This paper appears in IEEE Transactions on Medical Robotics and Bionics, 2024.
https://doi.org/10.1109/TMRB.2024.3464110

Hands Collaboration Evaluation for Surgical Skills Assessment: An Information Theoretical Approach

Abed Soleymani[†,1], Mahdi Tavakoli[1,2], Farzad Aghazadeh[3], Yafei Ou[1], Hossein Rouhani[3], Bin Zheng[4], and Xingyu Li[1]

***Abstract*—Bimanual tasks, where the brain must simultaneously control and plan the movements of both hands, such as needle passing and tissue cutting, commonly exist in surgeries, e.g., robot-assisted minimally invasive surgery. In this study, we present a novel approach for quantifying the quality of hands coordination and correspondence in bimanual tasks by utilizing information theory concepts to build a mathematical framework for measuring the collaboration strength between the two hands. The introduced method makes no assumption about the dynamics and couplings within the robotic platform, executive task, or human motor control. We implemented the proposed approach on MEELS and JIGSAWS datasets, corresponding to conventional minimally invasive surgery (MIS) and robot-assisted MIS, respectively. We analyzed the advantages of hands collaboration features in the skills assessment and style recognition of robotic surgery tasks. Furthermore, we demonstrated that incorporating intuitive domain knowledge of bimanual tasks potentially paves the way for other complex applications, including, but not limited to, autonomous surgery with a high level of model explainability and interpretability. Finally, we presented preliminary results to argue that incorporating hands collaboration features in deep learning-based classifiers reduces uncertainty, improves accuracy, and enhances the out-of-distribution robustness of the final model.**

***Index Terms*—Information Theory, Deep Learning, Hands Collaboration Evaluation, Style Recognition, Surgical Skills Assessment.**

I. Introduction

Robot-assisted minimally invasive surgery (RAMIS) is becoming popular in modern clinical practice. To ensure the safety and quality of RAMIS-based operations, surgeons must acquire a variety of skills [1]. Conventional RAMIS skills assessment methods, which rely on structured checklists and rating scales [2], need expert observation and extensive time, making them subjective and less efficient. Moreover, the observational nature of conventional skills assessment methods makes them less sensitive to small but potentially important improvements in the trainee (e.g., a minor reduction in hand tremor or slight improvement in performing smooth motions [3]) and may fail to reveal the core reasons for unfavorable surgical outcomes. Due to the availability of surgical data in RAMIS platforms, automated surgical skills evaluation methods (e.g., artificial intelligence (AI)-based systems) are gaining traction as they are time and cost-efficient and do not need laborious human interventions. Additionally, automated surgical skills evaluation approaches also pave the way for robot-assisted training (e.g., applying virtual fixtures for training guidance) and autonomous surgery by providing knowledge about task performance [4], [5].

[†] Corresponding author: zsoleymani@ualberta.ca
Authors are with the [1]Electrical and Computer Engineering Department, [2]Biomedical Engineering Department, [3]Mechanical Engineering Department, [4]Surgery Department, University of Alberta, Edmonton, Alberta, Canada.

[‡] This research was supported by the Canada Foundation for Innovation (CFI), the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), Alberta Innovates, and the Alberta Jobs, Economy, and Trade Ministry's Major Initiatives Fund A-Medico.

These methods usually utilize convolutional neural networks (CNNs) [4], [6] and/or recurrent neural networks (RNNs) [7] to extract skills-related features and rate skills level of the user from kinematics or endoscopic video data [8]–[10] in surgical training datasets. Another category of studies such as Tao et al. [11] utilizes hidden Markov models (HMMs) to break up surgical trajectories into predefined building blocks called gestures to perform skills assessment.

Apart from the mentioned data-driven methods, the feature-based models extract understandable, meaningful, and clinically-proven metrics such as movement total path length [12], trajectory motion jerk [13], task execution time [12], [13], or even more complicated metrics including trajectory smoothness, fluidity of movements, and economy of motion [3].

The core limitation of the aforementioned surgical tasks which, based on our field knowledge, are bimanual tasks is that they treat each hand's trajectory as a package of independent data with no regard for possible *collaboration* between two hands. However, what defines a person as an *expert* surgeon is not just what they perform by each individual hand but what they plan for the next step by executing a complex sequence of correlated actions by making collaboration between both hands [14].

Hand collaboration is a complex cognitive ability that integrates the motor skills of both hands, and the majority of daily activities require some level of this ability [15]. Our brain is well-developed to manage hands collaboration; however, there is a considerable difference between the hands collaboration level in a task that we are doing for the first few times and that of the task we have mastered. Both intuition and prior research suggest that there is a meaningful connection between the level of hands collaboration and dexterity (i.e., skillful manipulation of the hands) [16]. Research conducted on patients who underwent surgical removal of the corpus callosum (i.e., the massive fibre tract that connects the two cerebral hemispheres in the brain) as a treatment against severe epilepsy reinforces the fact that gaining skills in performing complex or even simple bimanual tasks requires delicate hands collaboration which is the result of generating sophisticated connections between the left and right hemispheres of the brain over time [17], [18]. In Fig. 1, single and cyclic arrows connecting left and right hemispheres are illustrating corpus callosum fibre tract and single and double dashed line arrows are demonstrating corticospinal neurons of the lateral region of the brain that are responsible for receiving hands information and controlling hands movements [19].

In our setting, due to intuitive and mathematical motivations that we will discuss in Sections II and III, we decouple the notion of hands collaboration into two concepts: *hands coordination* and *hands correspondence*. Hand coordination measures the synchronized skill of the real-time movements of two hands working together towards parallel motion goals (e.g., the interaction between the

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2024.3464110

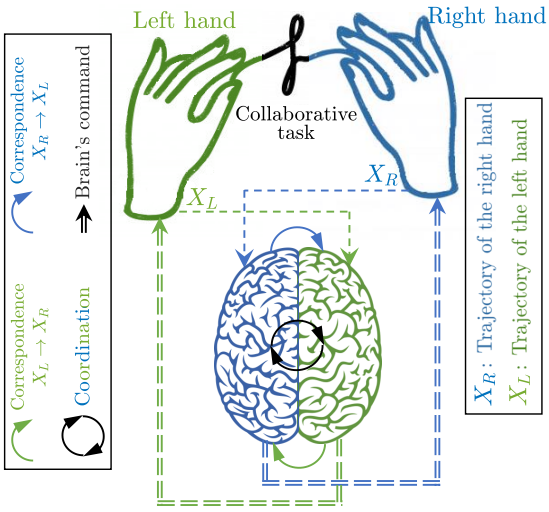

Figure 1: The illustration of the hands coordination and correspondence in bimanual tasks. Please note that, for example, the correspondence $\boldsymbol{X}_L \rightarrow \boldsymbol{X}_R$ represents the lead-follower relationship between the hands when the right hand is the leader.

vertical displacement of the right hand and that of the left hand in a bimanual lifting task). A good example of hands coordination in surgical tasks is when two hands are working together to pull out the two ends of the suture to tie a knot. Hands correspondence means how well one hand is serially followed by the motion step or intention of the other hand which allows us to capture the *leader-follower* relationship between the sequence of actions in human hands' displacements (e.g., grabbing the suture from the left hand (leader) with the right hand (follower)). An illustration for the concepts of correspondence and coordination in a collaborative task is provided in Fig. 1.

Previously, the bimanual dexterity between two hands was quantified using the correlation metric between velocities of surgical instruments held in hands [20]. However, the correlation metric is sensitive to time shifts in motion data. Aghazadeh et al. [21] implemented the dynamic time warping (DTW) algorithm to characterize bimanual dexterity to address the aforementioned shortcoming since the DTW algorithm is robust to time shifts in motion data of hands/surgical instruments, i.e., non-simultaneous movements. Despite the advantages of using the DTW metric to characterize the coordination of hands, this metric evaluates the coordination for the whole task and not dynamically. Therefore, quantitative and dynamic assessment of bimanual task performance has not yet been considered in prior surgical skill assessment approaches. This is because defining an exact and sufficiently comprehensive mathematical formulation for hands collaboration and especially capturing causality between their trajectories is problematic [22]. Another important factor that makes causality or coupling detection hard is that these relationships are often investigated in situations which are influenced by the user uncertainty [23] (e.g., cardio-respiratory interaction detection and synchronization of neural signals [24]). Moreover, the ultimate goal is not only to detect synchronized states or drive–response relationships but also to measure or quantify the relative strengths of these relationships. Under these circumstances, probability theory and information theory seem to be promising mathematical approaches with intuitive and fruitful results in detecting synchronized states and causal relationships in time series for various applications [25], [26].

