Improving Technical Proficiency in Robot-mediated Surgery Through Counterfactual Inquiry - Literature Review

Group 13: Hao, Ding

Mentor: Dr. Mathias Unberath

Brief Problem Summary

- **Motivation**

  Surgeon skill is among the strongest and most direct predictors for patient outcome. Using the powerful deep learning (DL) algorithm to assist novice surgeons deserves a try. However, DL algorithm suffers generalization issues and low interpretability. Incorporating causality into DL algorithm is one promising way to address these problems.

- **Goals**
  - Ultimate Goal:

    Build a robust system to assist surgeons especially novice surgeons perform at expert-level
  - Practical Subgoals:
    - Data: Build a suitable dataset for study
    - Understandings: Explore the difference between novice- and expert- level commands
    - Algorithm: Explore methods to incorporate causal inference mechanisms into deep learning algorithms

Paper Selected

- **Tittle**

  Learning Invariant Representation of Tasks for Robust Surgical State Estimation.

- **Brief Introduction**
  - Authors: Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, Mahdi Azizian
  - Source: https://arxiv.org/abs/2102.09119
  - Released date: 18 Feb 2021
  - Institutions: Intuitive Surgical Inc; Mechanical and Civil Engineering, Caltech.
  - Relation to our projects:
    - Work on Robot-Assisted Surgeries (RAS) data.
    - Aiming at improving robustness of the algorithm.
    - Their invariant representation shares some ideas with causal factors.

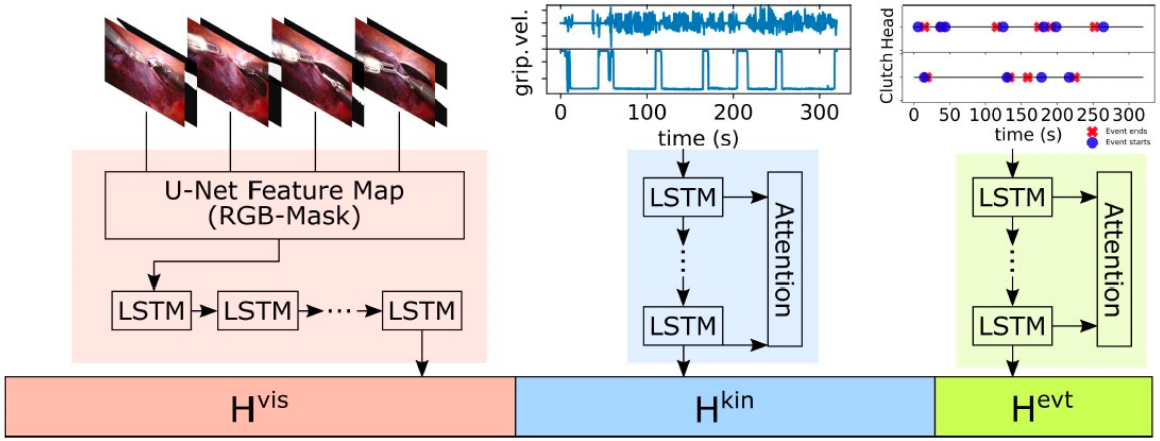
Introduction

- **Background**
  - Autonomy is the trend, and real-time estimation of the current surgical state is a key prerequisite
  - Prior surgical state estimators relied heavily on RAS datasets for model fitting/training, which leads to overfitting, especially to some nuisance factors like endoscope lighting.
  - Invariant representation learning (IRL) has been an active research topic in computer vision, where robustness is achieved through invariance induction.

- **Contribution - StiseNet**
  - An adversarial model design that promotes invariance to nuisance and surgical technique factors in RAS data.
  - 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.

