HERNIA-20 were 9, 7, and 4,
respectively. The temporal clustering process was repeated to
ensure reproducibility due to the randomness in initialization.
Section IV described how $k$ is determined in these datasets.

StiseNet is trained end-to-end with the minimax objectives
(Eq.s4 and 7). We used the categorical cross-entropy loss for
$L_M$ and $L_D$. $L_f$ and $L_R$ are mean squared error loss. $\psi$ is a
dropout [35] with the rate of 0.4, 0.1, and 0.4 for JIGSAWS,
RIOUS+, and HERNIA-20, respectively. To effectively train
the adversarial model, we applied a scheduled adversarial
optimizer [28], in which a training batch is passed to either
$P1$ or $P2$ while the other component’s weights are frozen.
The alternating schedule was found by grid search to be 1:5.

III. EXPERIMENTAL EVALUATION

We evaluated StiseNet’s performance on the JIGSAWS
suturing [15], RIOUS+ [6], and a newly collected HERNIA-
20 dataset, respectively. These datasets were annotated with
manually determined lists of fine-grained states (Table II).

A. Datasets

JIGSAWS: The JIGSAWS [15] bench-top suturing task
includes 39 trials by eight surgeons partaking in nine surgical
actions. We used the endoscopic vision and USM’s kine-
matics (gripper angle, translational and rotational positions
and velocities) data. There was no system events data. The
tooltips’ orientation matrices were converted to Euler angles.

RIOUS+: The RIOUS+ dataset, introduced in [6], [24],
captures 40 trials of an ultrasound scanning task on a da
Vinci Xi® Surgical System by five users in a mixture of
bench-top (27) and OR (13) trials. Eight states represent user
actions or environmental changes. Endoscopic vision, USM
kinematics, and six binary system events serve as inputs—see
[6]. A finite state machine model of the task was determined
prior to data collection. The operators were instructed to
strictly follow this predetermined task workflow and to
ignore environmental disruptions. The action sequences and
techniques are therefore highly structured and similar across
trials. While it includes more realistic RAS elements, such
as OR settings and endoscope movements, RIOUS+ lacks
the behavioral variability of real-world RAS data.

HENRIA-20: The HERNIA-20 dataset contains 10 fully
anonymized real-world robotic transabdominal preperitoneal
inguinal hernia repair procedures performed by surgeons on
da Vinci Xi® Surgical Systems. For performance evalua-
tion, we selected a running suturing task performed to re-
approximate the peritoneum, which contains 11 states. The
endoscopic vision, USM kinematics, and system events are
used as inputs. Because HERNIA-20 captures real-world
RAS performed on patients, the robustness of surgical state
estimation models can be fully examined.

B. Metrics

The quality of the learned invariant representations of sur-
gical states $e_1$ and other information $e_2$ is visually examined.
Arrays of $e_1$ and $e_2$ in each state instance (a consecutive
block of time frames of the same surgical state) are embed-
ded in 2D space using the Uniform Manifold Approximation
and Projection (UMAP) algorithm [36] - a widely-adopted
dimension reduction and visualization method that preserves
more of the global structure of the data.

We used the percentage of accurately identified frames in
a test set to evaluate each model’s surgical state estimation
accuracy. Model performance was evaluated in non-causal
and causal settings. In a non-causal setting, the model can

Fig. 3: StiseNet training architecture. Symbols for the estimator components $P1 = \{E, M, R\}$ are red, the adversarial
component $P2 = \{f_1, f_2, D\}$ is blue, and training loss calculations are black. $P2$ implements invariance to nuisance (yellow
shading) and surgical techniques (pink shading). RAS data features $H$ are divided into information essential for state
estimation, $e_1$, and other information $e_2$. $H$ is reconstructed from $\psi(e_1)$ and $e_2$, where $\psi$ is dropout.
| Gesture | JIGSAWS Suturing Dataset | Duration (s) |
|---------|--------------------------|--------------|
| G1      | Reaching for the needle with right hand | 2.2 |
| G2      | Positioning the tip of the needle | 3.4 |
| G3      | Pushing needle through the tissue | 9.0 |
| G4      | Transferring needle from left to right | 4.5 |
| G5      | Moving to center with needle in grip | 3.0 |
| G6      | Pulling suture with left hand | 4.8 |
| G7      | Orienting needle | 7.7 |
| G8      | Using right hand to help tighten suture | 3.1 |
| G9      | Dropping suture and moving to end points | 3.1 |

