

Graph-Based Force Estimation from Explicit Displacement Fields for Robot-Assisted Surgery

Abstract—Vision-based force estimation could provide force feedback in robot-assisted surgery without an instrument-mounted force sensor. A central challenge is to extract contact-induced tissue deformation from visual observations, yet existing methods regress force from appearance or geometry and leave this deformation implicit. We introduce an explicit residual displacement field as an intermediate representation: rigid motion is removed by point-cloud alignment, and the non-rigid residual is stored for force estimation. We use this field in two ways. PCD-IMG-Disp adds the field to image–point-cloud fusion only as a force-space residual. On our benchmark dataset, it has lower mean error than fusion at every label fraction under the reported schedules. In few-shot experiments on a stereo adaptation dataset, it also reduces mean error across indenters, from the nominal medium hardness label to the nominal hard label, and between two same-compound kidney phantoms of different shape. The proposed Ours-GNN aggregates displacement along tetrahedral organ-graph edges and estimates 3-D contact force in one forward pass. Without any image or raw point cloud, Ours-GNN achieves the lowest mean relative error among compared methods at every label fraction on the benchmark dataset, reaching $9.70 \pm 0.85\%$ with the full training set, 34% lower than image–point-cloud fusion. These results indicate that the residual displacement field left after rigid registration carries contact-force information and can serve as an intermediate representation between visual observation and force regression, a basis for vision-based force estimation on real tissue and in endoscopic surgery.

Index Terms—Force and Tactile Sensing, Medical Robots and Systems, Computer Vision for Medical Robotics.

I. INTRODUCTION

Accurate tool–tissue contact-force estimation could support future vision-based force feedback in robot-assisted minimally invasive surgery [1] without integrating a physical sensor into the instrument. Although direct force feedback has begun to emerge in recent commercial robotic systems [2], [3], broader deployment of instrument-integrated sensing remains difficult. Sensors at the instrument tip, wrist, shaft, or joints face stringent constraints on sterilization, miniaturization, biocompatibility, and mechanical integration [4]–[7]. In the absence of direct feedback, surgeons infer contact in part from visually perceived tissue deformation, a form of pseudo-haptic perception [8]. If this visual inference misjudges contact, excessive force can tear or shear soft tissue [2], [9].

Sensorless force estimation from visual observations has therefore been widely studied [4], [5]. Networks using only red–green–blue (RGB) images [9], RGB–point-cloud temporal convolutional networks [10], and recurrent models that

recover 3-D deformable structure [11] map appearance or geometry directly to force. Subsequent work adds robot-state inputs, structured-light geometry, and cross-dataset training [12]–[14]. These studies show that visual observations carry force information. A camera, however, observes deformation rather than force, and in these methods that deformation remains entangled with appearance, viewpoint, and rigid motion, leaving the estimator to separate these factors implicitly while learning the mechanics.

Deformation has been made explicit mainly for tracking and manipulation, through 3-D deformation models, structure-preserving non-rigid registration, and structured deformable-state representations [15]–[17]. Physically grounded estimators instead rely on a mechanical model: stereo-constrained biomechanical maps infer contact from tool boundary conditions [19], U-Mesh maps a prescribed load to displacement [20], and visual servoing has been coupled with online finite element method (FEM) simulation [21]. The deformation–force relation is thus given by the model, which may require boundary conditions or prescribed loads, repeated registration, or online computation.

Learned visual estimators thus treat the organ as an unstructured set of pixels or points, whereas physically grounded estimators make deformation explicit at the cost of these requirements. Force and displacement are linked through tissue stiffness, and soft-tissue deformation is nonlinear and topology-dependent [18]. An explicit displacement field on an organ graph is therefore a natural input for a learned inverse mapping from deformation to force, evaluated in one forward pass. Two questions follow: whether a dense, reference-relative displacement field adds force information beyond image–point-cloud fusion, and whether that field alone supports force estimation when placed on organ topology.

To address both questions, we separate the problem into a geometric step and a mechanical step: registration recovers where the organ deforms, and a learned estimator maps that deformation to force. Iterative closest point (ICP) removes rigid motion [22]. An offline-trained Free Point Transformer (FPT) predicts the non-rigid residual [23]. For a direct comparison, displacement-residual fusion keeps image–point-cloud fusion fixed and uses the field only as a force-space residual; because the organ graph cannot be transferred to another phantom, it is also the model used for few-shot adaptation on the stereo adaptation dataset. The proposed estimator, Ours-GNN, then takes the stored field as its only

input on a tetrahedral organ graph. Only the surface is observed, so the graph is completed from a watertight Poisson surface and includes interior nodes [25]. Mesh and particle graph networks pass messages for deformable dynamics [26], [27]. Those models usually simulate the next deformation from a load or state. Here the observed displacement is already on the graph, and the output is contact force.

The contributions are:

- An explicit residual displacement field as the intermediate representation for vision-based force estimation. Rigid motion is removed first and a frozen non-rigid residual is stored, so the force model receives contact-induced deformation explicitly. Under the reported schedules, adding this field only in force space lowers error relative to image–point-cloud fusion at every label fraction.
- A GNN-based force estimation method (Ours-GNN) that uses the field as its only input on a tetrahedral organ graph with interior nodes beyond the observed surface. On the benchmark dataset we collected, its mean relative error is substantially lower than image–point-cloud fusion at every label fraction, and lower than the other reported regressors. Its node weights concentrate near the indentation site without any contact-location label.

II. METHOD

A. Problem Formulation and Overview

We estimate the contact force from synchronized visual observations acquired during robotic indentation. Each indentation cycle contains $T=10$ frames. For a target time t , the inputs are an RGB image I_t and a reconstructed point cloud $\mathbf{P}_t = \{\mathbf{p}_i^t\}_{i=1}^{N_t}$, $\mathbf{p}_i^t \in \mathbb{R}^3$, where N_t is the number of points. The output is the 3-D force $\mathbf{f}_t \in \mathbb{R}^3$. The force network sees a four-frame context: the cycle reference and the previous, current, and next frames. Because the context includes the next frame, each estimate waits one frame interval; at endoscopic video rates this delay is a single camera frame, provided the model is trained at that frame spacing. $\mathbf{P}_t$ is a structured-light scan in the benchmark dataset and a cloud reconstructed from a calibrated stereo pair in the adaptation dataset, where no ranging sensor is used.

