2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)
September 30 – October 2, 2025 | COEX, Seoul, Korea

Bidirectional Mapping Between Physical Contacts and Visual Tactile Images for Physics-Based Simulation

Taehwa Hong[1] and Yong-Lae Park[1]

2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids) | 979-8-3315-9869-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/HUMANOIDS65713.2025.11203137

Abstract— **This paper introduces a simulation framework for the vision-based tactile (ViTac) sensor through accurate modeling of contact deformation and pressure. A finite element model replicates a ViTac sensor to simulate contact events and generate high-resolution surface deformation and force. These simulated contacts are translated into tactile RGB images using data-driven mapping, enabling large-scale synthetic data generation without requiring real measurements. The framework also infers depth and contact force from real RGB images using simulated contact as supervision, reversely. The resulting bidirectional mapping connects simulated and real tactile domains, supporting both synthetic data generation and the addition of physical annotations to existing datasets. This framework is applicable to learning-based tactile perception tasks where high-quality paired data are limited or difficult to collect.**

I. INTRODUCTION

Recent advances in large-scale datasets have greatly improved robotic policy learning and enabled larger tasks [1]–[3]. In addition to diverse state observations, the rich force feedback from the end effector is crucial for precision and robustness in contact-rich manipulation [4], [5], especially for humanoid platforms requiring safe, adaptive physical interaction.

In-hand manipulation, where a gripper makes multiple simultaneous contacts, requires high-resolution spatially dense tactile sensing to accurately capture both contact forces and geometry [6]–[8]. Conventional proprioceptive sensors, such as joint torque encoders, provide only coarse contact information [9], [10] and often fail to capture local contact points, shapes, or forces.

Given the limitations of proprioceptive sensing, recent advances have focused on the development of vision-based tactile (ViTac) sensors [11]–[13], which embed internal cameras to capture deformations of an elastic surface. These sensors have demonstrated superior capabilities in interpreting visual signals to extract physically meaningful contact information, including contact location [14], surface geometry [15], and force distribution [16]. As a result, they enable fine-grained control in manipulation tasks such as object reorientation [8] and slip detection [17].

Vision-based tactile (ViTac) sensor has improved downstream tasks such as grasp outcome prediction [18], material recognition [19], force estimation [20], and contact-rich manipulation via end-to-end reinforcement learning [21], [22]. These advances rely on large volumes of high-quality tactile data [16], [19], [23], motivating simulation environments that can reproduce high-fidelity ViTac outputs for scalable data augmentation and policy training [24].

A variety of simulation-based approaches have been developed to meet this demand. These include realistic tactile image generation [25]–[27], physics-informed perception [8], [16], [28], rendering in physics-based simulators [23], [29], and learning-based tactile perception [30], [31]. While these methods have advanced tactile rendering and representation learning, most cannot produce physically grounded tactile data that jointly capture force distribution, contact shape, and surface deformation. Finite element methods (FEM) can model these interactions [32], [33], but are often too computationally expensive for real-time control or accelerated reinforcement learning. Moreover, existing work typically focuses on perception rather than generation, and only a few methods render realistic RGB outputs that closely match actual sensor responses [23], [25], [34], [35].

This study proposes a simulation rendering framework based on an optimized FEM implementation to enable physically accurate and high-resolution modeling of ViTac sensors. The simulator computes contact-induced surface deformation and force distribution, producing large-scale data sets with physically grounded annotations. A calibration procedure aligns simulated responses with real sensor measurements across various contact shapes and loading conditions.

Using this calibrated model, we construct paired data sets by matching simulated nodal outputs with real RGB images from mirrored indentation experiments. These data train two bidirectional networks: (i) a perception model that estimates deformation and force from RGB input, and (ii) a rendering model that synthesizes realistic tactile images from physical-state representations. Together, they enable interpretable annotation of real sensor data, physics-consistent RGB generation from simulation, and strong generalization to unseen contact types, supporting both scalable data augmentation and simulation-based learning.

II. SIMULATION AND LEARNING FRAMEWORK

Fig. 1 shows the proposed framework, which integrates FEM-based simulation with bidirectional mapping between real and simulated tactile signals. The DIGIT sensor [6] was calibrated in SOFA [36] using minimal real-world indentation data to replicate the hyperelastic behavior of its

This work was supported by Industrial Strategic Technology Development Program (No. 00423940) funded by the Ministry of Trade, Industry & Energy(MOTIE, Korea).

[1] Taehwa Hong and Yong-Lae Park are with the Department of Mechanical Engineering, Institute of Advanced Machines and Design, Seoul National University, Republic of Korea. ndolphin93@gmail.com, ylpark@snu.ac.kr