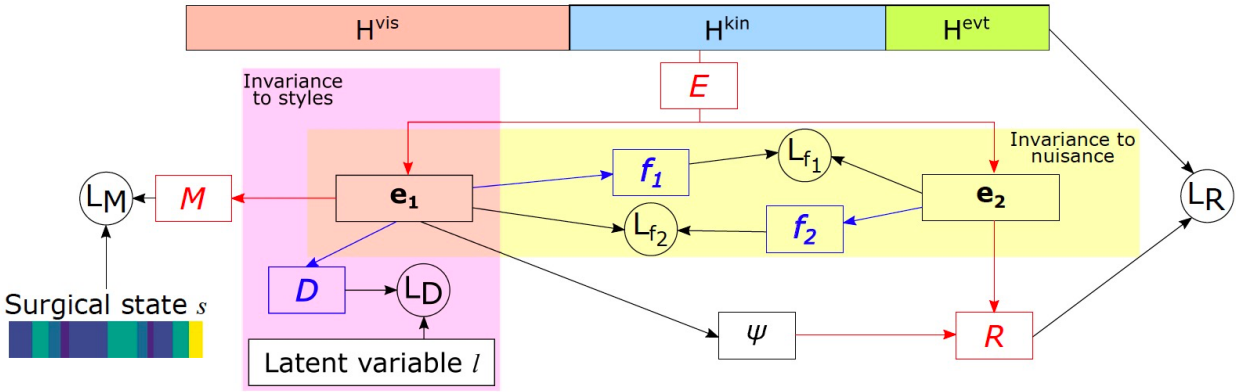
Methods

- **Feature extraction**
  - Extract **visual, kinematic, system event** feature simultaneously
  - RGB image + pretrained segmentation mask.
  - Using LSTM to extract visual feature.
  - Using LSTM with Attention extract kinematic and system event feature.



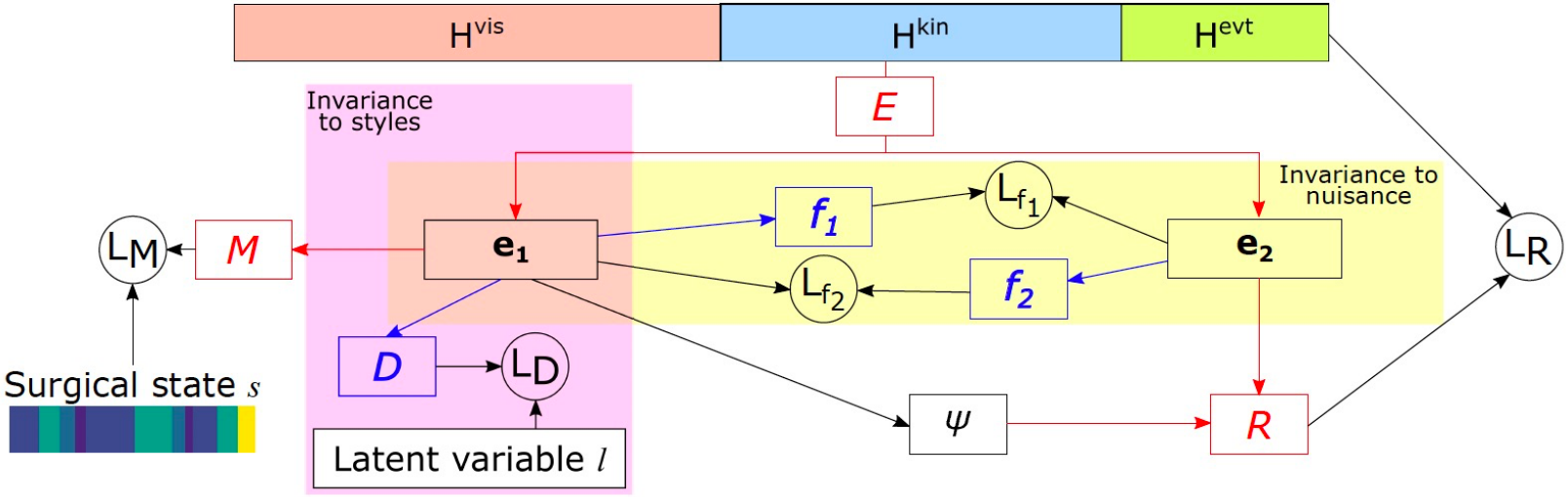
Methods

- Learning an invariant representation - components
  - Encoder E is a function trained to partition H: [e1, e2] =E(H) by two fc layers.
  - The surgical state s at time t is estimated from the history of the useful signal {e1,t−Tobs+1, . . . , e1,t} using an LSTM decoder M
  - Reconstructor R attempts to reconstruct from the separated signals.
  - Dropout ψ is added to e1 to make it an unreliable source to reconstruct H. (To ensure e1 is not trivially all information)
  - Two FC layers f1 and f2 are implemented as disentanglers. f1 attempts to infer e2 from e1, while f2 infers e1 from e2
  - discriminator D : e1 -> l for surgical technique invariance.



Methods

- Learning an invariant representation - optimization
  - L_M is minimized w,r,t. M and E to make better prediction for surgical states
  - L_R is minimized w.r.t R and E to make e2 contains necessary information to reconstruct feature.
  - L_{f_1}, L_{f_2} are maximized w.r.t f1 and f2 to make the e1 and e2 mutually exclusive.
  - L_D is maximized w.r.t D to make e1 can't be discriminable.