Contact force is mediated by tissue deformation, but raw RGB and unordered geometry mix that deformation with rigid camera–organ motion and with appearance that is only indirectly related to load. We therefore factor the estimator into an explicit residual displacement field followed by force estimation from that field. Let $\mathbf{P}_0 = \{\mathbf{p}_i^0\}_{i=1}^{N_0}$ denote the reference cloud of a cycle, where N_0 is its number of points. After rigid alignment, the non-rigid residual of each reference point is a vector $\mathbf{d}_i^t \in \mathbb{R}^3$, and the stored field is

$$\mathbf{D}_t = \{\mathbf{d}_i^t\}_{i=1}^{N_0}. \quad (1)$$

The field is computed *offline* and stored with the dataset. The registration network is never trained with the force models.

Fig. 1 summarizes the pipeline. Point clouds pass through offline ICP and frozen FPT to form the stored residual

displacement field $\mathbf{D}_t$ (Section II-B), which is then used in two ways. In PCD-IMG-Disp (Section II-C), RGB and point-cloud features produce a base force and $\mathbf{D}_t$ supplies a force-space residual, for a direct comparison; this model is also used for few-shot adaptation on the adaptation dataset, because the organ graph cannot be transferred to another phantom. Ours-GNN (Section II-D) observes only this field on a tetrahedral organ graph. Section II-E gives the training objectives.

B. Explicit Residual Displacement Field

A single learned warp from $\mathbf{P}_0$ to $\mathbf{P}_t$ would have to absorb both rigid motion and non-rigid displacement. To separate the two, a rigid transform $\mathbf{T}_t \in \text{SE}(3)$ is first estimated by ICP [22],

$$\mathbf{T}_t = \arg \min_{\mathbf{T} \in \text{SE}(3)} \sum_{i=1}^{N_0} \|\mathbf{T}\mathbf{p}_i^0 - \pi_t(\mathbf{T}\mathbf{p}_i^0)\|_2^2, \quad (2)$$

where $\pi_t(\cdot)$ denotes closest-point correspondence on $\mathbf{P}_t$. The aligned reference $\mathbf{T}_t\mathbf{P}_0$ still differs from $\mathbf{P}_t$ by a non-rigid residual. A Free Point Transformer U-Net [23], trained once and then frozen, predicts that residual on the aligned reference:

$$\mathbf{d}_i^t = \text{FPT}(\mathbf{T}_t\mathbf{p}_i^0, \mathbf{P}_t), \quad \hat{\mathbf{p}}_i^t = \mathbf{T}_t\mathbf{p}_i^0 + \mathbf{d}_i^t. \quad (3)$$

Only $\mathbf{D}_t$ is written into the dataset and used by the downstream models. Surface markers, when present, are never an input to (3). The FPT that produced the stored field was trained with a marker term on all 100 cycles. Table IV audits a second FPT trained without markers or test cycles. Fig. 2 shows the stored residual for the rod, bar, and cross.

C. Field as a Force-Space Residual (PCD-IMG-Disp)

To test whether the explicit residual displacement field supplies force information that image–point-cloud fusion does not already extract, we keep the image–point-cloud pathway identical to the PCD-IMG baseline [10] and add the field only as a force-space correction. For each of the four context frames, a VGG16 image encoder [29] and a PointNet++ geometric encoder [30] produce per-frame features; a temporal block then maps the four-frame stack to a base force $\hat{\mathbf{f}}_t^{\text{base}}$. In parallel, a shared displacement encoder ϕ reads each frame’s residual field as an unordered set of 3-D vectors. The four encoded vectors are concatenated as $\Phi_t = [\phi(\mathbf{D}_{\text{ref}}), \phi(\mathbf{D}_{\text{prev}}), \phi(\mathbf{D}_t), \phi(\mathbf{D}_{\text{next}})]$ and passed through a residual head h_{disp} that predicts a force correction. The fused prediction is

$$\hat{\mathbf{f}}_t = \hat{\mathbf{f}}_t^{\text{base}} + \lambda(e) h_{\text{disp}}(\Phi_t), \quad (4)$$

where $\lambda(e)$ ramps linearly to 0.3 over the first five epochs and is then held constant.

D. Field as the Sole Input: Graph-Based Force Estimation (Ours-GNN)

PCD-IMG-Disp still relies on appearance and raw geometry. Ours-GNN instead estimates force from the field alone, defined on an organ graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Load is transmitted

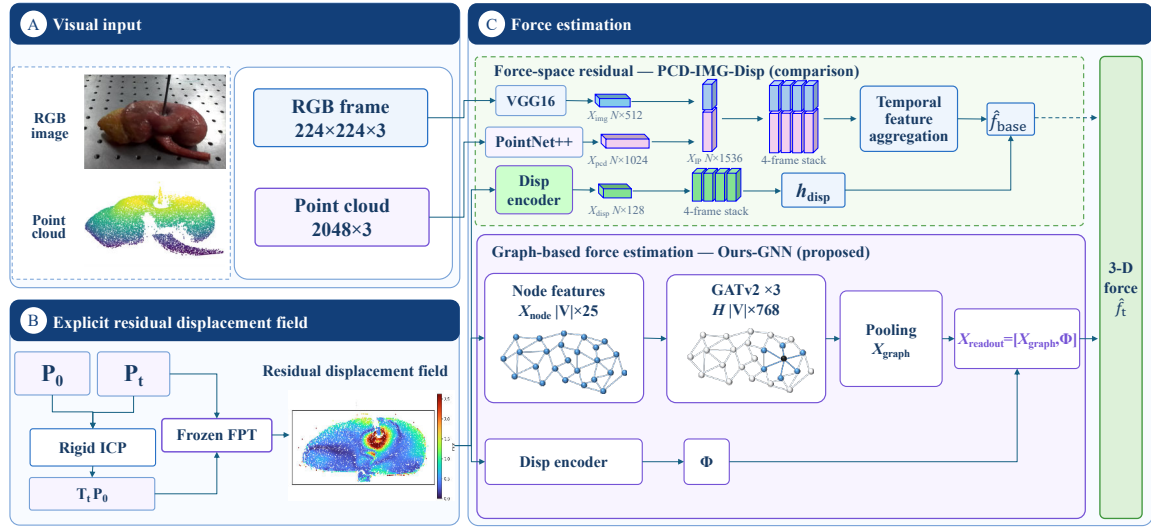


Fig. 1. Overview of the proposed method.