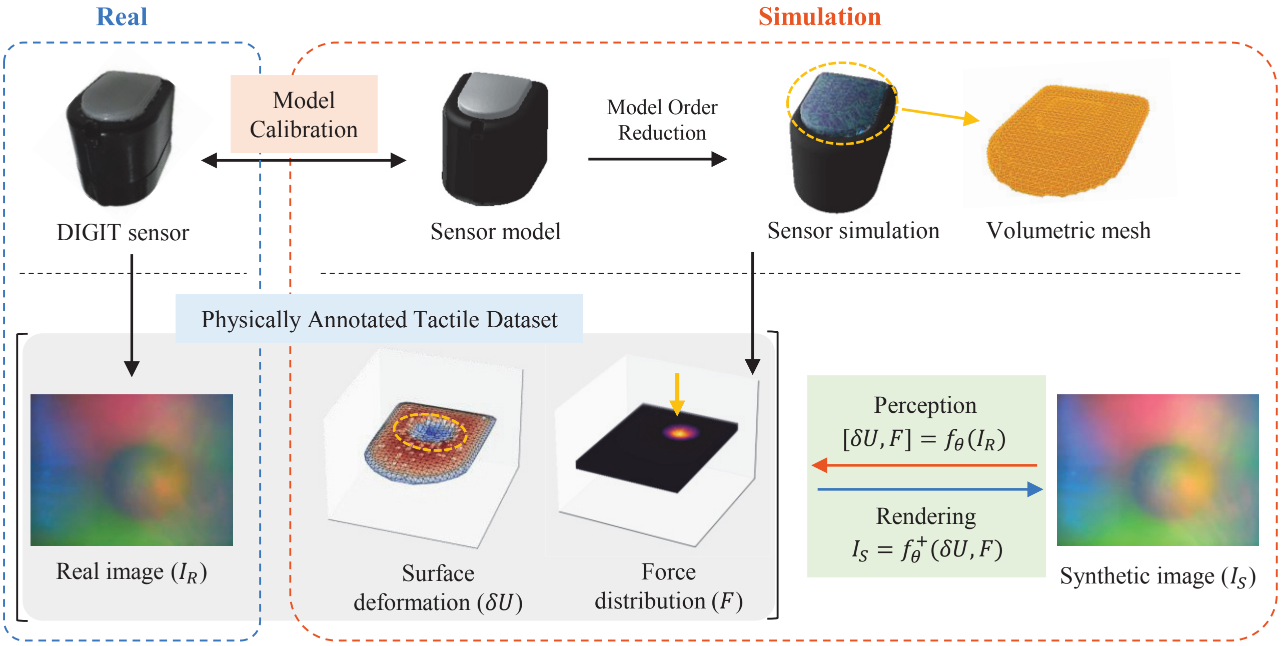

Fig. 1. Structure of the bidirectional framework linking real and simulated outputs from vision-based tactile sensors. A mirrored setup was used to collect paired data between physical indentation and simulation (gray region). Using this dataset, a perception model f_θ was trained to produce physically grounded annotations from RGB input, while a rendering model f_θ^+ was trained to render realistic RGB images from simulated physical data.

silicone gel tip. The sensor model employs a fine-resolution tetrahedral mesh to capture nodal displacements and contact forces at the spatial resolution of the physical device. To enable fast execution, model order reduction (MOR) [37], [38] was applied, achieving significant acceleration without loss of physical accuracy.

Identical contact conditions were reproduced in both the physical setup and the simulation to construct a paired dataset of real RGB images (I_R) and simulated physical annotations, including 3D surface deformation (U) and force distribution (F). This dataset was used to train two networks: a perception model (f_θ) that infers physically grounded states from I_R, and a rendering model (f_θ^+) that generates realistic RGB tactile images from simulated contact states. Together, these components enable real-time, bidirectional translation between physical and visual tactile domains, supporting physics-informed annotation and high-fidelity tactile image synthesis in accelerated simulation environments.

A. Sensor Modeling

Accurate FEM simulation of a ViTac sensor required precise modeling of both the deformable gel tip and its deformation under contact. Calibration was performed by collecting force–displacement measurements under quasi-static loading using a universal testing machine.

The 3D CAD models of the sensor and indenter were tetrahedralized to generate simulation meshes. The initial contact surface region was refined to improve numerical stability and reduce surface noise, and iterative Laplacian smoothing [39] was applied to the mesh vertices prior to simulation. The final remeshed geometry was used for FEM simulation.

To enable real-time execution without compromising physical accuracy, model order reduction (MOR) was applied using proper orthogonal decomposition (POD) [38]. Let $\mathbf{u}_z(t) \in \mathbb{R}^{N\times 1}$ denote the full-order normal displacement at time t, where N is the total number of FEM nodes, governed by:

$$\mathbf{K}(\mathbf{u_z}) = \mathbf{f}, \tag{1}$$

where $\mathbf{K}$ is the stiffness operator and $\mathbf{f}$ is the external force vector. A sequence of snapshot vectors $\mathbf{u}z, 0, \ldots, \mathbf{u}_{z,T-1}$ was assembled into:

$$\mathbf{U} = [\mathbf{u}_{z,0}, \ldots, \mathbf{u}_{z,T-1}] \in \mathbb{R}^{N\times T}. \tag{2}$$

Using singular value decomposition (SVD), the snapshot matrix was factorized as:

$$\mathbf{U} = \mathbf{\Phi}\mathbf{\Sigma}\mathbf{V}^\top, \tag{3}$$

and the top-r basis vectors in $\mathbf{\Phi} \in \mathbb{R}^{N\times r}$ were retained to form the reduced subspace. The displacement field was then approximated as:

$$\mathbf{u_z} \approx \mathbf{\Phi}\mathbf{q}, \tag{4}$$

where $\mathbf{q} \in \mathbb{R}^r$ is the reduced coordinate vector. Substituting into the original system yielded the reduced-order form:

$$\mathbf{\Phi}^\top\mathbf{K}\mathbf{\Phi}\mathbf{q} = \mathbf{\Phi}^\top\mathbf{f}. \tag{5}$$

This reduced system preserved dominant deformation behavior while significantly reducing computational cost.

