https://doi.org/10.1109/LRA.2022.3186769

# A Domain-Adapted Machine Learning Approach for Visual Evaluation and Interpretation of Robot-Assisted Surgery Skills

Abed Soleymani, Xingyu Li, *Member, IEEE*, and Mahdi Tavakoli, *Senior Member, IEEE*

***Abstract*—In this study, we present an intuitive machine learning-based approach to evaluate and interpret surgical skills level of a participant working with robotic platforms. The proposed method is domain-adapted, i.e., jointly utilizes an end-to-end learning approach for smoothness detection and domain knowledge-based metrics such as fluidity and economy of motion for extracting skills-related features within a given trajectory. An advantage of our approach compared to similar stochastic or deep learning models is its intuitive and transparent manner for extraction and visualization of skills-related features within the data. We illustrate the performance of our proposed method on trials of the JIGSAWS data set as well as our own experimental data gathered from Phantom Premium 1.5A Haptic Device. This approach utilized $t$-SNE technique and provides visualized low-dimensional representation for different trials that highlights nuanced information within the executive task and returns unusual or faulty trials as outliers far away from their normal skill or participant clusters. This information regarding the input trajectory can be used for evaluation and education applications such as learning curve analysis in surgical assessment and training programs.**

***Index Terms*—Machine Learning, Surgical Skills Evaluation, Ensemble Models, Contrastive Principal Component Analysis ($c$PCA), $t$-distributed Stochastic Neighbor Embedding ($t$-SNE).**

## I. Introduction

Robot-assisted minimally invasive surgery (RAMIS) is gaining traction in modern clinical practice. To do RAMIS safely and effectively, surgeons must acquire a variety of skills [1]. To assist surgical trainees, reliable surgical assessment methods with informative and instructive feedback would be helpful.

Conventionally, RAMIS skills assessment has been carried out via outcome-based analysis, structured checklists, and rating scales [2]. These qualitative assessment methods need extensive expert monitoring, time, and manual ratings that make them less efficient as well as less reliable because of human bias and variability in human interpretation about similar events. Moreover, these scoring systems due to their observational nature can be insensitive to small but important improvements in the skills level of the trainee and fail to provide insights into the core reasons for surgical failures.

Automated RAMIS skills assessment techniques, on the other hand, bridge these gaps, save time and money, and provide targeted feedback to inexperienced surgeons during their learning phase [3]. Thanks to surgical robot technologies, surgical procedural data are becoming more available and have the potential to pave the way for artificial intelligence (AI) based systems such as machine learning and deep learning models to be deployed in surgical skills assessment.

Autonomous robotic surgery assessment approaches use two main categories of AI models: data-driven (or inductive) models and feature-based (or domain knowledge-based) models [4]. The overwhelming philosophy behind the data-driven models is to use end-to-end models with minimal use of domain knowledge to prevent introducing user bias into the learning procedure. These approaches let the model learn features, choose its structure, and tune its hyperparameters mostly from the input data.

The feature-based models, on the other hand, do not rely on a model to learn features that are already known according to human intuition or dynamical equations of the system. The model of operator's skill is too complex to be captured by a limited amount of training data and always there are model uncertainties and unmodeled dynamics involved. Incorporating domain knowledge as priors reduces uncertainties rendering the modeling problems easier to solve with fewer training data points [5].

### A. Data-Driven Models

There is a rich body of literature including papers from our research group addressing the autonomous robotic skills evaluation problem using data-driven models. For instance, [6], [7] incorporate convolutional neural networks (CNNs) to discover skills-related temporal patterns of kinematic data in the motions of participants performing robotic surgery. Some work tries to combine CNN with recurrent neural networks (RNNs) to systematically classify various levels of expertise in surgical training data sets [8]. Other work such as [3], [9], [10] incorporate CNN or RNN-based spatial attention models to extract skills-related temporal features from endoscopic video frames and predict users' surgical skills level.

Although the majority of these work have reported a relatively low skill misclassification rate, the feature extraction and decision-making procedures in these black-box models are unknown and sometimes unreliable. Moreover, these models due to their high capacity are always prone to be overfitted on small data sets, especially in the field of robotic surgery where the scarcity of qualified human participants and expensive experimentation limit access to standard and large data sets. The mentioned limitations negatively affect the transparency and generalization of any performance feedback for the trainee since the feedback is relied on the model's learned parameters and its confidence about the predicted outcome (see [6]).

Another category of data-driven research in the area of skills assessment rely on utilizing hidden Markov models (HMMs) to

Manuscript received: February 24, 2022; Revised May 17, 2022; Accepted June 13, 2022. This paper was recommended for publication by Editor Jessica Burgner-Kahrs upon evaluation of the Associate Editor and Reviewers' comments.

This research was supported by the Canada Foundation for Innovation (CFI), UAlberta Huawei-ECE Research Initiative (HERI), the Government of Alberta, the Natural Sciences and Engineering Research Council (NSERC) of Canada, the Canadian Institutes of Health Research (CIHR), and the Alberta Economic Development, Trade and Tourism Ministry's grant to Centre for Autonomous Systems in Strengthening Future Communities.

All authors are with the Electrical and Computer Engineering Department, University of Alberta, Edmonton, Alberta, Canada. {zsoleymani, xingyu, mahdi.tavakoli}@ualberta.ca

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

segment a surgical task into its pre-defined building blocks so-called *gestures* to classify subjects according to their skills level [11]. In addition to the fact that these methods mainly suffer from limited recognition rates and challenges for finding the optimal number of hidden states, they require a large number of manually made gesture annotations, which would be very laborious. Moreover, HMMs project trajectories into another space defined by static descriptors that increases the chance of losing important temporal information.

### B. Feature-Based Models

Feature-based approaches for skills evaluation calculate meaningful features as evaluation metrics including total path length [12], motion jerk [13], execution time [12], [13], etc., and run descriptive statistic analysis between human participants on a single metric at a time. Since such metrics have a lot of statistical variation intra and between participants, usually there are considerable overlaps between skill classes and there is little statistically significant difference between different human users. This is because some features such as motion jerk are too noisy or other ones such as total path length or execution time are not informative enough as a single factor for revealing the skills level of the user. Additionally, these papers just consider global metrics across the entire procedure and do not take care of detailed events within the sub-task level of the operation. For instance, these work neglect challenging temporal metrics such as trajectory smoothness (i.e., lack of random motions such as hand tremor or uncontrolled fast actions) as an important contributing factor in skills evaluation of a given trajectory. Different metrics complement each other and could be considered together in a high-dimensional space to find a meaningful and expressive representation of surgical performance for comparison. These methods remain ad-hoc and non-generalizable for new trajectories due to the fact that they neglect temporal patterns and do not build a model for skills assessment.