| Gesture | RIOUS+ Dataset | Duration (s) |
|---------|----------------|--------------|
| H1      | Probe released, out of endoscopic view | 6.3 |
| S2      | Probe released, in endoscopic view | 7.6 |
| S3      | Reaching for probe | 3.1 |
| S4      | Grasping probe | 1.1 |
| S5      | Lifting probe up | 2.4 |
| S6      | Carrying probe to tissue surface | 2.3 |
| S7      | Sweeping | 5.1 |
| S8      | Releasing probe | 1.7 |

| Gesture | HERNIA-20 Dataset | Duration (s) |
|---------|------------------|--------------|
| S1      | Reaching for the needle | 3.9 |
| S2      | Positioning the tip of the needle | 3.3 |
| S3      | Pushing needle through the tissue | 4.2 |
| S4      | Pulling tissue with left hand | 3.6 |
| S5      | Transferring needle from left to right | 3.7 |
| S6      | Orienting needle | 6.6 |
| S7      | Pulling suture with left hand | 5.8 |
| S8      | Pulling suture with right hand | 4.8 |
| S9      | Transferring needle from right to left | 4.6 |
| S10     | Using right hand to tighten suture | 4.3 |
| S11     | Adjusting endoscope | 3.8 |

TABLE II: Datasets State Descriptions and Mean Duration

use information from future time frames, which is suitable for post-operative analyses. In causal settings, the model only has access to the current and preceding time frames. Surgical state estimation is harder in the causal setting; however, it is a more useful evaluation metric for real-time applications.

We used the source code provided by the authors of the comparison methods when the model performance of a particular setting or dataset was not available [11], [12] and performed training and evaluation ourselves. JIGSAWS suturing and RIOUS+ datasets were evaluated using Leave One User Out (LOUO) [15], while HERNIA-20 was evaluated using 5-fold cross validation, since each trial’s surgeon ID is not available due to privacy protection.

C. Ablation Study

We compared StiseNet against its two ablated versions: StiseNet-Non Adversarial (StiseNet-NA) and StiseNet-Nuisance Only (StiseNet-NO). StiseNet-NA omits the adversarial component P2 entirely (the yellow and pink-shaded areas in Fig. [3]) and uses H for estimation with Estimator $M : H_{t} \rightarrow s_{t}$. StiseNet-NO separates useful information and nuisance factors, but excludes the invariance to surgical techniques (pink-shaded area in Fig. [3]). The ablation study demonstrates the necessity of the adversarial model design and individual contributions of each model component towards a more accurate surgical state estimation.

IV. RESULTS AND DISCUSSIONS

Fig. [4] plots for each dataset the total inertia and the mean silhouette coefficient $\bar{k}$ as functions of the number of clusters $k$. Fig. [5] shows the UMAP visualizations of $e_1$ and $e_2$ for all surgical states. We compare both the non-causal (Table III) and causal (Table IV) performance of StiseNet with its ablated versions and prior methods. Fig. [6] shows the variability in HERNIA-20 data through sample sequences from three technique clusters, each performed in a distinctively different style with environmental variances. Invariance of StiseNet to nuisances and surgical techniques is shown by its accurate surgical state estimations in the presence of visibly diverse input data. Fig. [7] shows a sample state sequence from HERNIA-20 and the causal state estimation results using multiple methods, including forward LSTM [12], Fusion-KVE [6], and the ablated and full versions of StiseNet.

As mentioned in Section II-C, the optimal number of clusters $k$ can be estimated from the elbow point of the inertia-$k$ curve, or the $k$ associated with the maximum mean silhouette coefficient $\bar{d}$. We implemented both methods and illustrate our choices of $k$ in Fig. [4]. The optimal $k$ is easily identifiable for JIGSAWS and HERNIA-20 (Fig. [4] and [4]), with the largest $\bar{d}$ occurs near the “elbow” of the inertia-$k$ curve. A peak in the RIOUS+ mean silhouette coefficient curve is less evident (Fig. [4]). The optimal number of clusters need not match the number of operators, as the inter-personal characteristics are not the only accountable factor for the variations among trials. Intra-personal variations can affect clustering. E.g., JIGSAWS contains metadata corresponding to expert ratings of each trial [15]: the ratings fluctuate among trials performed by the same surgeon. The optimal $k$ determined by kinematics data is somewhat robust against patient anatomy; however, a highly unique patient anatomy can lead surgeons to modify their maneuvers significantly. Such a trial could fall into a different technique cluster.

