Look-to-Touch: A Vision-Enhanced Proximity and Tactile Sensor for Distance and Geometry Perception in Robotic Manipulation

Yueshi Dong, Jieji Ren, Zhenle Liu, Zhanxuan Peng, Zihao Yuan, Ningbin Zhang, and Guoying Gu, *Senior Member, IEEE*

***Abstract*—Camera-based tactile sensors provide robots with a high-performance tactile sensing approach for environment perception and dexterous manipulation. However, achieving comprehensive environmental perception still requires cooperation with additional sensors, which makes the system bulky and limits its adaptability to unstructured environments. In this work, we present a vision-enhanced, camera-based dual-modality sensor that integrates proximity and tactile sensing, enabling accurate long-distance proximity sensing while simultaneously maintaining ultra-high-resolution texture sensing and reconstruction capabilities. Unlike conventional designs with fixed opaque gel layers, our sensor features a partially transparent sliding window, enabling mechanical switching between tactile and visual modes. For each sensing mode, a dynamic distance sensing model and a contact geometry reconstruction model are proposed. Through integration with soft robotic fingers, we systematically evaluate the performance of each mode, as well as in their synergistic operation. Experimental results show robust distance tracking across various speeds, nanometer-scale roughness detection, and submillimeter 3D texture reconstruction. The combination of both modalities improves the robot's efficiency in executing grasping tasks. Furthermore, the embedded mechanical transmission in sensor allows for fine-grained intrahand adjustments and precise manipulation, unlocking new capabilities for soft robotic hands.**

***Index Terms*—Dual-modality sensor, fine-texture reconstruction, in-hand manipulation, intelligent grasping, proximity sensing, soft robotic hand, tactile sensing.**

I. Introduction

AMONG various exteroceptive modalities, tactile sensors play a crucial role in enabling robots to perceive contact-related information, such as pressure, surface texture, and object geometry [1], [2], [3]. When embedded in robotic fingertips or palms, these sensors allow robots to adapt to uncertainties in object shape, surface friction, and contact force during manipulation tasks [4]. Over the past decade, a wide variety of tactile sensors have been developed based on diverse transduction mechanisms, including capacitive [5], [6], piezoelectric [7], [8] and fiber Bragg grating [9], etc. While offering high sensitivity and resolution, their practical deployment is often hindered by challenges, such as complex wiring, signal crosstalk, and limited scalability, especially in soft or highly integrated robotic systems [10], [11]. In response to these limitations, vision-based tactile sensing (VBTS) has emerged as a promising alternative [12]. By embedding a camera within an enclosed space and illumination of structured or directional lighting, VBTS enables the extraction of high-fidelity tactile information on a soft contact layer by interpreting visual changes in the contact surface. This approach allows simultaneous acquisition of surface geometry, contact force distribution, and fine textures [13], [14], [15], [16], making VBTS one of the most widely adopted tactile sensing methods in the field of dexterous manipulation.

While VBTS provides detailed contact information, a single modality is often insufficient for achieving comprehensive environmental perception in robotic operations. To overcome this limitation, robots integrate data from peripheral sensors—such as proximity sensors and external cameras—to enhance spatial awareness and task adaptability. Among them, capacitive [17], [18] proximity sensors are widely used for short-range detection and dynamic distance estimation. However, their sensing performance is highly susceptible to environmental factors, such as the refractive index and ambient light, often leading to compromised accuracy and stability. Cameras, on the other hand, offer a more robust and versatile sensing modality—including object color, contour, and depth—across a broader field of view. Nevertheless, visual occlusion remains a critical challenge [17].

Received 26 March 2025; revised 1 August 2025 and 14 October 2025; accepted 11 December 2025. Recommended by Technical Editor M. Selvaggio and Senior Editor K. Oldham. This work was supported in part by the National Natural Science Foundation of China under Grant 52025057, Grant 52305029, and Grant 52505029 and in part by the Science and Technology Commission of Shanghai Municipality under Grant 24511103400 and Grant 25ZR1401191. *(Corresponding author: Guoying Gu.)*

The authors are with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China, and also with the Shanghai Key Laboratory of Intelligent Robotics, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: dongys525@sjtu.edu.cn; jiejiren@sjtu.edu.cn; extraordinary@sjtu.edu.cn; steven2606@sjtu.edu.cn; yuanzihao@sjtu.edu.cn; zhangnb@sjtu.edu.cn; guguoying@sjtu.edu.cn).

