# Learning to Estimate Palpation Forces in Robotic Surgery From Visual-Inertial Data

Young-Eun Lee, Haliza Mat Husin, *Member, IEEE*, Maria-Paola Forte, *Graduate Student Member, IEEE*, Seong-Whan Lee, *Fellow, IEEE*, and Katherine J. Kuchenbecker, *Fellow, IEEE*

***Abstract*—Surgeons cannot directly touch the patient's tissue in robot-assisted minimally invasive procedures. Instead, they must palpate using instruments inserted into the body through trocars. This way of operating largely prevents surgeons from using haptic cues to localize visually undetectable structures such as tumors and blood vessels, motivating research on direct and indirect force sensing. We propose an indirect force-sensing method that combines monocular images of the operating field with measurements from IMUs attached externally to the instrument shafts. Our method is thus suitable for various robotic surgery systems as well as laparoscopic surgery. We collected a new dataset using a da Vinci Si robot, a force sensor, and four different phantom tissue samples. The dataset includes 230 one-minute-long recordings of repeated bimanual palpation tasks performed by four lay operators. We evaluated several network architectures and investigated the role of the network inputs. Using the DenseNet vision model and including inertial data best-predicted palpation forces (lowest average root-mean-square error and highest average coefficient of determination). Ablation studies revealed that video frames carry significantly more information than inertial signals. Finally, we demonstrated the model's ability to generalize to unseen tissue and predict shear contact forces.**

***Index Terms*—Force estimation, indirect force sensing, robot-assisted minimally invasive surgery, visual-inertial input, deep learning.**

## I. Introduction

PALPATION of tissues and organs is crucial to localize visually undetectable tumors and buried blood vessels in open surgery [1]. However, surgeons performing minimally invasive surgery (MIS) and robot-assisted minimally invasive surgery (RMIS) cannot contact patient tissue with their fingertips and therefore cannot leverage these valuable tactile and kinesthetic cues [2]. In particular, the friction and transversal moments imposed on the surgical instruments by the trocar compromise the haptic cues in MIS [3], [4], [5], [6]. The situation is even worse in RMIS as no haptic feedback is provided to the surgeon operating the robot from the console [7]. Hence, surgeons must learn to adjust their applied forces by relying on vision [8].

In particular, surgeons learn to translate visual cues into artificial tactile cues subconsciously by using their experience in non-robotic surgery [8]; for example, a mental picture of bowel being grasped is connected with the corresponding haptic sensation [8]. Typical visual cues are the deformation and discoloration of the tissue and the appearance of the suturing material [8]. However, visual cue compensation alone highly depends on the surgeon's experience [8], makes tissue characterization via palpation immensely difficult [9], increases the operation time, and leads to excessive contact forces [10], [11]. Exerting excessive forces might cause tissue damage and trauma [12], whereas applying forces that are too low might lead to a delay or incomplete tasks [12]. Therefore, being able to exert the optimal force is essential for accomplishing a surgical task efficiently and safely.

Prior work directly measured the exerted force and provided it as haptic cues (vibrotactile and/or force feedback) during blunt dissection [13], needle insertion, suturing, palpation [14], [15], and surgical training tasks [16]. Across these many examples, providing haptic feedback improved tissue discrimination [15], shortened the operation time [15], [17], and reduced the magnitude of contact forces [13], [18]. These previous studies focused on one main application scenario: the online estimation of forces for providing real-time haptic feedback. However, estimating the forces applied to the tissue can also provide an objective metric for training and assessment of surgical performance [12], [19], [20], [21].

To measure forces, novel force sensors have been integrated into the shafts of surgical instruments [22]. However, due to their position, these sensors need to be biocompatible, sterilized, miniaturized, and robust [22]. Furthermore, in RMIS, the cables required to drive the instrument's distal degrees of freedom must typically act across such a force sensor, making contact forces difficult to distinguish. To overcome these limitations, recent research has emerged

Manuscript received 15 September 2022; revised 10 April 2023; accepted 20 June 2023. Date of publication 13 July 2023; date of current version 9 August 2023. This article was recommended for publication by Associate Editor P. Fiorini and Editor P. Dario upon evaluation of the reviewers' comments. This work was supported by the Max Planck Society and three Institute of Information and Communications Technology Planning and Evaluation (IITP) Grants funded by the Korean Government through the Global Internship Program for Developing Human Resources in Artificial Intelligence under Grant 2019-0-01605, through the Artificial Intelligence Graduate School Program, Korea University under Grant 2019-0-00079, and through the Artificial Intelligence Innovation Hub under Grant 2021-0-02068. *(Young-Eun Lee and Haliza Mat Husin contributed equally to this work.) (Corresponding authors: Seong-Whan Lee; Katherine J. Kuchenbecker.)*

Young-Eun Lee is with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea (e-mail: ye_lee@korea.ac.kr).

Haliza Mat Husin, Maria-Paola Forte, and Katherine J. Kuchenbecker are with the Haptic Intelligence Department, Max Planck Institute for Intelligent Systems, 70569 Stuttgart, Germany (e-mail: hmh@is.mpg.de; forte@is.mpg.de; kjk@is.mpg.de).

Seong-Whan Lee is with the Department of Artificial Intelligence, Korea University, Seoul 02841, South Korea (e-mail: sw.lee@korea.ac.kr).

This article has supplementary downloadable material available at https://doi.org/10.1109/TMRB.2023.3295008, provided by the authors.

Digital Object Identifier 10.1109/TMRB.2023.3295008

on *indirect force-sensing methods*. Some researchers have investigated the use of motor current measurements [23], [24], [25], [26], [27] or have mechanically modeled the organs and tissues on which the force is exerted [28], [29]. Others have utilized visual information captured from the endoscope, such as tissue deformation, to train deep-learning architectures to estimate contact forces. Specifically, similar work used vision models (*e.g.,* VGG [30], [31], [32], ResNet [33], InceptionResNet [34]) either alone or with temporal architectures (*e.g.,* LSTM [31], [32], [34], RNN [33]). Besides the tissue deformation, cameras can also see the trajectory of each instrument tip. Its utility in the force estimation problem has been previously demonstrated [31], but this information can also be captured with other sensors. Combining visual information with the instrument position given by either the manipulators [35], [36] or the robotic joints [37] has shown higher accuracy in predicting the contact forces than using only vision and, more specifically, stereoscopic vision. However, while there is an upward trend in using stereo endoscopes in MIS, 2D laparoscopy remains the predominant technique due to its lower costs, widespread accessibility, and familiarity to surgeons [38]. So far, only a few works have combined kinematic information with monocular images [31], [33], [34]. Importantly, these efforts used kinematic information from the robot, limiting their utility to RMIS and the specific robotic platform used. In contrast, we aim to develop a method whose kinematic information is not obtained from a robot to facilitate transfer to other platforms as well as traditional MIS.

Given the limitations of current direct and indirect force-sensing methods, we propose a *deep-learning architecture that estimates tissue contact forces* by combining *monocular visual images* with inertial information obtained with *tiny inertial measurement units (IMUs)* attached to the shafts of the instruments outside the patient's body. This sensor suite is commercially available, has high robustness, and avoids issues with biocompatibility and sterilization.

We collected visual-inertial data of palpation actions performed on phantom tissues with either the left or the right instrument. The visual information captures the tissue deformations and the motions of the instrument tips. The three-axis accelerometer, three-axis gyroscope, and three-axis magnetometer in each IMU respond to the acceleration, angular velocity, and orientation of each instrument during the palpation. Since palpation is commonly performed to localize tumors and blood vessels, we manufactured four different phantom tissue samples whose average stiffnesses matched those of healthy and cancerous tissues. We hypothesized that knowing the average tissue stiffness, *i.e.*, the Young's modulus, could facilitate the interpretation of tissue deformations and, in turn, enable more accurate force prediction. Even though measuring the stiffness intraoperatively is not straightforward, we included this information to test our hypothesis in a controlled lab setting.

We extensively analyzed several deep-learning architectures for estimating palpation forces; in particular, we compared seven pre-trained ImageNet models on the visual data and investigated the effects of the temporal dimension in the prediction. This comprehensive investigation offers valuable insights that combine and expand prior related work. We then performed an ablation study to analyze the contribution of each network input (video frames, inertial data, and Young's modulus). Finally, we tested the capability of our network to estimate normal forces on unseen tissue and also predict shear contact forces.

## II. Materials and Methods

### A. Experimental Setup

Fig. 1 shows the experimental setup. The palpation tasks were executed on phantom tissue samples by four operators who controlled an Intuitive Surgical da Vinci Si surgical robot. The robot was equipped with two 8-mm-diameter needle-driver instruments that passed through trocars without sealing gaskets to reach the phantom tissue.

*1) Stereo Endoscope:* The robot is equipped with a 0° stereo endoscope. A laptop computer connected to the da Vinci TilePro video output recorded the stereo images at 30 frames per second and with a resolution of 1280×1024 pixels. The screen annotations and day-to-day changes in camera mounting prohibited us from obtaining good stereo calibration; thus, we used only the left channel of the acquired stereo endoscope to simulate a monocular setup.

*2) Inertial Measurement Units:* To obtain kinematic information, we used two nine-axis IMU sensors (TDK-Invensense ICM-20948); each IMU is inserted into a 3D-printed bracket that is rigidly attached to the shaft of the left or right robotic instrument close to the instrument housing. This IMU contains a three-axis accelerometer, a three-axis gyroscope, and a three-axis magnetometer. The output of the accelerometer depends on the instrument's orientation relative to gravity, its translational acceleration, and the vibrations it experiences. The IMU's three-axis gyroscope measures the angular velocity vector of the instrument shaft, and its three-axis magnetometer measures the orientation and strength of the local magnetic field. These inertial data were recorded at a sampling rate of 500 Hz for each robotic instrument using the same laptop computer to achieve temporal synchronization with the stereo images.