### C. Contributions

In this work, we bridge the gap between data-driven models and feature-based models and introduce a *domain-adapted model* that simultaneously incorporates manually engineered metrics as well as temporal features learnt from the input data for skills assessment purposes in surgical training programs in which there is more focus on generic trajectory-based skills of surgeons than patient's surgical outcome. Our proposed novel approach extracts smoothness features from the input data under the context of data-driven learning and utilizes global metrics such as fluidity and economy of motions as approved features for detecting the skills level of the executed trajectory [14]. We adopt contrastive principal component analysis ($c$PCA) technique [15] to tackle the challenging problem of smoothness/noise detection within trajectories.

By ensembling these clinically meaningful features, we will achieve an expressive high-dimensional feature space that meaningfully reflects the user's skills level and highlights nuanced information within the surgery. We use $t$-distributed stochastic neighbor embedding ($t$-SNE) as an unsupervised technique [16] to visualize the high-dimensional feature space in a three-dimensional embedding space and investigate the performance of our proposed model.

Leveraging domain knowledge together with data in an unsupervised learning mode reduces the reliance of our method on large data sets and benefits its generalizability, transparency, and reliability compared to other black-box models including deep learning-based models. These features make our approach have a potentially good impact in robotic surgery, where demands for enhanced reliability and explainability in educational and assessment procedures are extremely strong. Note that since this paper proposes general criteria for the surgical skills assessment, it cannot provide detailed feedback about the sub-task performance of the trainee. However, abnormal behaviors will implicitly show themselves in the final global metric, and faulty trajectories will be mapped as outliers far from the more skillful clusters.

The outline of the paper is organized as follows: In Section II, basic concepts and motivations will be discussed. In Section III, experimental results with Phantom Premium Haptic Device will be provided. In Section IV, the quality and performance of the proposed approach will be investigated using JIGSAWS data set. Concluding remarks are provided in Section V.

## II. Methodology

It has been proven that a casual observer can discover and rate the skills levels of a surgeon just by looking at pre-recorded endoscopic videos of surgical tasks with comparable accuracy to an experienced surgical mentor [17]. The intuition behind this observation is that the human understanding of *skill* can be more intuitive and intrinsic than sophisticated assessment [18]. A possible hypothesis about this interesting result is that the observer does not care about extreme details and miniscale translations/rotations inside the surgical trajectories. He/she mainly focuses on hand movement and tool maneuver skills of the user such as smoothness, fluidity, the economy of motion, and so on.

We conclude that as smoothness, fluidity, and energy economy of a given trajectory improve, the assigned level of expertise to the user should increase as well. Inspired by these facts, we develop intuitive and transparent domain adapted sub-models that separately capture each and every one of these important factors. Later on, we will show that by ensembling these sub-models we can achieve a practical feature extraction model, suitable for any downstream task such as classification or data visualization.

First, we elaborate on the smoothness detection technique in the following sections since it is the most challenging part of developing our sub-models. The challenge arises from the fact that the smoothness is a temporal feature that can happen anywhere inside a given trajectory and can be masked by other temporal patterns such as general trend and seasonal patterns. Moreover, there is no straightforward definition or method to search or detect the non-smooth behavior across the time series.

### A. Trajectory Smoothness

*1) Preliminaries:* Exploration of high-dimensional data is always considered as an omnipresent challenge across different fields of applications. Conventional techniques such as principal component analysis (PCA) [19] or multidimensional scaling (MDS) [20] aims to identify dominant trends inside the data. To capture nonlinear trends inside the more complex data sets, other methods including locally linear embedding (LLE) [21] and Isomap [22] were developed to find the *sub-manifold* that preserves local relations between data points in the original space.

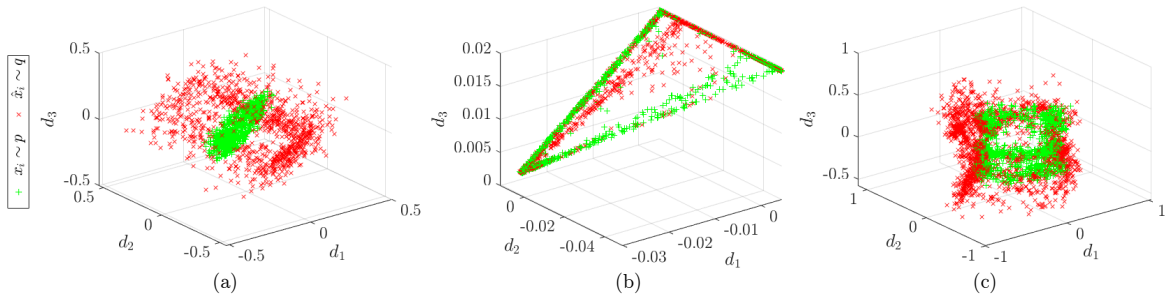



Fig. 1: Visualizing time series in $\boldsymbol{\mathcal{X}}_s$ and $\boldsymbol{\mathcal{X}}_n$ ($n=2000$) in the three dimensional space using (a) MDS and nonlinear manifold learning methods, (b) LLE, and (c) Isomap. These manifold learning methods try to discover the lower dimensional (in this case, three-dimensional) internal structure of each time series for illustration purposes. As it is clear, non of these techniques capture salient features that distiguish fundamental differences between two probability distributions $p$ and $q$ that $\boldsymbol{\mathcal{X}}_s$ and $\boldsymbol{\mathcal{X}}_n$ are sampled from.

However, in many settings we are interested in identifying trends or patterns that are *rich* in some specific features. For instance, in a data set of gene-expression measurements from intact and cancerous individuals, we want to highlight cancer-related variations that purely distinguish the two clusters of individuals. If we directly apply PCA or MDS, it is very likely that the top principal components capture the demographic variations of the individuals such as gene features related to skin color, age, or gender instead of gene features related to the cause of the cancer. Similarly, for nonlinear manifold learning methods such as LLE and Isomap, features of interest may not appear as dominant latent factors, or may be entangled with other prominent ones in the low-dimensional embedding.