This article has supplementary material provided by the authors and color versions of one or more figures available at https://doi.org/10.1109/TMECH.2025.3645239.

Digital Object Identifier 10.1109/TMECH.2025.3645239

TABLE I
COMPARISON OF REPRESENTATIVE PROXIMITY–TACTILE DUAL-MODALITY SENSORS

Related Works	Mode Switching	Proximity Method	Tactile Method	Notes
Shimononura *et al.* [22]	Passive	Binocular cameras	Infrared TIR	At least three cameras
Spectac [23]	Passive	MC	UV fluorescent markers	UV LED keeps ON/OFF
Finger-STS [24]	Passive	MC (VF) - Object detection	Optical flow (object + markers)	Task-specific training model
TIRgel [25]	Passive	MC (VF) - Object detection	Visible-light TIR	LED-controlled brightness contrast
This work	Active (servo motor)	MC (VF) - (Segmentation + MDE)	RGB PS	High-quality dual-modal sensing

Abbreviations: MC: Monocular cameras; VF: Variable focus; MDE: Monocular Depth Estimation; TIR: Total internal reflection; PS: Photometric stereo.

To mitigate this, researchers have proposed multicamera setups with optimized placement strategies [19], [20], [21]. As a result, such large-scale perception systems often struggle to adapt to unstructured environments, limiting the practical deployment in real-world robotic tasks. These challenges highlight the pressing need for a compact, multimodal sensing framework.

Therefore, several recent studies have attempted to enhance the perception capability of a single VBTS unit, as summarized in Table I. These efforts aim to unify proximity and tactile sensing within a single hardware platform to reduce system complexity. For instance, Shimonomura et al. [22] incorporated an infrared total internal reflection (TIR) setup within a transparent flexible contact layer, determining the sensing mode based on the presence of infrared patterns in raw camera images, and enabled adaptive grasping with a 6-degree of freedom (DOF) robotic arm. Wang et al. [23] introduced a design with continuous UV lamp switching, identifying whether the sensor was in contact with the target by comparing the movement of marker points across each frame of the video stream. However, both designs rely on passive mode switching triggered by contact, limiting their ability to perform proactive perception in complex environments. Hogan et al. [24] trained a network to enable the measurement of "approaching distance" for a specific target and effectively separated two modalities in continuous image frames captured by the camera. While this design allows simultaneous perception of external object depth and internal contact force, it sacrifices detailed surface characterization due to optical complexity. As another representative work, Zhang et al. [25] proposed TIRgel, a vision-based tactile sensor that leverages TIR within a transparent elastomer and achieves modality switching via focus adjustment. However, it remains limited by the lack of explicit distance estimation capability and passive mode conversion based on optical parameters. The design of the dual-mode sensor that integrates long-distance proximity sensing and high-quality contact sensing has long remained a persistent challenge.

In this article, we propose a vision-enhanced proximity–tactile sensor that achieves active and unified dual-modality perception through a mechanically driven rotatable belt mechanism (see Fig. 1). By incorporating a partially opaque sliding window, it can actively extend or retract an elastic opaque layer as needed. During tactile sensing mode, the elastic layer will be screwed out to serve as the outermost contact interface. By leveraging a photometric stereo (PS) algorithm [26], the detailed contact surface geometry can be reconstructed. In the proximity sensing mode, the exposed zoom camera captures visual data and enables robust long-range target tracking through the

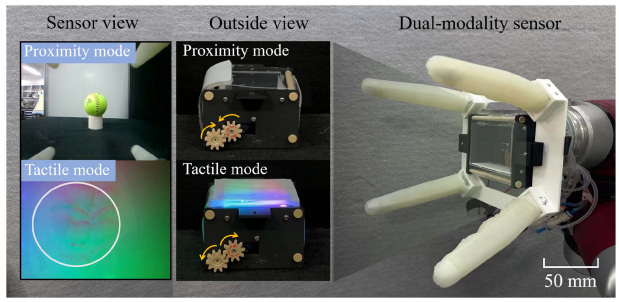

Fig. 1. Sensing modes and integrated application of the proposed sensor. Left: sensor outputs and configurations under two sensing modes. Right: integration with soft robotic fingers as a visual-tactile palm (V-T PALM).

combination of monocular depth estimation [27] and segmentation, while avoiding the need for multicamera calibration and complex occlusion processing. By tightly integrating high-fidelity tactile reconstruction and long-range proximity perception into a single compact module, our design offers a scalable, efficient, and adaptable sensing solution for real-world robotic grasping and interaction tasks.