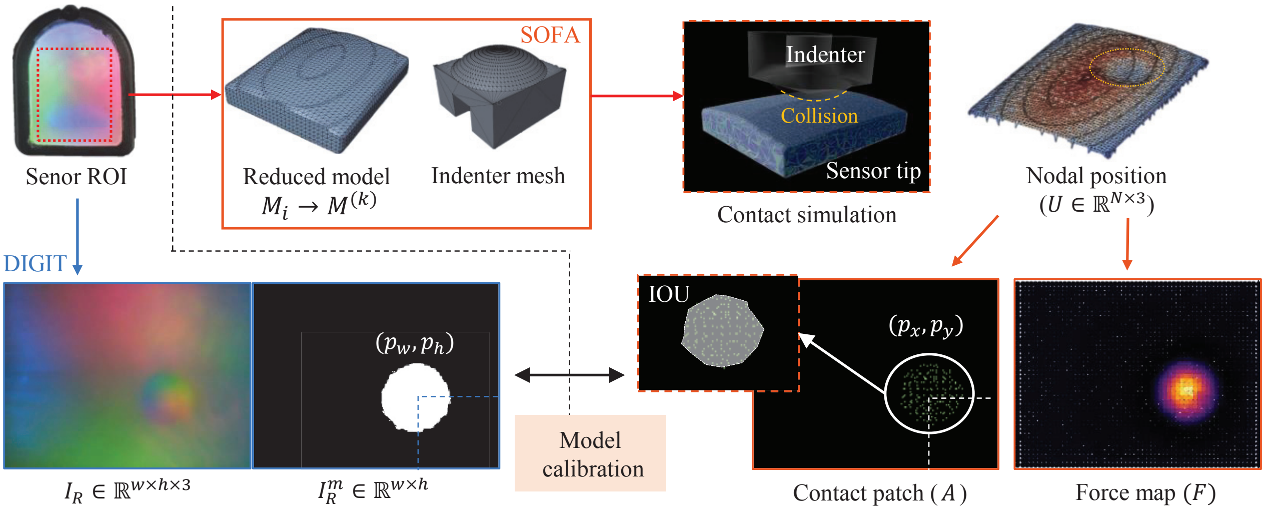

Fig. 2. Overview of the FEM model calibration process and evaluation metrics. Simulation in SOFA replicates real-world contact to generate surface deformation (U) and corresponding force distribution (F) for model calibration. From the simulated nodal displacement (U), the contact location (p), contact patch (A), and force map (F) are extracted and compared with the real sensor measurements.

As shown in Fig. 2, a custom SOFA simulation scene was constructed to reproduce the contact-induced deformation of a vision-based tactile sensor using the reduced FEM model. The sensor and indenter were represented as separate tetrahedral meshes, and the contact region of the sensor surface was remeshed to enhance resolution and stability. The reduced mesh ($M^{(k)}$) was derived via POD, enabling efficient simulation while preserving dominant deformation modes. During simulation, the indenter followed a perpendicular trajectory toward the sensor and the contact was resolved using collision detection and friction response. The resulting nodal displacement field (U) was used to extract contact location, contact patch (A), and force distribution ($F(U)$) to calibrate against real data.

B. *Calibration Metrics*

Calibration was performed by reproducing identical contact conditions with the physical DIGIT sensor and the FEM simulation, enabling a direct comparison of the geometric and physical responses.

1) Nodal Position Error: The FEM nodal positions were compared with the real measurements obtained from indentation experiments. The RMSE nodal position was computed as the mean Euclidean distance in millimeters between the corresponding node coordinates.

2) Contact Patch Geometry: The similarity of the contact shape was quantified by using the intersection over union (IoU) between binary contact masks as shown in Fig. 2. The real mask was obtained by subtracting the background and thresholding of I_R, and the simulated mask by thresholding the vertical displacement δU_z at the surface nodes. The contact centroid error was computed as the Euclidean distance between the centroids of the real and simulated masks, and the rotation error was computed as the angular difference between their normal z-axis.

3) Force-Depth Correspondence: The consistency of global contact force was evaluated from the force displacement curves measured in indentation trials. In simulation, the z-direction nodal contact force was computed from the calibrated Neo-Hookean material model:

$$W = \frac{\mu}{2}(\bar{I}_1 - 3) + \frac{\kappa}{2}(J-1)^2, \tag{6}$$

$$\sigma(U) = \frac{\mu}{J}(B - \mathbf{I}) + \kappa(J-1)\mathbf{I} \text{ and} \tag{7}$$

$$f_i^z(U) = \left[\sigma(U) \cdot \mathbf{n}_i\right]_z A_i, \tag{8}$$

where A_i is the lumped nodal surface area, $\mathbf{n}_i$ the outward surface normal, and μ, κ the shear and bulk moduli derived from E and ν. The reported metric is the RMSE between the global contact forces simulated and measured.

III. Bidirectional Networks

A. *Perception Network: Visual-to-Physical Displacement*

The perception model (Fig. 3-(a)) estimates a dense physical displacement field from an input RGB tactile observation. First, a background image (I_R^0) is subtracted from each tactile frame to remove variations in sensor-specific appearance, as the data set was collected from multiple DIGIT sensors. The resulting image $\Delta I \in \mathbb{R}^{3\times 240\times 320}$ is aligned with the fixed ROI of the sensor. For contact evaluation, a binary contact mask was derived from the FEM simulation by thresholding the vertical displacement of the nodes (δU_z) at the sensor surface. The contact mask provided a precise definition of the contact region.

The network architecture is a compact UNet encoder–decoder with three downsampling stages and symmetric upsampling via skip connections, using ReLU activations and batch normalization. The spatial resolution is progressively reduced during encoding to decrease the number of parameters and increase the receptive field, then