Moreover, methods that embed data based on preserving pairwise distances (e.g., MDS) or local patterns of the original data (e.g., LLE and Isomap) provide mappings only for those given training points with no straightforward and deterministic extension for out-of-sample examples [23]. In other words, if we want to embed new test points, we have to run these methods from scratch for the new training set, which is the old training set plus new query points. In addition to significant computational cost, we have no clear boundary between training and test sets for verifying the generalization of the method in classification tasks.

*2) Motivation:* We face a similar problem when we want to extract salient features related to hand tremor or in general noise inside a given trajectory as an informative factor in skills classification of surgical tasks. Consider a fabricated data set $\boldsymbol{\mathcal{X}} = \boldsymbol{\mathcal{X}}_s \cup \boldsymbol{\mathcal{X}}_n$ containing two different types of trajectories: randomly generated smooth trajectories $\boldsymbol{\mathcal{X}}_s$ and noisy trajectories $\boldsymbol{\mathcal{X}}_n$ to model the hand tremor (not sensor or process noises)

$$\boldsymbol{\mathcal{X}}_s = \{\boldsymbol{x}_i\}_{i=1}^{n} \in \mathbb{R}^{d\times n}, \quad \boldsymbol{\mathcal{X}}_n = \{\hat{\boldsymbol{x}}_i\}_{i=1}^{n} \in \mathbb{R}^{d\times n} \tag{1}$$

and we define each noisy sample $\hat{\boldsymbol{x}}_i$ as

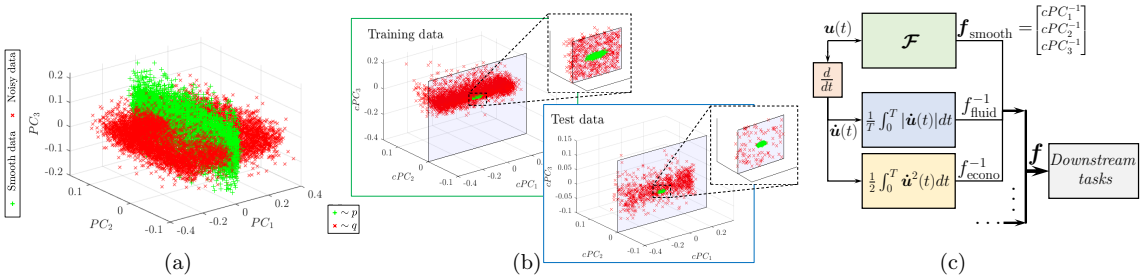
$$\hat{\boldsymbol{x}}_i = \boldsymbol{x}_i + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} = [\varepsilon_1,...,\varepsilon_d]^\top \tag{2}$$

where $\varepsilon_{l\neq\tau k} = 0$, and $\varepsilon_{\tau k} \sim \mathcal{N}(0,\sigma^2)$ in which $\tau$ is the tremor period and $0 \leq k \leq \frac{d}{\tau}$. The variance $\sigma^2$ controls the tremor intensity of samples within $\boldsymbol{\mathcal{X}}_n$ with three different values of 1 mm, 2 mm, and 3 mm. The reason behind such modeling is that physiological tremor is approximately a linear and Gaussian random process with the frequency ranging from 2.5 Hz to 13 Hz [24]. For the example case of $\tau = 3$, since the sampling frequency is 30 Hz, we add random tremor to time-stamps that are multiples of three (i.e., $\varepsilon_{3k} \sim \mathcal{N}(0,\sigma^2)$) to fabricate a 10 Hz tremor signal. We repeat the same procedure to generate tremors with other valid frequencies to make a realistic and intense noisy data set $\boldsymbol{\mathcal{X}}_n$ for the contrastive learning paradigm which will be discussed in Section II-A3. In a broader sense, we assume that $\boldsymbol{\mathcal{X}}_s \sim p$ and $\boldsymbol{\mathcal{X}}_n \sim q$ where $p$ and $q$ are probability distributions corresponding to smooth and non-smooth time series, respectively. Real-world trajectories $\boldsymbol{\mathcal{X}}_h$ do not fall into one of these two groups; they are sampled from another probability distribution $r$.

Since $\boldsymbol{\mathcal{X}}_h$, $\boldsymbol{\mathcal{X}}_s$, and $\boldsymbol{\mathcal{X}}_n$ share substantial temporal features (e.g., general trend or seasonal patterns such as surgical gestures) and these features are entangled with each other, there is no meaningful and straightforward relation between $r$, $p$, and $q$ in the original space. However, if these data points are mapped to a specific embedding space that each dimension identifies and exhibits the most interesting difference between samples from $p$ and $q$, real-world samples lie somewhere between these two ultimate boundaries. In other words, $r$ can be expressed as a linear combination of $p$ and $q$.

Time series are high-dimensional data, mostly with a lot of correlated and redundant variations inside. Dimensionality reduction techniques allow us to extract and analyze only the high-impact information to make insightful data-driven decisions. Fig. 1 and Fig. 2a illustrate four different dimensionality reduction and manifold learning methods that try to uncover the low-dimensional intrinsic structure of each time series. Each point in the three-dimensional embedding space represents a high-dimensional original time series after detecting and deleting unnecessary correlations and dimensions. As it is illustrated in Fig. 1 and Fig. 2a, applying conventional methods do not return a low-dimensional map with enhanced noise-related features for the fabricated training sets $\boldsymbol{\mathcal{X}}_s$ and $\boldsymbol{\mathcal{X}}_n$ since there is no clear boundary between mapped data points of the two sets. In fact, noise-related features are masked due to their entanglement with other dominant temporal features inside trajectories. In addition to the computational simplicity of PCA, non of the above-mentioned techniques are capable to return a mapping function $\boldsymbol{\mathcal{F}}: \mathbb{R}^d \mapsto \mathbb{R}^e$ for out-of-sample points which acts similar to a trained model where $e$ is the intrinsic dimensionality of the embedding space. Another motivation for utilizing PCA is its compatibility with contrastive learning paradigm which will be elaborated in Section II-A3.

