IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Image-to-Force Estimation for Soft Tissue Interaction in Robotic-Assisted Surgery Using Structured Light

Jiayin Wang, Mingfeng Yao, Yanran Wei, Xiaoyu Guo, *Member, IEEE*, Ayong Zheng, and Weidong Zhao

***Abstract*—For Minimally Invasive Surgical (MIS) robots, accurate haptic interaction force feedback is essential for ensuring the safety of interacting with soft tissue. However, the majority of existing MIS robotic systems cannot facilitate direct measurement of the interaction force with hardware sensors due to space limitations. This letter introduces an effective vision-based scheme that utilizes a One-Shot structured light projection with a designed pattern on soft tissue coupled with haptic information processing through a trained image-to-force neural network. The images captured from the endoscopic stereo camera are analyzed to reconstruct high-resolution 3D point clouds for soft tissue deformation. The proposed methodology involves a modified PointNet-based force estimation method, which has demonstrated proficiency in accurately representing the intricate mechanical properties of soft tissue. To validate the efficacy of the proposed methodology, numerical force interaction experiments were conducted on three silicon materials with varying stiffness levels. The experimental results substantiate the efficacy of the proposed methodology.**

***Index Terms*—Force estimation, haptics, surgical robots, vision-based measurements, deformable objects.**

I. Introduction

MINIMALLY Invasive Surgery (MIS) robotic systems represent a formidable frontier in contemporary medicine, offering reduced tissue trauma and improved operational safety [1]. However, these systems often prohibit direct haptic sensing between the surgeon and soft tissues, thereby increasing the risk associated with real-time force interactions. Therefore, haptic force sensing in such scenarios has become an essential requirement [2], [3].

The primary methods for haptic force sensing include [4]: additive force sensor-based measurement and sensorless force estimation. In [5], [6], [7], force sensors are mounted on the end-effectors of surgical robots to directly measure interaction forces. Alternatively, in [8], sensors are affixed to the surface of the tissue itself. While these methods provide intuitive operation and high measurement accuracy, their clinical application remains hindered by challenges such as cost constraints, limited installation space, and inadequate resistance to high temperatures and corrosion [9].

To address these limitations, sensorless force estimation methods have been developed. In previous work [10], the dynamic information of the robot was utilized to estimate external interaction forces. In [11], the mechanical properties of the deformable environment were integrated with the robot dynamics to improve the accuracy of the estimation. Although these dynamic models-based robotic methods are both effective and non-reliant on additive sensors, they inherently rely on precise modeling of the dynamics of the surgical robot. Another indirect force estimation approach is vision-based force estimation (VBFE) methods which refers the force from the model of deformable objects and the displacement of the surface. In [12], a method to predict surface force and friction coefficients by embedding marked elastomers in silicone membranes was proposed. However, model-based VBFE methods are not appropriate for real-time applications due to the requirement of inaccessible a priori knowledge of the reference shape and the mechanics information [13], [14].

Utilizing the versatility of deep learning methods to model complex deformation, learning-based VBFE methods are developed [15], [16], [17]. In [18], a force estimation method is proposed via time-delayed neural networks and Gaussian processes based on dynamic vision sensors. In [19], the surface deformation is modeled using cubic B-splines combined with an energy minimization strategy, while the visual-geometric-force relationship is learned through a recurring neural network (RNN). However, the aforementioned methods are primarily limited to push actions, overlooking the more complex force estimation required for pull (traction) tasks, which represent a particularly challenging scenario in MIS systems.Tissues under tension exhibit limited deformation texture, hindering traditional image-based methods from capturing sufficient surface features for accurate force prediction.This study aims to propose a method that effectively predicts both tensile and compressive forces on tissues during surgery, capturing the overall force distribution rather than localized forces.

Towards the goal of practical vision-based force estimation in MIS systems, two major technical challenges arise: 1) How to establish a vision-based force estimation framework suitable for scenarios where a surgical robot interacts with texture-deficient soft tissues, particularly during pulling tasks, which are both more challenging and common in surgical procedures (e.g., suturing, cutting); 2) How to model the complex displacement-force relationship and train it using a high-quality custom dataset

Received 8 January 2025; accepted 19 May 2025. Date of publication 13 June 2025; date of current version 23 June 2025. This article was recommended for publication by Associate Editor M. Bianchi and Editor K.-U. Kyung upon evaluation of the reviewers' comments. *(Corresponding authors: Yanran Wei; Weidong Zhao.)*

Jiayin Wang is with the School of Computer Science and Technology, Tongji University, Shanghai 200092, ChinaMicroPort MedBot (Group) Company Ltd., Shanghai 201203, China (e-mail: jaryerwang@tongji.edu.cn).

Mingfeng Yao and Ayong Zheng are with the MicroPort MedBot (Group) Company Ltd., Shanghai 201203, China (e-mail: MingFeng.Yao@microport.com; ayzheng@microport.com).

Yanran Wei is with the College of Engineering, Peking University, Beijing 100871, China (e-mail: yanranwei@pku.edu.cn).

Xiaoyu Guo is with the Department of Biomedical Engineering, City University of Hong Kong, Kowloon 999077 Hong Kong (e-mail: xiaoyguo@cityu.edu.hk).

Weidong Zhao is with the School of Computer Science and Technology, Tongji University, Shanghai 200092, China (e-mail: wd@tongji.edu.cn).

Digital Object Identifier 10.1109/LRA.2025.3579640

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

specifically designed for this task, acknowledging the well-known difficulty in collecting datasets for physical interaction in medical contexts.