*3) Phantom Tissue Samples:* The surgical environment is made of a custom laparoscopic box trainer containing a piece of simulated tissue in the shape of a torus; the entire assembly is attached to a tilting table [39]. Following Forte et al. [39], we prepared four circular phantom tissues with a diameter of 11 cm and a thickness of 2 cm to be placed at the center of the torus for palpation interactions. Each tissue sample is fabricated on top of a rigid plate with a diameter of 11 cm to allow the transfer of palpation forces to the force sensor below independent of where the force is applied. Two tissue samples were prepared with Smooth-On Soma Foama 25 (T270 and T250) and two with Smooth-On Soma Foama 15 (T150 and T120). Smooth-On Soma Foama is a soft two-component platinum-cure silicone casting foam; different stiffness values were obtained by modifying the percentage of the two components. Each sample has a different average stiffness and significant variations in stiffness across its surface to simulate

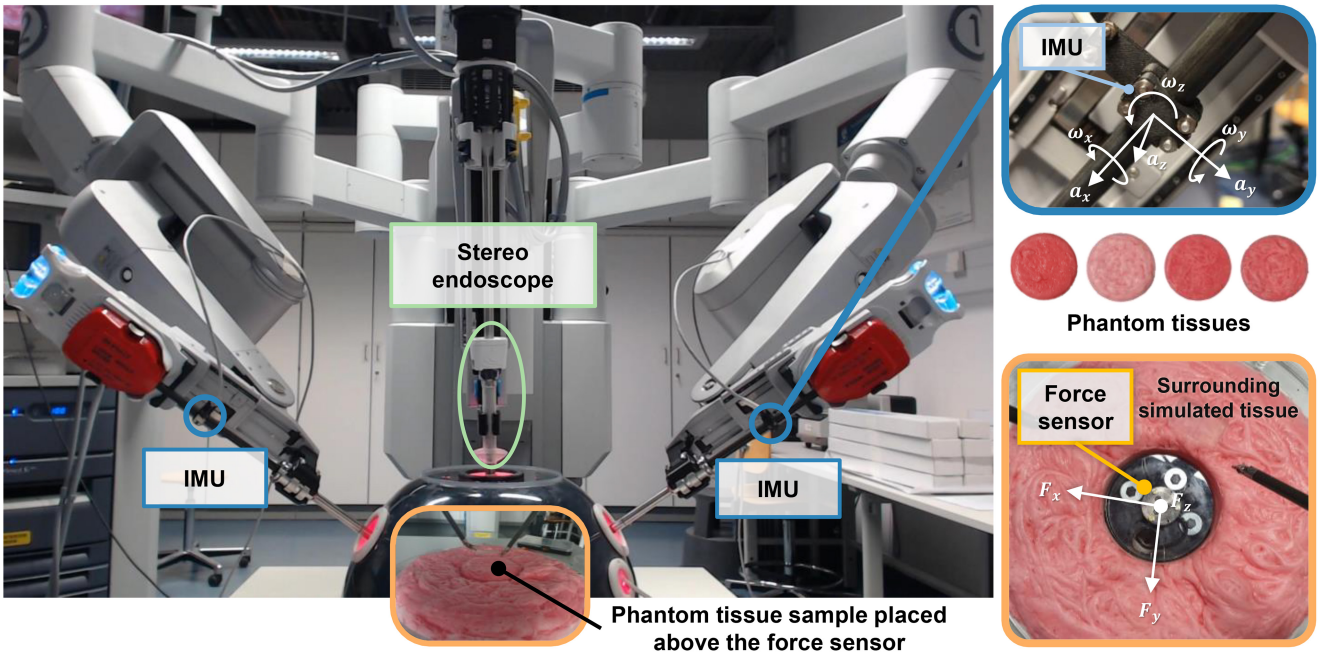


Fig. 1. Experimental setup. The patient cart of our da Vinci Si surgical system is composed of three robotic arms: one arm moves the stereo endoscope, and the other two move the surgical instruments. Each IMU is inserted into a 3D-printed bracket that is rigidly attached to the instrument shaft close to the instrument housing. The surgical environment consists of a phantom tissue sample placed on the force sensor at the center of a toroidal piece of simulated tissue; these items are covered by a custom laparoscopic box trainer. The images at right depict the coordinate frames of the IMU and the force sensor.

healthy and cancerous human tissues [40], [41]. They also have somewhat different colors and textures.

After fabrication, we measured the Young's modulus of each circular tissue sample at five locations: one in the center and four at the extremities of axis-aligned diameters. In each location, we manually indented the tissue at a range of depths using a flat-bottomed 11-mm-diameter rod (chosen to match the size of MIS and RMIS instruments), and we recorded the applied force with a digital scale. This procedure was performed three times for indentation depths ranging from 1 mm to 10 mm at 1 mm increments, and the force measurements at each depth were averaged. For each location, we obtained the Young's modulus $E$ using the following equation:

$$E = \frac{\bar{F}/A}{\Delta l/L}, \tag{1}$$

where $\bar{F}$ denotes the average force measured by the digital scale, $A$ indicates the contact area between the rod and the tissue sample (a circular area with a diameter of 11 mm), $\Delta l$ corresponds to the indentation depth (change in tissue sample thickness), and $L$ is the initial sample thickness (20 mm). The ranges, means, and standard deviations of these Young's modulus measurements are reported in Table I.

*4) Force Sensor:* Below the sample and attached to the tilted table, we placed a three-axis force sensor (ATI Mini40) to obtain the ground-truth forces ($F_x$, $F_y$, $F_z$). The rigid plastic plate of the tissue sample being palpated is fixed to the upper surface of the force sensor. The data from the force sensor were recorded using a data acquisition module (National Instruments USB6361) sampling at 1000 Hz. Table I shows the means and standard deviations of the maximum per-recording force magnitude in each orthogonal direction.

TABLE I
THE RANGES, MEANS, AND STANDARD DEVIATIONS (SDS) OF THE YOUNG'S MODULUS MEASURED AT FIVE LOCATIONS ON EACH TISSUE SAMPLE, ALONG WITH THE MEANS AND SDS OF THE MAXIMUM FORCES EXERTED IN EACH DIRECTION ON EACH OF THE FOUR PHANTOM TISSUE SAMPLES DURING DATA COLLECTION

| | | T270 | T250 | T150 | T120 |
|---|---|---|---|---|---|
| | |  | | | |
| Young's Modulus (kPa) | Range | 189 – 297 | 205 – 274 | 64 – 294 | 40 – 186 |
| | Mean ± SD | 267 ± 43 | 250 ± 26 | 147 ± 81 | 122 ± 51 |
| Max Force (N) | $F_x$ | 2.56 ± 1.10 | 2.56 ± 1.26 | 1.66 ± 0.80 | 1.13 ± 0.56 |
| | $F_y$ | 2.54 ± 0.93 | 2.62 ± 1.22 | 1.74 ± 0.62 | 1.01 ± 0.52 |
| | $F_z$ | 15.54 ± 6.48 | 14.76 ± 4.34 | 9.40 ± 4.43 | 3.95 ± 2.30 |

## B. Synchronization

We synchronized the endoscopic images with the measurements from the IMUs and force sensor by aligning the start times of all recordings. Inertial and force data were collected with two MATLAB instances, and the videos were recorded using the software program OBS Studio, all running on the same computer. Synchronization between the two MATLAB instances was achieved through serial communication, followed by a command to initiate video recording via a key-press event.

## C. Dataset

The dataset consists of three palpation tasks performed by four operators. The tasks are:

- Right-hand palpation: the operator used the right instrument to pressdown repeatedly on the phantom tissue, while the left instrument was held in a mostly static position not in contact with the tissue.