This study introduces an innovative and streamlined application of information theory to quantify the efficiency and synergy of hand movements, namely hands coordination and correspondence, in bimanual tasks, specifically within surgical training programs. The analysis focuses on basic, specific, and controlled educational tasks, offering a novel perspective on evaluating surgical skills. We utilized mutual information and information flow between trajectories of two hands as expressive metrics of coordination and correspondence in collaborative tasks, respectively. Energy-based metrics were utilized as informative features for capturing hands coordination and correspondence efficiency in given trajectories. Our approach needs no prior knowledge or assumption about the dynamics and models within the robotic platform, executive task, or human motor control. After testing the performance of the proposed method in surgical skills assessment using human data, we demonstrated that our approach is a promising direction for the development of more intelligent bimanual manipulation and more realistic human-robot collaboration. Our analysis of the MEELS (Motion, Eye, EMG Data in Laparoscopic Surgery) and JIGSAWS datasets reveals significant insights into the dynamics of bimanual coordination, demonstrating the method's effectiveness in identifying skill levels among surgeons. Finally, we showed that our method is a promising approach for reducing the model's uncertainty and improving accuracy. The proposed framework not only advances surgical skills assessment but also opens new avenues for understanding the underlying mechanisms of bimanual task execution, with implications for training and robotic surgery design. Please note that the presented idea is different from other work such as [27] which incorporates approximate entropy (ApEn) to measure the fluency and predictability of motion in surgical trajectories for skills assessment purposes. In addition to the elementary application of information theory in [27], authors considered the ApEn of each hand separately and assumed no correlation or interaction between hands.

The paper is organized as follows: In Section II, basic definitions and methods that led us to our contributions will be discussed. In Section III, we use information theory concepts to measure hands collaboration in bimanual tasks. In Section IV, the proposed method is justified in surgical skills assessment and style recognition in robotic surgery applications using the JIGSAWS dataset. In Section VI-A, the proposed metrics were applied to the MEELS database [21], [28] to evaluate users' skills in performing simulated laparoscopic surgical tasks. In Section VI, the advantages of the proposed metrics in training an intelligent reinforcement learning (RL) agent for autonomous human-robot/robot-robot collaboration and developing a surgical skills classification model will be discussed. Concluding remarks are provided in Section VII.

II. Preliminaries

Throughout this paper, we will denote a given trajectory $\boldsymbol{X} = [x_1, x_2, \cdots, x_n]$ as a discrete random variable drawn from the sample space $\boldsymbol{X}$ with the empirical probability of $p(x_i)$ which defines the probability of observing the state x_i in the i^{th} time-stamp of the time series $\boldsymbol{X}$. Moreover, $\boldsymbol{x}_i^{(k)} = \{x_{i-k+1}, x_{i-k+2}, \cdots, x_i\}$ is the k-histories (also known as k-blocks) of $\boldsymbol{X}$ with the occurrence empirical probability of $p(\boldsymbol{x}_i^{(k)})$. Before reading this section, we encourage readers to see Appendix A for gaining/refreshing the knowledge of basic information theory concepts.

A. Conditional Entropy

The conditional entropy $\mathcal{H}(\boldsymbol{Y}|\boldsymbol{X})$ is the averaged remaining amount of uncertainty in predicting trajectory $\boldsymbol{Y}$ after seeing trajectory $\boldsymbol{X}$ over all possible values of x_i:

$$\mathcal{H}(\boldsymbol{Y}|\boldsymbol{X}) := \mathbb{E}_{p(x_i)}[\mathcal{H}(p(y_j|x_i)] = \sum_{x_i} p(x_i)\, \mathcal{H}(p(y_j|x_i)) \quad (1)$$

where $\mathbb{E}$ indicates the mean or the expected value.

***Property** 1:* $\mathcal{H}(\boldsymbol{Y}|\boldsymbol{X}) = \mathcal{H}(\boldsymbol{Y}, \boldsymbol{X}) - \mathcal{H}(\boldsymbol{X})$ [29]. □

According to Property 1, if trajectories $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent variables, observing $\boldsymbol{X}$ will not give particular information about $\boldsymbol{Y}$ and hence $\mathcal{H}(\boldsymbol{Y}|\boldsymbol{X}) = \mathcal{H}(\boldsymbol{Y})$. On the other hand, if trajectory $\boldsymbol{Y}$ is a deterministic function of $\boldsymbol{X}$ (i.e., $\mathcal{H}(\boldsymbol{X}) = \mathcal{H}(\boldsymbol{Y}) = \mathcal{H}(\boldsymbol{X},\boldsymbol{Y})$), knowing $\boldsymbol{X}$ completely removes any uncertainties about $\boldsymbol{Y}$ and hence $\mathcal{H}(\boldsymbol{Y}|\boldsymbol{X}) = 0$. The interpretation of Property 1 is that *stochastically* (i.e., not for any single observation), conditioning on the random variable $\boldsymbol{X}$ (i.e., treating $\boldsymbol{X}$ as a known trajectory of the occurred event) never increases the uncertainty about another event $\boldsymbol{Y}$.

B. Mutual Information

The mutual information between random variables $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as

$$\begin{aligned}\mathcal{I}(\boldsymbol{X}; \boldsymbol{Y}) &:= D_{\mathbb{KL}}(p(x_i, y_j) \,||\, p(x_i)\, p(y_j)) \\ &= \sum_{x_i,\, y_j} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(x_i)\, p(y_j)} \geq 0 \end{aligned} \quad (2)$$

where $D_{\mathbb{KL}}$ indicates the Kullback-Leible (KL) divergence (see [29]). The mutual information measures the *information gain* if we update a model that treats time series $\boldsymbol{X}$ and $\boldsymbol{Y}$ as independent trajectories (considering $p(x_i)p(y_j)$ as the probability distribution) to a more accurate model that considers their true probability distribution $p(x_i, y_j)$ with all possible correlations.

C. Transfer Entropy

The transfer entropy ($\mathcal{T}$) quantifies the information transfer between a *source* and a *destination*, conditioning out common history effects. In other words, $\mathcal{T}$ measures the $\mathcal{I}$ between the past states of the source trajectory $\boldsymbol{X}$ and the future state of the target trajectory y_{t+1} while conditioning on the past states of target trajectory $\boldsymbol{Y}$. The transfer entropy from source discrete trajectory $\boldsymbol{X}$ to destination discrete trajectory $\boldsymbol{Y}$ is defined by

$$\begin{aligned}\mathcal{T}_{\boldsymbol{X}\to\boldsymbol{Y}}(k, l) &= \mathcal{I}\Big(\boldsymbol{x}_t^{(l)}; y_{t+1}|\boldsymbol{y}_t^{(k)}\Big) \\ &= \sum_{\boldsymbol{x}_t^{(l)},\, \boldsymbol{y}_t^{(k)},\, y_{t+1}} p\Big(y_{t+1}, \boldsymbol{y}_t^{(k)}, \boldsymbol{x}_t^{(l)}\Big) \log_2 \frac{p\Big(y_{t+1}|\boldsymbol{y}_t^{(k)}, \boldsymbol{x}_t^{(l)}\Big)}{p\Big(y_{t+1}|\boldsymbol{y}_t^{(k)}\Big)}. \end{aligned} \quad (3)$$

Transfer entropy takes into account the dynamics of the information flow with minimal assumptions about the dynamics and couplings within the system [30]. According to (3), transfer entropy can be interpreted as additional information gain about predicting y_{t+1} conditioned on both l-histories of $\boldsymbol{X}$ and k-histories of $\boldsymbol{Y}$ compared to conditioned on k-histories of $\boldsymbol{Y}$ only. While mutual information quantifies the information gain from modelling trajectories $\boldsymbol{X}$ and $\boldsymbol{Y}$ as independent random variables to a more accurate model that considers their correlation, transfer entropy quantifies the added accuracy to the prediction of the future of $\boldsymbol{Y}$ from being estimated by its own history only versus having an auxiliary history of source signal $\boldsymbol{X}$ (i.e., detecting leader-follower relationship). Transfer entropy unlike mutual information is not symmetric since it measures the dependency of $\boldsymbol{Y}$ on $\boldsymbol{X}$ and not vice versa.

III. Relevance to Hands Coordination and Correspondence

The proposed concepts in Section II can be viewed from the perspective of human hands collaboration. For instance, we can condition the data of the left hand with respect to that of the right hand and according to Property 1, we will have fewer uncertainties about the future behaviours of the left hand or in the worst case, we will gain no information about the upcoming trajectories of the left hand. Since hands coordination and correspondence reflect the extent to which the user can proactively and simultaneously predict, plan, and control the trajectory of one hand in response to the current (or past) activity of the other hand, hands with good collaboration yield a higher reduction in the uncertainties of our predictions.

A. Collaboration Metrics

The mutual information between the data of two hands returns the added accuracy in our modelling when we consider the coordination of two hands in our problem formulations rather than modelling them separately. In other words, mutual information can be thought of as the reduction in uncertainty about one hand's movement given knowledge of the other's, akin to predicting one hand's actions based on observing the other in a surgical task, such as when the surgeon must simultaneously navigate an endoscope with one hand and manipulate surgical instruments with the other. Mutual information is a perfect metric for capturing lumped instantaneous relationships between the executive actions of two hands in particular tasks such as bimanual lifting tasks. A high mutual information value suggests a high degree of coordination, indicating that the movements of one hand are predictive of the movements of the other, a trait often seen in more experienced surgeons. Another notable feature of mutual information that makes it appropriate for evaluating of hands coordination is the symmetric relationship between entropy and mutual information:

***Property** 2:* $\mathcal{I}(\boldsymbol{X}; \boldsymbol{Y}) = \mathcal{H}(\boldsymbol{X}) - \mathcal{H}(\boldsymbol{X}|\boldsymbol{Y}) = \mathcal{H}(\boldsymbol{Y}) - \mathcal{H}(\boldsymbol{Y}|\boldsymbol{X}) = \mathcal{I}(\boldsymbol{Y}; \boldsymbol{X})$. □

Proof: See Appendix B. ■

The interpretation of Property 2 in hands coordination evaluation is that the amount of reduction in uncertainty about the right-hand future trajectory after observing the data of the left hand is the same as the reduction in uncertainty about the left-hand future trajectory after observing the data of the right hand.