During surgical procedures, information on the forces exerted on tissues is essential for protecting patients. However, due to constraints in assembly space and high costs, most current surgical robots are not equipped with high-precision force sensors. Currently, numerous methods can estimate the interaction force between robots and tissues using image and joint information. These methods provide interaction force data during surgical actions such as grasping and pressing, aiding in the protection of the target tissue [20], [21], [22], [23]. However, most of these studies focus on predicting forces when tissues are under compression, with little attention given to predicting tensile forces, despite tissues frequently experiencing tension during surgery. Meanwhile, some vision-based approaches can effectively predict tensile forces on tissues [24], but these methods rely on passive vision reconstruction, whose accuracy is limited by the inherently low-texture characteristics of soft tissue surfaces. As a result, these methods typically estimate forces only at instrument contact points—where localized deformation provides visual features—while failing to capture the global deformation of smooth soft tissues, thereby limiting their applicability to real-world MIS scenarios. In addition, passive vision-based approaches are highly sensitive to camera viewpoint changes, with estimation errors increasing by up to 20% when the viewing angle shifts [24]. Furthermore, methods such as [21] require precise camera calibration and are sensitive to system variations. These limitations hinder their application to laparoscopic systems using stereo endoscope configurations. To address the limitations of existing methods in robot-assisted laparoscopic surgery, we propose a novel VBFE framework that actively reconstructs the 3D surface of soft tissues under texture-deficiency and frame-rate constraints. The main contributions of this work are summarized as follows:

1) A custom-designed one-shot fringe pattern based on a DeBruijn sequence encodes spatially unique structured light information, enabling dense and reliable 3D point cloud generation from a single stereo RGB frame.
2) Then, a modified PointNet-based regression network is developed to learn the mapping from deformation to force. Architectural improvements on the network including the Exponential Linear Unit (ELU) activation function, a regression-oriented output layer, and the Nadam optimizer—enhance training efficiency and predictive accuracy.
3) Finally, a custom dataset of paired 3D point clouds and force measurements is constructed for training, and extensive experiments on a commercial laparoscopic MIS robot platform validate the effectiveness of the proposed framework across soft tissues with varying stiffness levels. Unlike existing temporal neural networks (TNN)-based methods [23], the proposed method eliminates the need for temporal information or high-speed sensors and enables dense point cloud recovery from a single image.

The outline of the letter is as follows. Section II outlines the necessary preliminaries and defines the problem. Section III details the proposed VBFE framework. Section IV presents experimental results from a real-world force interaction task conducted with the Toumai laparoscopic MIS robots. Finally, Section V concludes this letter.

II. Preliminaries

The VBFE methods typically consist of three coupled modules: force-deformation modeling, 3D visual reconstruction, and model learning. Each component involves a variety of existing algorithms. This section introduces the general VBFE framework and provides the relevant background and foundational algorithms for each component, along with a summary of the limitations of current methods in laparoscopic MIS platforms, as discussed in the subsequent problem formulation.

A. Force Model for Deformable Tissue

Classical constitutive models are used to describe the deformation behavior of different materials under external interaction forces, as follows

$$\sigma = f(\epsilon), \tag{1}$$

where σ and ϵ represent the stress and strain of the deformable object, respectively. Constitutive models based on the mechanical assumptions of the materials are used to infer the interaction force estimate including the elastic models, hyperelastic models, and viscoelastic models.

A kind of accurate model for capturing the time-varying behavior of the soft tissues is viscoelastic models including the Maxwell model and the Kelvin-Voigt model. The two models can effectively describe the stress relaxation behavior and the creep behavior of viscoelastic materials, respectively. These two models are formulated as Eq. (2) and Eq. (3).

$$\frac{d\epsilon}{dt} = \frac{1}{E} \times \frac{d\sigma}{dt} + \frac{\sigma}{\eta}, \tag{2}$$

$$\sigma(t) = E\epsilon + \eta\frac{d\epsilon}{dt}, \tag{3}$$

where η and E are the viscosity coefficient and elastic modulus, respectively. In surgical simulations, Finite Element Analysis (FEA) tools formulate constitutive models to analyze soft tissue deformations by decomposing displacement fields. However, FEA's complexity and high computational cost hinder its use in real-time VBFE for surgical robotics. Moreover, model-based force estimation relies heavily on prior knowledge of mechanical parameters, which are often difficult to acquire accurately.

B. Vision Reconstruction

The methods presented in this letter for VBFE rely on image-based vision reconstruction. Dense point clouds, generated via active or passive sensing, enable high-resolution 3D reconstruction, capturing detailed shapes of deformable objects. In most laparoscopic surgeries, endoscopes use standard stereo cameras, such as the commercial DFVision endoscope featured in this study. Different from traditional structured light methods, the proposed approach does not require utilizing the matching relationship between the projector and the camera; instead, they capture images of the same scene from two different viewpoints (i.e., left and right) using a stereo camera configuration [25]. Depth information is then obtained by identifying corresponding feature points in both views and applying triangulation. This method is well-suited for scenes with rich texture; however, in texture-deficient environments—such as smooth soft tissue surfaces—depth estimation may be unreliable due to challenges in establishing accurate point correspondences.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Structured light-based 3D reconstruction methods are robust active techniques for recovering object surfaces [26]. By projecting predefined light patterns, these methods generate artificial features that enhance surface texture. Structured light systems offer high reconstruction accuracy and, with appropriate algorithmic optimization, can support real-time depth estimation, particularly in short-range and low-velocity applications.

C. Problem Formulation

In laparoscopic surgery, the task of applying a unidirectional, low-speed pulling force is challenging and should be prioritized to minimize the risk of unexpected soft tissue damage [27]. For simplicity, and without loss of generalization, this scenario makes two fundamental assumptions about the deformable object: 1) its mechanical properties are isotropic, and 2) its geometric properties are uniform, disregarding minor geometric irregularities.

The workflow of classical methods of VBFE for deformable objects using binocular stereo vision is as follows: Initially, a disparity map is generated through stereo matching. The disparity map is used to reconstruct 3D point clouds of the object's surface, followed by force estimation via FEA based on known material properties. However, these methods fall short due to: (1) the nonlinear mechanics of soft tissue, making accurate modeling difficult; (2) texture-deficient surfaces, which degrade reconstruction accuracy and lead to force estimation errors; and (3) insufficient real-time performance for surgical applications.