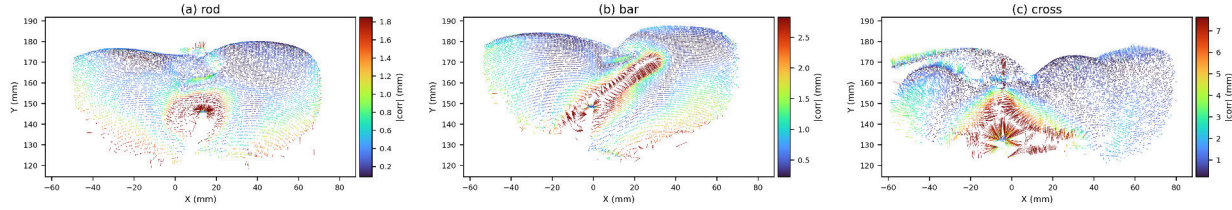


Fig. 2. Representative stored FPT residual fields (arrows at true scale). These frames are illustrative, not a controlled comparison, because both indenter and hardness label change: (a) rod, hard; (b) bar, medium; and (c) cross, soft. A. Points are colored by $|d|$; arrows run from rigidly aligned reference points to FPT-warped locations.

through the volume, but only the visible surface is observed. We therefore complete the scan into a watertight solid and lift $\mathbf{D}_t$ onto a tetrahedral graph with interior nodes (Fig. 3).

a) From surface field to organ graph: The scanned surface is completed into a watertight mesh by Poisson reconstruction [25] with a supporting table cap. Interior nodes are sampled inside the solid, and a tetrahedral mesh is built by 3-D Delaunay partitioning with exterior tetrahedra removed. Graph edges $\mathcal{E}$ are the unique tetrahedral edges. The stored field is defined on the point cloud, not on the graph. Each node j at position $\mathbf{q}_j$, on the surface or in the interior, receives the residual of its nearest stored point,

$$\hat{\mathbf{d}}_j^t = \mathbf{d}_{i^*(j)}^t, \quad i^*(j) = \arg \min_i \|\mathbf{q}_j - \mathbf{p}_i^0\|_2, \quad (5)$$

not by a second learned interpolator.

b) Message passing on displacement: Node coordinates $\mathbf{q}_j$ are used to build the tetrahedral edges and to assign residuals in (5). They are not concatenated into the GNN feature, so the network sees only the copied residual field. Writing $\mathbf{u}_i^{(\tau)}$ for the copied residual at node i in context frame $\tau \in \{0, 1, 2, 3\}$ (reference, previous, current, next), let $\mathbf{U}_i = [\mathbf{u}_i^{(0)}, \mathbf{u}_i^{(1)}, \mathbf{u}_i^{(2)}, \mathbf{u}_i^{(3)}] \in \mathbb{R}^{12}$ and let $\mathbf{m}_i$ collect the four Euclidean norms. The three temporal differences $\mathbf{u}_i^{(2)} - \mathbf{u}_i^{(0)}$, $\mathbf{u}_i^{(2)} - \mathbf{u}_i^{(1)}$, and $\mathbf{u}_i^{(3)} - \mathbf{u}_i^{(2)}$ are stacked as $\Delta \mathbf{U}_i$.

The node feature is

$$\mathbf{x}_i = [\mathbf{U}_i, \Delta \mathbf{U}_i, \mathbf{m}_i] \in \mathbb{R}^{25}. \quad (6)$$

A stack of GATv2 layers [28], [31] passes messages along the tetrahedral edges $\mathcal{E}$. For an edge $(i, j) \in \mathcal{E}$, the implemented pairwise score is

$$e_{ij} = \mathbf{a}^\top \text{LeakyReLU}(\mathbf{W}_s \mathbf{h}_i + \mathbf{W}_n \mathbf{h}_j), \quad (7)$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k \in \mathcal{N}(i)} \exp(e_{ik})},$$

and the updated embedding is $\mathbf{h}_i' = \sum_{j \in \mathcal{N}(i)} \alpha_{ij} \mathbf{W}_n \mathbf{h}_j$. The outputs of the attention heads are concatenated and passed through a second attention layer; three such residual blocks with hidden width 768 and eight heads are used in the reported model. The tetrahedral graph organizes displacement features by location and adjacency: message passing lets each node representation aggregate the deformation at that node and its neighbors, and a global readout then regresses the net contact force.

c) Displacement-gated force estimation: Contact is spatially localized, so a uniform mean over $\mathcal{V}$ alone would dilute nodes with large residual. After the last attention block we form node logits from the embedding and the current-

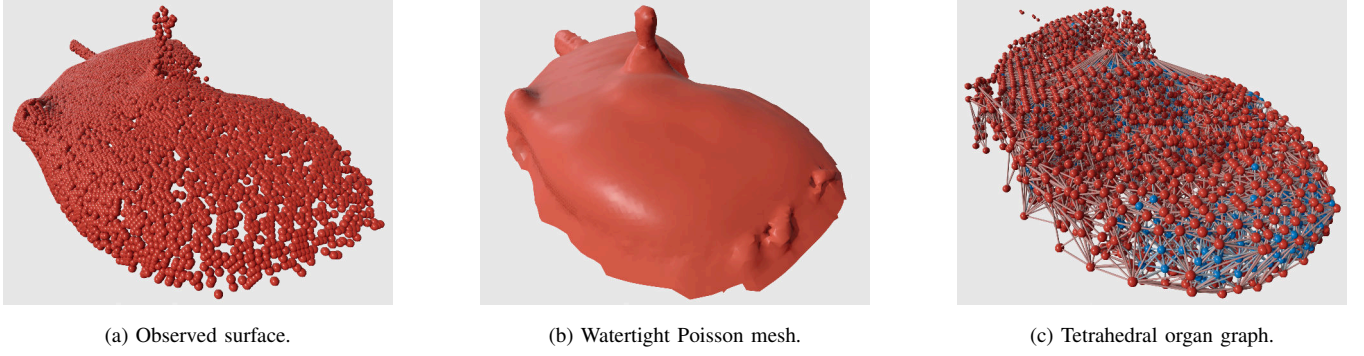


Fig. 3. Organ-graph construction on the kidney phantom, shown from one camera pose. (a) Observed surface scan. (b) Watertight Poisson mesh. (c) Tetrahedral graph used by Ours-GNN ($|\mathcal{V}|=5759$, $|\mathcal{E}|=28758$): red spheres are surface nodes, blue spheres are interior nodes, and gray segments are tetrahedral edges. The graph is built once from one rest shape.

frame residual magnitude,

$$a_i = \psi(\mathbf{h}_i) + \beta \|\mathbf{u}_i^{(2)}\|_2, \quad w_i = \frac{\exp(a_i)}{\sum_{k \in \mathcal{V}} \exp(a_k)}, \quad (8)$$

where ψ is a two-layer gate and β is a learned scalar. The graph descriptor concatenates the gate-weighted sum, a max pool, and a mean pool, and is fused with a shared displacement encoder of the four-frame residual field before a multilayer readout to $\hat{\mathbf{f}}_t$.