Although mutual information offers insight toward hands coordination, it cannot show the direction of information flow (i.e., leader-follower relationship) between hands. For instance, during passing the suturing needle from the right hand to the left hand, the information flow is mostly from the right-hand discrete trajectory ($\boldsymbol{X}_R$) to the left-hand discrete trajectory ($\boldsymbol{X}_L$). This is because the left hand is following or responding to the intention of the right hand. More importantly, if a time delay is introduced to one of the observations, mutual information fails to accurately distinguish the information that is generated from a common response to the input

signal or external factor. This issue can be crucial in detecting hand-eye coordination, which plays an indispensable role in minimally invasive surgeries and has been shown to exhibit different patterns across surgical skill levels [31]. This is because the reaction time of the human eye gaze and hand motor system is different in response to external stimulus [32]. Transfer entropy between two hands will address all of these shortcomings. For instance, $\mathcal{T}_{\boldsymbol{X}_R \to \boldsymbol{X}_L}$ means how much knowing the history of the right-hand data is useful in forecasting the future behaviour of the left-hand data.

B. Energy of a Time Series

Energy and entropy, despite their historical connection originating from statistical mechanics and subsequent information theory concepts, play distinct roles in the realm of signal processing. While energy quantifies the overall strength for executing the task that generates the trajectory, metrics derived from entropy (such as those introduced in this paper) specifically gauge the intrinsic unpredictability within the trajectory. In the context of bimanual performance evaluation, the energy of the extracted time series of mutual information and transfer entropy can be thought of as the *coordination and correspondence efficiency*, which reveals important skills-related insights in user's motions [3]. The energy of a given time series (or a discrete signal from the signal processing perspective) is defined as

$$\mathcal{E}_1 = \sum_{t=1}^{n} |\boldsymbol{X}[t]|^2 = \sum_{t=1}^{n} |x_t|^2 \tag{4}$$

where n is the total number of time-samples of $\boldsymbol{X}[t]$.

Another possible way to approximate the energy efficiency is to calculate the total *kinetic energy* within a given trajectory which is an important factor in surgical performance analysis [33]. This is because, on average, a user's high energy consumption reflects skills deficiency and uncontrolled motions. Moreover, in RAMIS, high energy consumption on the surgeon-side robot results in high energy injection to the patient-side robot, which is one of the main sources of trauma in an operation [3]. The total kinetic energy of the patient-side robot with given discrete trajectory $\boldsymbol{X}[t]$ is defined by

$$\mathcal{KE} = \frac{1}{2}\sum_{t=1}^{n} \mathbf{m}_t |\dot{\boldsymbol{X}}[t]|^2 = \frac{1}{2\,T_s^2}\sum_{t=1}^{n-1} \mathbf{m}_t |x_{t+1} - x_t|^2 \tag{5}$$

where T_s is the sampling time and $\mathbf{m}_t$ is the lumped mass of the patient-side robot in the given configuration at time t. Due to the design and tiny movements of the patient-side robot in delicate RAMIS, $\mathbf{m}_t$ remains quite the same for different surgical configurations, i.e., $\mathbf{m}_t \approx \mathbf{m}$ [3]. Under this assumption, we can write

$$\mathcal{E}_2 = \frac{\mathcal{KE}}{\mathbf{m}} = \frac{1}{2}\sum_{t=1}^{n} |\dot{\boldsymbol{X}}[t]|^2 = \frac{1}{2\,T_s^2}\sum_{t=1}^{n-1} |x_{t+1} - x_t|^2. \tag{6}$$

The energy content of a given trajectory shown by vector $\boldsymbol{\mathcal{E}} = [\mathcal{E}_1, \mathcal{E}_2]^\top$ can be fed to the downstream tasks such as data visualization model (see Fig. 3(a)).

C. Ensembling Sub-models

Ensembles of models have been empirically shown to be a promising approach for improving the accuracy, uncertainty, and out-of-distribution robustness of DL models since they tend to sample and explore from different parts of the feature space [3]. Each one of the above-mentioned metrics is not comprehensive enough to reflect the true skill levels of the user during the execution of the task. A possible solution is to concatenate all of these metrics and feed them to downstream data analysis tasks (see Fig. 3(a)). This makes our approach more understandable relative to other similar black-box machine learning (ML) or DL approaches, including but not limited to [4], [6]–[11]. This is particularly significant given the data scarcity in the field of robotic surgery [3].

IV. Method Justification on JIGSAWS Dataset

In this section, we will assess the validity of the proposed approach for detecting coordination and correspondence between a user's two hands in the complex task of suturing, offering a detailed view of skill-related dynamics. Our analysis reveals that higher expertise levels correlate with superior hand coordination and correspondence. Notably, expert surgeons demonstrated significantly enhanced mutual information and transfer entropy values, indicating more synchronized and efficient hand movements. We utilized the JIGSAWS dataset [34], which comprises data from minimally invasive surgical tasks performed by seven surgeons with three different skill levels using the *da Vinci* Surgical System. The tasks included suturing, knot-tying, and needle-passing. Each surgeon according to their experience of working with the robotic platform will be classified into three levels of expertise: novice, intermediate, and expert. JIGSAWS contains three linear motions along x, y, and z axes of the Cartesian coordinate system as well as the nine elements of rotation matrix $\boldsymbol{R} \in \mathbb{R}^{3\times3}$ of both surgeon-side and patient-side robots. Note that nine elements of $\boldsymbol{R}$ can be reduced to three rotation angles *roll* (Φ), *pitch* (Θ), and *yaw* (Ψ) as formulated in [3].

A. Illustrative Evaluation

For calculating $\mathcal{I}$ and $\mathcal{T}$ time series, we concatenated local $\mathcal{I}$ and $\mathcal{T}$ of a sliding window of a given surgical trajectory with the size of 20 sample (~ 0.6 second), which is short and informative enough for downstream tasks based on our observations. The evaluation approach in this section is primarily illustrative and intuitive and will be supported by endoscopic videos of trials provided in the supplementary video file. In Sections IV-B and VI, we will investigate the functionality and advantages of the proposed methods for surgical skills assessment and style recognition.

According to [34], given that all users were right-handed, each suturing trial is composed of four sequential gestures as follows:

- G_2: Positioning the needle above the tissue with the right hand,
- G_3: Pushing the needle through the tissue with the right hand,
- G_6: Pulling the needle with the left hand,
- G_4: Transferring the needle from the left hand to the right hand.

These gestures are illustrated in Fig. 2(a), and their boundaries were determined by an expert surgical mentor in the JIGSAWS dataset. For a better explanation, we split G_3 and G_4 into two different phases as shown in Fig. 2(b) and Fig. 2(c). As explained in the supplementary video, in G_2, when the user moves the needle above the tissue, the left hand stands still and as expected, there is no mutual information (i.e., coordination) nor transfer entropy (i.e., correspondence) for this gesture in the top row of Fig. 2. In G_3, as the users insert the needle into the tissue, their left hand begins to move towards the approximate point where the needle tip will emerge. In this gesture, there is minor coordination and slight