Accurate and efficient force estimation requires One-Shot pattern projection for dense, absolute, and high-precision reconstruction. Additionally, a deep learning-based displacement–force model is employed as an alternative to Eq. (1), directly processing point cloud data without relying on structured formats like voxels or meshes.

III. The Presented Scheme

In this section, the proposed VBFE scheme is detailed. In Section III-A, we present a stereo vision 3D reconstruction method for deformable tissue in MIS using a designed One-Shot absolute structured light projection. In Section III-B, a modified PoinNet-based force estimation network is presented.

A. 3D Reconstruction With One-Shot Structured Light

The 3D reconstruction process in this scheme involves designing a specialized One-Shot absolute structured light pattern, performing stereo vision matching using the semi-global block matching (SGBM) algorithm with pattern projection, and generating a real-time dense 3D point cloud of the object surface.

1) Structured Light Pattern Creation: To achieve time-efficient reconstruction, a One-Shot absolute pattern is employed for structured light encoding. As shown in Fig. 1, the pattern consists of a set of colored sinusoidal fringes generated in Hue, Saturation, and Value (HSV) space, with the H channel encoded using a DeBruijn sequence. A DeBruijn sequence is a circular sequence in which each element belongs to an alphabet of n symbols. This sequence can be directly constructed from the Hamiltonian or Eulerian paths of an n-dimensional DeBruijn graph [28]. A key property is that any substring of length m appears exactly once, making it ideal for generating the colored fringe sequence in the H channel. This ensures each projected unit corresponds to a unique decoding, improving stereo matching accuracy.

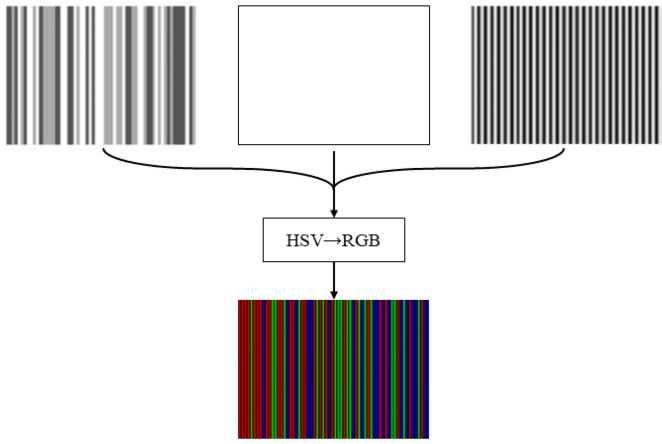
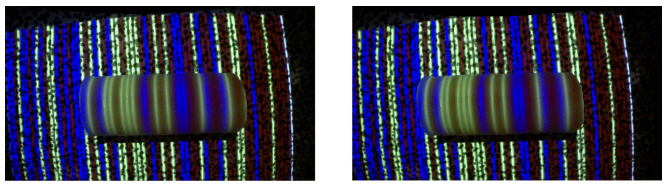


Fig. 1. Pattern creation. H channel pattern generated by De Bruijn sequence (Top-left); S channel with constant maxima (Top-middle); Sinusoidal intensity pattern in V channel (Top-right); The result RGB pattern.



(a) (b)

Fig. 2. Images of structured light projection captured by the stereo camera system: (a) left camera image, (b) right camera image.

In our setup, the alphabet for encoding includes $n = 3$ colors: red, green, and blue. A fringe sequence of length 64 is selected based on projector resolution, and $m = 4$, generating a DeBruijn sequence of length 81 to satisfy the condition $n^m > 64$. Saturation is fixed at 1 (maximum) for all pixels. Vertically, each colored fringe has a sinusoidal light intensity, encoded in the V channel as:

$$I(i) = 0.5 + 0.5 \cdot \cos\left(2\pi f \cdot i\right),\ i = 1 \ldots N, \tag{4}$$

where i denotes the column index, $N = 64$ is the maximum horizontal resolution of the projector, and f is the frequency given by $f = \frac{64}{N}$.

The HSV pattern is converted to RGB and projected onto a silicone model of a humanoid intestine. Fig. 2, shows the stereo camera's captured images. This structured light is then used for stereo matching.

2) Pattern Recovery and Stereo Matching: Before decoding, object base color is removed and pixel correction is applied using the Caspi model. To improve stereo matching, a two-step approach is used: initial matching followed by refined verification.

For initial matching, the SGBM algorithm generates candidate pairs and their encoded data. These candidate match pairs typically contain salient visual features and are referred to as points of interest (POIs) in this paper. In the refinement step, DeBruijn subsequences of each POI pair are compared between stereo views. Mismatches trigger a search in the SGBM

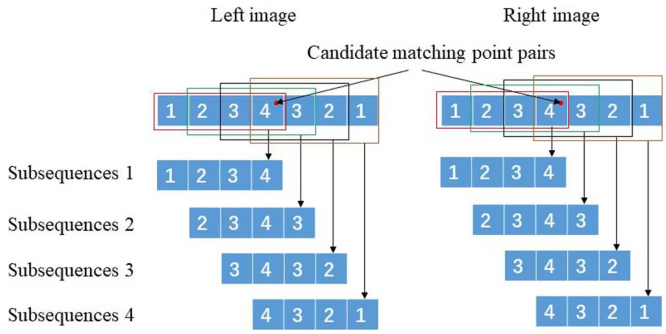


Fig. 3. DeBruijn analysis of refined matching verification.

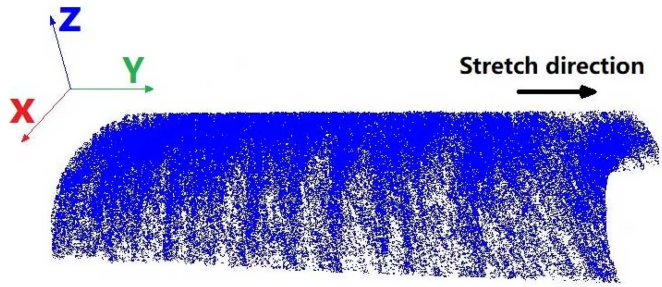


Fig. 4. The generated 3D point cloud of a deformable silicone object's surface.

candidate pool and neighborhood to locate matching points with identical encoding. If the DeBruijn sequence decoding fails to validate the candidate matching pair, the corresponding pair is discarded to ensure reliable disparity estimation.