E. Training Objectives

Forces are standardized per axis, and all force models minimize a Smooth- L_1 loss ρ ($\delta=0.05$) on the standardized force,

$$\mathcal{L}_{\text{force}} = \frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \rho(\hat{\mathbf{f}}_t - \mathbf{f}_t), \quad (9)$$

where $\mathcal{B}$ is a mini-batch. Ours-GNN also adds a relative-error term on frames above 0.5 N, weighted by 0.25.

III. EXPERIMENTS

A. Experimental Setup

a) Hardware and phantoms: Data are collected with a FAIRINO FR3 arm and a Sunrise Instruments (SRI) M3703B six-axis force sensor. Indenters are a rounded-tip rod, a transverse bar, and a cross-shaped tip. Silicone hardness was not instrumented; we refer to the three levels by the nominal labels hard, medium, and soft. Two phantoms at the soft level, soft A and soft B, use the same compound and differ only in kidney shape (Fig. 4).

We collect two datasets and name them by their use. The *benchmark dataset* supports method comparison, label efficiency, and displacement-field validation. It is recorded on the nominal-medium phantom with the rod indenter, using a Zivid 2+ M60 structured-light scanner. From cycle to cycle the contact point is randomized within ± 5 mm in the horizontal plane. Each cycle indents from contact through a 10 mm stroke, and the ten frames are recorded along that stroke. The dataset contains 100 indentation cycles and 1000 synchronized frames, with contact force mainly along the indentation axis and magnitude up to 8.7 N (896 frames above 0.5 N). Surface dots on this phantom are detected with

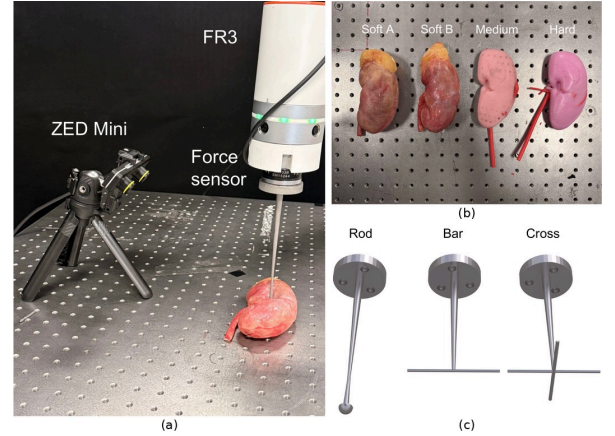


Fig. 4. Experimental hardware and phantoms. (a) Platform of the adaptation dataset, a stereo proxy for endoscopic sensing. A ZED Mini supplies the left-right pair; FoundationStereo reconstructs the cloud, without native ZED depth. The benchmark dataset uses a Zivid 2+ M60 and is not shown. (b) Four silicone kidney phantoms. Soft A and soft B share a compound and differ in shape. Hard, medium, and soft are nominal labels; hardness was not instrumented. (c) Rod, bar, and cross indenters.

YOLO and lifted to 3-D. They are not a network input; Fig. 6 shows them on the stored field, and Section II-B states how they enter FPT training and the audit in Table IV. The *adaptation dataset* supports the few-shot adaptation experiments. In robot-assisted surgery, the scene is typically observed through a stereo endoscope, so this dataset uses a ZED Mini (Fig. 4). FoundationStereo reconstructs a metric cloud from the left-right pair [32], and the same offline ICP and FPT then store the residual field. The open-field stereo camera is a proxy for endoscopic stereo input, not a laparoscope.

The adaptation dataset covers three few-shot adaptation experiments: cross-indenter adaptation at the nominal medium label (rod $\rightarrow$ bar and rod $\rightarrow$ cross), cross-hardness adaptation with the rod (medium $\rightarrow$ hard), and cross-shape adaptation between same-compound phantoms (soft A $\rightarrow$ soft B). Each recorded combination contains 100 cycles and 1000 frames.

b) Protocol: Cycles are split into 70/15/15 train/validation/test partitions with a fixed split seed.

Label fractions keep the first 70, 35, 18, or 7 training cycles of a fixed permutation; validation and test stay at 15 cycles. On the benchmark dataset, every reported configuration is trained three times with independent random seeds, which change network initialization and data order. The same $n=3$ protocol is used for the adaptation experiments. Tables I–III report test mean $\pm$ s.d. over the three runs, with checkpoints selected on validation. Training schedules were fixed on validation from a single run before the multi-seed evaluation: image-only, point-cloud-only, and image–point-cloud fusion use 50 epochs with plateau scheduling; PCD-IMG-Disp uses 80 epochs with a cosine schedule; the GNN uses the fixed tetrahedral graph ($|\mathcal{V}| = 5759$, $|\mathcal{E}| = 28758$) and 80 epochs with a cosine schedule. For the GNN, SWA and EMA weights are tracked, and the checkpoint with the lowest validation relative error is reported; in every reported run this is the non-averaged checkpoint.

c) *Metrics*: We report held-out test relative error

$$E_{\text{rel}} = \frac{\|\hat{\mathbf{f}}_t - \mathbf{f}_t\|_2}{\|\mathbf{f}_t\|_2} \quad \text{if} \quad \|\mathbf{f}_t\|_2 > 0.5 N, \quad (10)$$

and absolute error $E_{\text{abs}} = \|\hat{\mathbf{f}}_t - \mathbf{f}_t\|_2$ (N). Relative error is the score reported for the benchmark dataset. Absolute error is also reported for the adaptation experiments, where zero-shot relative error exceeds 100%; the two are interpreted together.

B. Compared Methods

Five contact-force estimators are compared on the benchmark dataset under the same split and four-frame context:

- **IMG**: RGB-only temporal regression, following image-based force estimators [9].
- **PCD**: PointNet++ cloud-only regression [30].
- **PCD-IMG**: temporal fusion of VGG16 image features and PointNet++ cloud features. This is our image–point-cloud fusion baseline [10], with feature widths adapted to our cloud size. Its training schedule is the one stated above.
- **PCD-IMG-Disp**: displacement-residual fusion, PCD-IMG plus the residual head (4). It adds the field only in force space and is the model used for few-shot adaptation.
- **Ours-GNN**: the proposed estimator, graph regression on the explicit displacement field only (6)–(8).

The adaptation dataset carries no markers. Because the Ours-GNN organ graph cannot be transferred to another phantom, the adaptation tables compare PCD-IMG and PCD-IMG-Disp.