The experimental results show an accurate distance measurement across various targets under different speeds in real-time conditions. Furthermore, it can achieve a nanometer-scale roughness resolution and sub-millimeter texture reconstruction. The combination of both modalities improves the efficiency of robot in executing grasping tasks. Meanwhile, the transmission mechanism endows the sensor with an additional DOF for in-hand manipulation, which was experimentally validated through a card-insertion task, demonstrating the feasibility of precise adjustments within a soft gripper. The contributions of this work are threefold as follows.

1) *Modular Dual-Modality Sensor:* A dual-modality sensing prototype is introduced, seamlessly integrating proximity and tactile sensing to support robotic operation scenarios.
2) *High-Quality Sensing Ability:* achieve long-distance proximity perception while maintaining high-resolution tactile texture reconstruction.
3) *Fine-tuning Operation Capability:* A sliding sensing window is coupled to the sensor, enabling real-time feedback and in-grasp pose adjustment for the soft robotic hand.

II. DESIGN AND FABRICATION

The overall design of the Visual–Tactile Palm (V-T PALM) (see Fig. 2) can be broadly divided into two main components: the modal switching module and the sensing module.

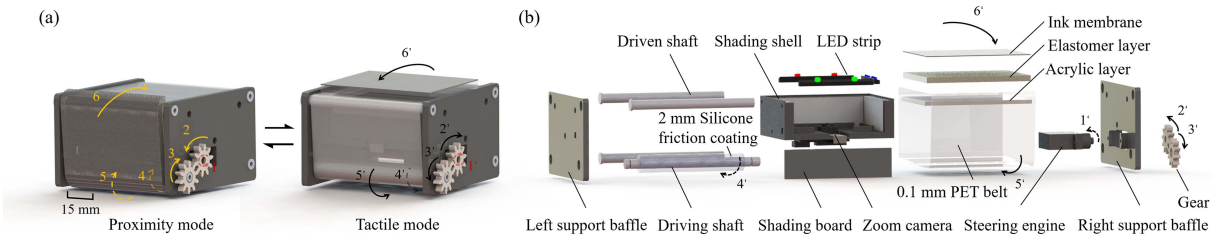

Fig. 2. Structural design of the V-T PALM sensor. (a) Assembly views in proximity and tactile modes, with annotated arrows indicating the internal actuation sequence during bidirectional mode switching. (b) Detailed exploded view illustrating the component layout of the sensor.

A. Dual-Mode Switching Module

The rolling switching module is a key focus of this work. Although several effective rolling plane design approaches have been proposed [3], [28], [29], ensuring that the transmission belt accommodates both opaque and fully transparent regions while maintaining stable frictional transmission remains a significant design challenge. The entire transmission mechanism consists of two baffles, a driving roller, three passive rolling shafts, a conveyor belt measuring 180 × 55 mm, a 360° steering gear, and two gears with a 1:1 transmission ratio. Most of these are fabricated using 3-D printing (X1Carbon, Bambu Lab). The conveyor belt is partially covered by an ink-membrane-coated elastomer layer, which is adhered by silicone adhesive (Silpoxy, Ecoflex). The driving shaft is coated with a 2-mm thick silicone layer (PDMS 0030) and driven by the steering gear. The adhesive friction and intermolecular forces generated by the silicone ensure strong and stable coupling with the belt, enabling consistent friction-based transmission. The transition between the two modes follows a symmetric rotational process, as shown in Fig. 2(a). Taking the transition from the proximity mode to the tactile mode as an example, when receiving a command: the internal servo motor rotates clockwise (1), driving a 1:1 gear train (2–3), which in turn rotates the driving shaft (4). The silicone coating on the driving shaft ensures no-slip rotation of the outermost belt (5), thereby driving the ink-membrane-covered elastomer layer into its new position (6). The entire process takes an average of 0.78 s.