Prior to matching, the horizontal Sobel operator is applied to both left and right images to enhance vertical edge features by computing the horizontal gradient:

$$Gu(u,v) = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} * I(u,v), \tag{5}$$

where $Gu(p)$ and $I(p)$ denote the horizontal gradient and intensity at pixel (p), respectively; $*$ denotes the convolution operation.

The pixel-wise matching cost $C(p,d)$ is computed by combining the Birchfield-Tomasi(BT) cost on the original image and that on the prepocessed Sobel image:

$$C(p,d) = C^{BT}_{original}(p,d) + C^{BT}_{sobel}(p,d), \tag{6}$$

where $C^{BT}_{original}(p,d)$ and $C^{BT}_{sobel}(p,d)$ denote the BT cost values in the original and Sobel-filtered images, respectively [25]. d is the disparity value. The features are stored in a disparity space image (DSI) matrix for subsequent matching.

Each candidate matching pixel pair undergoes DeBruijn analysis. As shown in Fig. 3, a sliding window of length four is used to sequentially verify the decoded DeBruijn subsequences of the matching point pairs. If the subsequences differ, the candidate matching pair is discarded.

The SGBM algorithm employed for stereo matching computes depth maps from stereo images by aggregating matching costs along multiple paths. This strategy achieves accurate disparity estimation with computational efficiency, making it suitable for real-time applications [25]. To ensure that the cost values accurately reflect the correlation between pixels, the SGBM algorithm employs path integration of the matching cost in stereo vision across multiple directions. Let L_r represent a path traversed in the direction r. The cost $L_r(p,d)$ for pixel p at disparity d is recursively defined as:

$$L_r(p,d) = C(p,d) + \min\left(L_1, L_2, L_3, L_4\right) - L_5, \tag{7}$$

where L_1, L_2, L_3, L_4, and L_5 are the path costs corresponding to the neighboring pixel, defined as:

$$L_1 = L_r\left(p-r, d-1\right) + P_1, \quad L_2 = L_r\left(p-r, d\right),$$
$$L_3 = L_r\left(p-r, d+1\right) + P_1,$$
$$L_4 = \min\left(L_r\left(p-r, i\right)\right) + P_2, \quad i = d_{\min} \ldots d_{\max},$$
$$L_5 = \min\left(L_r\left(p-r, k\right)\right), \quad k = d_{\min} \ldots d_{\max}.$$

Here, P_1 and P_2 are smoothness penalty coefficients. The term L_5, representing the minimum path cost of the previous pixel, is introduced to prevent excessively large values in the calculations and ensure numerical stability.

The aggregated matching cost is defined as:

$$S(p,d) = \sum_r L_r(p,d), \tag{8}$$

where $S(p,d)$ represents the total matching cost aggregated across all paths r. The final disparity value $D(p)$ is then determined by minimizing the aggregated cost:

$$D(p) = \arg\min_d S(p,d). \tag{9}$$

Conventional passive stereo matching techniques often struggle with weakly textured objects, producing disparity maps of suboptimal quality [25]. In contrast, integrating active structured light projection into stereo matching significantly enriches the texture of object surfaces and leverages structured light encoding information to yield highly accurate disparity maps.

3) Point Cloud Generation: Based on the disparity map, the 3D point cloud of the object's surface can be constructed, as illustrated in Fig. 4. Specifically, for each pixel p with coordinates (u,v) in the field of view, the 3D coordinates (x_i, y_i, z_i) can be computed using the focal length f, the optical center coordinates (c_x, c_y), the baseline distance B, and the disparity $D(p)$ from the left and right cameras as follows:

$$z_i = \frac{f \cdot B}{D(p)}, \; x_i = \frac{(u-c_x)\cdot z_i}{f}, \; y_i = \frac{(v-c_y)\cdot z_i}{f}. \tag{10}$$

By iterating over all pixels in the disparity map $D(u,v)$, the 3D spatial point cloud of the surface is generated.

A labeled dataset is thus constructed, where each data pair consists of the reconstructed 3D point cloud (obtained from each RGB frame) and the corresponding interaction force measurement (obtained using the force-sensing trocar). This dataset is used in the subsequent subsection to train the force estimation model.

B. Modified PointNet-Based Force Estimation

The PointNet framework [29] demonstrates exceptional capability in processing unstructured point cloud data. By independently applying multilayer perceptrons (MLPs) to individual points, it efficiently extracts local features, while leveraging

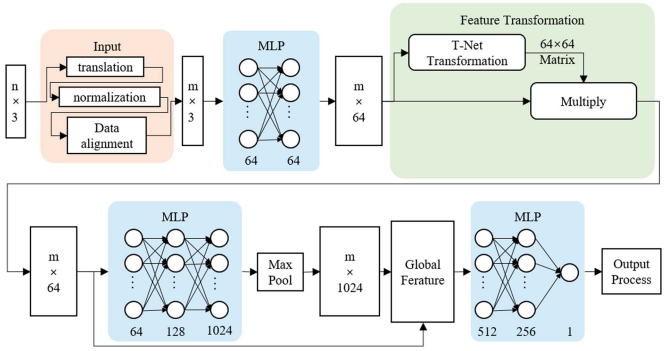


Fig. 5. The modified PointNet network architecture.

global max pooling to capture global deformation characteristics. This design ensures invariance to the order of the input points. In particular, PointNet boasts an architecturally simple, computationally efficient, and highly flexible structure. However, PointNet was originally designed for classification tasks, rendering it inapplicable for force estimation, a fundamentally regression-oriented problem. Surgical force estimation, in particular, demands not only exceptional accuracy but also real-time performance to deliver high-quality haptic feedback to the operating surgeon. Additionally, the training and deployment of accurate force estimation models using the high-density point cloud data generated in Section III-A can be computationally intensive and time-consuming.