*3) Solution:* $c$PCA is a machine learning technique designed to fill the above-mentioned gaps that distinguishes noisy trajectories from smooth ones. $c$PCA generates a mapping function $\boldsymbol{\mathcal{F}}$ that




(a) (b) (c)

Fig. 2: Visualizing data points in $\boldsymbol{\mathcal{X}}_s$ and $\boldsymbol{\mathcal{X}}_n$ ($n=10{,}000$) in the three dimensional space using (a) PCA (8,000 training data), (b) $c$PCA in the three dimensional sub-space (with 8,000 training data and 2,000 test data), and (c) illustration of ensemble model used in this research. $c$PCA technique successfully managed to separate noisy and smooth data (for both test and training) in the three dimensional space.

projects $\boldsymbol{\mathcal{X}}$ to a low-dimensional map where each contrastive principal component ($cPC$) exhibits the most expressive differences between $\boldsymbol{\mathcal{X}}_s$ and $\boldsymbol{\mathcal{X}}_n$ that are related to the smoothness.

$c$PCA takes a *target* data set that has some patterns of interest within as the first input and a *background* data set that does not contain the patterns that we are looking for as the second input. $c$PCA returns sub-spaces that capture variations that are large in the target, but little in the background. In our problem formulations, target patterns are the tremor inside the trajectories. As a result, our target data set is $\boldsymbol{\mathcal{X}}_n$ and the background data set is $\boldsymbol{\mathcal{X}}_s$. At first, we calculate $\boldsymbol{C}_t$ and $\boldsymbol{C}_b$, which are the target and background covariance matrices

$$\boldsymbol{C}_t = \boldsymbol{\mathcal{X}}_n \boldsymbol{\mathcal{X}}_n^\top, \quad \boldsymbol{C}_b = \boldsymbol{\mathcal{X}}_s \boldsymbol{\mathcal{X}}_s^\top. \tag{3}$$

Then, we calculate the so-called *contrastive* covariance matrix $\boldsymbol{C}_c$ and its eigen decomposition

$$\boldsymbol{C}_c = \boldsymbol{C}_t - \alpha \boldsymbol{C}_b = \boldsymbol{W}_c \boldsymbol{\Lambda} \boldsymbol{W}_c^\top \tag{4}$$

where the *contrastive strength* parameter $\alpha$ represents our desire to magnify interesting target variances and attenuate the irrelevant background variance [15]. For $\alpha = 0$, $c$PCA reduces to PCA applied on the target dataset and as $\alpha$ increases, directions with smaller background variance become more significant. For very large $\alpha$, $c$PCA first projects the target data onto the null space of the background data, and then applies conventional PCA on the projected data. This is because when $\alpha$ increases in (4), the penalty for any direction not in the null space of the background data increases accordingly. Hence, the mapping function of $c$PCA technique to the $e$- dimensional embedding for a given data set $\boldsymbol{X}$ is:

$$\boldsymbol{\mathcal{F}}(\boldsymbol{X}) = {}_e\boldsymbol{W}_c^\top \boldsymbol{X} \tag{5}$$

where ${}_e\boldsymbol{W}_c$ is the matrix composed of top first $e$ columns of $\boldsymbol{W}_c$.

We incorporate $c$PCA technique to extract noise-related features from real surgical trajectories. To produce a general and unbiased smoothness detector mapping function $\boldsymbol{\mathcal{F}}$, we fabricate $10,000$ randomly generated smooth trajectories $\boldsymbol{x}_i$ for the smooth trajectory set $\boldsymbol{\mathcal{X}}_s$ and add each of them with randomly generated noise vector $\boldsymbol{\varepsilon}$ to generate noisy trajectories $\hat{\boldsymbol{x}}_i$ for $\boldsymbol{\mathcal{X}}_n$ with three different values for $\sigma^2$ in (2) to have a more realistic noisy data set. In this setting, we choose $\boldsymbol{\mathcal{X}}_n$ as the target and $\boldsymbol{\mathcal{X}}_s$ as the background data set in the $c$PCA formulations. As it is shown in Fig. 2b, $c$PCA with $\alpha \geq 30$ gives us a good embedding space for which noisy and smooth samples are well separated from each other. Since the performance of the approach is the same for a large range of parameter $\alpha$, the sensitivity of the data separation is very low to this particular hyperparameter. The calculated mapping function $\boldsymbol{\mathcal{F}}$ in (5) successfully mapped $8{,}000$ training data points into $e=3$ dimensional embedding space in such a way that all smooth samples are densely clustered around origin and as the noise intensity increases, the embedded representation of the trajectory goes further away from $\vec{\boldsymbol{0}}$ (note that a possible approach for measuring the density of each cluster is calculating the variance inverse of each class). As a result, three $cPC$s in this space form a feature vector $\boldsymbol{f}_{\text{smooth}} = [cPC_1^{-1}, cPC_2^{-1}, cPC_3^{-1}]^\top$ that reflects the smooth behavior of a given trajectory. Using the trained mapping function $\boldsymbol{\mathcal{F}}$ expressed in (5), we mapped other $2{,}000$ test points to the latent space and get the same embedding similar to Fig. 2b (i.e., all smooth data points separately clustered near origin with a scatter of noisy trajectories' representation around them). $\boldsymbol{f}_{\text{smooth}}$ returns higher values for smoother trajectories and low values for jittery motions or trajectories with a lot of hand tremors.