C. Label Efficiency on the Benchmark Dataset

Table I and Fig. 5(a) separate two uses of the explicit field at label fractions $\{1.0, 0.5, 0.25, 0.1\}$. Under the schedules reported above, PCD-IMG-Disp has lower relative error than image–point-cloud fusion at every fraction. Ours-GNN has the lowest mean relative error, $9.70 \pm 0.85\%$ with the full training set and $17.62 \pm 0.43\%$ with 10% of the training set, reductions of 34% and 35% from PCD-IMG. Relative to PCD-IMG-Disp, the ± 1 s.d. bands separate at 10%; at

TABLE I
RELATIVE ERROR (%) AT 100%, 50%, 25%, AND 10% OF THE TRAINING CYCLES (MEAN $\pm$ S.D., $n=3$ INDEPENDENT RUNS; BEST IN EACH COLUMN IN BOLD).

Method	100%	50%	25%	10%
IMG	15.12 $\pm$ 1.41	19.03 $\pm$ 1.30	21.49 $\pm$ 1.34	29.04 $\pm$ 3.60
PCD	24.86 $\pm$ 2.96	43.75 $\pm$ 5.63	47.94 $\pm$ 5.50	72.70 $\pm$ 5.08
PCD-IMG	14.60 $\pm$ 1.41	21.66 $\pm$ 0.60	22.06 $\pm$ 1.75	27.13 $\pm$ 2.21
PCD-IMG-Disp	11.00 $\pm$ 1.38	15.59 $\pm$ 1.60	16.53 $\pm$ 1.63	21.54 $\pm$ 0.48
Ours-GNN	9.70$\pm$0.85	13.09$\pm$1.03	15.34$\pm$1.28	17.62$\pm$0.43

TABLE II
CROSS-INDENTER ADAPTATION ON THE ADAPTATION DATASET AT THE NOMINAL MEDIUM LABEL (MEAN $\pm$ S.D., $n=3$ INDEPENDENT RUNS; BEST PER METRIC IN EACH ROW IN BOLD).

K	Relative error (%)		Absolute error (N)	
	PCD-IMG	PCD-IMG-Disp	PCD-IMG	PCD-IMG-Disp
rod $\rightarrow$ bar				
0	154.12 $\pm$ 6.20	129.22$\pm$21.07	2.510 $\pm$ 0.091	2.131$\pm$0.291
1	60.58 $\pm$ 0.93	42.50$\pm$6.17	1.375 $\pm$ 0.093	1.155$\pm$0.089
3	39.61 $\pm$ 3.26	33.90$\pm$6.47	1.161 $\pm$ 0.125	0.996$\pm$0.100
5	29.48 $\pm$ 1.98	25.59$\pm$2.99	0.742 $\pm$ 0.063	0.660$\pm$0.027
rod $\rightarrow$ cross				
0	157.36 $\pm$ 3.48	137.92$\pm$17.68	2.287 $\pm$ 0.043	2.035$\pm$0.209
1	74.14 $\pm$ 7.39	61.89$\pm$4.78	1.321 $\pm$ 0.037	1.087$\pm$0.088
3	49.17 $\pm$ 2.30	42.95$\pm$2.52	0.904 $\pm$ 0.038	0.777$\pm$0.073
5	34.41 $\pm$ 0.55	28.62$\pm$3.36	0.688 $\pm$ 0.006	0.594$\pm$0.020

100%, 50%, and 25% the difference is a mean ordering, and no paired test is reported. These results show that a graph network on the displacement field alone regresses contact force effectively across label fractions.

D. Cross-Indenter Adaptation

Table II and Fig. 5(b)–(c) report few-shot adaptation from the rod to the bar and the cross at the nominal medium label on the adaptation dataset, with $K \in \{0, 1, 3, 5\}$ labeled target training cycles. Source models are trained on the rod recording of the adaptation dataset. For $K > 0$, both models fine-tune all parameters at one fifth of the source learning rate on those K cycles, which are fixed and nested across K . The stopping epoch is selected on 15 additional labeled target validation cycles, which are not included in K . For $K = 0$, the source checkpoint is tested directly. PCD-IMG-Disp has the lower mean relative error, 12–30% below PCD-IMG, and the lower mean absolute error in all eight K –target cells. The field thus aids few-shot adaptation when the indenter changes.

E. Cross-Hardness and Cross-Shape Adaptation

Table III and Fig. 5(d)–(e) keep the rod and change either the nominal hardness label (medium $\rightarrow$ hard) or kidney shape at the same soft compound (soft A $\rightarrow$ soft B). PCD-IMG-Disp has the lower mean relative error, 11–24% below PCD-IMG, and the lower mean absolute error in all eight K –target cells. One zero-shot run on medium $\rightarrow$ hard still reverses the ranking (49.50% / 0.890 N versus 36.43% / 0.777 N).

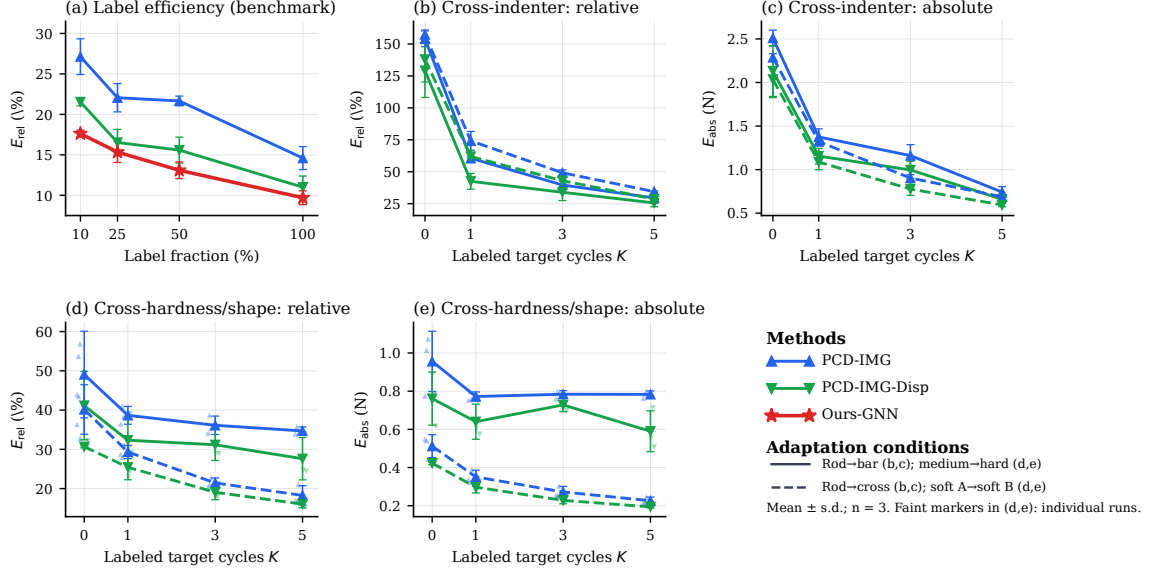


Fig. 5. Force-estimation results (mean $\pm$ s.d., $n=3$ independent runs). (a) Benchmark dataset, the three methods central to the displacement-field analysis; Table I adds IMG and PCD. (b)–(c) Cross-indenter adaptation: solid rod $\rightarrow$ bar, dashed rod $\rightarrow$ cross. (d)–(e) Cross-hardness and cross-shape adaptation: solid medium $\rightarrow$ hard, dashed soft A $\rightarrow$ soft B; faint markers show individual runs.