To overcome these challenges, targeted modifications and optimizations are implemented in both the network architecture and the training algorithm, ensuring the framework aligns with the stringent requirements of surgical force estimation tasks.

1) Input Preprocessing and Network Modifications: To facilitate consistent input representation for the network, a preprocessing step is introduced for the point-cloud data. This step involves normalizing each point cloud individually by translating its centroid to the origin of the coordinate system and scaling the point cloud to fit within a unit sphere.

The centroid coordinates (x_c, y_c, z_c) of a point cloud are computed as follows:

$$x_c = \frac{\sum_{i=1}^{n} m_i x_i}{\sum_{i=1}^{n} m_i}, y_c = \frac{\sum_{i=1}^{n} m_i y_i}{\sum_{i=1}^{n} m_i}, z_c = \frac{\sum_{i=1}^{n} m_i z_i}{\sum_{i=1}^{n} m_i}, \tag{11}$$

where m_i represents the weight of the i-th point in the cloud, and n denotes the total number of points in the cloud.

By translating and normalizing, the coordinates of the point cloud are transformed as follows:

$$\tilde{x}_i = \frac{x_i - x_c}{D_{\max}}, \tilde{y}_i = \frac{y_i - y_c}{D_{\max}}, \tilde{z}_i = \frac{z_i - z_c}{D_{\max}}, \tag{12}$$

where $D_{\max}$ represents the maximum bounding envelope distance of the point cloud, which is the largest Euclidean distance between any two points within the point cloud.

The architecture of the proposed force estimation network is shown in Fig. 5, comprising an input layer, convolutional layers, pooling layers, a feature aggregation layer, activation layers, MLPs, dropout layers, and an output layer. The output layer is modified to a fully connected network layer to accommodate the continuous regression problem.

The original PointNet network employs ReLU as the activation function, which outputs zero for negative inputs, potentially leading to inactive neurons ("dying ReLU"). This limits the network's ability to model complex nonlinearities, critical for capturing the mechanical properties of soft tissues.

To address this, the ELU is utilized, defined as:

$$\mathrm{ELU}(\epsilon) = \begin{cases} \epsilon, & \epsilon > 0 \\ \alpha(\exp(\epsilon) - 1), & \epsilon \leq 0 \end{cases}$$

where the hyperparameter $\alpha > 0$ for the ELU activation function is set to 1.0, following the guideline in reference [30]. ELU is continuous and differentiable, mitigating the vanishing gradient problem. For $\epsilon > 0$, it resembles ReLU, and for $\epsilon \leq 0$, it behaves like sigmoid/tanh, effectively combining their strengths. This adaptation improves the network's capacity to capture soft tissue mechanics, enhancing its suitability for force estimation tasks.

2) Optimization and Adaptation of the Training Algorithm: The proposed model is trained using the Nadam optimizer, an enhancement of the Adam optimizer. While Adam's adaptive learning rate adjustment is effective in many scenarios, it may struggle with slow convergence or fail to precisely locate the optimal point. Nadam addresses these issues by integrating Nesterov momentum, which combines Adam's adaptive learning rates with a lookahead mechanism to accelerate convergence and improve accuracy.

The original Adam updates for momentum and velocity are defined as:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t, \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2, \end{aligned} \tag{13}$$

where g_t is the gradient at time t, β_1 and β_2 are the decay rates, and m_t and v_t represent the momentum and velocity updates, respectively.

In Nadam, Nesterov momentum introduces a refined momentum update, given by:

$$\tilde{m}_t = \beta_1 m_t + (1 - \beta_1) g_t.$$

The bias-corrected estimates of momentum and velocity are computed as:

$$\hat{m}_t = \frac{\tilde{m}_t}{1 - \beta_1^t}, \ \hat{v}_t = \frac{v_t}{1 - \beta_2^t}, \tag{14}$$

where β_1^t and β_2^t represent β_1 and β_2 to the power of t.

The parameter update rule for Nadam is:

$$\theta_{t+1} = \theta_t - \eta \cdot \frac{1}{\sqrt{\hat{v}_t} + \varepsilon} \left(\beta_1 \tilde{m}_t + \frac{(1 - \beta_1) g_t}{1 - \beta_1^t} \right),$$

where η is the learning rate, and ε is a small constant to ensure numerical stability. By leveraging the benefits of both adaptive learning rates and Nesterov momentum, Nadam enhances optimization efficiency and achieves better performance for the force estimation network. For this regression task, the root mean squared error (RMSE) is employed as the loss function.

IV. Experimental Validation

In this section, to validate the effectiveness of the proposed VBFE scheme, traction tests were performed on a platform developed consisting of silicone intestinal models with three different stiffness levels and a commercial surgical robot.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

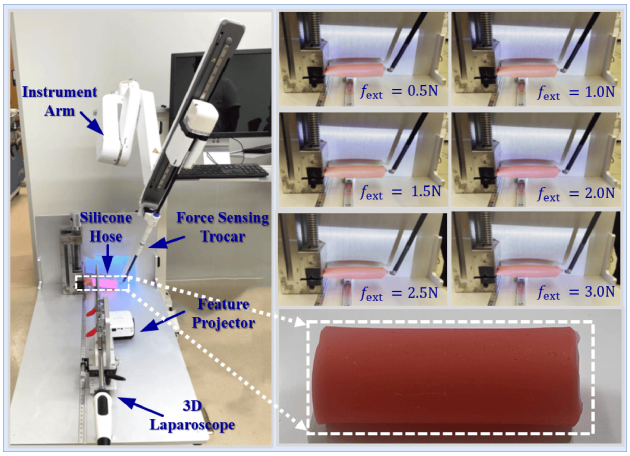


Fig. 6. The experimental setup. The experimental platform is shown on the left. The upper right illustrates snapshots of the pull-hold process under varying external forces (f_{ext}), ranging from 0.5 N to 3.0 N. The silicone hose, with calibrated stiffness, is displayed in the lower right.