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2024.3464110

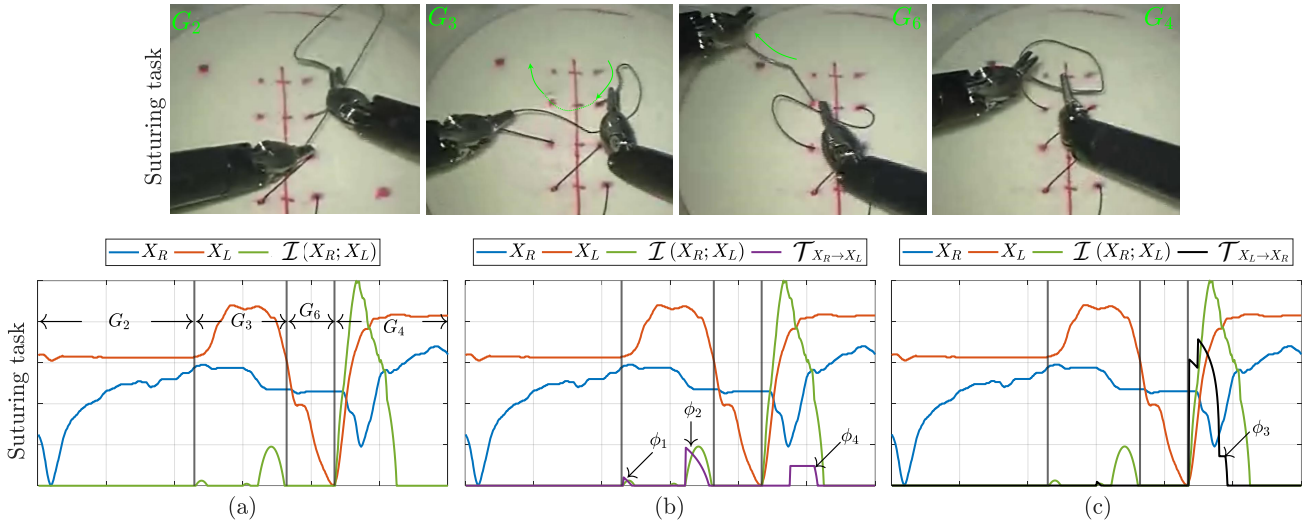

Figure 2: Hands coordination analysis for the right-hand data ($\boldsymbol{X}_R$) and left-hand data ($\boldsymbol{X}_L$) of a suturing trial in the JIGSAWS dataset. (a) Mutual information ($\mathcal{I}$) diagram and boundaries of four gestures of suturing task according to [34], (b) Transfer entropy ($\mathcal{T}$) diagram with $\boldsymbol{X}_R$ as the source and $\boldsymbol{X}_L$ as the destination and its three important variations ϕ_1, ϕ_2, and ϕ_4 (ϕ_i's are for the sake of clarification and do not convey particular meanings in surgery), and (c) Transfer entropy diagram with $\boldsymbol{X}_L$ as the source and $\boldsymbol{X}_R$ as the destination and its main variation ϕ_3. Note: the diagrams of $\mathcal{I}$ and $\mathcal{T}$ are magnified to be visible near to hands position data $\boldsymbol{X}_R$ and $\boldsymbol{X}_L$ with the unit of meter.

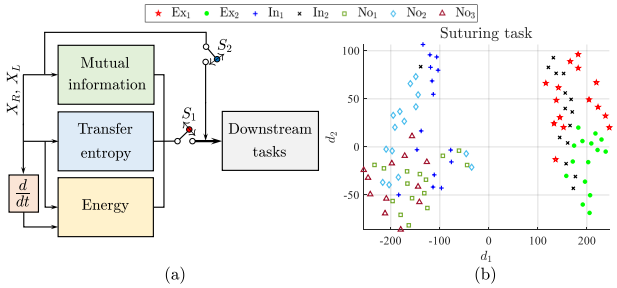

Figure 3: (a) Ensembling trajectory data of the right and the left hands for downstream tasks with selection switches S_1 and S_2. (b) t-SNE visualization of coordination/correspondence-related features (i.e., $S_1=1$ and $S_2=0$) of suturing trials in JIGSAWS dataset.

correspondence of the left hand to the intention of the right hand. As illustrated in Fig. 2(a) and Fig. 2(b), there is a slight activity in $\mathcal{I}$ and $\mathcal{T}_{\boldsymbol{X}_R\to\boldsymbol{X}_L}$ pointed as ϕ_1 and ϕ_2. Please note that the ϕ_i's in Fig. 2 are used for clarification purposes and do not have specific meanings in surgical tasks. While the user is pulling out the needle in G_6, the right hand stands still and there is no $\mathcal{I}$ nor $\mathcal{T}$ in this gesture. Finally, during passing the needle from the left hand to the right there is a considerable amount of $\mathcal{I}$ and $\mathcal{T}$ in G_4. The noteable point about this observation in Fig. 2(b) and Fig. 2(c) is that ϕ_3 happens before ϕ_4 in time. This is because at the beginning of G_4, the right hand approaches the left hand to grab the needle (i.e., early increase in $\mathcal{T}_{\boldsymbol{X}_L\to\boldsymbol{X}_R}$) and then the left hand follows the right hand for delicate hand-over completion (i.e., late increase in $\mathcal{T}_{\boldsymbol{X}_R\to\boldsymbol{X}_L}$).

B. Application to Surgical Skills Evaluation

One intriguing observation in complex collaborative tasks, such as suturing, is that individuals with higher levels of expertise exhibit superior hand coordination and correspondence compared to novices [12], [21], [35]. Building upon this insight, we explored the benefits of the proposed metrics for evaluating skills and recognizing styles in robot-assisted surgery trials, which are inherently highly complicated collaborative tasks.

In this section, we aim to explore whether ensembling the coordination and correspondence-related features of hands, as described in Section II (see Fig. 3(a)), provides insights into the user's skill level and style in suturing as a good example of highly collaborative task. We extract the information and energy content from all possible combinations of translational data (i.e., x, y, and z) and rotational data (i.e., Φ, Θ, and Ψ), as well as the tool-tip distance from the origin $\mathcal{P}=\sqrt{x^2+y^2+z^2}$, for both hands of all participants in the JIGSAWS dataset. This is conducted after normalizing each trajectory and unifying the length of each trial. Our hypothesis is that the union of these features together (i.e., when $S_1=1$ and $S_2=0$ in Fig. 3(a)) is very expressive to capture salient features related to the style and skills level of each participant.

One approach to test this hypothesis involves feeding the generated feature vectors (i.e., total energy, $\mathcal{I}$, and $\mathcal{T}$ time series) to a classifier network and evaluating its accuracy on the test dataset. However, supervised discriminative models, designed to penalize misclassification rates based on labels, struggle to distinguish between different trials of the same class and may fail to capture outlier points. Furthermore, classification techniques, being supervised learning methods, are prone to overfitting when dealing with small datasets. This characteristic is unfavorable in tasks like robotic surgery, where there are stringent requirements for the safety and explainability of intelligent algorithms. To address this issue, one approach is to create an interpretable graphical representation of high-dimensional data using unsupervised data visualization techniques such as t-SNE [36], which provides insight into the arrangement of data in a high-dimensional space. In this way, anomalies and outlier points will be distanced from their normal clusters in the

visualization space. In brief, t-SNE is a nonlinear and unsupervised clustering-based data visualization technique that maintains the proximity of neighboring points from the high-dimensional space in the corresponding low-dimensional map. Using this technique, trials with similar features will be mapped to close-by clusters in the t-SNE visualization space (authors fully justified the choice of t-SNE for such visualizations in [3]). Please note that throughout this paper, the use of t-SNE is not intended for clustering users but for visualizing the relative position of a given trainee in comparison to other registered users, spanning from novices to experts. This insight can be valuable for interpretable skills assessment [3], learning curve analysis [3], and diagnosing abnormal user performance [37].

Now, we apply the t-SNE technique on the extracted feature vectors of several suturing tasks to reach 2D visualizations (see Fig. 3(b)). Please note that the axes d_1 and d_2 in Fig. 3(b) are not meant to have specific interpretations in terms of the axes/units of the original high-dimensional data. While the exact positions of points on the axes are not meaningful per se, the overall layout, clustering, and relative distances between points do. Clusters in a t-SNE plot represent groups of similar data points, and the quantity of the distances between clusters can provide insights into the relationships between different groups. Last but not least, the chosen hyperparameters of t-SNE remained consistent with the default settings of the t-SNE toolbox in MATLAB, with the exception of perplexity, which we set to the number of trials for each user in the given task (e.g., 12 for the suturing task).

In Fig. 3(b), the visualization of extracted features from the suturing task is illustrated. It is evident that trials from two expert users and the intermediate user In_2 cluster near each other, forming the *good* cluster. The goodness of each group is determined by the user's skill level as specified in the JIGSAWS dataset; expert trials are more likely to be classified as 'good.' Moreover, our investigation into endoscopic data for detecting mid-task failures and restarts will affect the goodness of each trial. The other intermediate user In_1 clusters close to novice trials (i.e., the *poor* cluster). This implies that In_2 exhibits more expert-like performance than In_1, and endoscopic videos confirm that In_2 performed exceptionally well compared to In_1 in the suturing task. In addition to our investigations, the global rating scores (GRS) assigned by a skilled gynecologic surgeon further confirm that In_2 performed expertly in the suturing task. For a more in-depth understanding of GRS, please refer to [3]. The mean GRS of In_2 is 3.1 ± 0.5 out of 5 which is similar to those of expert trials (i.e., 2.6 ± 0.4 for Ex_1 and 3.2 ± 0.3 for Ex_2) and are evidently higher than mean GRS of In_1 (i.e., 2 ± 0.5) and other novice participants (i.e., 1.7 ± 1.1 for No_1, 1.6 ± 0.3 for No_2, and 2.8 ± 0.8 for No_3).

The presented results indicate that manual annotations with multiple labels are often coarse-grained and prone to bias. However, our approach is capable of addressing these limitations, as expert users might perform poorly, or novice users might exhibit expert performance. This raises questions about the validity and generalization of end-to-end learning methods in studies using JIGSAWS, trained based on skill labels [38]. No related supervised learning study has uncovered such insightful label-free information; instead, they merely conducted classification on each data point based on the originally assigned label [4], [6]–[11].