TABLE III

CROSS-HARDNESS AND CROSS-SHAPE ADAPTATION ON THE ADAPTATION DATASET WITH THE ROD INDENTER (MEAN $\pm$ S.D., $n=3$ INDEPENDENT RUNS; BEST PER METRIC IN EACH ROW IN BOLD).

K	Relative error (%)		Absolute error (N)	
	PCD-IMG	PCD-IMG-Disp	PCD-IMG	PCD-IMG-Disp
medium $\rightarrow$ hard				
0	49.05 $\pm$ 11.04	41.13$\pm$8.71	0.956 $\pm$ 0.158	0.761$\pm$0.139
1	38.66 $\pm$ 2.27	32.32$\pm$5.87	0.772 $\pm$ 0.024	0.640$\pm$0.092
3	36.12 $\pm$ 2.34	31.16$\pm$4.00	0.784 $\pm$ 0.019	0.728$\pm$0.035
5	34.68 $\pm$ 1.02	27.61$\pm$5.39	0.783 $\pm$ 0.018	0.590$\pm$0.107
soft A $\rightarrow$ soft B				
0	40.14 $\pm$ 6.30	30.67$\pm$0.28	0.512 $\pm$ 0.060	0.423$\pm$0.008
1	29.32 $\pm$ 1.68	25.43$\pm$3.19	0.351 $\pm$ 0.035	0.297$\pm$0.030
3	21.42 $\pm$ 1.29	19.05$\pm$1.88	0.272 $\pm$ 0.029	0.228$\pm$0.018
5	18.25 $\pm$ 2.47	16.03$\pm$0.92	0.226 $\pm$ 0.019	0.194$\pm$0.005

TABLE IV

DISPLACEMENT AUDIT ON THE 15 FORCE-MODEL TEST CYCLES (MM; $\downarrow$ LOWER IS BETTER). TEST: 135 NON-REFERENCE FRAMES; CONTACT: $|F| > 0.5$ N AND STORED $p_{95}|d| < 10$ mm. FPT SAW THESE FRAMES AND MARKERS IN TRAINING; FPT-NM DROPS BOTH.

Subset	n	Marker EPE (mm) $\downarrow$				Chamfer (mm) $\downarrow$			
		ICP	CPD	FPT	FPT-NM	ICP	CPD	FPT	FPT-NM
test	135	0.545	0.809	0.425	0.414	0.448	0.805	0.257	0.203
contact	130	0.442	0.713	0.321	0.310	0.411	0.774	0.252	0.201

The field thus also aids few-shot adaptation across nominal hardness labels and phantom shapes.

F. Displacement-Field Validation on Held-Out Cycles

Fig. 6 shows the stored-field markers and relates contact-force magnitude to the 95th percentile of $|d|$ over the

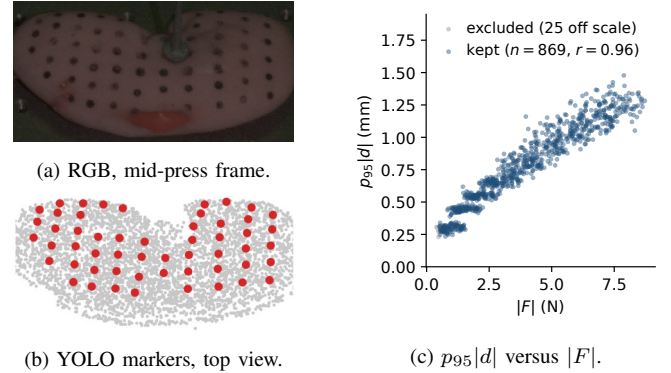


Fig. 6. Displacement-field audit on the benchmark dataset. (a) RGB crop of a mid-press frame. (b) Top view of the scan with the YOLO surface markers. (c) Per-frame $p_{95}|d|$ versus $|F|$ over the 900 non-reference frames. Navy points satisfy $|F| > 0.5$ N and $p_{95}|d| < 10$ mm ($r = 0.96$, $n = 869$); gray points lie outside this contact subset, and 25 field outliers with $p_{95}|d| > 24$ mm are outside the plotted range.

900 non-reference frames. The raw correlation is near zero (-0.003) for two reasons. Frames below 0.5 N carry little force, and 25 contact frames (2.8%) are residual outliers whose $p_{95}|d|$ exceeds 24 mm, well separated from the bulk below 10 mm; these frames remain in the force datasets. On the remaining 869 contact frames ($|F| > 0.5$ N, $p_{95}|d| < 10$ mm), displacement magnitude tracks $|F|$ ($r = 0.96$).

Table IV compares four warps of the non-reference frames in the 15 force-model test cycles, after the same ICP used to build the benchmark dataset: rigid ICP, unsupervised CPD [24] (1536 source / 4096 target points, no marker supervision), the original FPT, and FPT-NM. On the 135 test frames, FPT-NM has lower marker endpoint error (0.414 mm) and Chamfer distance (0.203 mm) than rigid ICP (0.545 mm and 0.448 mm). CPD remains worse than rigid

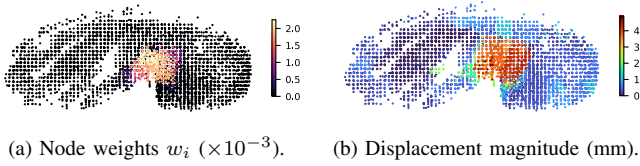


Fig. 7. Displacement-gated node weights on one held-out end-of-press frame ($|F| = 8.44$ N; inference only). (a) Node weights w_i from (8) over the tetrahedral graph (top view). (b) Current-frame displacement magnitude.

ICP in this implementation. The FPT-NM and stored fields differ by 0.063 mm (pointwise 95th percentile), so the force models keep the stored field. A learned non-rigid residual thus tracks marker motion more closely than rigid alignment.