Experiments

- Accuracy:

Non-causal setting: information from future time frames

Causal setting: current and preceding time frames

StiseNet-NO separates useful information and nuisance factors, but excludes the invariance to surgical techniques

StiseNet-NA omits the adversarial component P2 entirely and uses H for estimation with Estimator M : Ht -> st.

Non-causal

	Input data	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**

Causal

	Input data	JIGSAWS	RIOUS+	HERNIA-20
TCN [11]	vis	76.8	54.8	58.3
TCN [11]	kin	72.4	78.4	68.1
Forward LSTM [12]	kin	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**

Experiments

- U-MAP:

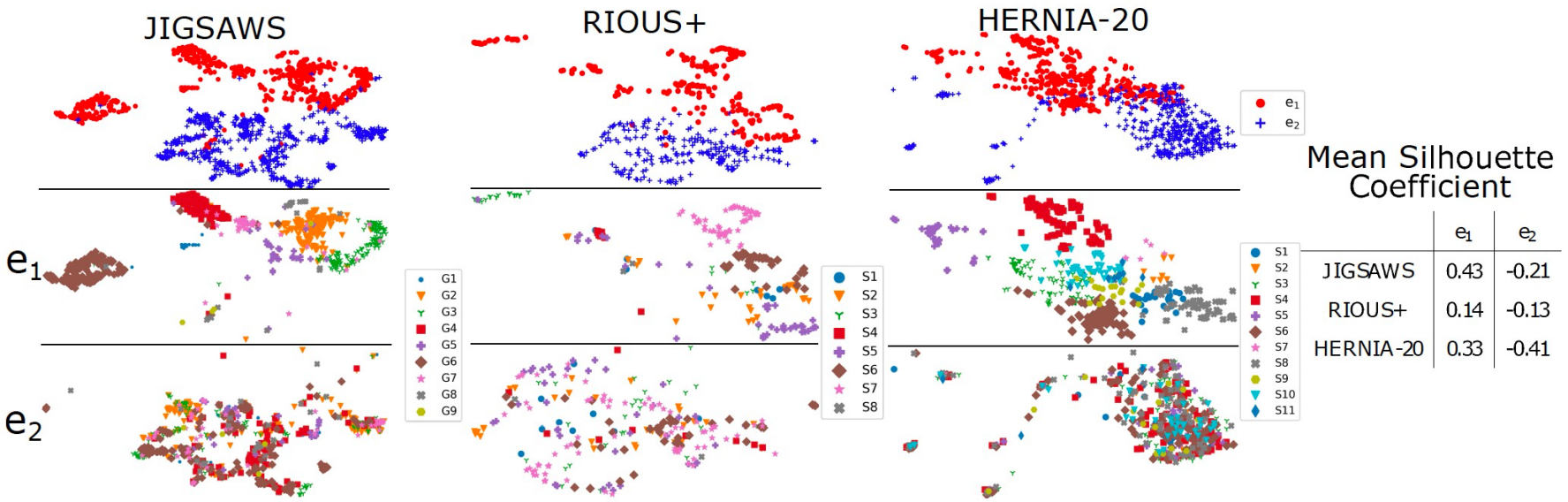


Fig. 5: 2D UMAP plots of information enclosed in $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ at each state instance. **Top row:** $\boldsymbol{e}_1$ and $\boldsymbol{e}_2$ segregates into distinguishable clusters, which indicates little overlap in information. **Middle row:** Information in $\boldsymbol{e}_1$ color-coded by surgical states clusters relatively neatly. **Bottom row:** Information in $\boldsymbol{e}_2$ is more intertwined and non-distinguishable by state. The mean silhouette coefficient $\overline{d}$ of each graph is shown, with a larger $\overline{d}$ indicating better clustering quality.

Assessment & Inspiration

- **Pros**
  - Provides a practical method for multi model feature extraction
  - Provides a good two-player game thoughts to learn invariant features
  - Optimizes a reconstructor to cut the trivial solution of e2.

- **Cons**
  - May made mistakes on the optimization design
  - Experiments results are not Steady

- **Inspirations**
  - The two-player game methods can be used to filter some nuisance factors in our project
  - The method also may help design causal learning.

References

- Yidan Qin, Max Allan, Yisong Yue, Joel W. Burdick, andMahdi Azizian. Learning invariant representation of tasks forrobust surgical state estimation, 2021