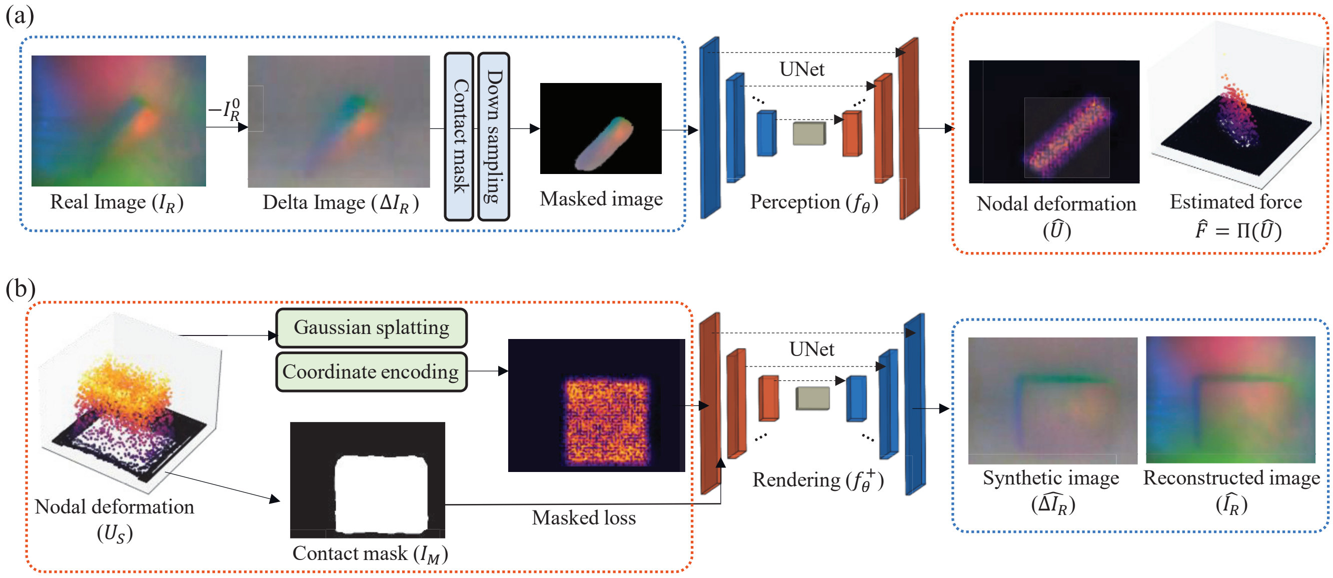
(a)
Real Image (I_R)
$-I_R^0$
Delta Image (ΔI_R)
Contact mask
Down sampling
Masked image
UNet
Perception (f_θ)
Nodal deformation ($\hat{U}$)
Estimated force $\hat{F} = \Pi(\hat{U})$
(b)
Gaussian splatting
Coordinate encoding
Nodal deformation (U_S)
Contact mask (I_M)
Masked loss
UNet
Rendering (f_θ^+)
Synthetic image ($\widehat{\Delta I}_R$)
Reconstructed image ($\hat{I}_R$)

Fig. 3. Mapping between physical states and RGB image representations. (a) Perception model (f_θ): given an input RGB image (I_R), the network predicts a nodal deformation field ($\hat{U}$), consisting of 3D displacement at each node. The output nodal deformation can be projected to form a spatial force map $\hat{F}$ using the sensor model. (b) Rendering model (f_θ^+): given a physical state input $U_S \in \mathbb{R}^{N\times 3}$, the network reconstructs a synthetic RGB image $\Delta\hat{I}_R$ that mimics the real sensor image.

restored in the decoder to produce a dense output. The model predicts a three-channel physical field $\hat{F} = [p_x, p_y, \tilde{d}_z]$, where p_x and p_y are normalized sensor coordinates and $\tilde{U}_z$ is the normalized out-of-plane displacement. The normal depth deformation was included in loss function using mean squared error (MSE):

$$\mathcal{L}_{\text{disp}} = \text{MSE}\big(\hat{U}_z, U_z\big), \tag{9}$$

while the p_x and p_y coordinate acted as an auxiliary structural context. The target displacement U_z is obtained by rasterizing the FEM nodal outputs onto the image grid using calibrated sensor bounds. Evaluation metrics for the perception network included contact IoU, contact centroid error, and contact rotation error.

B. Rendering Network: Physical-to-Visual Reconstruction

The rendering model (Fig. 3-(b)) generates an RGB delta image $\Delta\hat{I}_S$ from a multichannel physical input that includes displacement and 19 positional Fourier features. The architecture is also a lightweight UNet with two encoder stages, a bottleneck, and a symmetric decoder, using GELU activations and group normalization. Training uses a mask-weighted L1 reconstruction loss:

$$\mathcal{L} = \mathcal{L}_c + \lambda_b \mathcal{L}_b + \lambda_n \mathcal{L}_n, \tag{10}$$

with contact mask I_R^M derived from simulated δU_z. Rendering quality is evaluated with: (i) L1, the mean absolute pixel difference between predicted and ground truth images; (ii) Peak Signal-to-Noise Ratio (PSNR), computed as

$$\text{PSNR} = 20 \log_{10}\left(\frac{\text{MAX}}{\sqrt{\text{MSE}}}\right),$$

where MAX is the intensity range of the image and MSE is the mean squared pixel error used internally for the PSNR calculation; and (iii) Structural Similarity Index Measure (SSIM), which compares luminance, contrast and structural similarity to the ground truth image. The predicted delta image $\Delta\hat{I}_R$ can be converted to the final tactile image $\hat{I}_R$ by adding the sensor-specific background image I_R^0. The hyperparameters of both networks are summarized in Table I.

TABLE I

HYPERPARAMETERS OF THE TWO NETWORKS.

Component	**Perception** f_θ	**Rendering** f_θ^+
Input	ΔI_R	$\hat{U_S}$ [x, y, dz]
Output	$U_S\ [x, y, dz]$	ΔI_S
Architecture	UNet	UNet
Activations / Norm	ReLU + BatchNorm	GELU + GroupNorm
Loss Function	MSE (dz)	$\alpha_1 L_1 + \alpha_2 (1 -$ SSIM)
Optimizer	Adam (1e$-$5)	Adam (1e$-$5)

IV. Experiments

A. Calibration

As shown in Fig. 4-(a), a planar stage was used to precisely align the indenter with the region of interest (ROI) on the sensor for force–displacement measurements. Indentation tests were performed at 100 uniformly distributed locations using a universal testing machine (34SC, Instron) in quasi-static mode, with a constant indentation speed of 0.01 mm/s, a maximum depth of 1 mm, force resolution of 10 mN, and displacement accuracy of 20 μm. At each location, a complete indentation cycle was recorded to obtain a force–displacement curve. Preliminary speed tests indicated that indentation speeds exceeding 0.5 mm/s introduced viscoelastic and dynamic effects; all calibration experiments were therefore conducted below this threshold.