A. Experimental Setup

The experimental platform was constructed using the Toumai research kits provided by Shanghai Microport Medbot (Group) Co., Ltd. [31]. As shown in Fig. 6, the platform includes a high-precision force sensing trocar (Model TRF85D) to measure the true interaction force and a surgical instrument arm (Model M0000339) equipped with a 3D electronic endoscope (Model EL824). The force-sensing trocar features a measurement range of ± 5 N with a precision of 0.1 N. The robotic arm is equipped with grasping forceps (Model IN803 A) that execute retraction and clamping operations in Cartesian space. These forceps are used to secure clamp one end of the silicone tube and perform the retraction operation. During the retraction process, the force sensor measures the interaction force applied to the silicone tube in real time, while the 3D electronic endoscope captures deformation images. These images are subsequently processed to generate a 3D point cloud for further analysis. The platform is equipped with a projector (Model L-mix) that projects structured light patterns onto the surface of the object. The entire experimental platform is designed to stably simulate the clinical operation environment, ensuring the reliability and repeatability of the experimental data. In this experiment, silicone tubes were custom-fabricated in the laboratory to simulate soft tissues by varying the base-to-curing-agent mixing ratios of liquid silicone. To ensure the accuracy of the stiffness characterization, each sample underwent multiple calibration procedures using a universal mechanical testing system. As a result, three representative samples with calibrated stiffness values of 40 N/m, 80 N/m, and 200 N/m were obtained, corresponding to soft, medium, and hard tissue analogs, respectively.

Remark 1: In most cases, it has been observed that stiffness variations in abdominal soft tissues due to individual factors such as age or gender are relatively limited [32], [33], [34]. However, in extreme cases in which these factors significantly impact the stiffness of tissues, the proposed method would require re-calibration with new data. An interesting future research topic is to simultaneously estimate the force and stiffness online, which can be achieved via expectation-maximization and could circumvent the aforementioned issue.

TABLE I
FORCE ESTIMATION RESULTS WITH THREE STIFFNESS

Stiffness of the silicone tube	MAE(σ)	RMSE(σ)
40N/m	0.4055 (0.2558)	0.5498 (0.3531)
80N/m	0.4748 (0.3593)	0.5953 (0.5047)
200N/m	0.8044 (0.6108)	1.01 (1.4735)

1) Force Model Training: A custom dataset has been collected using the Toumai commercial laparoscopic surgical robot affixed with a force-sensing trocar, along with the DFVision medical stereo endoscope and a silicone hose. This dataset has been open source on GitHub and is accessible at: https://github.com/CrisYaoMF/Force-Estimation-for-Soft-Tissue. As the manipulator stretches the silicone tube, a force-sensing trocar measures the applied force in real time, while an endoscope captures deformation images. The 3D laparoscope generates 3D point cloud data at 30 Hz. These paired data train the proposed interaction force estimation network, implemented in Keras with the Nadam optimizer to boost training speed and convergence. Throughout the process, the network weights were iteratively adjusted to minimize the RMSE between the force estimates and the actual measurements recorded by the force-sensing trocar, ensuring accurate model performance. In this study, the dataset for network training was split into a training set and a validation set at a 7:3 ratio. Training was terminated when the loss function value fell below 0.2 or after 200 epochs. During the training process, we observed signs of overfitting, which may partially explain the suboptimal performance of the high-stiffness experimental group discussed later.

Remark 2: Due to the absence of temporal information, existing methods such as TNN-based methods cannot be directly implemented or fairly compared on the laparoscopic MIS platform. As such, a quantitative comparison with these architectures is not feasible within the scope of this study. Integrating such approaches with our active VBFE framework is a promising direction for future work.

2) Traction Experiment: In the traction experiment, the manipulator gradually applies a force from 0 N to 3 N, holding the force steady once reached, with the entire process lasting 5 seconds. To ensure a uniform variation of force over time, different pulling speeds are employed for silicone tubes of varying stiffness: objects with stiffness levels of 40 N/m, 80 N/m, and 200 N/m are stretched at a predetermined speeds, respectively. During the experiment, as the manipulator stretches the silicone tube, the 3D laparoscope continuously captures deformation images of the silicone tube and generates the corresponding 3D point cloud at a sampling rate of 30 Hz. Each experiment is repeated 40 times for each type of material to evaluate reliability and robustness.

B. Experimental Results

To evaluate the proposed force estimation scheme, the force estimation error statistics were quantified using the mean absolute error (MAE), RMSE and standard deviation (SD), denoted σ, as shown in Table I. The experimental results demonstrate a high level of accuracy in force prediction, particularly for lower stiffness conditions (40 N/m and 80 N/m), where prediction errors are relatively small, highlighting robust predictive precision.

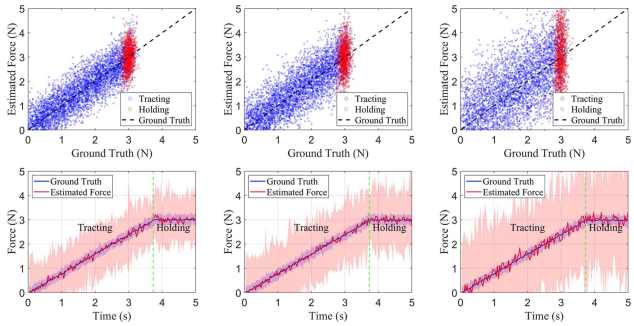


Fig. 7. The subplots in the upper row present scatter plots of estimated forces during the traction phase (blue) and holding phase (red), in comparison with the ground truth indicated by the black dashed line. The subplots in the lower row illustrate the mean force estimates (red) compared to the measurements (blue), with shaded regions indicating data intervals. The subplots, arranged from left to right, correspond to silicone samples with low, medium, and high stiffness levels, respectively.