G. Spatial Distribution of Ours-GNN Node Weights

Fig. 7 examines where Ours-GNN places its node weights. On the shown end-of-press frame, the node weights of (8) concentrate near the indentation site ($r = 0.76$ between w_i and $\|\mathbf{u}_i^{(2)}\|$; the top-5% weight and top-5% displacement nodes overlap by 32.8%). Over the 150 test frames, the correlation is 0.47 ± 0.27 . It is stronger on loaded frames and weaker at rest or early press. Ours-GNN thus weights nodes near the contact site without a contact-location label.

IV. DISCUSSION

A. Mechanical Basis of Contact-Induced Tissue Deformation

A contact load reaches the camera only as tissue deformation around the contact site. In a discretized linear-elastic body, $\mathbf{f} = \mathbf{K}\mathbf{u}$, and the stiffness matrix $\mathbf{K}$ couples only nodes that share an element, so the contact force is set by the displacement of the surrounding tissue; vision can therefore sense force indirectly, through deformation. Raw images and clouds also contain rigid camera–organ motion and appearance. Rigid motion lies in the null space of $\mathbf{K}$ and produces no elastic force, so removing it by registration leaves the non-rigid residual, the part of the displacement produced by the load. The residual field, not the raw observation, is thus the mechanically relevant quantity. Soft-tissue deformation is nonlinear and topology-dependent [18], so the field-to-force mapping is learned rather than prescribed.

B. Residual-Displacement Graph Network Mechanism for Vision-Based Force Estimation

Each Ours-GNN node carries only its copied residual over the four context frames, with temporal differences and magnitudes (6); coordinates, appearance, and raw geometry are excluded. Message passing along tetrahedral edges follows the locality of $\mathbf{K}$, so each node representation summarizes the deformation at that node and its neighborhood, and interior nodes link surface regions through the loaded volume. Because contact is localized, the displacement-gated readout (8) emphasizes nodes with a large current residual, and the global readout regresses the net 3-D contact force in one forward pass. The analogy is limited: the network sees neither element geometry nor material parameters, its node values are nearest-neighbor copies, and it outputs net force rather than nodal forces. It is a learned inverse mapping from deformation to force, not an online FEM solve.

C. Evidence from Vision-Based Force Estimation and Cross-Condition Adaptation

The experiments support this account at three levels. For the field, an FPT trained without markers tracks held-out markers more closely than rigid ICP (Table IV), and the field magnitude tracks contact force on filtered contact frames ($r = 0.96$), though not over all frames. For the model, Ours-GNN sees only the field yet stays about 35% below PCD-IMG from the full training set to 10% of it, whereas PCD rises from 24.86% to 72.70%, and its node weights concentrate where displacement is largest. Relative to PCD-IMG-Disp, the ± 1 s.d. bands separate only at 10%, and the gain of PCD-IMG-Disp over fusion is a system comparison, because the schedules differ. For the conditions, the field is tested through the non-graph PCD-IMG-Disp, which has lower mean error in all sixteen cells, twelve few-shot ($K \geq 1$) and four zero-shot, across indenter, nominal hardness, and kidney shape. The force information in the field thus persists across these changes, and a few labeled cycles recalibrate the field-to-force mapping; one zero-shot reversal on medium→hard limits the claim to few-shot adaptation.

D. Extension Toward Real Tissue and Endoscopic Surgery

These results suggest that an explicit residual field can serve as an intermediate representation between visual observation and force regression. Silicone indentation gives repeatable reference force and deformation, with hardness and shape set independently and reproducible across sessions, as adaptation requires. Benchmark forces reach 8.7 N, mainly along the indentation axis; because the field and the predicted force are both 3-D, the formulation can in principle be applied to oblique and sliding contact. Extension to tissue and endoscopy can follow the method’s two steps. The geometric step needs only a metric point cloud, and our two datasets already feed different sensors into the same field construction, so an endoscope-specific stereo model can replace FoundationStereo; a short baseline adds reconstruction error, and physiological motion such as respiration enters the residual directly and must be separated from contact deformation, for example by respiratory gating. The mechanical step absorbs tissue properties: we expect viscoelasticity and anisotropy to change mainly the field-to-force mapping, which few-shot recalibration can update, as in cross-hardness and cross-shape adaptation. Ours-GNN adds one constraint: its graph is built from one phantom’s rest shape, so a new organ requires a new graph and retraining. Once trained at video frame spacing and validated on tissue, the estimate could serve as a force-limit alert, visual or haptic feedback, or a force signal in autonomous robotic surgery [1], and offline could force-annotate recorded procedures for training and skill assessment.

V. CONCLUSION

Vision observes deformation, not force. This letter proposed a vision-based force estimation method built on an explicit residual displacement field: rigid motion is removed by point-cloud alignment, and the non-rigid residual is stored for force estimation. On this field we built Ours-GNN, an

organ-graph network that takes only the field as input using a tetrahedral graph with interior nodes. Ours-GNN outperforms image–point-cloud fusion at all benchmark label fractions, with $9.70 \pm 0.85\%$ relative error on the full training set and $17.62 \pm 0.43\%$ on 10% of it. As a force-space residual in this fusion, the field lowers mean error at every fraction under unmatched schedules and in few-shot adaptation across indenter, hardness, and shape. The results indicate that the field captures contact-induced deformation, and the tetrahedral graph aggregates local displacements via spatial connectivity into a structured mapping to contact force. The method may support force-limit alerts, visual force feedback, force-aware autonomous surgery, and force annotation of recordings. Future work will evaluate ex vivo tissue, endoscopic stereo input, and online tests during robot-assisted manipulation.