To identify the optimal transmission belt material, we conducted a set of comparative experiments using polyethylene terephthalate (PET) belts (0.1 mm, 0.2 mm, 0.3 mm) and thermoplastic polyurethane (TPU) belts (0.3 mm, 0.5 mm, 0.65mm). All samples were mounted on the same sensor prototype to 20 full mode-switching cycles. Belt slippage was quantified by measuring the cumulative deviation of a marked reference point from its expected position. As shown in Fig. 4(a), the 0.1-mm PET belt showed the best performance with minimal slippage under repeated switching cycles, and was thus selected as the base material for the conveyor belt in this design. In addition, a hook-and-loop fastener is installed at the end of the strip, allowing for convenient retensioning of the conveyor belt as needed. When the distance between the target and the palm is detected as less than the threshold (e.g., 10 cm), a command is executed to rotate the elastic contact module above the camera, activating the tactile measurement mode. Conversely, when the module is retracted, the high transparency of the PET material ensures that the camera's ability to capture the external environment remains unaffected, enabling the proximity measurement mode. The operating principles of the two sensing modes are illustrated in Fig. 3. The power supply control of the LED lamp and the steering gear's control signals are coordinated and transmitted by an Arduino board (Arduino Mega 2560, Arduino SRL), following the output commands from the algorithm running on the host computer.

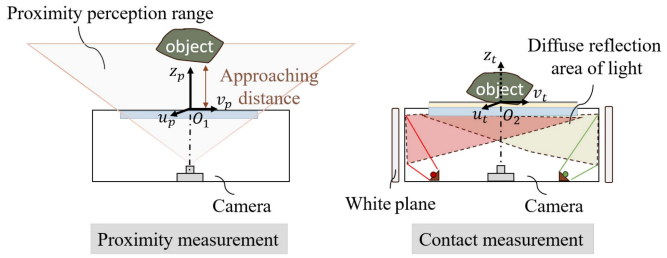

Fig. 3. Measurement principle. Proximity mode captures external target images for distance sensing; tactile mode employs internal illumination to extract contact information.

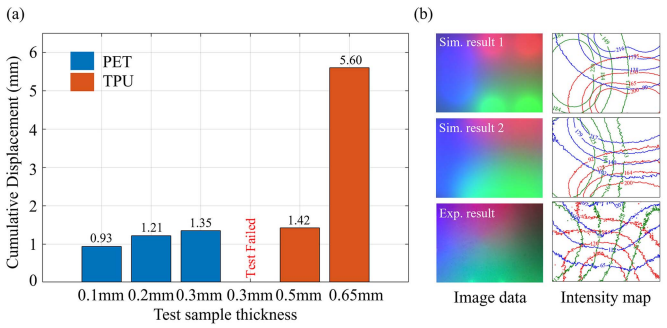

Fig. 4. Experimental validation of key design parameters for the sensor system. (a) Displacement evaluation of PET and TPU belts under repeated switching cycles. (b) Comparison of different illumination methods: the first two columns show the light intensity distribution of direct and diffuse reflection schemes rendered in Blender; the third column illustrates the actual illumination setup of the sensor.

B. Sensor Module

Both measurement modes share the same design foundation. A monocular zoom camera (OV5640) with an effective viewing angle of 120 degrees is installed inside the sensor, capturing real-time RGB images at a resolution of 5 megapixels and a frame rate of 30 frames per second (fps). The zoom lens features both active and passive focusing capabilities, enabling clear

imaging of target objects at varying distances from the camera. On the top of the sensor is a transparent acrylic plate with a thickness of 3 mm, which provides a reliable support strength of 0–200 kPa for the flexible contact measurement layer of the sensor in addition to providing a light refractive index close to air and PET material, provides a reliable support strength for the stable grasping task. The inner cavity of the sensor is 3-D printed by black Polylactic Acid material, which provides a stable inner space for the optical coding of tactile information. A strip structure embedded with RGB LED strip is designed at the lower edge of the bottom near the three side walls so that the plane normal vector of the lamp strip directly faces the three sides. Referring to the principle of diffuse reflection, we applied an even coating of white paint to the three walls. This design enables the red, green, and blue LED point light sources to transform into surface light sources through reflection, thereby reducing the influence of concentrated beam refraction between different light-transmitting materials and creating more uniform lighting conditions for tactile imaging. Using Blender simulations, we compared our method with the lighting effects produced by direct point light source illumination, as shown in Fig. 4(b). The results demonstrate that the light intensity distribution across the lens is significantly more uniform under the proposed illuminated solution, the effective sensing area is also further expanded, which enhances the quality of raw sensing image data.