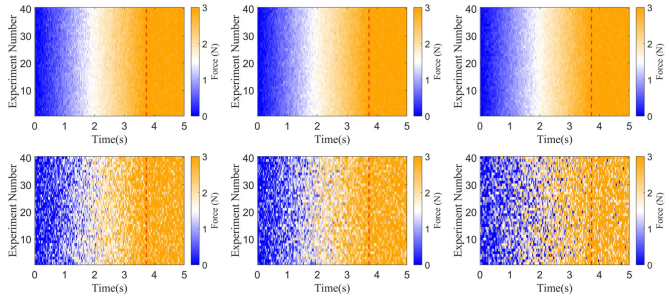


Fig. 8. The top row displays the ground truth force measurements obtained from the force-sensing trocar, while the bottom row shows the corresponding force estimates derived from visual deformation data. The three columns correspond to silicone objects with soft, medium, and hard stiffness levels, respectively. Each row represents a single experiment. In each subplot, the color gradient represents force values in Newtons (N), ranging from low contact force (blue) to medium contact force (white) and high contact force (orange). The red dotted lines differentiate between the traction and holding phases.

However, as the stiffness of the silicone material increases, the force estimation error also rises. For the object with a stiffness of 200 N/m, the MAE and RMSE reach 0.8044 and 1.0201, respectively, with SD (σ) values of 0.6108 and 1.4735, significantly exceeding the errors observed for materials with stiffness levels of 40 N/m and 80 N/m. This increase in error can be attributed to the amplification of complex nonlinear characteristics in high-stiffness materials during deformation. Subtle errors in the deformation captured by visual data disproportionately propagate to interaction force estimates [35].

The ensuing discussion will focus on the results of the force estimation for the complete interaction operation, as illustrated in Fig. 7. The figure demonstrates that the proposed method achieves highly accurate force estimation during both the traction and holding phases. The performance of the proposed force estimation method is consistent across these phases in terms of mean values and uncertainties, highlighting its robustness across different operational stages. As illustrated in the upper subplots, the force estimates demonstrate a strong linear correlation with the ground truth across a range of material stiffness levels. High-stiffness materials introduce greater uncertainties due to their smaller deformations under identical forces compared to softer materials. These minimal deformations amplify errors in visual data, imposing stricter requirements on the vision-based force estimation workflow, including training data quality, the network's capacity to model nonlinearity, and imaging system resolution. Notably, in the context of MIS, the stiffness of human tissues typically ranges from 10 N/m to 100 N/m [34], [36], which falls well within the effective operating range of the proposed method.

To validate the consistency of the force estimation algorithm, the force measurements and estimates from 40 repeated traction experiments are illustrated in Fig. 8. The experimental results show that the proposed force estimation algorithm performs well for materials with soft (40 N/m) and medium (80 N/m) stiffness, where the estimated forces closely follow the trends of the ground truth across 40 repeated experiments, demonstrating good stability and repeatability. A video of the experiments performed can be found at the following link https://youtu.be/9c4ebP97gh0.

Remark 3: For high-stiffness materials (200 N/m), force estimates exhibit increased variability and reduced consistency but still capture the overall force trend effectively. However, the method's performance on specific tissues, such as fascia and blood vessels, remains untested, and it cannot estimate forces on occluded objects. Despite these limitations, it offers substantial value and potential for minimally invasive surgery applications.

V. Conclusion

In this letter, we propose a novel and effective binocular vision-based VBFE framework for soft tissue interaction in robot-assisted surgeries. The framework employs a one-shot structured light with a custom pattern and a two-step stereo vision method to generate dense, pixel-wise 3D point clouds of smooth, texture-deficient tissue surfaces. A modified PointNet-based model is introduced for force estimation, featuring optimized activation functions, loss design, training strategies, and output layers for learning complex nonlinear force mappings. Traction experiments were performed on a platform built with a commercial surgical robot and silicone samples of varying stiffness. Results demonstrate the method's accuracy and consistency, highlighting its potential for force feedback in MIS. Compared to traditional approaches, this framework reduces hardware dependency. Future work will focus on hardware integration and miniaturization to support practical deployment in commercial surgical systems. Force estimation under significant stiffness variability presents a valuable direction for future research particularly for applications extending beyond laparoscopic robotic surgery.

References

[1] E. Abdi, D. Kulić, and E. Croft, "Haptics in teleoperated medical interventions: Force measurement, haptic interfaces and their influence on user's performance," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 12, pp. 3438–3451, Dec. 2020.

[2] Y.-Y. Juo et al., "Center for advanced surgical and interventional technology multimodal haptic feedback for robotic surgery," in *Handbook of Robotic and Image-Guided Surgery*. New York, NY, USA: Elsevier, 2020, pp. 285–301.

[3] J.-J. Cabibihan, A. Y. Alhaddad, T. Gulrez, and W. J. Yoon, "Influence of visual and haptic feedback on the detection of threshold forces in a surgical grasping task," *IEEE Robot. Automat. Lett.*, vol. 6, no. 3, pp. 5525–5532, Jul. 2021.

[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] A. H. Hadi Hosseinabadi and S. E. Salcudean, "Force sensing in robot-assisted keyhole endoscopy: A systematic survey," *Int. J. Robot. Res.*, vol. 41, no. 2, pp. 136–162, 2022.

[6] U. Kim, Y. B. Kim, D.-Y. Seok, J. So, and H. R. Choi, "Development of surgical forceps integrated with a multi-axial force sensor for minimally invasive robotic surgery," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2016, pp. 3684–3689.

[7] J. Rosen, B. Hannaford, M. P. MacFarlane, and M. N. Sinanan, "Force controlled and teleoperated endoscopic grasper for minimally invasive surgery-experimental performance evaluation," *IEEE Trans. Biomed. Eng.*, vol. 46, no. 10, pp. 1212–1221, Oct. 1999.

[8] L. Bahar, Y. Sharon, and I. Nisky, "Surgeon-centered analysis of robot-assisted needle driving under different force feedback conditions," *Front. Neurorobotics*, vol. 13, p. 108, 2020.