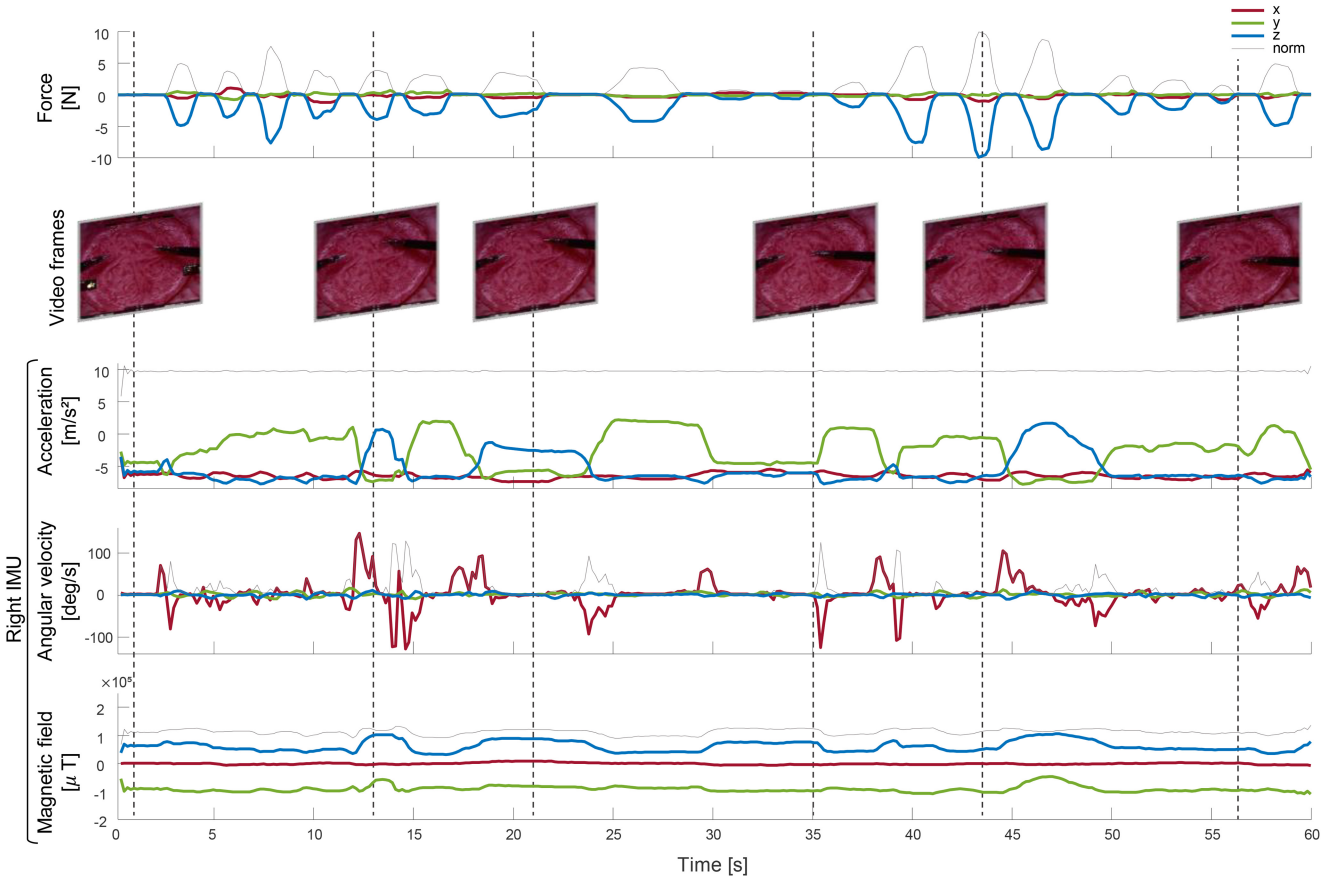


Fig. 2. Data captured during a right-hand palpation task on T120. The first row shows the ground-truth forces collected from the three-axis force sensor. The second row depicts sample frames captured by the left channel of the stereo endoscope. The other three rows depict the inertial data from the nine-axis IMU attached to the right instrument. The *x*-, *y*-, and *z*-axes are colored in red, green, and blue, respectively. In all subplots, the solid thin gray line presents the square root of the sum of squares of the three axes. The six gray dashed lines correspond to the six instants of the sample frames, temporally aligned with the measurements; the first, fourth, and sixth points represent non-contact conditions, and the others capture interactions between the right instrument and the phantom tissue sample. As instructed, the operator held the left instrument mostly static and not in contact with the tissue while the right instrument moved to many positions to palpate the phantom tissue.

- Left-hand palpation: the operator used the left instrument to press down repeatedly on the phantom tissue, while the right instrument was held in a mostly static position not in contact with the tissue.
- Bimanual palpation: The operator alternately used the right and left instruments to press down on the phantom tissue.

The supplemental video associated with this article shows a sample recording of each of our three palpation tasks.

The four operators have no clinical background but are familiar with controlling the da Vinci robot. They palpated the presented sample tissue by repeatedly applying a force approximately perpendicular to the surface. We did not give any instructions about the number of palpations to perform, the palpation force, or the locations to palpate. During each 60-second-long recording, the operator palpated the sample multiple times, with different force levels (each peak of Fig. 2 represents a palpation event), and in different locations. Each recording contains repetitions of only one of our three palpation tasks, and each operator performed multiple 60-second-long recordings for the same tissue. In total, we collected 20 recordings of each palpation task on each tissue sample, with the only exception being T150, on which we collected only 10 bimanual-palpation recordings. Thus, in total, our dataset encompasses 80 recordings of right-hand palpation, 80 of left-hand palpation, and 70 of bimanual palpation. We plan to publicly share this full dataset upon publication.

We downsampled the input data using the Fourier method to decrease the computational time required to train our network and, thus, facilitate the evaluation of multiple architectures, different inputs, and generalization ability. We explored various sampling-rate reductions from the video's raw rate of 30 Hz, ultimately selecting 6 Hz as it demonstrated minimal performance degradation. Consequently, in every recording, we have 360 data points for each input.

Figure 2 shows sample data collected during a right-hand palpation task. Among the six temporal points highlighted with gray dashed lines, the first, fourth, and sixth points represent non-contact conditions, while the other three points correspond to timestamps with instrument-tissue interactions.

### *D. Network Architecture*

Figure 3 displays the network architecture used to predict the palpation forces. Our full architecture has three inputs

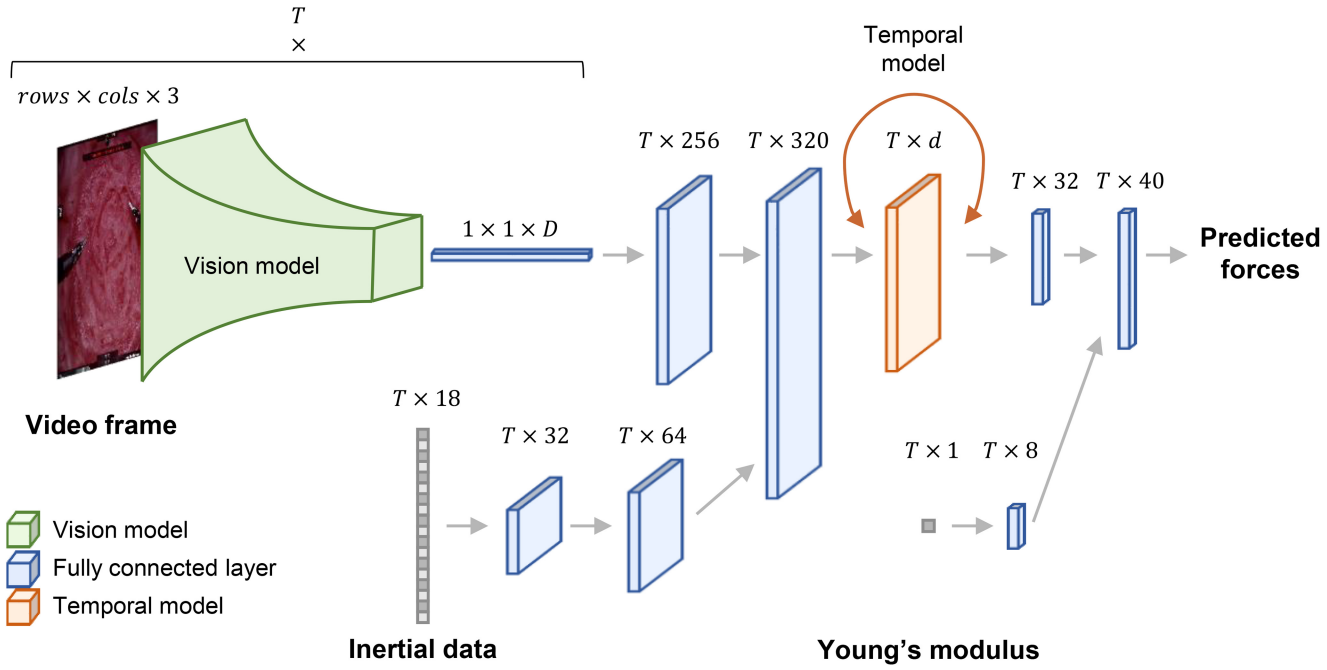


Fig. 3. Network architecture. The inputs of the network are the video frames obtained from the left channel of the stereo endoscope, the left and right inertial data captured by the IMUs, and the average value of the Young's modulus measured on the tissue sample. The green block represents the vision model. The input size of the video frame (row×cols×3) follows each vision model, *e.g.*, 224×224×3 for VGG, DenseNet, and ViT. $T$ refers to the number of time points in a recording, *i.e.*, 360, and $D$ represents the output dimension of the second-to-last layer of each vision model. The blue blocks are the fully connected layers. The orange block is the temporal model, and $d$ refers to its output dimension.

(left-channel video frames, inertial data, Young's modulus) added at different stages of the network.

For the visual input, as a pre-processing step, we resized the RGB video frames obtained from the left channel of the stereo endoscope to match the different input sizes of the pre-trained models, *e.g.*, 224×224×3 for VGG, DenseNet, and ViT. The resized frames were then fed into a pre-trained ImageNet model. Since the last two layers of a pre-trained ImageNet model are designed for object classification, we removed them and used only the portion of the architecture that extracts image features. Keeping in mind that the performance of vision models greatly depends on the task, we performed a comprehensive evaluation of widely used image-classification methods, *i.e.*, VGG16 [42], ResNet-v2 [43], Inception-v3 [44], InceptionResNet-v2 [45], NASNet [46], [47], ViT-Large [48], DenseNet-201 [49], MobileNet [50], and EfficientNet [51]. From this initial study, we chose to continue investigating models that either had a foundation in prior related research, *i.e.*, VGG [30], [31], [32], ResNet [33], and InceptionResNet [34] or achieved superior performance in our preliminary analysis, *i.e.*, DenseNet, Inception, NASNet, and ViT.

The eighteen channels of inertial data (nine for the right instrument and nine for the left instrument) are passed through a two-layer neural network to increase the non-linearity of the inertial data, allowing our system to model more complex input-output relationships. Furthermore, this step projects the inertial measurements and the visual features into a common feature space where they are then concatenated.