III. Method

To achieve high-quality single object perception in both sensing states, we proposed a distance measurement framework and a contact texture reconstruction approach. Specifically, for proximity sensing, a nonlinear mapping was established between real-world and image-space coordinates, allowing the system to estimate the object's approaching distance. In tactile sensing, we established a correlation between the geometric characteristics of surface deformations and the internal light field distribution, enabling high-resolution perception.

A. Nonlinear Dynamic Mapping Model of Proximity Sensing

In the proximity sensing mode, we utilize a pretrained monocular depth estimation algorithm DepthAnythingV2 [27] to convert raw RGB images into dense, pixelwise depth maps. This method employs a vision transformer backbone to extract a diverse set of high-level semantic features from a single RGB image—including texture sharpness, shading and lighting variations, and contextual semantics—rather than depending solely on perspective scaling. Such rich representations are then decoded into per-pixel depth estimations by a multiscale regression head, ensuring robust performance across objects of varying sizes and shapes. As the model is trained without stereo or LiDAR supervision, the predicted depth values are relative to the camera coordinate system. Higher predicted values correspond to regions that are closer to the camera. After target segmentation, the depth value is estimated under image coordinate. To map these estimates to real-world distances, a nonlinear

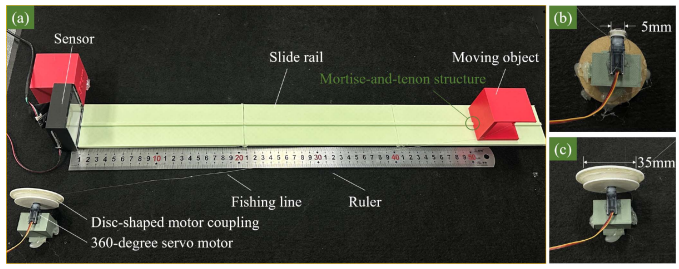

Fig. 5. Experimental platform setup of proximity sensing. (a) Detailed composition of the test platform. (b), (c) Two kinds of disc steering gear couplings with inner diameters of 5 and 35 mm, which realize the movement of the target object at different speeds.

calibration is performed using ground-truth measurements, as shown in Fig. 5. We let a calibration block move along the central axis of the camera-captured area from far to near within a distance of $D \in [0, 50]$ cm under randomly various velocities. The setup of the platform will be detailed in Section IV. To ensure the highest calibration accuracy, we trained a U-Net-based mask segmentation network for the calibration blocks to locate the target pixel regions. We constructed a synthetic dataset consisting of red square blocks of varying sizes, positions, and brightness, overlaid on a background image captured from the experimental platform. Each image was paired with a binary mask, where the target region was labeled as 1 and the background as 0, yielding 100 RGB image–mask pairs for training. We adopted a standard U-Net with a symmetric encoder–decoder structure. The encoder is composed of four convolutional blocks followed by max-pooling layers to extract multiscale features. The decoder mirrors the encoder with transposed convolutions and skip connections to progressively recover spatial resolution and preserve fine details. The network takes an RGB image as input and produces a single-channel probability map representing pixelwise likelihoods of the target region. A threshold of 0.5 is applied to convert the probability map into a binary mask. The model was trained using binary cross-entropy loss and the Adam optimizer with a learning rate of 1×10^{-3}, and converged within 10 epochs with a training loss of 0.0124.

For each group of distance measurement data, we extract video frames corresponding to the target world distance of $D = \{3, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50\}$cm, calculate the predicted distance of the target with the U-NET segmentation network and the pretrained monocular depth estimation algorithm. As shown in Fig. 6(a), the data exhibit a two-stage distribution trend. Considering the practical application scenario and the required imaging distance of the camera, the interval $D \in [10, 50]$ cm was used to construct the mapping model. Within this interval, the predicted value exhibit a monotonic increase as the target object approaches. Several inverse models were evaluated, and the double exponential decay model was ultimately selected as the best fit, achieving a coefficient of determination of $R^2 = 0.9697$ and a mean squared error (MSE) of 2.2462 [see Fig. 6(b)]. Its mathematical form is defined as