B. Data Acquisition

Real-world indentation tests were conducted with 10 different indenter geometries as shown in Fig. 4-(b). For each contact, indentation was performed in 100 depth increments of 10 μm, up to a maximum depth of 1 mm. The maximum depth was chosen to ensure the contact force remained within the safe operating limit of the sensor, up to 15 N. To increase dataset diversity, asymmetric indenters (e.g., maze, pyramid, square) were rotated between trials to produce varied contact patterns, while symmetric indenters (e.g., point, donut, sphere) were varied only in contact position, as shown in Fig. 4-(b). To account for sensor-to-sensor variations, the experiments were performed using four different DIGIT sensors. In total, 15,000 real contact images were acquired for training. Corresponding FEM simulations were run under identical contact positions and orientations, yielding a paired dataset in which each real RGB image is matched with a physically annotated label containing nodal deformation and contact pressure. A total of 15 percent of the data was separated solely for the validation test used for inference.

V. Results

A. Calibration precision

To validate the realism and fidelity of the FEM simulation of the DIGIT sensor, calibration experiments were conducted by reproducing identical indentation sequences in both the physical and simulated setups. The calibration results in Table II report *Force RMSE* and *Nodal position error* across the entire ROI for complete indentation experiments.

The global force error between the simulated and measured force–displacement curves was 0.20 N on average, with a maximum deviation not exceeding 0.30 N. The errors were smallest in the central region of the sensor contact area and increased slightly toward the edges, which is attributed to the convex geometry of the silicone gel tip. The nodal position RMSE, computed as the average Euclidean distance between the simulated and measured nodal coordinates on the ROI, was 0.192 mm. All reported values are averaged over the ten indenter geometries, with the mean and standard deviation calculated from the set of 100 uniformly distributed indentation points per indenter. The SOFA-based reduced-order FEM simulation runs at approximately 30 frame per second (FPS) on a 32 core 64 thread CPU.

B. Perception Network Validation

The perception network was evaluated on the held-out validation dataset across all four sensors and ten indenter geometries. The metrics included both contact location and orientation errors. The similarity of the contact shape was quantified using the IoU metric calculated between the real and simulated contact patches, yielding a mean IoU of 0.85 ± 0.05 across all indenters. The centroid position error in the contact was 0.458 ± 0.078 mm and the rotation error in the contact shape was 0.047 ± 0.01 rad.

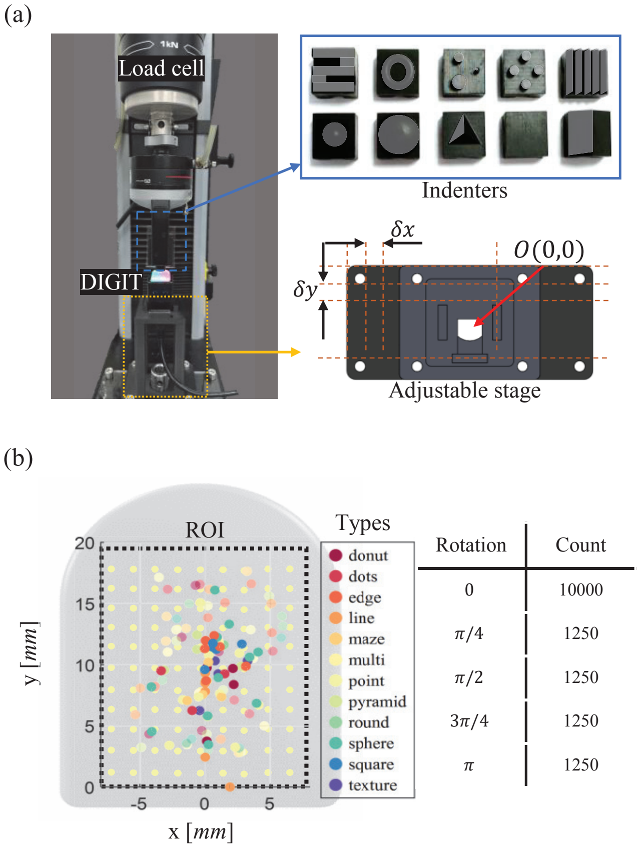

Fig. 4. (a) Experimental setup using a universal testing machine and ten different indenter tips to press against the vision-based tactile sensor. The tips represent a variety of geometric shapes. (b) Contact ROI on the sensor surface, along with the number of rotated indenter configurations used for data collection. For asymmetric indenters, rotations were also applied to collect various contact patch.

TABLE II

Summary of Results

Domain	Metric	Value
FEM Calibration	Force RMSE [N]↓ Nodal position [mm]↓	0.20 ± 0.030 0.192 ± 0.017
Network Perception	Contact IoU↑ Contact centroid [mm]↓ Contact rotation [rad]↓ Surface force [N]↓	0.85 ± 0.055 0.458 ± 0.078 0.047 ± 0.014 0.11 ± 0.037
Rendering Quality	L1↓ SSIM↑ PSNR↑	13.59 ± 1.03 0.971 ± 0.02 39.13 ± 1.81

Although the network does not directly measure force, the predicted nodal displacements can be used with the calibrated sensor model (Eq. 8) to compute the corresponding contact forces. Using this approach, the surface force RMSE was 0.11 ± 0.03 N.