Based on the presented results, we can conclude that there is a strong correlation between the surgeon's dexterity level and their hands' coordination and correspondence while performing collaborative tasks in surgical operations. The majority of the presented findings align with those in [37]. Particularly, the observed patterns in Fig. 3(b) regarding the proximity of In_2 to other expert trials are similar to the patterns observed in [37] for the suturing task. The visual and conceptual similarity between the results presented in [3] and this paper, as well as those in [37], is noteworthy. Particularly, [37] demonstrates the embedding space representation of its proposed method without relying on data visualization techniques, such as t-SNE.

Despite the incorporation of high-capacity end-to-end learning models in related studies, this paper takes a different approach by generating a small number of skill-related features and reducing their dimensionality using t-SNE, making them ready for human analysis and interpretation. The acquired interpretability of our model is achieved at the expense of high prediction accuracy, as it necessitates the generation of a high-dimensional feature space. As a result, a numerical comparison of our approach with other high-capacity models in terms of classification accuracy may not be entirely fair. The primary merit of this paper lies in revealing *label-free* insights and providing reliable information for trainees, which, to the best of our knowledge, none of the high classification accuracy DL models can offer. Among related deep learning-based skills evaluation studies, only [6] utilized class activation maps (CAM) [39] to offer insights into the quality of surgeon task execution based on the deep network's confidence in the predicted label. This method is susceptible to *confirmation bias* [40], stemming from its heavy reliance on the parameters of the non-transparent deep network, which are learned based on biased labels as discussed earlier (for further explanations, please refer to [3]).

Please note that the methodology presented in this paper is not a universal solution for surgical skill assessment due to the complexity of surgical tasks. As previously mentioned, studies such as [3], [37] demonstrate similar performance and can complement this approach. These studies contribute to a broader framework for comprehensive surgical skill assessment by combining multiple interpretable metrics with a robust feature extraction and classification network, thereby enhancing the generalization of the final product.

V. Method Evaluation on MEELS dataset

In this section, we explored the MEELS dataset, introduced in [21], to complement our evaluation using the JIGSAWS dataset. While the JIGSAWS dataset provides a benchmark for robot-assisted surgical tasks, MEELS offers a unique perspective on manual laparoscopic surgeries, which are pivotal for assessing hand collaboration and correspondence. This dataset is especially suitable for evaluating our metrics because of its focus on the nuanced interactions between a surgeon's two hands during laparoscopic surgeries. MEELS's design allows us to assess our metrics in scenarios that demand varying degrees of coordination and correspondence, tailored to different skill levels. Incorporating MEELS ensures that our methodology is robustly validated across diverse surgical techniques, enhancing the generalizability and relevance of our findings. MEELS dataset comprises nineteen participants categorized into three groups: five seasoned general surgeons forming the expert group, five surgical residents making up the intermediate group, and nine participants in the novice group who had no prior experience with laparoscopic surgery. Participants in the MEELS dataset utilize two laparoscopic graspers (Ethicon Endo-Surgery, Cincinnati, OH)

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2024.3464110

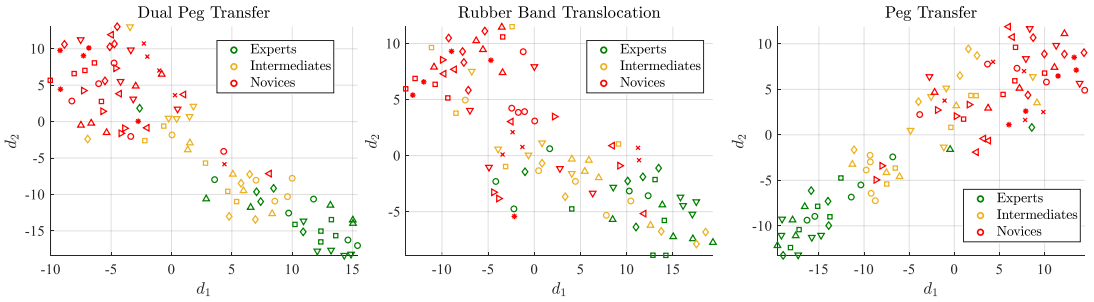

Figure 4: t-SNE visualization depicting collaboration-related features across three laparoscopic tasks in the MEELS dataset. Given the extensive number of participants and trials, each expertise group's trials share the same color, with distinct shapes for individual participants.

to perform three different tasks: peg transfer task, dual transfer task, and rubber band translocation task. The peg transfer task involves transferring pegs between the grasper of the non-dominant hand and that in the dominant hand, and transferring them to a target positioned 10 cm away, followed by repeating the task in reverse. The dual transfer task involves simultaneously moving two pegs from one side to the other side and returning them to their original position, testing bimanual coordination challenges. The rubber band translocation task involves grasping both ends of a rubber band, moving it to distal pins (5.5 cm away), releasing the rubber band, re-grasping it, and moving it back to the original position to capture tool-tissue interaction and coordinated movements between graspers.

The same procedure employed on the JIGSAWS dataset was executed on the MEELS dataset, and the resulting graphical representation using t-SNE is illustrated in Fig. 4. As depicted in the Fig. 4, a clear separation exists among distinct expertise groups, highlighting the effectiveness of our proposed method in uncovering essential skills-related features. Similar to the JIGSAWS dataset, our method demonstrates robust performance across entirely different tasks in practical scenarios. Upon reviewing the video files, we found that certain existing outliers, such as expert trials in the novice region during peg and dual peg transfer tasks, were meaningful. These trials represented faulty attempts where users experienced mid-task failures, notably dropping the peg within the surgical field. The presence of expert-level outliers in novice areas during specific tasks underscores the nuanced nature of skill assessment, suggesting areas where even experienced surgeons may benefit from targeted training.

Further investigation shows that novices consistently underperform across all tasks, signaling a foundational gap in understanding requisite behaviors. Moreover, as task complexity increases from peg transfer to dual peg transfer, and then to rubber band translocation, a notable pattern emerges. Experts exhibit a decrement in performance, aligning more closely with intermediate levels and narrowing the gap to novices. This observation is underscored by the clustering analysis, which shows a convergence of performance levels under heightened difficulty, suggesting that increased task complexity introduces challenges that diminish the performance distinction between expertise levels. Clinical observations of heightened frustration among experts with more difficult tasks provide context to these findings, indicating the potential for adaptive training strategies to enhance skill across all levels of expertise.

In line with the discussions in Section IV, these observations further support the notion that our method effectively captures label-free information, thereby boosting the generalizability of the as well as enhancing the interpretability and reliability of the final results. These aspects are challenging for supervised methods to capture, as they inherently rely solely on assigned labels.

In addition to the expert mistakes explained in Section IV-B that result in outlier trials, there is a mild overlap between skill classes shown in the plots of Fig. 4. These overlaps can be attributed to labeling inaccuracies [38], where, for instance, intermediates might perform near the expert level or exhibit behaviors similar to novices, depending on the task or other random factors. Specifically for the MEELS dataset, since peg transfer, dual transfer, and rubber band translocation are basic surgical tasks with relatively low complexity, some overlap between intermediates and experts is expected. The highest overlap occurred in the rubber band translocation task, where participants pulled the rubber band to an arbitrary position to wrap it around the pins before releasing it. Such overlaps or outliers could potentially provide valuable insights into training improvements, allowing participants to learn from both their mistakes and strengths.

The observed overlap between skill classes in Fig. 4 prevents the final clusters of the data from being linearly separable, making the decision boundaries complex and less generalizable. As elaborated in Section IV, our goal is not to solve the problem of surgical skill assessment solely with the presented approach, recognizing that surgical skills depend on multiple factors beyond hand coordination/correspondence. However, as discussed in Section VI-C, once the presented approach is integrated with interpretable [3], [37] and/or end-to-end learned features [4], [8], it can transform the complex decision boundaries of the classifier into simple 2D lines, thus enhancing the generalization and reliability of the final model.

VI. Discussions

A. Applications in Robot Learning and Control

In this section, we will show that our methodology, leveraging mutual information as an additional reward, significantly expedited the RL agent's learning curve in a simulated human-robot collaboration task, as evidenced by improved performance metrics over traditional reward structures. Using RL to achieve a higher level of automation in bimanual robot manipulation or human-robot collaboration tasks, such as robot-robot or human-robot object handover has gained increasing attention in robotics community [41]–[44]. In such complex and highly collaborative tasks, it is impossible to design an informative dense reward to guide the learning process. In fact, sparse and delayed rewards that only indicate whether the goal has been achieved are usually implemented for such complex learning tasks (e.g., bimanual peg transfer task

in which the reward $\mathcal{R} \in \{-1,\, 0\}$ is always -1 unless the block is transferred to the goal position [45]). In such cases, it is very challenging for the RL agent to learn with this type of reward signal.

Potential-based reward shaping (PBRS) approaches [46] can address the aforementioned problem by adding an informative shaping function F to the original reward while ensuring the policy invariance property. However, handcrafting F requires high-level domain knowledge about the environment and/or task. As a result, heuristic reward design is still a common procedure in real practice in which the policy invariance may get violated [47]. As previously mentioned, skills evaluation can be applied in task autonomy including autonomous surgery by incorporating performance information into the design or learning of the control strategy. In this section, we will show that coordination information $\mathcal{I}$ which acts as a powerful skills evaluation metric, can be incorporated as an additional heuristic reward for the learning procedure of the RL agent through a complex collaborative environment without deviating from the original optimal policy.