This concatenated vector is fed into a temporal model, *e.g.*, gated recurrent units (GRU), long short-term memory network (LSTM). Considering the limited size of our dataset, we evaluated GRU but found that they under-performed compared to LSTMs. Furthermore, the selection of the temporal model depends on the application scenario; whereas LSTM relies only on past data and can be thus used for the real-time application scenario, offline force estimation can also benefit from future data and thus use BiLSTM. Our implementations of LSTM and BiLSTM consist of six units, each containing one second of data; thus, they both output six dimensions from past data, and BiLSTM also outputs six dimensions from future data.

Finally, the tissue's average Young's modulus is concatenated and fed into the last layer before predicting the palpation forces. To localize tumors and buried blood vessels, the palpation direction is defined as perpendicular to the tissue surface [52], which in our case is the $z$-axis of the force sensor (see Fig. 1). The standard version of our network thus seeks to use the camera images, IMU measurements, and Young's modulus to predict the force applied to the tissue along the $z$-axis, *i.e.*, $F_z$.

### *E. Training Details and Evaluation Metrics*

The model loss function was defined as the mean squared error between the predicted forces and the ground-truth forces. We used the Adam optimizer with a learning rate of $10^{-4}$ and a weight decay of $10^{-4}$ to minimize the model loss and train our network. We trained the network for $10^4$ epochs.

To thoroughly evaluate our model, we used a five-fold cross-validation method. The dataset was split into five subsets that contain 46 recordings each. The subsets are balanced with

respect to the three palpation tasks, the four phantom tissue samples, and the four operators.

We then used two evaluation metrics: the root-mean-squared error (RMSE) and the coefficient of determination ($R^2$). The RMSE represents the square root of the average squared difference between the predicted and actual force values; it is commonly used for evaluating force estimation [31], [32], [33]. The $R^2$ value shows the percentage of the variation in the predicted force values that can be explained by the ground-truth force values; this metric is often used to evaluate nonlinear regression models in many fields [53]. Low RMSE and high $R^2$ represent different aspects of good performance.

### F. Ablation Study

We performed an ablation study to analyze the importance of each of the three input modalities, *i.e.*, the video frames, the inertial data of the left and right instruments, and the average Young's modulus of the phantom tissue being palpated. The training parameters, validation method, and evaluation metrics are the same as described in Section II-E.

### G. Model Generalization

To establish how well our learned network generalizes, we first evaluated its ability to predict palpation forces on an unseen tissue sample; an unseen sample has a different average Young's modulus and, in turn, different instrument-tissue interactions. A leave-one-out validation was used; we split the dataset into four subsets, each containing all the recordings from the same phantom tissue. For each of the four different phantom tissues, we then trained the force-estimation model using the three subsets from the other three tissue samples and tested it on the data from the left-out sample. The training parameters and evaluation metrics are the same as in Section II-E.

Second, we investigated the performance of our model at estimating the contact forces along all three axes, *i.e.*, $F_x$, $F_y$, and $F_z$, since palpating the tissue also generated small contact forces in the shear directions, which might be clinically relevant for some applications.

## III. Results

### A. Comparison of Network Architectures

We first focused on the visual input and investigated the performance of seven different vision models: six convolutional neural networks, *i.e.*, VGG, ResNet, Inception, NASNet, DenseNet, and one transformer, *i.e.*, ViT. As a temporal model, we used BiLSTM for all architectures. The results are reported in Table II as the means and standard deviations of the two evaluation metrics across the five-fold cross-validation: Dense-BiLSTM showed the best performance, *i.e.*, the lowest RMSE and the highest $R^2$ for almost all tissue samples in all three palpation tasks, followed by InceptionRes-BiLSTM. ViT-BiLSTM consistently showed the worst results. Figure 4 shows sample prediction results for each model compared with the ground truth; it is possible to qualitatively observe that Dense-BiLSTM follows the ground truth most accurately.

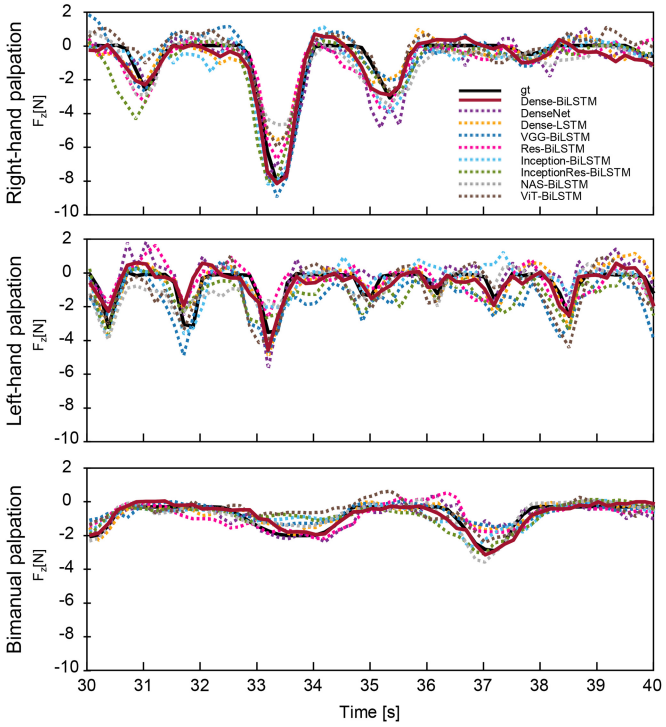


Fig. 4. Sample force signals from our comparison of network architectures. Ground-truth palpation forces (gt) and palpation forces predicted from the tested architectures, i.e., Dense-BiLSTM, DenseNet, Dense-LSTM, VGG-BiLSTM, Res-BiLSTM, Inception-BiLSTM, InceptionRes-BiLSTM, NAS-BiLSTM, and ViT-BiLSTM. Row 1 shows a ten-second right-hand palpation task performed on sample T270 by Operator 1; Row 2 shows a left-hand palpation task performed on sample T250 by Operator 1; Row 3 shows a bimanual palpation task performed on sample T120 by Operator 2. These ten-second force signals were cropped from 60-second recordings.

We then investigated the effect of the temporal dimension (time and sequence order) applied to the visual-inertial feature vector. We used DenseNet as the vision model since it proved to be the best among the seven tested. In particular, we compared BiLSTM against LSTM and the vanilla version of DenseNet. Unlike BiLSTM, LSTM does not rely on future information and can thus provide real-time predictions. The inference time further decreases when no past information is used (vanilla DenseNet). The last three rows of each section of Table II compare Dense-BiLSTM, vanilla DenseNet, and Dense-LSTM. As expected, Dense-BiLSTM has the lowest RMSE and the highest $R^2$, followed by Dense-LSTM, showing that the temporal information improves predictions.

Using SPSS, we performed a repeated-measures analysis of variance (ANOVA) with a Bonferroni post-hoc correction for pairwise comparisons to statistically compare the six lower-performing model architectures, *i.e.*, VGG-BiLSTM, Res-BiLSTM Inception-BiLSTM, InceptionRes-BiLSTM NAS-BiLSTM, ViT-BiLSTM, Dense-BiLSTM, and DenseNet, with the overall winning model, *i.e.*, Dense-BiLSTM. Results with $p < 0.05$ were regarded as statistically significant. As reported in Table II, among the different model architectures, ViT-BiLSTM had, overall, a significantly higher RMSE and lower $R^2$ compared to Dense-BiLSTM for all tissues in all tasks.

For the temporal model, the analysis revealed an overall significantly higher RMSE and lower $R^2$ of vanilla DenseNet

TABLE II
RMSE AND $R^2$ OF THE PREDICTED PALPATION FORCES (MEANS AND STANDARD DEVIATIONS OF THE FIVE-FOLD CROSS-VALIDATION) ACROSS DIFFERENT DEEP-LEARNING ARCHITECTURES. LOW RMSE (↓) REPRESENTS BETTER PERFORMANCE, WHILE HIGHER $R^2$ (↑) INDICATES SUPERIOR PERFORMANCE. BOLD FONT DENOTES THE BEST PERFORMANCE FOR EACH PALPATION TASK, WHILE THE ASTERISKS INDICATE STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE OVERALL BEST MODEL, I.E., DENSE-BILSTM