These results show that the perception network can reliably recover both the spatial structure and the physical re-

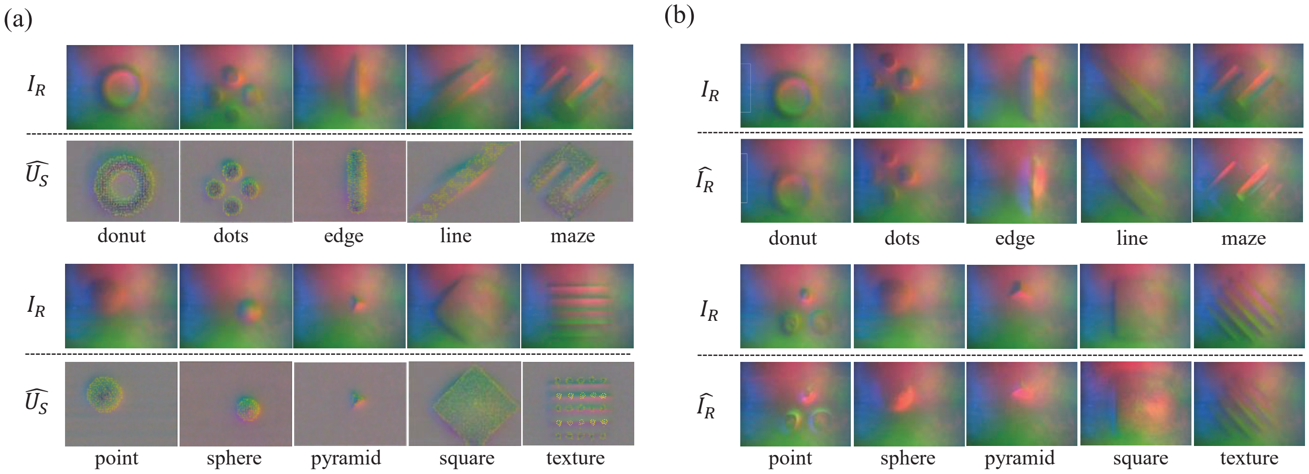

Fig. 5. (a) Visual examples of perception network. The network predicted deformation field $\hat{U}$ and corresponding force distribution for various indenters. These positions and rotations were not included data in the training. (b) Visual examples of rendering results showing synthetic tactile images $\hat{I}_R$ compared with the real image I_R generated from deformation inputs $\hat{U}_S$ for various indenter geometries.

sponse of real-world contacts from visual tactile input. Fig. 5-(a) illustrates example predictions, showing the reconstructed nodal displacement fields and corresponding force distribution for unseen dataset. The model reproduces both global contact geometry and the localized pressure distribution, as well as the accurate force estimation. The perception model processes over 1,000 images per second on the RTX 5090 (32G VRAM), allowing rapid batch inference or real-time deployment in high-throughput tactile perception pipelines.

C. Rendering Network Validation

The rendering network was evaluated on the validation data set using paired FEM-RGB data. Given multi-channel physical inputs from the FEM simulation, the network synthesized delta RGB images ($\Delta\hat{I}_R$) of the contact imprint.

Quantitative evaluation was performed using three standard image similarity metrics. The mean absolute error (L1) between the predicted and ground-truth images was 13.59±1.03 (pixel intensity scale 0–255). The SSIM reached 0.971 ± 0.02, indicating very high similarity in luminance, contrast, and structural features. The PSNR was 39.13 ± 1.81 dB, reflecting low overall reconstruction error and high perceptual quality.

Fig. 5-(b) presents representative examples of rendered images alongside their ground truth, showing that the network accurately reproduces both the global contact shape and fine local texture details of the tactile imprint. These results demonstrate that the rendering network can generate high-fidelity synthetic tactile images from physically grounded FEM outputs, enabling realistic visual augmentation for simulation-based learning. The rendering network achieves an inference speed of approximately 220 images per second on the same GPU, supporting fast generation of high-fidelity tactile images for simulation-based training.

D. Generalization to Unseen Data

Beyond controlled quantitative evaluation, both networks were tested in scenarios that differed from training conditions to assess their generalization capabilities (Fig. 6).

1) Perception network on external dataset: The perception model was evaluated on samples from the YCB object set [40], provided through the TACTO simulation framework [23]. In this setting, the lighting and shading conditions differed from those in the training dataset, and no background image I_R^0 was available for subtraction. As shown in Fig. 6-(a), despite these differences, the network successfully segmented the contact region and reconstructed the corresponding height map of the nodal deformation, producing plausible 3D surface estimates directly from raw RGB input.

2) Rendering network on unseen indenters: The rendering model was tested with complex indenter geometries that were not present in the training set. These indenters were created as 3D meshes and simulated in the FEM environment, producing nodal deformation fields without any corresponding real images. As shown in Fig. 6-(b), the rendering network synthesized realistic delta RGB images for these novel contact shapes, accurately capturing their global geometry. Although minor artifacts were visible in certain fine-scale details, overall shapes were rendered consistently, demonstrating the potential for generating tactile imagery from purely virtual contact scenarios.

3) Rendering sensor in Physics-based simulation: As illustrated in Fig. 7, the rendering network can be seamlessly integrated into physics-based robotic simulation pipelines. Given a simulation mesh of the ViTac sensor, local contact information can be projected onto the nodal grid and passed through the rendering network to generate realistic DIGIT sensor images in real time. This capability enables scalable augmentation of training data for contact-rich manipulation,

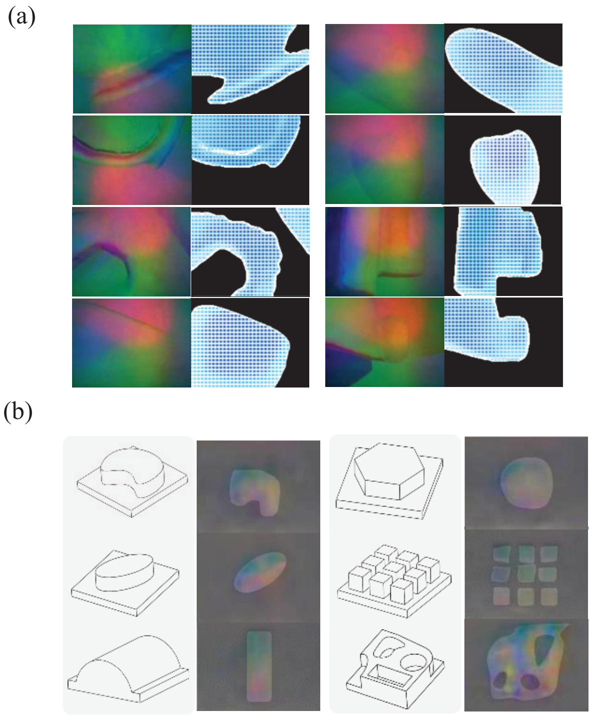

Fig. 6. Applications of the learned bidirectional model beyond quantitative evaluation. (a) Perception network inference of surface deformation using RGB images collected from real-world objects not seen during training. (b) Rendering results from text-based indentation experiments, showing high-fidelity RGB outputs for various indenter shapes.