### B. Fluidity of Movements

Fluidity of the movement is another informative skills-related feature that can be extracted from the translational or rotational trajectories. Fluidity reflects how quick and accurate a task is done in transnational or rotational space and can be derived from the time derivative of the input trajectory [25]. A good metric to capture fluidity for a given time series $\boldsymbol{u}(t)$ is calculating the inverse of relative total path length during the execution time. $\boldsymbol{u}(t)$ can be robot's end effector motions along $x$, $y$, and $z$ axis of the Cartesian coordinate system or its rotations around roll, pitch, and yaw directions. This feature $f_{\text{fluid}}$ for both continuous and discrete time series can be calculated as

$$f_{\text{fluid}} = \left(\frac{1}{T}\int_{t=0}^{T} |\dot{\boldsymbol{u}}(t)|dt\right)^{-1}, \text{ or } \quad f_{\text{fluid}} = \left(\frac{1}{N}\sum_{t=0}^{N} |\dot{\boldsymbol{u}}[t]|\right)^{-1} \tag{6}$$

where $T$ is the execution time of $\boldsymbol{u}(t)$, $\dot{\boldsymbol{u}}(t)$ is the time derivative of $\boldsymbol{u}(t)$, and $N$ is the total number of time samples of $\dot{\boldsymbol{u}}[t]$. In this way, $f_{\text{fluid}}$ returns high values for fluent trajectories and low values for faulty, non-accurate, and suddenly generated paths (i.e., happening frequent mid-task failures and restarts). Note that (6) can be vulnerable to sensor noise and return low values even for fluid motions. To address this problem, we filtered all frequencies above 25 Hz which is beyond the human working frequency range [24] to exclude any measurement noise while preserving important human motion variations. Moreover, for the fast but fluid tasks since we have the total execution time $T$ or $N$ in the denominator, (6) will not return incorrect low values for expert trials.

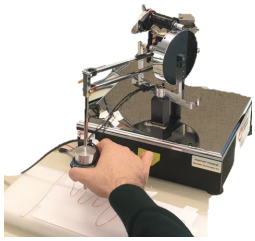


Fig. 3: Phantom Premium 1.5A Haptic Device in the squiggly line tracking task.

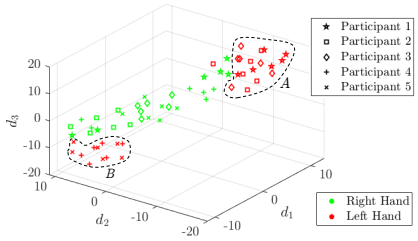



Fig. 4: Three dimensional visualization of Phantom robot experimental trajectories using $t$-SNE technique. Cluster boundaries (i.e., dashed lines) are for the demonstration purpose.

### C. Economy of Motion

Economy of motion reflects the total energy demand for accomplishing a particular task and is a generally accepted factor contributing in skills assessment of various activities [26]. In addition to the fact that in RAMIS applications, user's high energy consumption reflects his/her skills deficiency, human errors (i.e., unintentional random events) typically result in higher execution velocity and higher energy injection to the patient-side robot which are the main sources of danger and trauma in an operation [27]. One possible way to approximate this important factor for RAMIS applications is calculating the total kinetic energy within a given trajectory. We do not consider potential energy because it is trajectory-independent and only depends on what the task requires to do so (i.e., potential energy change for displacing from point $P_1$ to $P_2$ is the same for all possible trajectories with any execution quality). Since the lumped mass of patient-side robot is quite the same for surgical configurations, we can neglect its mass in calculating the inverse of kinetic energy and define $f_{\text{econo}}$ as:

$$f_{\text{econo}} = \left(\tfrac{1}{2}\int_{t=0}^{T} \dot{\boldsymbol{u}}^2(t)dt\right)^{-1}, \text{ or } \quad f_{\text{econo}} = \left(\tfrac{1}{2}\sum_{t=0}^{N} \dot{\boldsymbol{u}}^2[t]\right)^{-1}. \quad (7)$$

In this context, $f_{\text{econo}}$ returns high values for energy economic and safe surgical motions and low values for unsafe and aggressive trajectories.

### D. Ensembling Sub-models

Each one of the above-mentioned metrics is not comprehensive enough to reflect the skills levels of the user during the execution of the trajectory. A possible solution is to concatenate all of these metrics and feed them to downstream data analysis (see Fig. 2c). This makes sense since generating plentiful uncorrelated and meaningful features is an essential part of training a model to make the feature space rich and expressive to achieve better accuracy and generalization.

One major advantage of using such ensemble model in skills assessment task is that the contribution of each factor in the final decision making procedure is transparent. In this way, the framework can give instructive feedback to each user about their performance, weaknesses, and strengths in different characteristics of surgery. This simple and transparent advantage makes our approach superior over other black-box machine learning or deep learning models in which the procedure of feature extraction and decision making is not clear, understandable, or reliable under the context of end-to-end learning over small data sets of surgical tasks.

One method of assessing the feature ensemble is to feed all generated features to a classifier and evaluate its performance. However, since this approach tries to minimize misclassification rate, it does not differentiate between different trials of the same class or a particular participant. In other words, we cannot discriminate outlier points and investigate the reasons for them. Additionally, classification techniques are supervised learning methods which are prone to overfitting over a small dataset. Such a model is not favorable in tasks such as robotic surgery where demands for safety and explainability of intelligent algorithms are very high.

Unsupervised clustering-based data visualization techniques, on the other hand, bridge these gaps and provides an understandable low-dimensional graphical representation of high-dimensional data. Since in this study we have a relatively high-dimensional feature space, linear data visualization techniques such as PCA may fail to capture relationships within the data that may be entangled with others (e.g., tangent, parallel, enveloping, or orthogonal features). Moreover, due to the limited and sparse data points in the feature space, capturing global structure of the data might fail to represent true local structure (cluster) of data points. In this study, we apply $t$-SNE for this purpose. Briefly, $t$-SNE is a nonlinear data visualization method that keeps neighboring points of the high dimensional space close to each other in the low-dimensional embedding [16]. The advantage of $t$-SNE over other nonlinear data visualization methods (e.g., LLE or Isomap) and the main reason for incorporating this method in our paper is the ability to preserve the local structure of the high-dimensional data and handle outliers while taking care of other dimensionality reduction challenges including crowding problem [16] that are neglected in other approaches which are typically slower than $t$-SNE in terms of training time. That is, similar ensemble features in our tasks will be mapped to close by representations (i.e., clusters) in the $t$-SNE visualization space that gives us a feel or intuition of how the data points are arranged in a high-dimensional space. The consistency between results and interpretations of this work and the sequel paper of this research line [28] (which incorporates a completely different skills evaluation and data visualization technique) reinforces the fact that the choice of $t$-SNE as a powerful dimensionality reduction and data visualization technique was right (more details are provided in Section IV).