| | | RMSE [N] ↓ | | | | $R^2$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | T270 | T250 | T150 | T120 | T270 | T250 | T150 | T120 |
| Right-hand palpation | VGG-BiLSTM | 1.60 ± 0.54 | 1.43 ± 0.29* | 0.89 ± 0.45 | 0.55 ± 0.23* | 0.76 ± 0.07 | 0.66 ± 0.13* | 0.67 ± 0.10* | 0.51 ± 0.10* |
| | Res-BiLSTM | 1.69 ± 0.53 | 1.17 ± 0.25 | 1.00 ± 0.58 | 0.52 ± 0.22 | 0.72 ± 0.12* | 0.75 ± 0.12 | 0.64 ± 0.12* | 0.58 ± 0.12 |
| | Inception-BiLSTM | 1.60 ± 0.46 | 1.26 ± 0.26 | 0.81 ± 0.45 | 0.50 ± 0.20 | 0.74 ± 0.09 | 0.74 ± 0.13 | 0.74 ± 0.07 | 0.58 ± 0.11 |
| | InceptionRes-BiLSTM | 1.63 ± 0.50* | **1.15 ± 0.24** | 0.90 ± 0.55 | 0.50 ± 0.21 | 0.73 ± 0.13 | 0.76 ± 0.11 | 0.72 ± 0.09 | 0.60 ± 0.14 |
| | NAS-BiLSTM | 1.52 ± 0.42 | 1.22 ± 0.32 | **0.80 ± 0.43** | 0.52 ± 0.25 | 0.76 ± 0.09 | 0.76 ± 0.11 | 0.75 ± 0.09 | 0.60 ± 0.07 |
| | ViT-BiLSTM | 1.71 ± 0.45* | 1.44 ± 0.29* | 0.94 ± 0.49 | 0.56 ± 0.21* | 0.69 ± 0.14* | 0.66 ± 0.14* | 0.64 ± 0.12* | 0.48 ± 0.13* |
| | **Dense-BiLSTM** | **1.37 ± 0.39** | 1.17 ± 0.28 | 0.81 ± 0.48 | **0.47 ± 0.19** | **0.79 ± 0.13** | **0.76 ± 0.12** | **0.75 ± 0.09** | **0.62 ± 0.11** |
| | DenseNet | 1.62 ± 0.42* | 1.30 ± 0.32 | 1.01 ± 0.51* | 0.57 ± 0.22* | 0.70 ± 0.19* | 0.72 ± 0.12 | 0.58 ± 0.13* | 0.45 ± 0.14* |
| | Dense-LSTM | 1.48 ± 0.39 | 1.25 ± 0.30 | 0.91 ± 0.51 | 0.51 ± 0.22 | 0.76 ± 0.14 | 0.74 ± 0.11 | 0.68 ± 0.09* | 0.58 ± 0.11 |
| Left-hand palpation | VGG-BiLSTM | 2.08 ± 0.88 | 1.69 ± 0.49* | 1.04 ± 0.46 | 0.60 ± 0.37 | 0.71 ± 0.13 | 0.71 ± 0.08* | 0.66 ± 0.11 | 0.50 ± 0.14 |
| | Res-BiLSTM | 1.74 ± 1.03 | 1.50 ± 0.41 | 0.95 ± 0.53 | 0.70 ± 0.35 | 0.67 ± 0.12 | 0.74 ± 0.09* | 0.66 ± 0.13 | 0.56 ± 0.12 |
| | Inception-BiLSTM | 1.89 ± 0.87 | 1.32 ± 0.31 | 0.95 ± 0.39 | **0.54 ± 0.30** | 0.77 ± 0.09 | 0.82 ± 0.07 | 0.70 ± 0.12 | 0.58 ± 0.14 |
| | InceptionRes-BiLSTM | **1.56 ± 0.94** | 1.31 ± 0.30 | **0.88 ± 0.49** | 0.61 ± 0.34 | 0.72 ± 0.11 | 0.80 ± 0.06 | 0.67 ± 0.17 | **0.65 ± 0.11** |
| | NAS-BiLSTM | 1.96 ± 0.80 | 1.35 ± 0.34* | 0.97 ± 0.43 | 0.55 ± 0.32 | 0.75 ± 0.08 | 0.81 ± 0.06* | 0.71 ± 0.08 | 0.58 ± 0.12 |
| | ViT-BiLSTM | 1.94 ± 0.75 | 1.58 ± 0.35* | 1.13 ± 0.51* | 0.61 ± 0.36* | 0.75 ± 0.10 | 0.74 ± 0.07* | 0.61 ± 0.08* | 0.46 ± 0.14* |
| | **Dense-BiLSTM** | 1.77 ± 0.69 | **1.22 ± 0.32** | 0.95 ± 0.48 | 0.55 ± 0.39 | **0.79 ± 0.08** | **0.85 ± 0.06** | **0.74 ± 0.07** | 0.62 ± 0.09 |
| | DenseNet | 2.00 ± 0.79* | 1.38 ± 0.34* | 1.16 ± 0.51* | 0.64 ± 0.42* | 0.73 ± 0.12 | 0.80 ± 0.07* | 0.59 ± 0.08* | 0.47 ± 0.12* |
| | Dense-LSTM | 1.97 ± 0.87 | 1.29 ± 0.37 | 1.06 ± 0.53* | 0.60 ± 0.43 | 0.76 ± 0.08 | 0.83 ± 0.05 | 0.68 ± 0.05* | 0.57 ± 0.07* |
| Bimanual palpation | VGG-BiLSTM | 2.03 ± 0.70* | 1.56 ± 0.57* | 1.45 ± 0.41 | 0.62 ± 0.22* | 0.59 ± 0.13* | 0.65 ± 0.14* | 0.62 ± 0.08 | 0.46 ± 0.13* |
| | Res-BiLSTM | 1.92 ± 0.48 | 1.59 ± 0.73 | 1.60 ± 0.96 | 0.90 ± 0.61 | 0.66 ± 0.10 | 0.63 ± 0.12 | 0.65 ± 0.12 | 0.59 ± 0.11 |
| | Inception-BiLSTM | 1.87 ± 0.73* | 1.39 ± 0.40 | 1.40 ± 0.40 | 0.55 ± 0.21 | 0.67 ± 0.08* | 0.70 ± 0.14 | 0.65 ± 0.08 | 0.58 ± 0.10 |
| | InceptionRes-BiLSTM | 1.72 ± 0.46* | 1.49 ± 0.69 | 1.41 ± 0.84 | 0.90 ± 0.62 | 0.72 ± 0.10 | 0.67 ± 0.10 | **0.71 ± 0.14** | 0.59 ± 0.12 |
| | NAS-BiLSTM | 1.76 ± 0.60 | 1.29 ± 0.46 | 1.39 ± 0.37 | 0.56 ± 0.23 | 0.70 ± 0.08 | **0.75 ± 0.11** | 0.65 ± 0.07 | 0.58 ± 0.08 |
| | ViT-BiLSTM | 1.98 ± 0.67* | 1.56 ± 0.45* | 1.53 ± 0.33* | 0.62 ± 0.24* | 0.61 ± 0.09* | 0.64 ± 0.13* | 0.57 ± 0.07* | 0.47 ± 0.14* |
| | **Dense-BiLSTM** | **1.70 ± 0.62** | **1.28 ± 0.45** | **1.28 ± 0.32** | **0.54 ± 0.23** | **0.72 ± 0.07** | 0.75 ± 0.12 | 0.70 ± 0.09 | **0.61 ± 0.08** |
| | DenseNet | 1.92 ± 0.62* | 1.45 ± 0.44* | 1.59 ± 0.39* | 0.64 ± 0.26* | 0.64 ± 0.08* | 0.67 ± 0.17* | 0.55 ± 0.09* | 0.45 ± 0.11* |
| | Dense-LSTM | 1.87 ± 0.72* | 1.32 ± 0.42 | 1.43 ± 0.37 | 0.56 ± 0.24 | 0.67 ± 0.08* | 0.73 ± 0.13 | 0.63 ± 0.07 | 0.58 ± 0.08 |

compared to Dense-BiLSTM, while only a few significant differences were found for the RMSE and $R^2$ of Dense-LSTM compared with the Dense-BiLSTM across all tasks. Independent of the vision model used, on average, the right- and left-hand palpation tasks obtained better results than the bimanual palpation task, and the estimation of $F_z$ in the right-hand palpation task outperformed the estimation of $F_z$ in the left-hand palpation task.

### B. Ablation Study

We investigated the importance of each network input (video frames, inertial data, and Young's modulus). As reported in Table III with the means and standard deviations of the two evaluation metrics across the five-fold cross-validation, removing either the inertial data or the Young's modulus did not impact the model performance; overall, the RMSE slightly increased, and the $R^2$ decreased moderately. In some cases, removing them produced the best results. On the contrary, removing the video frames greatly deteriorated the predictions. Nonetheless, as visible in Fig. 5, the network could still predict forces reasonably well even when the video frames were not provided.

Similar to the previous analysis, we used a repeated-measures ANOVA with a Bonferroni correction to compare a model using all three inputs ("All") with the model's ablated versions. The repeated-measures ANOVA showed a significant main effect of the network inputs on the RMSE and $R^2$ for all tissue samples across tasks. A paired t-test revealed a significantly higher RMSE and lower $R^2$ of the model without the video compared to "All" (RMSE: $p < 0.01$, $R^2$: $p < 0.01$). No other comparisons were significant.

### C. Predictions on Unseen Tissue and of Three-Axis Forces

Table IV presents the means and standard deviations of both performance metrics for our generalization experiments. We conducted paired t-tests to investigate the model's ability to predict forces on unseen tissues. As expected, estimating forces on a new tissue usually showed significantly worse performance than our baseline method that trained and tested on all tissue samples. Furthermore, unlike the previous analyses, no differences were observed among the palpation tasks for this more challenging test.

In addition, we report the results of Dense-BiLSTM when estimating the three-axis contact forces, recalling that the $x$ and $y$ shear forces were much smaller than the normal forces (Tab. I). $F_x$ and $F_y$ showed both a lower RMSE (better performance) and a lower $R^2$ (worse performance) than $F_z$.

We then performed a paired t-test to compare the prediction of $F_z$ when it was estimated independently ("$F_z$ baseline" in Table IV) with the prediction of $F_z$ when it was inferred together with $F_x$ and $F_y$ ("$F_z$ combined"). As reported in Table IV, the RMSE and $R^2$ values were significantly different between the two predictions for only about one third of the tested palpation tasks and sample tissues, with some instances of each prediction performing better.