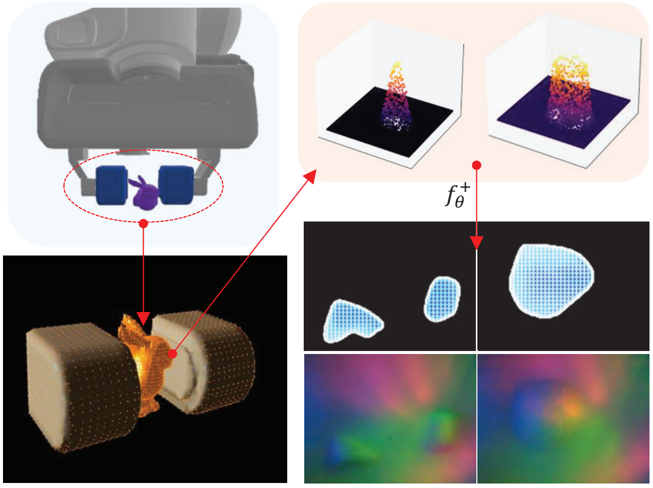

Fig. 7. Rendering based on high-resolution contact data from a physics-based simulator. Deployment on a robot arm equipped with dual DIGIT sensors in environment. The rendering network synthesizes realistic RGB outputs during dynamic grasping

such as in-hand control, tool use or fine assembly tasks. By preserving the physical grounding of contact geometry and force distribution, the generated images can be used directly for training perception or control policies, significantly reducing the need for labor-intensive real-world data collection.

VI. Discussion

This work presents a bidirectional framework that integrates a calibrated finite element simulation with UNet-based networks to reconcile real ViTac measurements with physics-based simulation. The framework includes a perception model that regresses to a dense three-dimensional deformation field $\hat{U}$ from high-resolution RGB and a rendering model that synthesizes realistic tactile images from simulated physical states. Coupling both directions within one dataset and an architecture enables large-scale synthetic data generation and automated physical annotation of unlabeled real images, creating a closed simulation–reality loop.

The objective of perception is to assess dense deformation regression, unlike Sim2Surf [41] for surface classification and GenForce [20] for the estimation of three axes of force. The closest objective is SimTacLS [28], which reconstructs skin shape from tactile images, while the present framework emphasizes a calibrated sim-to-real pipeline. On the simulation side, TACTO [23] and TacSL [31] prioritize scalability by approximating soft contact with rigid-body or penalty methods, which do not explicitly resolve nonlinear elastomer deformation. An FEM-based approach, similar to DiffTactile [29], captures calibrated hyperelastic behavior from real data.

The FEM model is calibrated with force–displacement measurements from a real sensor to produce paired states that are visually and physically representative, reducing the gap before learning so that training uses physically grounded supervision.

From $\hat{U_S}$, the contact geometry and forces are derived and evaluated against the calibrated ground truth. Under the stated protocol, the results include the 0.458 mm centroid error, the 0.047 rad rotation error, and the 0.11 N force RMSE. The mean errors, confidence intervals, coordinate frames, alignment, and sample counts are documented in the Github page description. Checkpoints and evaluation scripts are provided for exact reproduction.

Rendering FEM states into RGB images achieves SSIM 0.971 and PSNR 39.13 on a held-out test set. Inference reaches 220 FPS on an RTX 5090 with batch size 8 at native resolution. Because prior work uses different sensors, datasets, and metrics, claims of superiority are limited to shared evaluations under identical conditions. For Taxim [25], comparisons use controlled reimplementations when SSIM or PSNR are unavailable on the present data.

The release materials include data acquisition scripts, FEM setup and calibration utilities, trained networks, paired datasets, and evaluation code at `https://github.com/ndolphin-github/DIGIT_simulation.git`. The current FEM assumes quasistatic contact and a fixed hyperelastic law; rate-dependent effects, sensor transfer with minimal recalibration, and edge deployment via compression remain important directions.

References

[1] A. O'Neill, A. Rehman, A. Maddukuri, A. Gupta, A. Padalkar, A. Lee, A. Pooley, A. Gupta, A. Mandlekar, A. Jain *et al.*, "Open

x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6892–6903.

[2] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison, "Rlbench: The robot learning benchmark & learning environment," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3019–3026, 2020.

[3] A. Brohan, N. Brown, J. Carbajal, Y. Chebotar, J. Dabis, C. Finn, K. Gopalakrishnan, K. Hausman, A. Herzog, J. Hsu *et al.*, "Rt-1: Robotics transformer for real-world control at scale," *arXiv preprint arXiv:2212.06817*, 2022.

[4] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" *arXiv preprint arXiv:1710.05512*, 2017.

[5] H. Van Hoof, T. Hermans, G. Neumann, and J. Peters, "Learning robot in-hand manipulation with tactile features," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2015, pp. 121–127.

[6] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[7] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—a review," *Sensors and Actuators A: physical*, vol. 167, no. 2, pp. 171–187, 2011.

[8] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess *et al.*, "Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation," *Science Robotics*, vol. 9, no. 96, p. eadl0628, 2024.

[9] C. Y. Wong and W. Suleiman, "Sensor observability index: Evaluating sensor alignment for task-space observability in robotic manipulators," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1276–1282.