Fig. [5] visualizes the 2D projections of $e_1$ and $e_2$. The first row shows that $e_1$ and $e_2$ separate neatly into two clusters for all datasets, validating the effectiveness of disentanglers $f_1$ and $f_2$ since $e_1$ and $e_2$ contain little overlapping information. Since $e_1$ contains useful information for state estimation, while $e_2$ does not, $e_1$ should be better segregated into clusters.

Table III: State estimation performance in a non-causal setting.

| Non-causal | JIGSAWS | RIOUS+ | HERNIA-20 |
|-----------|---------|--------|-----------|
| TCN [11]  | kin     | 79.6   | 82.0      | 72.1      |
| TCN [11]  | vis     | 81.4   | 62.7      | 61.5      |
| Bidir. LSTM [12] | kin | 83.3   | 80.3      | 73.8      |
| LC-SC-CRF [14] | vis+kin | 83.5 | -         | -         |
| 3D-CNN [13] | vis     | 84.3   | -         | -         |
| Fusion-KVE [6] | vis+kin+evt | 86.3 | 93.8      | 78.0      |
| StiseNet-NA | vis+kin+evt | 86.5 | 93.1      | 80.0      |
| StiseNet-NO | vis+kin+evt | 87.9 | 90.3      | 83.2      |
| StiseNet  | vis+kin+evt | 90.2 | 92.5      | 84.1      |

Table IV: State estimation performance in a causal setting.

| Causal | JIGSAWS | RIOUS+ | HERNIA-20 |
|--------|---------|--------|-----------|
| TCN [11]  | vis     | 76.8   | 54.8      | 59.3      |
| TCN [11]  | kin     | 72.4   | 78.4      | 68.1      |
| Forward LSTM [12] | vis | 80.5   | 72.2      | 69.8      |
| 3D-CNN [13] | vis     | 81.8   | -         | -         |
| Fusion-KVE [6] | vis+kin+evt | 82.7 | 89.4      | 75.7      |
| StiseNet-NA | vis+kin+evt | 83.4 | 88.9      | 77.3      |
| StiseNet-NO | vis+kin+evt | 84.1 | 88.9      | 81.0      |
| StiseNet  | vis+kin+evt | 85.6 | 89.5      | 82.7      |
Fig. 4: Normalized inertia (with respect to the maximum value) and mean silhouette coefficient as functions of the number of clusters $k$ for each dataset. The vertical dotted line indicates the optimal $k$ (the maximum mean silhouette coefficient).

Fig. 5: 2D UMAP plots of information enclosed in $e_1$ and $e_2$ at each state instance. **Top row:** $e_1$ and $e_2$ segregates into distinguishable clusters, which indicates little overlap in information. **Middle row:** Information in $e_1$ color-coded by surgical states clusters relatively neatly. **Bottom row:** Information in $e_2$ is more intertwined and non-distinguishable by state. The mean silhouette coefficient $d$ of each graph is shown, with a larger $d$ indicating better clustering quality.

Table III and IV show non-causal and causal surgical state estimation performance of recently proposed methods and StiseNet (and its ablated versions). Both StiseNet and StiseNet-NO yield an improvement in frame-wise surgical state estimation accuracy for JIGSAWS suturing (up to 3.9%) and HERNIA-20 (up to 7%) under both settings, which shows the necessity and effectiveness of the adversarial model design. The non-causal performance of StiseNet on RIOUS+ is slightly worse compared to our Fusion-KVE method [6], which does not dissociate nuisance or style variables. This result can be explained by StiseNet’s model design and training scheme. The added robustness of StiseNet against variations in background, surgical techniques, etc. comes at the cost of the increased training complexity associated with adversarial loss functions and minimax training. Surgeon techniques and styles vary in JIGSAWS, and more significantly in HERNIA-20. Nuisance factors (tissue deformations, endoscopic lighting conditions and viewing angles, etc.) also vary considerably among trials and users in HERNIA-20. However, since RIOUS+ users were instructed to strictly follow a predetermined workflow, there are few nuisance and technique factors. The disentanglement between essential information $e_1$ and other information $e_2$ is therefore less effective. This hypothesis is supported by the observation that the dropout rate required for StiseNet training convergence is 0.1 for RIOUS+, whereas JIGSAWS and HERNIA-20 training converged with a dropout rate of 0.4. A lower dropout rate indicates that $e_2$ contains little information despite the dropout’s effort to avoid the trivial solution. Additionally, the uniformity across RIOUS+ participants results in a nearly constant mean silhouette coefficient (Fig. 4b). Hence, StiseNet’s invariance properties cannot be fully harnessed, explaining its less competitive performance in RIOUS+ as compared to the real-world RAS data of HERNIA-20.