## III. Preliminary Analysis and Discussion

To evaluate and observe the performance of the proposed method in practice, we gathered a data set resulting from the collaboration of the user and Phantom Premium 1.5A Haptic Device, Geomagic Inc (see Fig. 3). The experiments were approved by the University of Alberta Research Ethics and Management Online under study

ID Pro00055825. Five right-handed users performed the task of tracking squiggly line in 6 trials (data, codes, and supplementary video are available here).

Since there is a significant difference between the performance of the trials done by the dominant hand and with ones performed by the non-dominant hand, we consider trajectories of the dominant hand of the users as expert data and trajectories coming from the other hand as novice data. This assumption is due to the fact that our dominant hand is fully trained to skillfully perform elaborate motions with low hand tremor and high precision compared to the non-dominant hand. We are not arguing that the differences between the left and right hand exactly correspond to the differences between the skills level of novices and experts, we are trying to create two conditions that are different in the performance level to test our approach on. To increase the dexterity of the dominant hand, each participant performed at least five trial sessions with his/her dominant hand until he/she feels is well-prepared before starting the actual experiment. In this experiment, we will show that our proposed method can deliver an expressive representation of skills in which all trajectories coming from the different hands have their own skill clusters. Moreover, we will see that hidden information such as mental concentration can be implicitly interpreted from the visualized representation and how can be affected by the level of expertise of the executing hand.

We feed translational data (i.e., motions executed along $x$, $y$, and $z$ axes of the Cartesian coordinate system) of a given trajectory to the feature extractor model described in Fig. 2c after normalizing each trajectory and unifying the length of each trial. In this paper, we resampled trajectories using a linear interpolation between two consecutive time stamps to bring the sequences from different trials at the same length (i.e., 600 samples) and then rescaled them between 0 and 1 before feeding into feature extractor sub-models. Note that in case of downsampling, the final sampling frequency is greater than 20 Hz since the maximum length of trials is less than 30 seconds. This preserves human-related variations within each time series since the final sampling frequency is higher than the upper limit of the frequency range of human hands movements. Moreover, according to our investigations, no data variation is removed during the resampling/rescaling since we have no sudden motion in minimally invasive surgical tasks. Since task execution time and total path length will change during the resampling and rescaling procedures, we can investigate these factors for our further investigations. To capture inter-channel dependencies between $x$, $y$, and $z$, we calculate robot's end effector's position $\mathcal{P}=\sqrt{x^2+y^2+z^2}$ and feed it to the feature extractor model as well as other transnational data to extract and concatenate all smoothness, fluidity, and economy of motion features for the downstream tasks. In this way, each trial will be represented in the 20-dimensional feature space. The 3D visualization of trials using $t$-SNE technique is illustrated in Fig. 4. As it is clear, novice trials (i.e., trials performed by the left hand) have two different dense clusters $A$ and $B$ and are completely separated from expert cluster with green scattered points. Cluster $A$ exclusively belongs to participants 1 to 3 and cluster $B$ exclusively belongs to participants 4 and 5. This separation can be attributed to the possible nonlinear nature of smoothness that the $c$PCA as a linear transformation cannot catch it properly. Our observations suggest that due to the crisp separation between different data clusters in the high dimensional space, generated 3D visualizations were not sensitive to changes in $t$-SNE hyperparameters (e.g., perplexity, early exaggeration, random state, etc. [16]) which can be attributed to the fact that proposed metrics in this work are expressive in term of capturing skills-related features within surgical trajectories.

Moreover, this separation can be attributed to the different levels of mental concentration between participants of clusters $A$ and $B$. Participants in cluster $A$ showed significant vertical variation and higher tracking error while performing the task compared to participants in cluster $B$ which we assume is related to their lower level of attention or concentration while performing the task. However, the right hand trajectories have quite the same behavior between all participants. This can be attributed to the high level of expertise of the dominant hand that even compensate the lack of concentration of the user. Although class label-free factors such as concentration are not part of our problem formulations, they leave their tracks in the embedding space and a knowledgeable person can interpret them from the visualized data.

Although the main focus of this paper is the visual evaluation and interpretation of task trajectories in terms of the skills level of the user, one may whether the meaningful separation between different skills clusters in Fig. 4 is due to the performance of $t$-SNE as a powerful data visualization technique. A complementary approach is to show that utilizing proposed features in Section II is beneficial for user's skills classification. We trained a primary machine learning-based support-vector machine (SVM) model [29] with a simple third-degree polynomial kernel to classify the right-hand data (i.e., experts) and the left-hand data (i.e., novices) from each other. Since we have a limited number of data samples, we performed a 10-fold cross-validation [10] for 500 times over original features set (i.e., without applying $t$-SNE). We achieved the skills classification accuracy of $96.27\%$ $(\pm 1.06)$ for the training set and that of $95.2\%$ $(\pm 2.98)$ for the test set. This result indicates that features presented in Section II are expressive, performant, and interpretable for the user's skills assessment purposes.

## IV. Full Analysis and Discussion

In addition to translational data (i.e., $x$, $y$, $z$, and $\mathcal{P}$) we also can feed rotational data (i.e., *roll* ($\Phi$), *pitch* ($\Theta$), and *yaw* ($\Psi$) angles) of a given trajectory to the feature extractor model described in Fig. 2c. To capture inter-channel dependencies between rotational data, we feed $\mathcal{R}=\sqrt{\Phi^2+\Theta^2+\Psi^2}$ to the feature extractor model as well.

All analyses in this section are based on the standard JIGSAWS data set [30] collected from surgical activities of eight surgeons in three different levels of expertise (i.e., novice, intermediate, and expert) performing suturing, knot-tying, and needle-passing tasks on the *da Vinci* Surgical System. JIGSAWS contains three Cartesian motions along $x$, $y$, and $z$ axes as well as 9 elements of rotation matrix $\boldsymbol{R}\in\mathbb{R}^{3\times 3}$ for both hands of the user. Note that, all 9 elements of rotation matrix $\boldsymbol{R}$ can be expressed as $\Phi$, $\Theta$, and $\Psi$ angles as follows

$$\Phi=\text{atan}\left(\frac{r_{21}}{r_{11}}\right),\ \Theta=\text{atan}\left(\frac{-r_{31}}{\sqrt{r_{32}^2+r_{33}^2}}\right),\ \Psi=\text{atan}\left(\frac{r_{32}}{r_{33}}\right)$$

where $r_{ij}$ is the element in the $i^{\text{th}}$ row and $j^{\text{th}}$ column of $\boldsymbol{R}$.