We design a 2D human-robot collaboration task in which a human moves their hand in a 2D plane following a randomly generated trajectory, and the agent should figure out to track the human hand, as shown in Fig. 5(a). The task simulates a human-robot collaborative situation where the robotic arm aims to fetch one object from the human hand which is performing an unknown task trajectory. This environment can be considered as a simplified version of dual-arm manipulation tasks, such as autonomous bimanual peg transfer and bimanual needle re-grasp which was shown to be quite impossible for an RL agent to learn [45]. This situation is also common in RAMIS tasks, such as autonomous suturing, needle re-grasping, tissue retraction, and tissue cutting. In the developed environment, the observations are the current locations of the human and the agent, and the actions are the movements of the agent along x and y directions. The reward is designed to be sparse:

$$\mathcal{R}_1 = \begin{cases} -1 & \text{if the current tracking error is less than } \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where $\epsilon = 1$ cm is the desired tracking accuracy. As an initial result, we observed that the value of $\mathcal{I}$, serving as a metric for collaboration between the RL agent trained using Proximal Policy Optimization (PPO) and the human hand, increases with the agent's improved performance throughout the training process (please see the trend of blue lines in Fig. 5(b) and Fig. 5(c)). Accordingly, we hypothesize that incorporating $\mathcal{I}$ as an additional reward may accelerate the agent's learning process. To test our hypothesis, an augmented reward signal $\mathcal{R}_2$ is designed as the weighted sum of the original sparse reward $\mathcal{R}_1$ and a function of mutual information between the agent and the human trajectory for the previous 10 steps

$$\mathcal{R}_2 = \alpha \mathcal{R}_1 + (1-\alpha)\mathcal{I}_{\mathrm{RL}} \tag{8}$$

where $\mathcal{I}_{\mathrm{RL}} = \mathcal{I}(\boldsymbol{X}^{(10)}_{\mathrm{agent}};\, \boldsymbol{X}^{(10)}_{\mathrm{human}})^2 - 1$, $\boldsymbol{X}^{(10)}_{\mathrm{agent}}$ and $\boldsymbol{X}^{(10)}_{\mathrm{human}}$ are the windowed trajectories of the agent and the human for the last 10 time-stamps, and $\alpha = 0.8$ is the weighting factor. The function $\mathcal{I}_{\mathrm{RL}}$ emphasizes higher mutual information over lower ones and scales the final value to $[-1, 0]$ to prevent gradient scaling for fair comparison (i.e., $-1 \leq \mathcal{R}_1, \mathcal{R}_2 \leq 0$). While violating the Markov property, this additional reward signal is largely dependent on the agent's current action and can actually punish lower and encourage higher mutual information throughout the whole trajectory. Additionally, as previously discussed, sparse and delayed reward that violates the Markov property commonly exists in multistage bimanual manipulation, human-robot collaboration, and surgical robot learning tasks due to the difficulty of designing a dense and immediate reward.

As a comparison, two agents are trained using the original reward $\mathcal{R}_1$ and the reward $\mathcal{R}_2$ defined in (8). As shown in Fig. 5(b), while both agents generate trajectories with increasing mutual information about the human throughout the training, it increases faster for the agent using $\mathcal{R}_2$ (for more details, please see the supplementary video file). Additionally, as shown in Fig. 5(c), the learning and convergence speed of the agent using $\mathcal{R}_2$ is faster than the one using $\mathcal{R}_1$. Both of aforementioned results validate our hypothesis and indicate the effectiveness of coordination information in helping the RL agent to learn collaborative tasks. Furthermore, both agents were able to converge to the optimal policy which tracks the given random human trajectory, that indicates the added reward term does not generate a sub-optimal policy and adding it facilitates the learning procedure without varying the learned policy.

B. *Explainability and Interpretability*

Although the majority of the prior AI-based approaches presented in Section I demonstrated high classification accuracy in surgical skills assessment task, the interpretability and explainability of the trained models still remain questionable. In this paper, we define *interpretability* and *explainability* as the extent to which the decision-making mechanism of a model is interpretable and explainable for a human [48]. Since DL and ML models operate as black-boxes, it is hard for the human to understand whether the outcome is based on relevant features or artifacts and biases in the training set. This reinforces the need to enhance the explainability and interpretability of the model to improve safety and meet the ethical requirements of skills assessment methods for robotic surgery [49].

Integrating hands coordination and correspondence metrics as informative domain knowledge in AI models not only improves the interpretability and explainability of the solution but also plays a crucial role in improving the learning performance, especially when training data is limited [50]. In data-scarce surgical tasks, utilizing domain knowledge as priors not only reduces uncertainties about the success of the operation but also makes the modeling problems easier to solve and more generalizable with fewer training data points [51]. Moreover, ensembling features presented in Section III-C benefits the explainability and interpretability of the skills assessment model since the contribution of each factor in the final decision-making process is transparent. The improved explainability and interpretability makes any possible application of our approach (ranging from skills assessment/transfer modules to autonomous surgical/bimanual tasks) more reliable and safe which eventually benefits surgeons and patients [3]. The emphasis on explainability and interpretability in our approach addresses crucial ethical concerns in AI for surgery, ensuring that our model's decision-making process remains transparent and justifiable, thus enhancing trust and safety in robotic surgery.

C. *Classification Accuracy and Representation Quality*

In this section, we incorporate a complementary approach to demonstrate that utilizing hands coordination and correspondence data significantly improves the accuracy of user skills classification

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2024.3464110

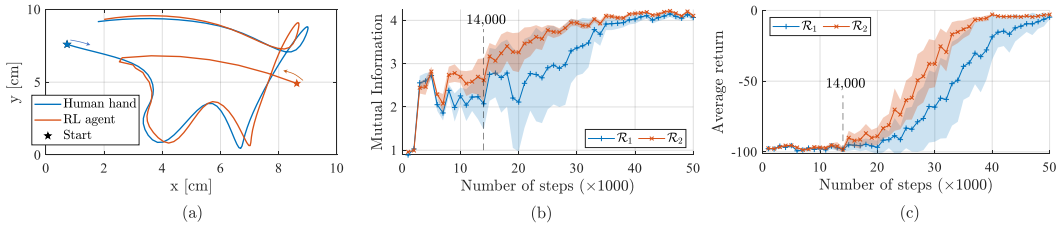

Figure 5: Application in reinforcement learning for collaborative robotic tasks. (a) Task environment where the agent should track a human hand that moves randomly following a random trajectory during each episode. (b) and (c) are the mutual information between the human and the agent trajectories and the average return evaluated every 1000 training steps when using different rewards $\mathcal{R}_i$, respectively. Solid lines are the mean values calculated out of 5 training instances using different random seeds, and the shaded areas represent the standard deviations.

and enhances the final model's generalization. To investigate the advantages of incorporating hands collaboration data in the skills assessment task, we trained two light convolutional neural networks, CNet_i, with two different inputs that are roughly identical in terms of the total number of training parameters ($\approx$179,000) and network architecture (three layers consisting of a Conv1D, batch normalization, and leaky ReLu with $\alpha = 0.3$). The first network, CNet_1, just uses hands' raw data as two separate inputs (i.e., X_R and X_L which is identical to $S_1 = 0$ and $S_2 = 1$ in Fig. 3(a); for more details please see [4]). The second one, CNet_2, in addition to hands' raw data, uses hands' collaboration data as three individual inputs (i.e., $S_1 = S_2 = 1$). The classification accuracy for CNet_2 is 92.89% ($\pm$ 0.87) and that of CNet_1 is 86.21% ($\pm$ 2.29).

In addition to the superior accuracy of CNet_2 comapred to CNet_1 due to the inclusion of informative hands coordination and correspondence data, CNet_2 can also extract more meaningful latent features for the classification head. As illustrated in Fig. 6(b), t-SNE visualization of the latent representation of CNet_2 has more clear boundaries between users compared to that of CNet_1 shown in Fig. 6(a). This clean latent space of CNet_2 not only benefits the model's accuracy, but also improves the model's generalization since the fully-connected layer in the classification head does not need to produce a highly nonlinear mapping from latent features to class labels (i.e., expert, intermediate, and novice). The plain area between participant clusters also reduces the model's uncertainty about unseen test data and improves the out-of-distribution robustness of the final model.

The separation between the clusters of different users gets even more noteworthy when we know that CNet_2 is trained to differentiate class labels and not individual users. This is because the test set is randomly separated from the shuffled data samples of all users and all trials, not by using other cross-validation methods such as leave-one-user-out (LOUO) that the model might tend to recognize the individual user during the training stage to infer the skill label. For instance, trials of Ex_1 and Ex_2 share the same class label 'expert' and CNet_2 has no information about which sample belongs to which expert user (the same story applies for In_1 and In_2 data samples). However, in the latent space, features of individual users cluster close to each other and are in the vicinity of the clusters with the same level of expertise. This proves the fact that incorporating hands collaboration data (and of course other domain knowledge-based metrics such as ones introduced in [3]) provides rich information regarding the skills level and style of each user. This observation enforces the fact that in the context of DL, an input feature is rarely important on its own, and incorporating the interactions between different input features can improve the learning quality of the entire network [52].