## IV. DISCUSSION

Palpation allows surgeons to localize visually undetectable tumors and buried blood vessels. Motivated by the increasing number of surgeries performed in a minimally invasive manner, research has focused on predicting the forces surgeons apply to the patient's tissues with surgical instruments using either direct or indirect sensing methods. This paper used

TABLE III
ABLATION STUDY FOR PREDICTION OF PALPATION FORCES (MEANS AND STANDARD DEVIATIONS OF THE FIVE-FOLD CROSS-VALIDATION) USING THE BEST MODEL, I.E., DENSE-BILSTM. BOLD FONT DENOTES THE BEST PERFORMANCE FOR EACH PALPATION TASK, WHILE THE ASTERISKS INDICATE STATISTICALLY SIGNIFICANT DIFFERENCES FROM THE PERFORMANCE USING ALL FEATURES, I.E., "ALL"

| | | RMSE [N] ↓ | | | | $R^2$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | T270 | T250 | T150 | T120 | T270 | T250 | T150 | T120 |
| Right-hand palpation | w/o Young's modulus | 1.38 ± 0.43 | 1.19 ± 0.25 | **0.80 ± 0.43** | 0.48 ± 0.21 | **0.80 ± 0.10** | 0.76 ± 0.12 | 0.74 ± 0.09 | 0.62 ± 0.11 |
| | w/o inertial data | 1.38 ± 0.40 | 1.19 ± 0.26 | 0.83 ± 0.45 | **0.47 ± 0.19** | 0.80 ± 0.10 | **0.77 ± 0.10** | 0.73 ± 0.09 | 0.62 ± 0.11 |
| | w/o video frames | 2.31 ± 0.92* | 1.92 ± 0.58* | 1.30 ± 0.71* | 0.69 ± 0.27* | 0.51 ± 0.14* | 0.44 ± 0.17* | 0.34 ± 0.20* | 0.21 ± 0.15* |
| | **All** | **1.37 ± 0.39** | **1.17 ± 0.28** | 0.81 ± 0.48 | 0.47 ± 0.19 | 0.79 ± 0.13 | 0.76 ± 0.12 | **0.75 ± 0.09** | **0.62 ± 0.11** |
| Left-hand palpation | w/o Young's modulus | 1.87 ± 0.77 | 1.27 ± 0.32 | 0.97 ± 0.47 | 0.54 ± 0.37 | 0.77 ± 0.08 | 0.83 ± 0.06 | 0.72 ± 0.07 | 0.61 ± 0.12 |
| | w/o inertial data | 1.86 ± 0.79 | 1.30 ± 0.33 | 0.99 ± 0.49 | **0.54 ± 0.39** | 0.78 ± 0.06 | 0.83 ± 0.07 | 0.71 ± 0.08 | **0.63 ± 0.08** |
| | w/o video frames | 2.72 ±1.46* | 2.25 ± 0.63* | 1.52 ± 0.70* | 0.76 ± 0.48* | 0.56 ± 0.15* | 0.50 ± 0.11* | 0.30 ± 0.18* | 0.24 ± 0.13* |
| | **All** | **1.77 ± 0.69** | **1.22 ± 0.32** | **0.95 ± 0.48** | 0.55 ± 0.39 | **0.79 ± 0.08** | **0.85 ± 0.06** | **0.74 ± 0.07** | 0.62 ± 0.09 |
| Bimanual palpation | w/o Young's modulus | 1.73 ± 0.61 | 1.30 ± 0.43 | 1.29 ± 0.35 | 0.55 ± 0.26 | 0.71 ± 0.07 | 0.74 ± 0.12 | 0.70 ± 0.09 | 0.61 ± 0.08 |
| | w/o inertial data | 1.71 ± 0.59 | 1.31 ± 0.42 | **1.26 ± 0.27** | **0.54 ± 0.22** | 0.71 ± 0.07 | 0.73 ± 0.13 | **0.71 ± 0.06** | 0.61 ± 0.09 |
| | w/o video frames | 2.73 ±1.07* | 2.32 ±1.08* | 2.15 ± 0.49* | 0.79 ± 0.36* | 0.30 ± 0.17* | 0.29 ± 0.22* | 0.16 ± 0.11* | 0.20 ± 0.09* |
| | **All** | **1.70 ± 0.62** | **1.28 ± 0.45** | 1.28 ± 0.32 | 0.54 ± 0.23 | **0.72 ± 0.07** | **0.75 ± 0.12** | 0.70 ± 0.09 | **0.61 ± 0.08** |

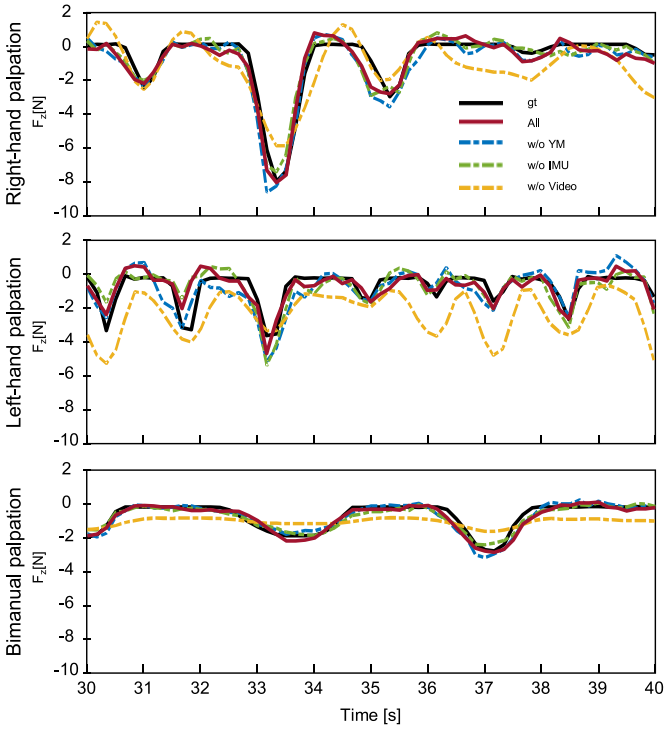


Fig. 5. Sample force signals from the ablation study. Ground-truth palpation forces (gt) and palpation forces predicted when each of the three inputs has been removed, i.e., without Young's modulus (w/o YM), without inertial data (w/o IMU), and without video frames (w/o Video). "All" represents the model when all inputs are used. Row 1 shows a ten-second right-hand palpation task performed on sample T270 by Operator 1; Row 2 shows a left-hand palpation task performed on sample T250 by Operator 1; Row 3 shows a bimanual palpation task performed on sample T120 by Operator 2. These ten-second force signals were cropped from 60-second recordings.

visual-inertial data and deep learning to indirectly estimate the palpation forces exerted on four different phantom tissues during three palpation tasks. We evaluated the prediction performance based on the RMSE and $R^2$ metrics. However, it is important to note that defining whether the force prediction is good enough for real applications is challenging, as, to the best of our knowledge, there are no quantitative metrics that correlate the force prediction to the surgeon's perception or to the quality of the skill assessment.

Among the tested state-of-the-art vision architectures, we found that DenseNet was the most promising vision model for estimating palpation forces. The higher performance of DenseNet could be explained by its high generalization during pre-training; DenseNet consists of deep residual modules and has a relatively small number of parameters, which can prevent overfitting and facilitate adaptation to images from other fields, such as surgery.

Furthermore, including past temporal information in the vanilla model increased the $R^2$ value by 10.3% (0.61 DenseNet vs. 0.68 Dense-LSTM), and also adding future information led to an additional 7.4% improvement (0.68 Dense-LSTM vs. 0.73 Dense-BiLSTM). However, the use of future information, *i.e.*, Dense-BiLSTM, is limited to offline force-estimation applications. For real-time force estimation, Dense-LSTM or vanilla DenseNet should be used; importantly, overall, Dense-LSTM did not show significantly worse results compared with Dense-BiLSTM and could thus be used to estimate the forces needed to provide real-time haptic feedback to surgeons.

The results of the ablation study highlighted the importance of visual information in learning the palpation forces. In contrast, removing the inertial data of the robotic instruments or the average tissue stiffness did not significantly affect the results. Vision alone might then suffice to capture changes in the instrument position (which we instead provided using IMUs) and deformations of the tissue (which we hinted through the Young's modulus). However, when vision is used alone, the choice of a powerful vision model is crucial. Interestingly, when we removed the video frames, the network trained on only inertial data and the Young's modulus was still able to predict the ground-truth force to some extent. Examination of sample predictions like those shown in Fig. 5 revealed that the predicted force peaks were somewhat smoothed. Furthermore, this force prediction never went to zero; during no-contact situations, this network tended to estimate a force in the opposite direction. These observations could indicate that tracking only the movement of the instruments might enable a system to estimate when an instrument approaches, touches, and retreats from the tissue but not the intensity of the resulting contact forces.