In real-world RAS, surgeons may use different techniques to accomplish the same task. Fig. 6 shows three HERNIA-20 trials with distinctive suturing geometries: suturing from left to right, from right to left, and back and forth along a vertical seam. These trials fall into three clusters. We show images from instances of states S3, S4, S5, S7 and S8 in each trial. These images of different instances of the same state vary greatly not just in technique and instrument layout, but also in nuisance factors such as brightness and endoscope angles. Yet, StiseNet accurately estimates the surgical states due to its invariant latent representation of the input data.

Fig. 7 demonstrates StiseNet’s robustness during rapid and
Fig. 6: Three HERNIA-20 trials from three technique clusters, and StiseNet’s performances compared to ground truth (GT). Instances of the same state in different trials are substantially and visibly different; however, StiseNet correctly estimates them. Variations across trials arise from both nuisances and techniques. Potential sources of nuisances include but are not limited to lighting conditions, presence of fat or blood, peritoneum color, endoscope movements, etc.

unpredictable state transitions in a real-world RAS suturing task. We compare the causal estimation performance of Forward-LSTM, Fusion-KVE, the ablated, and full versions of StiseNet against ground truth. Forward-LSTM, which only uses kinematics data, has a block of errors from 20s to 30s since it cannot recognize the "adjusting endoscope" state due to a lack of visual and event inputs. When those inputs are added, Fusion-KVE and StiseNet recognize this state. Fusion-KVE still shows a greater error rate due to limited training data with high environmental diversity, which reflects Fusion-KVE’s vulnerability to nuisance and various surgical techniques. StiseNet-NO shows fewer error blocks: yet it is still affected by different technique types. The higher estimation accuracy of StiseNet shows its technique-agnostic robustness in real-world RAS, even with a small training dataset that contains behavioral and environmental diversity.

V. CONCLUSIONS AND FUTURE WORK
This paper focused on improving the accuracy of surgical state estimation in real-world RAS tasks learned from limited amounts of data with high behavioral and environmental diversity. We proposed *StiseNet*: an adversarial learning model with an invariant latent representation of RAS data. StiseNet was evaluated on three datasets, including a real-world RAS dataset that includes different surgical techniques carried out in highly diverse environments. StiseNet improves the state-of-the-art performance by up to 7%. The improvement is significant for the real-world running suture tasks, which benefit greatly from invariance to surgical techniques, environments, and patient anatomy. Ablation studies showed the effectiveness of the adversarial model design and the necessity of invariance inductions to both nuisance and technique factors. StiseNet training does not require additional annotation apart from the surgical states. We plan to further investigate alternative labelling methods of surgical techniques and the invariance induction to other latent variables such as surgeon ID, surgeon levels of expertise, etc. Due to the limited data availability, StiseNet has only been evaluated on small datasets. Adding more trials to HERNIA-20 will allow us to evaluate StiseNet more comprehensively. To further improve estimation accuracy, StiseNet’s neural network architectures may be further optimized for a better learning of temporal correlations within data. We also plan to incorporate longer-term context information [37], [38]. StiseNet’s accurate and robust surgical state estimation could also aide the development of surgeon-assisting functionalities and shared control systems in RAS.

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

REFERENCES
[1] 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.
Fig. 7: Example HERNIA-20 surgical state estimation results by forward LSTM [12], Fusion-KVE [6], StiseNet-NO, and StiseNet, compared to ground truth. State estimation results (top) and discrepancies with ground truth in red (bottom) are shown in each block bar.
This figure "kin_d.png" is available in "png" format from:

http://arxiv.org/ps/2102.09119v1