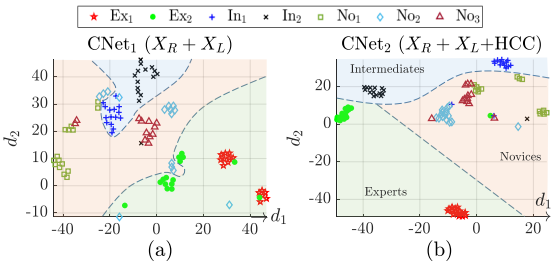

Figure 6: t-SNE 2D visualization of learned representations for surgical skills classification task in a convolutional neural network. (a) Hands' raw data. (b) Incorporating hands coordination and correspondence data as an extra input to case (a).

D. A Deep Learning Perspective

In this section, we will show that incorporating coordination and correspondence features in DL applications is aligned with neural network training concepts and discuss why such features benefit the latent representation quality and skills classification accuracy. DL models use cross-entropy loss with softmax output as a standard choice to train a classifier network. It can be proven that optimizing such network parameters using backpropagation maximizes the mutual information between the input data and output labels [53]. This gives us an alternative view of DL models as mutual information estimators. In a broader sense, maximizing the mutual information between inputs and outputs is a simple idea to train a representation-learning function (i.e. an encoder network). It turns out that, maximizing *global* $\mathcal{I}$ (i.e., training an ordinary skills evaluation network such as models discussed in Section I, which maximizes the mutual information between the output and entire input, e.g., stacked data of both hands) is often insufficient for learning meaningful representations [54]. This is because the network will be biased toward learning features that are unrelated since their sum has more lumped information than rare relevant features.

On the other hand, maximizing the *average* $\mathcal{I}$ between outputs and local patches of inputs (e.g., different data channels of two hands) considerably improves the representation quality for downstream tasks such as classification [54]. This encourages the model to prefer information that is shared across the input patches including hands coordination and correspondence information. Since DL models rely on large-sized datasets in the training stage, they may fail to extract meaningful shared information across the

This article has been accepted for publication in IEEE Transactions on Medical Robotics and Bionics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMRB.2024.3464110

input in data-scarcity healthcare applications. In our setting, instead of forcing the model to detect the shared information across the input patches with limited training samples in a self-supervision end-to-end learning paradigm, we simply feed the hands coordination and correspondence information of different data channels which eventually improves the accuracy and generalization of the model and the quality of its extracted representations.

E. Evaluating Proposed Methodology Under Noisy Data

To evaluate the robustness and practical applicability of our methodology in scenarios where data integrity is compromised by less precise measurement tools, we conducted a study examining the impact of low precision data on our proposed model's performance in extracting features related to surgical skills. This analysis is crucial for understanding the real-world applicability of the proposed method, especially in settings where high-end equipments such as the da Vinci Surgical Robots used in the JIGSAWS dataset or the OptiTrack motion capture system employed in the MEELS dataset are not available.

By introducing white noise to our dataset, we simulated lower precision environments and observed the performance of our skill assessment technique. Our findings show that the method maintains accuracy up to a noise level of $\sim 3\%$ of the maximum tajectory magnitude. The resilience to measurement error depends on how distinctly participants' skill-related features are differentiated for a given task. For instance, trajectories in peg transfer and dual peg transfer tasks were more resilient to noise than in rubber band translocation. As shown in Fig. 4, the clusters in the first two tasks are more separated than in the latter. Notably, expert skill features demonstrated resilience, tolerating noise levels up to $\sim 5\%$ across all tasks. This suggests that, in environments with less accurate measurement devices, it is advisable to primarily distinguish between expert and non-expert participants, as the intermediate class blends more with novices, potentially introducing artifacts and reducing model generalization. These results underscore the practicality of our approach in diverse surgical environments and emphasize the potential for broad adoption across different settings.

VII. Conclusions

An information theoretical approach for quantifying the coordination and correspondence between a user's two hands in bimanual surgical tasks was presented in this paper. Our investigations on tool trajectories of basic surgical tasks on da Vinci Surgical Systems and a simulated laparoscopic surgical setting proved the functionality and accuracy of the proposed approach in detecting surgical skills-related features from intuitive, statistical, and mathematical perspectives. We also showed that this method is performant for robot learning and control in collaborative tasks and has the potential to be further applied in complicated bimanual manipulation, human-robot collaboration, and autonomous surgery situations. Finally, we proved that utilizing the proposed features as an informative extra input in a DL skills classifier reduces uncertainty, boosts accuracy, and improves the out-of-distribution robustness of the model. Looking ahead, the integration of multi-modal data, especially stereo camera images from the JIGSAW dataset, represents a promising area for future research that could significantly enhance our methodologies for surgical skills assessment. By incorporating stereo image data, we can add depth and spatial context, enabling a more detailed analysis of surgical techniques and hand coordination. This promising direction aims to improve both the accuracy and robustness of our evaluations, offering a deeper understanding of complex surgical tasks. Our approach also paves the way for new paradigms beyond surgery, involving technologies such as humanoid robots and exoskeletons, where such objective and data-driven metrics can enhance training, control, and performance evaluation.

Appendix

A. Joint Entropy

Entropy can be thought of as the lack of predictability associated with a random trajectory $\boldsymbol{X}$ drawn from a given probability distribution $p(x_i)$. The entropy of trajectory $\boldsymbol{X}$ is defined as

$$\mathcal{H}(\boldsymbol{X}) := -\sum_{x_i} p(x_i)\log_2 p(x_i) \geq 0. \tag{A.1}$$

Since we use log base 2 in (A.1), the unit of $\mathcal{H}(\boldsymbol{X})$ will be *bits*. For instance, the entropy of three independent consecutive flips of a perfect coin as a sample trajectory is $\mathcal{H}(\boldsymbol{X})=3$. Accordingly, the result of these flips can be communicated with just 3 bits; each bit is allocated for the result of one flip. If the coin is unfair (i.e., a coin which gives a higher chance to either heads or tails being the outcome of a flip), the entropy for each flip will be less than 1 since we have an indication of which result is more likely to occur.

Inspired by (A.1), the joint entropy $\mathcal{H}(\boldsymbol{X},\boldsymbol{Y})$ which is the total lack of predictability associated with a pair of discrete trajectories $(\boldsymbol{X},\boldsymbol{Y})$ with a joint distribution $p(x_i, y_j)$ is defined by

$$\mathcal{H}(\boldsymbol{X},\boldsymbol{Y}) := -\sum_{x_i}\sum_{y_j} p(x_i, y_j)\log_2 p(x_i, y_j). \tag{A.2}$$

$\mathcal{H}(\boldsymbol{X},\boldsymbol{Y})$ can be interpreted as the overall uncertainty assigned to the system generating trajectories $\boldsymbol{X}$ and $\boldsymbol{Y}$. The largest value of $\mathcal{H}(\boldsymbol{X},\boldsymbol{Y})$ in (A.2) holds when trajectories $\boldsymbol{X}$ and $\boldsymbol{Y}$ are independent, i.e., $p(x_i, y_j) = p(x_i)p(y_j)$. The lowest value for $\mathcal{H}(\boldsymbol{X},\boldsymbol{Y})$ holds when one trajectory of $\boldsymbol{X}$ or $\boldsymbol{Y}$ is a deterministic function of the other one. In other words, we have

$$\mathcal{H}(\boldsymbol{X})+\mathcal{H}(\boldsymbol{Y}) \geq \mathcal{H}(\boldsymbol{X},\boldsymbol{Y}) \geq \max\{\mathcal{H}(\boldsymbol{X}),\mathcal{H}(\boldsymbol{Y})\} \geq 0. \tag{A.3}$$

(A.3) makes intuitive sense; the correlation (i.e., statistical relationship) between two trajectories reduces the overall uncertainty associated with the system generating these time series, and therefore, decreases the total amount of entropy within the process.

B. Proof of Property 2

According to (2)

$$\begin{aligned}
\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) &= \sum_{x_i}\sum_{y_j} p(x_i, y_j)\log_2 \frac{p(x_i, y_j)}{p(x_i)\,p(y_j)} \\
&\overset{*}{=} \sum_{x_i,y_j} p(x_i, y_j)\log_2 \frac{p(x_i|y_j)}{p(x_i)} \\
&= -\sum_{x_i,y_j} p(x_i, y_j)\log_2 p(x_i) + \sum_{x_i,y_j} p(x_i, y_j)\log_2 p(x_i|y_j) \\
&= -\sum_{x_i} p(x_i)\log_2 p(x_i) - \left[-\sum_{x_i,y_j} p(x_i, y_j)\log_2 p(x_i|y_j)\right] \\
&= \mathcal{H}(\boldsymbol{X}) - \mathcal{H}(\boldsymbol{X}|\boldsymbol{Y}).
\end{aligned}$$

If we apply $p(x_i,\, y_j) = p(x_i)\; p(y_j|x_i)$ in equality *, similarly we will prove that $\mathcal{I}(\boldsymbol{X};\boldsymbol{Y}) = \mathcal{H}(\boldsymbol{Y}) - \mathcal{H}(\boldsymbol{Y}|\boldsymbol{X})$. ■

REFERENCES

[1] A. Soleymani, X. Li, and M. Tavakoli, "Artificial intelligence in robot-assisted surgery: Applications to surgical skills assessment and transfer," *Medical and Healthcare Robotics*, pp. 183–200, 2023.