We expected the results on the unseen phantom tissue samples to be worse than those on phantom tissues that were

TABLE IV
PREDICTION OF THE PALPATION FORCES (MEANS AND STANDARD DEVIATIONS OF THE LEAVE-ONE-OUT VALIDATION) ON AN UNSEEN TISSUE SAMPLE AND PREDICTION OF THE THREE-AXIS FORCES (MEANS AND STANDARD DEVIATIONS OF THE FIVE-FOLD CROSS-VALIDATION). THE ASTERISKS INDICATE STATISTICALLY SIGNIFICANT DIFFERENCES IN PREDICTING $F_z$ COMPARED TO THE OVERALL BEST MODEL, I.E., "$F_z$ BASELINE"

| | | RMSE [N] ↓ | | | | $R^2$ ↑ | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | T270 | T250 | T150 | T120 | T270 | T250 | T150 | T120 |
| Right-hand palpation | $F_z$ baseline | 1.37 ± 0.39 | 1.17 ± 0.28 | 0.81 ± 0.48 | 0.47 ± 0.19 | 0.79 ± 0.13 | 0.76 ± 0.12 | 0.75 ± 0.09 | 0.62 ± 0.11 |
| | $F_z$ unseen tissue | 0.96 ± 0.67* | 1.59 ± 0.31* | 1.67 ± 0.46* | 0.54 ± 0.23 | 0.55 ± 0.12* | 0.63 ± 0.09* | 0.72 ± 0.11 | 0.52 ± 0.14* |
| | $F_z$ combined | 1.35 ± 0.44 | 1.06 ± 0.25 | 0.93 ± 0.53 | 0.47 ± 0.20 | 0.81 ± 0.11 | 0.80 ± 0.10 | 0.68 ± 0.10* | 0.64 ± 0.13 |
| | $F_y$ combined | 0.36 ± 0.09 | 0.36 ± 0.14 | 0.26 ± 0.10 | 0.23 ± 0.13 | 0.43 ± 0.18 | 0.43 ± 0.27 | 0.20 ± 0.18 | 0.28 ± 0.20 |
| | $F_x$ combined | 0.39 ± 0.14 | 0.32 ± 0.09 | 0.28 ± 0.09 | 0.20 ± 0.08 | 0.08 ± 0.06 | 0.16 ± 0.15 | 0.17 ± 0.13 | 0.14 ± 0.08 |
| Left-hand palpation | $F_z$ baseline | 1.77 ± 0.69 | 1.22 ± 0.32 | 0.95 ± 0.48 | 0.55 ± 0.39 | 0.79 ± 0.08 | 0.85 ± 0.06 | 0.74 ± 0.07 | 0.62 ± 0.09 |
| | $F_z$ unseen tissue | 1.61 ± 0.56 | 1.58 ± 0.50* | 0.57 ± 0.36* | 0.96 ± 0.49* | 0.62 ± 0.15* | 0.73 ± 0.10* | 0.57 ± 0.09* | 0.63 ± 0.11 |
| | $F_z$ combined | 1.47 ± 0.85* | 1.30 ± 0.35 | 0.90 ± 0.55 | 0.63 ± 0.34 | 0.75 ± 0.10 | 0.80 ± 0.08* | 0.71 ± 0.09 | 0.65 ± 0.10 |
| | $F_y$ combined | 0.37 ± 0.15 | 0.39 ± 0.13 | 0.25 ± 0.11 | 0.23 ± 0.11 | 0.23 ± 0.15 | 0.17 ± 0.11 | 0.14 ± 0.10 | 0.08 ± 0.05 |
| | $F_x$ combined | 0.43 ± 0.19 | 0.40 ± 0.14 | 0.26 ± 0.12 | 0.23 ± 0.11 | 0.40 ± 0.15 | 0.36 ± 0.18 | 0.17 ± 0.14 | 0.16 ± 0.15 |
| Bimanual palpation | $F_z$ baseline | 1.70 ± 0.62 | 1.28 ± 0.45 | 1.28 ± 0.32 | 0.54 ± 0.23 | 0.72 ± 0.07 | 0.75 ± 0.12 | 0.70 ± 0.09 | 0.61 ± 0.08 |
| | $F_z$ unseen tissue | 2.18 ± 0.89* | 1.94 ±1.33* | 0.71 ± 0.24* | 1.52 ± 0.25* | 0.56 ± 0.10* | 0.65 ± 0.10* | 0.61 ± 0.09* | 0.61 ± 0.17 |
| | $F_z$ combined | 1.65 ± 0.42 | 1.46 ± 0.67 | 0.92 ± 0.59* | 1.12 ± 0.78* | 0.74 ± 0.10 | 0.68 ± 0.10* | 0.61 ± 0.10* | 0.67 ± 0.12* |
| | $F_y$ combined | 0.50 ± 0.12 | 0.40 ± 0.25 | 0.29 ± 0.20 | 0.36 ± 0.17 | 0.50 ± 0.16 | 0.39 ± 0.20 | 0.25 ± 0.18 | 0.23 ± 0.19 |
| | $F_x$ combined | 0.49 ± 0.14 | 0.40 ± 0.21 | 0.31 ± 0.17 | 0.40 ± 0.17 | 0.35 ± 0.10 | 0.29 ± 0.21 | 0.37 ± 0.14 | 0.25 ± 0.13 |

used to generate the training data. However, although the predictions were on average less accurate, the results were promising; we even observed a few cases in which the unseen tissue performed better. This result indicates that most but not all of the extracted features are tissue-dependent. We attribute this initially unexpected result to the fact that each tissue sample has a range of Young's modulus values, and each operator moved in somewhat different ways. Thus, it was sometimes possible that the network had seen similar enough examples to achieve good predictions on an interaction with a tissue sample that was not seen in training.

The palpation task was mainly performed in the $z$-axis direction. When estimating the three-axis forces, we observed a lower RMSE in the $x$ and $y$ directions. We believe this finding is due to the low range of forces applied in these shear directions compared to those exerted along the $z$-axis. Overall, we did not find significant differences between estimating $F_z$ independently or together with $F_x$ and $F_y$, indicating that these prediction tasks may be relatively independent from one another, or that the benefits and drawbacks of the more complex model roughly cancel out.

Operators were instructed to perform three tasks: right-hand palpation, left-hand palpation, and bimanual palpation. Overall, one can observe that the right-hand palpation achieved slightly better performance compared with the left-hand palpation. The reason could be twofold. First, all four operators were right-handed and thus might have palpated in more predictable ways with this hand. Second, we used the left channel of the stereo endoscope, and since the two cameras are toed in [39], the left camera captured more visual information on the right side of the surgical field, where the right instrument usually was. Using both channels of the stereo endoscope could thus improve prediction performance. Our bimanual palpation recordings are more similar to what happens during real surgeries. This task created more-complex force signals and, as expected, achieved the worst results of the three tasks.

Inspection of the results also reveals the systematic differences in performance across the four different tissue samples. Overall, Dense-BiLSTM gave better predictions on tissues with higher stiffnesses (T270 and T250). We hypothesize that the higher forces applied to these tissue samples made prediction easier.

Based on our results and insights, we believe that future work should investigate the use of stereo images and explore different IMU positioning. In particular, when each IMU is attached to the instrument shaft (Fig. 1), the range of motion of the instrument is slightly reduced in the insertion direction. This choice did not impact our work in a dry-lab environment, but other positioning options should be investigated for applications in real surgery. Furthermore, we downsampled our input data due to the high computational cost; more efficient algorithms, such as temporal convolutional networks, could instead be investigated. While we considered only normal and shear contact-force values for the studied palpation task, future investigations of more-complex manipulation scenarios could also seek to estimate the torques applied to the tissue. Finally, our work is based on deep-learning techniques. At the current stage, we offered our hypotheses and insights about which features the network focuses on during learning. However, in the future, it would be desirable to use explainable AI to have a more profound understanding and interpretation of the system's predictions.

## V. CONCLUSION

This article presented a deep-learning model that predicts palpation forces using visual and inertial data captured through a sensor suite that can be applied to different surgical robots and traditional MIS. First, we collected a dataset that includes different operators, tissue samples, and palpation tasks. This novel dataset allowed us to create a model that could generalize to a range of palpation conditions involving contact by either instrument at arbitrary locations on the tissue surface. Second, we offered an extensive comparative analysis of several network architectures; in particular, we compared seven vision models, investigated the role of

temporal information depending on the application scenario, and investigated the role of each network input (video frames, inertial data, and Young's modulus). We then evaluated how our model generalized to unseen tissues and shear forces.

Even though we tested our method with phantom tissues, we enriched prior work [31], [33] by having multiple human operators palpating the tissues. Furthermore, operators interacted with the tissue samples with both hands, as opposed to performing only one-handed palpation [31], [33], [35]. Finally, differently from prior studies, *e.g.*, Marban et al. [31] and Jung et al. [34], we compared several state-of-the-art vision methods.

Nonetheless, several steps are still necessary to move closer to the surgical environment. In particular, future datasets of this type should include a wider range of tissue properties (Young's modulus and colors), various instruments, and a broader set of more clinically relevant palpation motions. A further step could include palpating animal tissues either ex vivo or in vivo, though ground-truth force measurements will then become more difficult to acquire. In a nutshell, our promising results showed the feasibility of detecting palpation forces with an indirect sensing method that could be applied to both RMIS and traditional MIS.

## ACKNOWLEDGMENT

The authors thank Bernard Javot for helping manufacture the experimental setup; Ravali Gourishetti, Ifat Gertler, and Mayumi Mohan for helping collect the dataset; and Joey Burns for IT support. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Maria-Paola Forte.

## REFERENCES

[1] R. F. Solodova et al., "Instrumental mechanoreceptoric palpation in gastrointestinal surgery," *Minimally Invasive Surg.*, vol. 2017, Dec. 2017, Art. no. 6481856.

[2] M. Beccani, C. Di Natali, M. E. Rentschler, and P. Valdastri, "Wireless tissue palpation: Proof of concept for a single degree of freedom," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2013, pp. 711–717.

[3] P. Puangmali, H. Liu, L. D. Seneviratne, P. Dasgupta, and K. Althoefer, "Miniature 3-axis distal force sensor for minimally invasive surgical palpation," *IEEE/ASME Trans. Mechatronics*, vol. 17, no. 4, pp. 646–656, Aug. 2011.

[4] P. Puangmali, K. Althoefer, L. D. Seneviratne, D. Murphy, and P. Dasgupta, "State-of-the-art in force and tactile sensing for minimally invasive surgery," *IEEE Sensors J.*, vol. 8, no. 4, pp. 371–381, Apr. 2008.

[5] S. Schostek, M. O. Schurr, and G. F. Buess, "Review on aspects of artificial tactile feedback in laparoscopic surgery," *Med. Eng. Phys.*, vol. 31, no. 8, pp. 887–898, 2009.

[6] A. Trejos, R. Patel, and M. Naish, "Force sensing and its application in minimally invasive surgery and therapy: A survey," *Proc. Inst. Mech. Eng. C J. Mech. Eng. Sci.*, vol. 224, no. 7, pp. 1435–1454, 2010.