We applied feature extraction method described in Fig. 2c on all 6 axes of translational and rotational data as well as $\mathcal{P}$ and $\mathcal{R}$ for both hands of all participants in the JIGSAWS data set. In total, we will have an 80-dimensional feature vector $\boldsymbol{f}$ for each trial inside the data

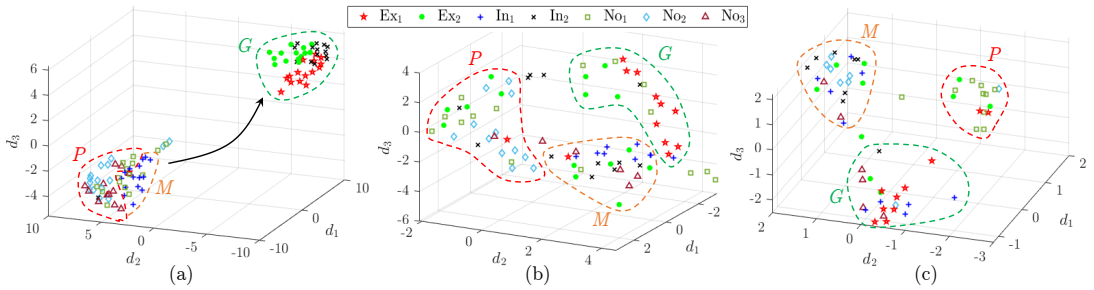



Fig. 5: $t$-SNE visualization of three different tasks of the JIGSAWS data set: (a) suturing task, (b) needle passing task, and (c) knot tying task.

set. Here, we will show that the ensemble of these features together is expressive enough in terms of capturing core features related to the skills level of each participant. Note that since the criteria discussed in this paper are very generic metrics in capturing trajectory-based skills-related features, they may not fully capture high-level hidden information within the trajectories such as surgeon style.

Now, we apply $t$-SNE technique on the extracted feature space of several surgical tasks to reach three-dimensional visualizations shown in Fig. 5. Fig. 5a illustrates the visualization of extracted features of the suturing task. As it is clear, trials of two expert users and one intermediate user cluster close to each other and create the *good* cluster $G$. Note that the level of goodness of each cluster is primarily determined by the class label of its trials (i.e., expert trials are more likely to be good) and our investigation into endoscopic data (i.e., trials with no or a very small number of mid-task failures and restarts are more probable to belong to the good cluster). Other novice and intermediate trials generate *poor* and *moderate* clusters $P$ and $M$, respectively, far from the cluster $G$. An important point regarding cluster $G$ is that each individual user has his/her own dense sub-cluster. In fact, participants with higher level of expertise perform each trial in a more consistent manner (i.e., denser sub-clusters) relative to less experienced ones. Such hidden information cannot be revealed by classification techniques or cannot be trusted in the data-driven models.

Another point about the cluster $G$ is existence of an intermediate user $\mathrm{In}_2$ near to expert users. Manual annotations with several labels are usually coarse-grained and our approach reveals their limitations. Based on the definition in [30], novice users have less than 10 hours of operation experience with the da Vinci Surgical Systems, intermediate users have between 10 and 100 hours, and expert users have more than 100 hours. It is clear that there is a big bound for intermediate users. Based on our investigation on endoscopic videos, $\mathrm{In}_2$ performed very well compared to expert users in the suturing task. It can be speculated that $\mathrm{In}_2$ have near to 100 hours of operation experience with the da Vinci Surgical Systems. This advantage of our method can be used to extract *learning curve* of a participant when he/she proceeds from the poor cluster towards the good cluster during the training program.

As it is clear in Fig. 5b and Fig. 5c, there is no sign of crisp boundaries between skills clusters and users' sub-clusters for needle passing and knot tying tasks, respectively. Additionally, each $G$, $M$, and $P$ clusters are not exclusive enough in terms of containing data points of their own users. This is because of frequent mid-task failures and restarts in many trials of all users which we found based on the recorded endoscopic video for each trials in the JIGSAWS data set. Under these circumstances, data points do not belong to specific sub-spaces or sub-manifolds and randomly occurring mistakes inside trajectories map them to random places in the feature space. In this situation, black-box models and supervised learning-based models fail to detect such information and are prone to be overfitted and unreliable. For further clarification of the above-mentioned discussions, a supplementary video is provided.

A majority of these notable results are well-supported in the sequel work of this research published by the same authors [28]. Surgical trials considered in this research are structured tasks and the main variations within each time series can be broken up into finite components namely *general trends* and *seasonal patterns*. According to this intuition, [28] took a completely different approach and proposed a novel dictionary factorization technique for approximate trajectory decomposition and surgical skills evaluation task. For example, the observed patterns in Fig. 5a (i.e., the neighborhood of $\mathrm{In}_2$ and expert trials in cluster $G$ and having trials of $\mathrm{In}_1$ and other novice trials near to each other) is similar to the patterns observed in [28] for suturing task. This result gets even more interesting when considering the fact that data illustration in [28] is according to the embedding space representation of the proposed method, not based on data visualization techniques such as $t$-SNE. This emphasizes the fact that first, the proposed approach in this paper despite its simplicity, is reliable and comparable to more sophisticated unsupervised techniques in revealing underlying skills-related features which might be forsaken by end-to-end learning methods. Secondly, the visualization similarity emphasizes that the $t$-SNE method is a reliable data dimensionality reduction technique for our particular skills evaluation task.