[2] N. Ahmidi *et al.*, "A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery," *IEEE Transactions on Biomedical Engineering*, 2017.

[3] A. Soleymani *et al.*, "A domain-adapted machine learning approach for visual evaluation and interpretation of robot-assisted surgery skills," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8202–8208, 2022.

[4] A. Soleymani, X. Li, and M. Tavakoli, "Deep neural skill assessment and transfer: Application to robotic surgery training," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8822–8829.

[5] A. A. Gumbs, I. Frigerio, G. Spolverato, R. Croner, A. Illanes, E. Chouillard, and E. Elyan, "Artificial intelligence surgery: How do we get to autonomous actions in surgery?" *Sensors*, vol. 21, no. 16, p. 5526, 2021.

[6] H. I. Fawaz *et al.*, "Accurate and interpretable evaluation of surgical skills from kinematic data using fully convolutional neural networks," *International journal of computer assisted radiology and surgery*, 2019.

[7] X. A. Nguyen *et al.*, "Surgical skill levels: Classification and analysis using deep neural network model and motion signals," *Computer methods and programs in biomedicine*, vol. 177, pp. 1–8, 2019.

[8] A. Soleymani *et al.*, "Surgical skill evaluation from robot-assisted surgery recordings," in *2021 International Symposium on Medical Robotics (ISMR)*.

[9] I. Funke *et al.*, "Video-based surgical skill assessment using 3d convolutional neural networks," *International journal of computer assisted radiology and surgery*, vol. 14, no. 7, pp. 1217–1225, 2019.

[10] H. Doughty *et al.*, "The pros and cons: Rank-aware temporal attention for skill determination in long videos," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7862–7871.

[11] L. Tao *et al.*, "Sparse hidden markov models for surgical gesture classification and skill evaluation," in *International conference on information processing in computer-assisted interventions*. Springer, 2012, pp. 167–177.

[12] T. N. Judkins *et al.*, "Objective evaluation of expert and novice performance during robotic surgical training tasks," *Surgical endoscopy*, 2009.

[13] K. Liang *et al.*, "Motion control skill assessment based on kinematic analysis of robotic end-effector movements," *The International Journal of Medical Robotics and Computer Assisted Surgery*, vol. 14, no. 1, p. e1845, 2018.

[14] S. S. Vedula *et al.*, "Task-level vs. segment-level quantitative metrics for surgical skill assessment," *Journal of surgical education*, 2016.

[15] J. Diedrichsen *et al.*, "Anticipatory adjustments in the unloading task: is an efference copy necessary for learning?" *Experimental Brain Research*, 2003.

[16] M. Wiesendanger and D. J. Serrien, "Toward a physiological understanding of human dexterity," *Physiology*, vol. 16, no. 5, pp. 228–233, 2001.

[17] B. F. Preilowski, "Possible contribution of the anterior forebrain commissures to bilateral motor coordination," *Neuropsychologia*, 1972.

[18] E. A. Franz *et al.*, "The effect of callosotomy on novel versus familiar bimanual actions: a neural dissociation between controlled and automatic processes?" *Psychological Science*, vol. 11, no. 1, pp. 82–85, 2000.

[19] E. P. Gardner, "Neural pathways for cognitive command and control of hand movements," *Proceedings of the National Academy of Sciences*, vol. 114, no. 16, pp. 4048–4050, 2017.

[20] F. P. Escamirosa *et al.*, "Face, content, and construct validity of the endovis training system for objective assessment of psychomotor skills of laparoscopic surgeons," *Surgical endoscopy*, vol. 29, pp. 3392–3403, 2015.

[21] F. Aghazadeh *et al.*, "Surgical tooltip motion metrics assessment using virtual marker: an objective approach to skill assessment for minimally invasive surgery," *International Journal of Computer Assisted Radiology and Surgery*, vol. 18, no. 12, pp. 2191–2202, 2023.

[22] B. Russell, "On the notion of cause," in *Proceedings of the Aristotelian society*, vol. 13. JSTOR, 1912, pp. 1–26.

[23] K. Hlaváčková-Schindler *et al.*, "Causality detection based on information-theoretic approaches in time series analysis," *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.

[24] M. G. Rosenblum *et al.*, "Synchronization approach to analysis of biological systems," in *The Random and Fluctuating World: Celebrating Two Decades of Fluctuation and Noise Letters*. World Scientific, 2022.

[25] M. Paluš *et al.*, "Synchronization as adjustment of information rates: Detection from bivariate time series," *Physical Review E*, vol. 63, no. 4, p. 046211, 2001.

[26] P. Manshour *et al.*, "Causality and information transfer between the solar wind and the magnetosphere–ionosphere system," *Entropy*, vol. 23, no. 4, p. 390, 2021.

[27] A. Zia and I. Essa, "Automated surgical skill assessment in rmis training," *International journal of computer assisted radiology and surgery*, 2018.

[28] F. Aghazadeh *et al.*, "Motion smoothness-based assessment of surgical expertise: The importance of selecting proper metrics," *Sensors*, vol. 23, no. 6, p. 3146, 2023.

[29] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022.

[30] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, p. 461, 2000.

[31] F. Aghazadeh *et al.*, "Experts employ a target-locking behaviour in laparoscopic surgery," 2022.

[32] H. Bekkering *et al.*, "Reaction time latencies of eye and hand movements in single-and dual-task conditions," *Experimental brain research*, 1994.

[33] D. P. Azari *et al.*, "Can surgical performance for varying experience be measured from hand motions?" in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. SAGE Publications, 2018.

[34] Y. Gao *et al.*, "Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling," in *Miccai workshop*, 2014.

[35] K. Narazaki *et al.*, "Robotic surgery training and performance," *Surgical Endoscopy and Other Interventional Techniques*, vol. 20, no. 1, pp. 96–103, 2006.

[36] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[37] A. Soleymani *et al.*, "Surgical Procedure Understanding, Evaluation, and Interpretation: A Dictionary Factorization Approach," *IEEE Transactions on Medical Robotics and Bionics*, 2022.

[38] A. Hendricks *et al.*, "Exploring the limitations and implications of the jigsaws dataset for robot-assisted surgery," *IEEE Robotics and Automation Letters*, 2024.

[39] B. Zhou *et al.*, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2921–2929.

[40] A. Tarvainen *et al.*, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems, volume=30, year=2017*.

[41] R. Tao *et al.*, "Modeling and emulating a physiotherapist's role in robot-assisted rehabilitation," *Advanced Intelligent Systems*, vol. 2, no. 7, p. 1900181, 2020.

[42] Z.-Y. Chiu, F. Richter, E. K. Funk, R. K. Orosco, and M. C. Yip, "Bimanual regrasping for suture needles using reinforcement learning for rapid motion planning," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 7737–7743.

[43] Y. Ou *et al.*, "Towards Safe and Efficient Reinforcement Learning for Surgical Robots Using Real-time Human Supervision and Demonstration," in *2023 International Symposium on Medical Robotics (ISMR)*.

[44] Y. Ou, A. Soleymani, X. Li, and M. Tavakoli, "Autonomous blood suction for robot-assisted surgery: A sim-to-real reinforcement learning approach," *IEEE Robotics and Automation Letters*, 2024.

[45] J. Xu, B. Li, B. Lu, Y.-H. Liu, Q. Dou, and P.-A. Heng, "Surrol: An open-source reinforcement learning centered and dvrk compatible platform for surgical robot learning," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 1821–1828.

[46] A. Y. Ng, D. Harada, and S. Russell, "hlPolicy invariance under reward transformations: Theory and application to reward shaping," in *Icml*, vol. 99. Citeseer, 1999, pp. 278–287.

[47] A. Gupta, A. Pacchiano, Y. Zhai, S. M. Kakade, and S. Levine, "Unpacking reward shaping: Understanding the benefits of reward engineering on sample complexity," *arXiv preprint arXiv:2210.09579*, 2022.

[48] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *STAT*, 2017.

[49] C. Molnar, *Interpretable machine learning*. Lulu. com, 2020.

[50] S. R. Islam *et al.*, "Explainable artificial intelligence approaches: A survey," *arXiv*, 2021.

[51] L. von Rueden *et al.*, "Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[52] J. Ish-Horowicz *et al.*, "Interpreting deep neural networks through variable importance," *arXiv preprint arXiv:1901.09839*, 2019.

[53] Z. Qin *et al.*, "Neural network classifier as mutual information evaluator," *arXiv preprint arXiv:2106.10471*, 2021.

[54] R. D. Hjelm *et al.*, "Learning deep representations by mutual information estimation and maximization," *International Conference on Learning Representations*, 2018.