[7] F. Amirabdollahian et al., "Prevalence of haptic feedback in robot-mediated surgery: A systematic review of literature," *J. Robot. Surg.*, vol. 12, no. 1, pp. 11–25, 2018.

[8] M. Hagen, J. Meehan, I. Inan, and P. Morel, "Visual clues act as a substitute for haptic feedback in robotic surgery," *Surg. Endoscopy*, vol. 22, no. 6, pp. 1505–1508, 2008.

[9] A. Abiri et al., "Artificial palpation in robotic surgery using haptic feedback," *Surg. Endoscopy*, vol. 33, no. 4, pp. 1252–1259, 2019.

[10] M. Tavakoli, R. Patel, and M. Moallem, "Haptic interaction in robot-assisted endoscopic surgery: A sensorized end-effector," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 1, no. 2, pp. 53–63, 2005.

[11] C. Lee et al., "A grip force model for the *davinci* end-effector to predict a compensation force," *Med. Biol. Eng. Comput.*, vol. 53, no. 3, pp. 253–261, 2015.

[12] A. K. Golahmadi, D. Z. Khan, G. P. Mylonas, and H. J. Marcus, "Tool-tissue forces in surgery: A systematic review," *Ann. Med. Surg.*, vol. 65, Mar. 2021, Art. no. 102268.

[13] C. R. Wagner, N. Stylopoulos, P. G. Jackson, and R. D. Howe, "The benefit of force feedback in surgery: Examination of blunt dissection," *Presence Teleoper. Virtual Environ.*, vol. 16, no. 3, pp. 252–262, 2007.

[14] C. Pacchierotti, D. Prattichizzo, and K. J. Kuchenbecker, "Cutaneous feedback of fingertip deformation and vibration for palpation in robotic surgery," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 2, pp. 278–287, Feb. 2016.

[15] L. Meli, C. Pacchierotti, and D. Prattichizzo, "Experimental evaluation of magnified haptic feedback for robot-assisted needle insertion and palpation," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 13, no. 4, 2017, Art. no. e1809.

[16] S. Machaca, E. Cao, A. Chi, G. Adrales, K. J. Kuchenbecker, and J. D. Brown, "Wrist-squeezing force feedback improves accuracy and speed in robotic surgery training," in *Proc. 9th IEEE RAS/EMBS Int. Conf. Biomed. Robot. Biomechatron. (BioRob)*, 2022, pp. 1–8.

[17] A. M. Okamura, "Haptic feedback in robot-assisted minimally invasive surgery," *Current Opin. Urol.*, vol. 19, no. 1, p. 102, 2009.

[18] Z. Chua, A. M. Jarc, S. M. Wren, I. Nisky, and A. M. Okamura, "Task dynamics of prior training influence visual force estimation ability during teleoperation," *IEEE Trans. Med. Robot. Bionics*, vol. 2, no. 4, pp. 586–597, Nov. 2020.

[19] J. D. Brown, C. E. O'Brien, S. C. Leung, K. R. Dumon, D. I. Lee, and K. J. Kuchenbecker, "Using contact forces and robot arm accelerations to automatically rate surgeon skill at peg transfer," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 9, pp. 2263–2275, Sep. 2017.

[20] A. L. Trejos, R. V. Patel, R. A. Malthaner, and C. M. Schlachta, "Development of force-based metrics for skills assessment in minimally invasive surgery," *Surg. Endoscopy*, vol. 28, no. 7, pp. 2106–2119, 2014.

[21] C. Richards, J. Rosen, B. Hannaford, C. Pellegrini, and M. Sinanan, "Skills evaluation in minimally invasive surgery using force/torque signatures," *Surg. Endoscopy*, vol. 14, no. 9, pp. 791–798, 2000.

[22] R. V. Patel, S. F. Atashzar, and M. Tavakoli, "Haptic feedback and force-based teleoperation in surgical robotics," *Proc. IEEE*, vol. 110, no. 7, pp. 1012–1027, 2022.

[23] M. Mahvash et al., "Force-feedback surgical teleoperator: Controller design and palpation experiments," in *Proc. IEEE Symp. Haptic Interfaces Virtual Environ. Teleoper. Syst.*, 2008, pp. 465–471.

[24] H. Sang, J. Yun, R. Monfaredi, E. Wilson, H. Fooladi, and K. Cleary, "External force estimation and implementation in robotically assisted minimally invasive surgery," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 13, no. 2, 2017, Art. no. e1824.

[25] S. Abeywardena et al., "Estimation of tool-tissue forces in robot-assisted minimally invasive surgery using neural networks," *Front. Robot. AI*, vol. 6, p. 56, Jul. 2019.

[26] B. Zhao and C. A. Nelson, "A sensorless force-feedback system for robot-assisted laparoscopic surgery," *Comput. Assist. Surg.*, vol. 24, no. s1, pp. 36–43, 2019.

[27] N. Tran, J. Y. Wu, A. Deguet, and P. Kazanzides, "A deep learning approach to intrinsic force sensing on the da vinci surgical robot," in *Proc. IEEE Int. Conf. Robot. Comput.*, 2020, pp. 25–32.

[28] W.-C. Lin and K.-T. Song, "Instrument contact force estimation using endoscopic image sequence and 3D reconstruction model," in *Proc. IEEE Int. Conf. Adv. Robot. Intell. Syst. (ARIS)*, 2016, pp. 1–6.

[29] N. Haouchine, W. Kuang, S. Cotin, and M. Yip, "Vision-based force feedback estimation for robot-assisted surgery using instrument-constrained biomechanical three-dimensional maps," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2160–2165, Jul. 2018.

[30] V. Kakani, X. Cui, M. Ma, and H. Kim, "Vision-based tactile sensor mechanism for the estimation of contact position and force distribution using deep learning," *Sensors*, vol. 21, no. 5, p. 1920, 2021.

[31] A. Marban, V. Srinivasan, W. Samek, J. Fernández, and A. Casals, "A recurrent convolutional neural network approach for sensorless force estimation in robotic surgery," *Biomed. Signal Process. Control*, vol. 50, pp. 134–150, Apr. 2019.

[32] H. Shin, H. Cho, D. Kim, D.-K. Ko, S.-C. Lim, and W. Hwang, "Sequential image-based attention network for inferring force estimation without haptic sensor," *IEEE Access*, vol. 7, pp. 150237–150246, 2019.

[33] Z. Chua, A. M. Jarc, and A. M. Okamura, "Toward force estimation in robot-assisted surgery using deep learning with vision and robot state," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021, pp. 12335–12341.

[34] W.-J. Jung, K.-S. Kwak, and S.-C. Lim, "Vision-based suture tensile force estimation in robotic surgery," *Sensors*, vol. 21, no. 1, p. 110, 2020.

[35] A. I. Aviles, A. Marban, P. Sobrevilla, J. Fernandez, and A. Casals, "A recurrent neural network approach for 3D vision-based force estimation," in *Proc. IEEE Int. Conf. Image Process. Theory Tools Appl. (IPTA)*, 2014, pp. 1–6.

[36] C. Gao, X. Liu, M. Peven, M. Unberath, and A. Reiter, "Learning to see forces: Surgical force prediction with RGB-point cloud temporal convolutional networks," in *Proc. OR Context-Aware Oper. Theaters Comput. Assist. Robot. Endoscopy Clin. Image Based Procedures Skin Image Anal.*, 2018, pp. 118–127.

[37] P. E. Edwards, E. Colleoni, A. Sridhar, J. D. Kelly, and D. Stoyanov, "Visual kinematic force estimation in robot-assisted surgery—Application to knot tying," *Comput. Methods Biomechan. Biomed. Eng. Imag. Visual.*, vol. 9, no. 4, pp. 414–420, 2021.

[38] L. Chen, W. Tang, N. W. John, T. R. Wan, and J. J. Zhang, "SLAM-based dense surface reconstruction in monocular minimally invasive surgery and its application to augmented reality," *Comput. Methods Programs Biomed.*, vol. 158, pp. 135–146, May 2018.

[39] M.-P. Forte, R. Gourishetti, B. Javot, T. Engler, E. D. Gomez, and K. J. Kuchenbecker, "Design of interactive augmented reality functions for robotic surgery and evaluation in dry-lab lymphadenectomy," *Int. J. Med. Robot. Comput. Assist. Surg.*, vol. 18, no. 2, 2022, Art. no. e2351.

[40] O. Rouvière et al., "Stiffness of benign and malignant prostate tissue measured by shear-wave elastography: A preliminary study," *Eur. Radiol.*, vol. 27, no. 5, pp. 1858–1866, 2017.

[41] C. F. Guimarães, L. Gasperini, A. P. Marques, and R. L. Reis, "The stiffness of living tissues and its implications for tissue engineering," *Nat. Rev. Mater.*, vol. 5, no. 5, pp. 351–370, 2020.

[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Comput. Vis. (ECCV) 14th Eur. Conf.*, Amsterdam, The Netherlands, Oct. 2016, pp. 630–645.

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 2818–2826.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4278–4284.

[46] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," 2016, *arXiv:1611.01578*.

[47] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 8697–8710.

[48] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[49] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.

[50] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[51] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[52] A. Talasaz, *Haptics-Enabled Teleoperation for Robotics-Assisted Minimally Invasive Surgery*. Univ. Western Ontario, London, ON, Canada, 2012.

[53] A. C. Cameron and F. A. Windmeijer, "An R-squared measure of goodness of fit for some common nonlinear regression models," *J. Econ.*, vol. 77, no. 2, pp. 329–342, 1997.