REFERENCES

- [1] S. A. Pedram, C. Shin, P. W. Ferguson, J. Ma, E. P. Dutson, and J. Rosen, "Autonomous suturing framework and quantification using a cable-driven surgical robot," *IEEE Trans. Robot.*, vol. 37, no. 2, pp. 404–417, Apr. 2021, doi: 10.1109/TRO.2020.3031236.
- [2] M. M. Awad, M. C. Raynor, M. Padmanabhan-Kabana, L. Y. Schumacher, and J. A. Blatnik, "Evaluation of forces applied to tissues during robotic-assisted surgical tasks using a novel force feedback technology," *Surg. Endosc.*, vol. 38, no. 10, pp. 6193–6202, Oct. 2024, doi: 10.1007/s00464-024-11131-z.
- [3] E. L. Servais, L. Rashidi, P. Porwal, M. Garibaldi, and A. J. Hung, "Novel force feedback technology improves suturing in robotic-assisted surgery: a pre-clinical study," *Surg. Endosc.*, vol. 39, no. 2, pp. 1217–1226, Feb. 2025, doi: 10.1007/s00464-024-11472-9.
- [4] W. Wang, J. Wang, Y. Luo, X. Wang, and H. Song, "A survey on force sensing techniques in robot-assisted minimally invasive surgery," *IEEE Trans. Haptics*, vol. 16, no. 4, pp. 702–718, Oct.–Dec. 2023, doi: 10.1109/TOH.2023.3329172.
- [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, Feb. 2022, doi: 10.1177/02783649211052067.
- [6] U. Kim, D.-H. Lee, W. J. Yoon, B. Hannaford, and H. R. Choi, "Force sensor integrated surgical forceps for minimally invasive robotic surgery," *IEEE Trans. Robot.*, vol. 31, no. 5, pp. 1214–1224, Oct. 2015, doi: 10.1109/TRO.2015.2473515.
- [7] T. Liu, X. Wang, J. Katupitiya, J. Wang, and L. Wu, "Automatic tissue traction using miniature force-sensing forceps for minimally invasive surgery," *IEEE Trans. Robot.*, vol. 40, pp. 4690–4704, 2024, doi: 10.1109/TRO.2024.3486177.
- [8] A. Lécuyer, S. Coquillart, A. Kheddar, P. Richard, and P. Coiffet, "Pseudo-haptic feedback: Can isometric input devices simulate force feedback?," in *Proc. IEEE Virtual Reality 2000*, 2000, pp. 83–90, doi: 10.1109/VR.2000.840369.
- [9] K. Masui, N. Kume, M. Nakao, T. Magaribuchi, A. Hamada, T. Kobayashi, and A. Sawada, "Vision-based estimation of manipulation forces by deep learning of laparoscopic surgical images obtained in a porcine excised kidney experiment," *Sci. Rep.*, vol. 14, no. 1, Art. no. 9686, 2024, doi: 10.1038/s41598-024-60574-w.
- [10] 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 *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, ser. LNCS, vol. 11041. Cham, Switzerland: Springer, 2018, pp. 118–127, doi: 10.1007/978-3-030-01201-4_14.
- [11] A. I. Aviles, S. M. Alsaleh, J. K. Hahn, and A. Casals, "Towards retrieving force feedback in robotic-assisted surgery: A supervised neuro-recurrent-vision approach," *IEEE Trans. Haptics*, vol. 10, no. 3, pp. 431–443, Jul. 2017, doi: 10.1109/TOH.2016.2640289.
- [12] 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, doi: 10.1109/ICRA48506.2021.9560945.
- [13] J. Wang, M. Yao, Y. Wei, X. Guo, A. Zheng, and W. Zhao, "Image-to-force estimation for soft tissue interaction in robotic-assisted surgery using structured light," *IEEE Robot. Autom. Lett.*, vol. 10, no. 8, pp. 7795–7802, 2025, doi: 10.1109/LRA.2025.3579640.
- [14] M. De Iturrate Reyzaal, M. Chen, W. Huang, S. Ourselin, and H. Liu, "DaFoEs: Mixing datasets towards the generalization of vision-state deep-learning force estimation in minimally invasive robotic surgery," *IEEE Robot. Autom. Lett.*, vol. 9, no. 3, pp. 2527–2534, 2024, doi: 10.1109/LRA.2024.3356984.
- [15] D. Navarro-Alarcon, H. M. Yip, Z. Wang, Y.-H. Liu, F. Zhong, T. Zhang, and P. Li, "Automatic 3-D manipulation of soft objects by robotic arms with an adaptive deformation model," *IEEE Trans. Robot.*, vol. 32, no. 2, pp. 429–441, Apr. 2016, doi: 10.1109/TRO.2016.2533639.
- [16] T. Tang and M. Tomizuka, "Track deformable objects from point clouds with structure preserved registration," *Int. J. Robot. Res.*, vol. 41, no. 6, pp. 599–614, May 2022, doi: 10.1177/0278364919841431.
- [17] B. Yang, C. Sui, F. Zhong, and Y.-H. Liu, "Modal-graph 3D shape servoing of deformable objects with raw point clouds," *Int. J. Robot. Res.*, vol. 42, no. 14, pp. 1213–1244, Sep. 2023, doi: 10.1177/02783649231198900.
- [18] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Sci. Robot.*, vol. 6, no. 54, Art. no. eabd8803, 2021, doi: 10.1126/scirobotics.abd8803.
- [19] 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, doi: 10.1109/LRA.2018.2810948.
- [20] A. Mendizabal, P. Márquez-Neila, and S. Cotin, "Simulation of hyper-elastic materials in real-time using deep learning," *Med. Image Anal.*, vol. 59, Art. no. 101569, 2020, doi: 10.1016/j.media.2019.101569.
- [21] C. Saghour, D. Navarro-Alarcón, P. Fraisse, and A. Cherubini, "Dual-arm shaping of soft objects in 3D based on visual servoing and online FEM simulations," *Int. J. Robot. Res.*, vol. 44, no. 7, pp. 1138–1155, Jun. 2025, doi: 10.1177/02783649241301076.
- [22] P. J. Besl and N. D. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, Feb. 1992, doi: 10.1109/34.121791.
- [23] Z. M. C. Baum, Y. Hu, and D. C. Barratt, "Multimodality biomedical image registration using free point transformer networks," in *Medical Ultrasound, and Preterm, Perinatal and Paediatric Image Analysis*, ser. LNCS, vol. 12437. Cham, Switzerland: Springer, 2020, pp. 116–125, doi: 10.1007/978-3-030-60334-2_12.
- [24] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 12, pp. 2262–2275, 2010.
- [25] M. Kazhdan, M. Bolitho, and H. Hoppe, "Poisson surface reconstruction," in *Proc. Eurographics Symp. Geometry Process.*, 2006, pp. 61–70.
- [26] T. Pfaff, M. Fortunato, A. Sanchez-Gonzalez, and P. W. Battaglia, "Learning mesh-based simulation with graph networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021.
- [27] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "RoboCraft: Learning to see, simulate, and shape elasto-plastic objects in 3D with graph networks," *Int. J. Robot. Res.*, vol. 43, no. 4, pp. 533–549, Apr. 2024, doi: 10.1177/02783649231219020.
- [28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
- [29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, arXiv:1409.1556.
- [30] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, 2017, pp. 5099–5108.
- [31] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2022.
- [32] B. Wen, M. Trepte, J. Aribido, J. Kautz, O. Gallo, and S. Birchfield, "FoundationStereo: Zero-shot stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2025, pp. 5249–5260, doi: 10.1109/CVPR52734.2025.00495.