[9] R. Calandra et al., "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robot. Automat. Lett.*, vol. 3, no. 4, pp. 3300–3307, Oct. 2018.

[10] Y. Wei, S. Lyu, W. Li, X. Yu, Z. Wang, and L. Guo, "Contact force estimation of robot manipulators with imperfect dynamic model: On Gaussian process adaptive disturbance Kalman filter," *IEEE Trans. Automat. Sci. Eng.*, vol. 21, no. 3, pp. 3524–3537, Jul. 2024.

[11] Y. Wei, J. Wang, W. Li, X. Du, X. Yu, and L. Guo, "Composite disturbance filtering for interaction force estimation with online environmental stiffness exploration," *IEEE/ASME Trans. Mechatron.*, early access, Aug. 30, 2024, doi: 10.1109/TMECH.2024.3443310.

[12] G. Obinata, A. Dutta, N. Watanabe, and N. Moriyama, "Vision based tactile sensor using transparent elastic fingertip for dexterous handling," in *Mobile Robots: Perception & Navigation*. London, U.K.: IntechOpen, 2007.

[13] A. A. Nazari, F. Janabi-Sharifi, and K. Zareinia, "Image-based force estimation in medical applications: A review," *IEEE Sensors J.*, vol. 21, no. 7, pp. 8805–8830, Apr. 2021.

[14] K. Vlack, T. Mizota, N. Kawakami, K. Kamiyama, H. Kajimoto, and S. Tachi, "GelForce: A vision-based traction field computer interface," in *Proc. CHI'05 Extended Abstr. Hum. Factors Comput. Syst.*, 2005, pp. 1154–1155.

[15] K. Mirniazy, "Supervised deep learning with finite element synthetic data for force estimation in robotic-assisted surgery," Ph.D. dissertation, Concordia Univ., Montreal, QC, Canada, 2022.

[16] K. Takahashi and J. Tan, "Deep visuo-tactile learning: Estimation of tactile properties from images," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 8951–8957.

[17] L. Pecyna, S. Dong, and S. Luo, "Visual-tactile multimodality for following deformable linear objects using reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2022, pp. 3987–3994.

[18] F. B. Naeini et al., "A novel dynamic-vision-based approach for tactile sensing applications," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 1881–1893, May 2020.

[19] A. I. Aviles, S. Alsaleh, P. Sobrevilla, and A. Casals, "Sensorless force estimation using a neuro-vision-based approach for robotic-assisted surgery," in *Proc. 7th Int. IEEE/EMBS Conf. Neural Eng.*, 2015, pp. 86–89.

[20] M. Neidhardt, R. Mieling, M. Bengs, and A. Schlaefer, "Optical force estimation for interactions between tool and soft tissues," *Sci. Reports*, vol. 13, no. 1, 2023, Art. no. 506.

[21] P. Sabique, P. Ganesh, and R. Sivaramakrishnan, "Stereovision based force estimation with stiffness mapping in surgical tool insertion using recurrent neural network," *J. Supercomputing*, vol. 78, no. 12, pp. 14648–14679, 2022.

[22] K. Masui et al., "Vision-based estimation of manipulation forces by deep learning of laparoscopic surgical images obtained in a porcine excised kidney experiment," *Sci. Reports*, vol. 14, no. 1, 2024, Art. no. 9686.

[23] F. Baghaei Naeini, D. Makris, D. Gan, and Y. Zweiri, "Dynamic-vision-based force measurements using convolutional recurrent neural networks," *Sensors*, vol. 20, no. 16, 2020, Art. no. 4469.

[24] 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. Automat.*, 2021, pp. 12335–12341.

[25] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, Feb. 2008.

[26] H. Rubinsztein-Dunlop et al., "Roadmap on structured light," *J. Opt.*, vol. 19, no. 1, 2016, Art. no. 013001.

[27] G. J. Shirk, A. Johns, and D. B. Redwine, "Complications of laparoscopic surgery: How to avoid them and how to repair them," *J. Minimally Invasive Gynecol.*, vol. 13, no. 4, pp. 352–359, 2006.

[28] H. Fredricksen, "A survey of full length nonlinear shift register cycle algorithms," *SIAM Rev.*, vol. 24, no. 2, pp. 195–221, 1982.

[29] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[30] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, "Fast and accurate deep network learning by exponential linear units (ELUs)," 2015, *arXiv:1511.07289*.

[31] J. Wang, L. Jiang, Z. Li, W. Ma, Y. Qiao, and W. Zhao, "Autosurg-research and implementation of automatic target resection key technologies via toumai surgical robot system," in *Proc. Int. Conf. Adv. Robot. Mechatron.*, 2023, pp. 1194–1198.

[32] J. Fung, C.-k. Lee, M. Chan, W.-k. Seto, D.-h. Wong, C.-l. Lai, and M.-f. Yuen, "Defining normal liver stiffness range in a normal healthy chinese population without liver disease," *PLoS One*, vol. 8, no. 12, 2013, Art. no. e85067.

[33] D. Roulot, S. Czernichow, H. Le Clésiau, J.-L. Costes, A.-C. Vergnaud, and M. Beaugrand, "Liver stiffness values in apparently healthy subjects: Influence of gender and metabolic syndrome," *J. Hepatol.*, vol. 48, no. 4, pp. 606–613, 2008.

[34] A. Chanda and G. Singh, *Mechanical Properties of Human Tissues*. Berlin, Germany: Springer, 2023.

[35] R. Penas, E. Balmes, and A. Gaudin, "A unified non-linear system model view of hyperelasticity, viscoelasticity and hysteresis exhibited by rubber," *Mech. Syst. Signal Process.*, vol. 170, 2022, Art. no. 108793.

[36] G. Singh and A. Chanda, "Mechanical properties of whole-body soft human tissues: A review," *Biomed. Materials*, vol. 16, no. 6, 2021, Art. no. 062004.