The method discussed in this paper unlike other high-capacity end-to-end models that address the same surgical skills evaluation problem fabricates a limited number of meaningful features and then $t$-SNE method further reduces the dimensionality of the feature space to make it easier for human analysis and interpretation. Since always there is a trade-off between prediction accuracy (high-dimensional feature space) and interpretability (low dimensional illustration), comparing our approach with other high-capacity models in terms of classification accuracy is not quite fair as we are not competing with existing methods in terms of predicting class labels (i.e., cross-validation accuracy). The merit of this paper is uncovering label-free information within trials and offering insightful correction hints (e.g., the user should improve his/her smoothness or avoid unnecessary movements) to the trainees which none of the end-to-end models with high classification accuracy are capable of. One example of label-free information that our method provided in this study is clustering $\mathrm{In}_2$ near to experts despite the

intermediate skill label assigned to that participant. The assigned global rating scores ($GRS$) by an experienced gynecologic surgeon using a modified objective structured assessments of technical skills (OSATS) approach [31] in suturing task to $\mathrm{In}_2$ (i.e., $3.1\pm0.57$ out of 5) are similar to those of expert trials (i.e., $2.64\pm0.47$ for $\mathrm{Ex}_1$ and $3.2\pm0.3$ for $\mathrm{Ex}_2$) and are significantly higher than $GRS$ of $\mathrm{In}_1$ (i.e., $2\pm0.54$) and other novice participants (i.e., $1.75\pm1.07$ for $\mathrm{No}_1$, $1.66\pm0.3$ for $\mathrm{No}_2$, and $2.8\pm0.84$ for $\mathrm{No}_3$) [30].

One similar work in this area that attempted to explain the internal behavior of the neural network is [6] in which authors used class activation maps (CAM) method [32] to visualize in which part of the input data, the deep model pays more attention that leads to the predicted outcome. This method heavily relies on the learned parameters of the deep network which makes it less transparent and explainable in human terms compared to our method. Moreover, the self-judgment paradigm in the CAM approach can suffer from an obvious problem: if the model generates incorrect or uncertain predictions for a given trajectory, the CAM method can be wrong and become progressively relied on artifacts that are generated by the network itself. This issue intensifies the *confirmation bias* [33] since the model is continually confirming its own incorrect or overestimated belief about the decision rule. Note that since we incorporated a large synthetic data set for the data-driven part of our solution (i.e., smoothness detection), we are less vulnerable to drawbacks of the CAM method discussed so far. Finally, CAM-based methods provide explanations in the trajectory domain without highlighting the skills deficiencies of the user.

## V. Conclusions

A novel domain-adapted approach for the evaluation, interpretation, and visualization of surgical maneuvers was presented in this paper. Inspired by domain knowledge, we defined clinically meaningful features such as fluidity and economy of motion as important metrics for evaluating the skills of surgical tasks. These metrics alongside the smoothness-related features captured by an end-to-end learning paradigm revealed salient features corresponding to the skills level of the user. The visualization of the ensemble of extracted features in the three-dimensional space using $t$-SNE technique for both experimental data gathered from the Phantom robot and JIGSAWS data set showed that our approach can meaningfully express skills level, abnormality, and hidden information within a given surgical trajectory. The explainability and transparency of our approach make it more reliable, safe, and interpretable compared to other state-of-the-art black-box models in the skills assessment of surgical tasks and learning curve analysis of intern users.

## References


[1] John D Birkmeyer et al. Surgical skill and complication rates after bariatric surgery. *New England Journal of Medicine*, 369(15):1434–1442, 2013.

[2] Narges Ahmidi et al. A dataset and benchmarks for segmentation and recognition of gestures in robotic surgery. *IEEE Transactions on Biomedical Engineering*, 64(9):2025–2041, 2017.

[3] Isabel Funke et al. Video-based surgical skill assessment using 3d convolutional neural networks. *International journal of computer assisted radiology and surgery*, 14(7):1217–1225, 2019.

[4] Nikhil Muralidhar et al. Incorporating prior domain knowledge into deep neural networks. In *2018 IEEE international conference on big data (big data)*, pages 36–45. IEEE, 2018.

[5] Laura von Rueden et al. Informed machine learning–a taxonomy and survey of integrating knowledge into learning systems. *arXiv*, 2019.

[6] Hassan Ismail 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] Abed Soleymani et al. 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.

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

[9] Hazel 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*, pages 7862–7871, 2019.

[10] Abed Soleymani and et al. Surgical skill evaluation from robot-assisted surgery recordings. In *2021 International Symposium on Medical Robotics (ISMR)*. IEEE, 2021.

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

[12] Timothy N Judkins et al. Objective evaluation of expert and novice performance during robotic surgical training tasks. *Surgical endoscopy*, 23(3):590, 2009.

[13] Ke 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*, 14(1):e1845, 2018.

[14] Ahmad Ghasemloonia et al. Surgical skill assessment using motion quality and smoothness. *Journal of surgical education*, 74(2):295–305, 2017.

[15] Abubakar Abid et al. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9(1):1–7, 2018.

[16] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[17] Carolyn Chen, et al. Crowd-sourced assessment of technical skills: a novel method to evaluate surgical performance. *Journal of surgical research*, 2014.

[18] Marzieh Ershad et al. Meaningful assessment of surgical expertise: Semantic labeling with data and crowds. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2016.

[19] Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

[20] Michael AA Cox and Trevor F Cox. Multidimensional scaling. In *Handbook of data visualization*, pages 315–347. Springer, 2008.

[21] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[22] Joshua B Tenenbaum et al. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

[23] Yoshua Bengio et al. Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering. *Advances in neural information processing systems*, 16:177–184, 2003.

[24] Jing Zhang and Fang Chu. Real-time modeling and prediction of physiological hand tremor. In *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing.*, volume 5, pages v–645. IEEE, 2005.

[25] Marzieh Ershad et al. Surgical skill level assessment using automatic feature extraction methods. In *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*, volume 10576, page 105762W. International Society for Optics and Photonics, 2018.

[26] David 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*, volume 62, pages 583–587. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[27] Bernadette McCrory, Chad A LaGrange, and MS Hallbeck. Quality and safety of minimally invasive surgery: past, present, and future. *Biomedical engineering and computational biology*, 6:BECB–S10967, 2014.

[28] Abed Soleymani et al. Surgical procedure understanding, evaluation, and interpretation: A dictionary factorization approach. *IEEE Transactions on Medical Robotics and Bionics*, 2022.

[29] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[30] Yixin Gao et al. Jhu-isi gesture and skill assessment working set (jigsaws): A surgical activity dataset for human motion modeling. In *Miccai workshop: M2cai*, volume 3, page 3, 2014.

[31] JA Martin et al. Objective structured assessment of technical skill (osats) for surgical residents. *British journal of surgery*, 84(2):273–278, 1997.

[32] Bolei Zhou et al. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016.

[33] Antti 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*, 30, 2017.