[10] I. Sorrentino, G. Romualdi, and D. Pucci, "Ukf-based sensor fusion for joint-torque sensorless humanoid robots," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 150–13 156.

[11] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[12] A. C. Abad and A. Ranasinghe, "Visuotactile sensors with emphasis on gelsight sensor: A review," *IEEE Sensors Journal*, vol. 20, no. 14, pp. 7628–7638, 2020.

[13] S. Zhang, Z. Chen, Y. Gao, W. Wan, J. Shan, H. Xue, F. Sun, Y. Yang, and B. Fang, "Hardware technology of vision-based tactile sensor: A review," *IEEE Sensors Journal*, vol. 22, no. 22, pp. 21 410–21 427, 2022.

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

[15] S. Wang, Y. She, B. Romero, and E. Adelson, "Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6468–6475.

[16] C. Higuera, A. Sharma, C. K. Bodduluri, T. Fan, P. Lancaster, M. Kalakrishnan, M. Kaess, B. Boots, M. Lambeta, T. Wu *et al.*, "Sparsh: Self-supervised touch representations for vision-based tactile sensing," *arXiv preprint arXiv:2410.24090*, 2024.

[17] R. Sui, L. Zhang, T. Li, and Y. Jiang, "Incipient slip detection method for soft objects with vision-based tactile sensor," *Measurement*, vol. 203, p. 111906, 2022.

[18] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, "Touch and go: Learning from human-collected vision and touch," *arXiv preprint arXiv:2211.12498*, 2022.

[19] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[20] Z. Chen, N. Ou, X. Zhang, Z. Wu, Y. Zhao, Y. Wang, N. Lepora, L. Jamone, J. Deng, and S. Luo, "General force sensation for tactile robot," *arXiv preprint arXiv:2503.01058*, 2025.

[21] E. Su, C. Jia, Y. Qin, W. Zhou, A. Macaluso, B. Huang, and X. Wang, "Sim2real manipulation on unknown objects with tactile-based reinforcement learning," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 9234–9241.

[22] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6437–6443.

[23] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022.

[24] Q. Liu, Y. Cui, Z. Sun, G. Li, J. Chen, and Q. Ye, "VTDexmanip: A dataset and benchmark for visual-tactile pretraining and dexterous manipulation with reinforcement learning," in *The Thirteenth International Conference on Learning Representations*, 2025.

[25] Z. Si and W. Yuan, "Taxim: An example-based simulation model for gelsight tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2361–2368, 2022.

[26] Y. Zhao, K. Qian, B. Duan, and S. Luo, "Fots: A fast optical tactile simulator for sim2real learning of tactile-motor robot manipulation skills," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5647–5654, 2024.

[27] D. H. Nguyen, T. Schneider, G. Duret, A. Kshirsagar, B. Belousov, and J. Peters, "Tacex: Gelsight tactile simulation in isaac sim–combining soft-body and visuotactile simulators," *arXiv preprint arXiv:2411.04776*, 2024.

[28] Q. K. Luu, N. H. Nguyen *et al.*, "Simulation, learning, and application of vision-based tactile sensing at large scale," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 2003–2019, 2023.

[29] Z. Si, G. Zhang, Q. Ben, B. Romero, Z. Xian, C. Liu, and C. Gan, "Difftactile: A physics-based differentiable tactile simulator for contact-rich robotic manipulation," *arXiv preprint arXiv:2403.08716*, 2024.

[30] M. Lambeta, H. Xu, J. Xu, P.-W. Chou, S. Wang, T. Darrell, and R. Calandra, "Pytouch: A machine learning library for touch processing," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 208–13 214.

[31] I. Akinola, J. Xu, J. Carius, D. Fox, and Y. Narang, "Tacsl: A library for visuotactile sensor simulation and learning," *arXiv preprint arXiv:2408.06506*, 2024.

[32] L. Van Duong *et al.*, "Large-scale vision-based tactile sensing for robot links: Design, modeling, and evaluation," *IEEE Transactions on Robotics*, vol. 37, no. 2, pp. 390–403, 2020.

[33] L. Zhang, T. Li, and Y. Jiang, "Improving the force reconstruction performance of vision-based tactile sensors by optimizing the elastic body," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 1109–1116, 2023.

[34] G. M. Caddeo, A. Maracani, P. D. Alfano, N. A. Piga, L. Rosasco, and L. Natale, "Sim2real bilevel adaptation for object surface classification using vision-based tactile sensors," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 15 128–15 134.

[35] C. Higuera, B. Boots, and M. Mukadam, "Learning to read braille: Bridging the tactile reality gap with diffusion models," *arXiv preprint arXiv:2304.01182*, 2023.

[36] F. Faure, C. Duriez, H. Delingette, J. Allard, B. Gilles, S. Marchesseau, H. Talbot, H. Courtecuisse, G. Bousquet, I. Peterlik *et al.*, "Sofa: A multi-model framework for interactive physical simulation," *Soft tissue biomechanical modeling for computer assisted surgery*, pp. 283–321, 2012.

[37] O. Goury, B. Carrez, and C. Duriez, "Real-time simulation for control of soft robots with self-collisions using model order reduction for contact forces," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3752–3759, 2021.

[38] O. Goury and C. Duriez, "Fast, generic, and reliable control and simulation of soft robots using model order reduction," *IEEE Transactions on Robotics*, vol. 34, no. 6, pp. 1565–1576, 2018.

[39] G. Taubin, "A signal processing approach to fair surface design," in *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, 1995, pp. 351–358.

[40] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*. IEEE, 2015, pp. 510–517.

[41] G. M. Caddeo, A. Maracani, P. D. Alfano, N. A. Piga, L. Rosasco, and L. Natale, "Sim2surf: A sim2real surface classifier for vision-based tactile sensors with a bilevel adaptation pipeline," *IEEE Sensors Journal*, 2025.