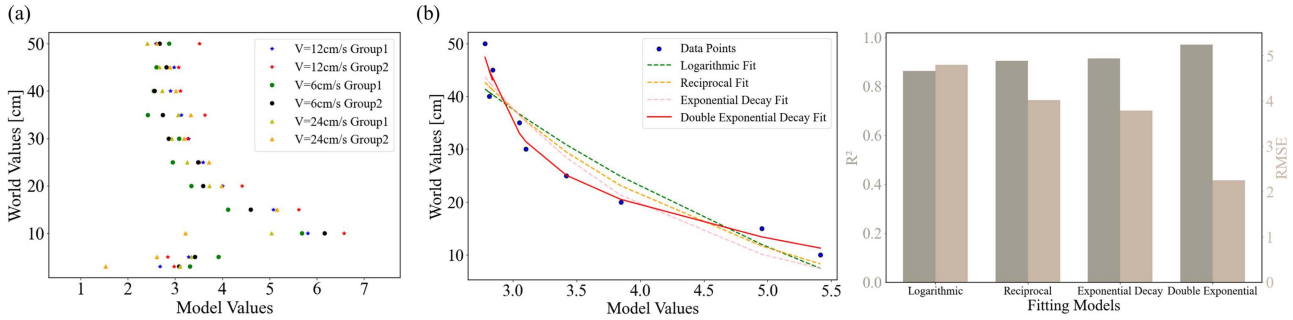

Fig. 6. Development of a nonlinear dynamic mapping model for proximity sensing modality. (a) Distribution of (Z_{img}, Z_{world}) across different experimental groups, all of which exhibit a similar distribution pattern. (b) Comparison of the fitting performance of different monotonic decreasing models on the experimental data. Among them, the double exponential decay model exhibits the best fitting accuracy, with $R^2 = 0.9697$ and RMSE $= 2.2462$.

follows:

$$y = a \cdot e^{-b \cdot x} + c \cdot e^{-d \cdot x} \tag{1}$$

where

1) y: represents the distance value Z_{world} in the real-world coordinate system;
2) x: represents the measured depth value Z_{img} in the camera coordinate system, which is equal to the mean value of image target segmentation μ_Ω

$$\mu_\Omega = \frac{\sum_{(u,v)\in\Omega} D(u,v)}{|\Omega|} \tag{2}$$

where

1) $\Omega = \{(u,v) \mid M(u,v) \neq 0\}$: represents the coordinate set of pixels in the nonzero region;
2) $D(u,v)$: denotes the dense, pixelwise depth map predicted by the algorithm.

B. Geometry Reconstruction of Tactile Sensing

Under tactile mode, we define the surface height as a function of pixel coordinates $z = f(u,v)$. The surface normal can then be expressed as follows:

$$\mathbf{N}(u,v) = \left(\frac{\partial f}{\partial u}, \frac{\partial f}{\partial v}, -1\right). \tag{3}$$

According to the principle of the PS algorithm [12], we add the local pixels position of lighting into mapping consideration to establish a local mapping relationship for each pixel by constructing a neural network, correlating the intensity values of the R, G, and B channels with the gradients along the *X*- and *Y*-axes, as represented by the following equation:

$$R_i(I_R, I_G, I_B, u, v) = (G_u, G_v) \tag{4}$$

where

1) R_i: represents the local pixel's mapping relationship between light intensity and geometry gradients;
2) (I_R, I_G, I_B): denotes the light intensity values of the red, green, and blue channels, respectively;
3) $(G_u, G_v) = (\frac{\partial f}{\partial u}, \frac{\partial f}{\partial v})$.

To obtain the training dataset, we randomly pressed a standard ball with a diameter $r = 5$ mm on the sensing surface, assuming the radius of circular indentation pattern is $r*$. Based on the geometric relationship, the distance h between the center of the sphere and the plane can be determined as

$$h = \sqrt{r^2 - r^{*2}}. \tag{5}$$

For the intersecting portion of the spherical surface, the equation of the sphere can be expressed as

$$z = h - \sqrt{r^2 - u^2 - v^2}. \tag{6}$$

Take the partial derivatives of u and v, respectively, we can get the geometric gradient expression of each point in the intersecting spherical part, where (u,v) satisfies the constraint conditions of intersecting circles

$$\nabla z = \left(\frac{u}{\sqrt{r^2 - u^2 - v^2}}, \frac{v}{\sqrt{r^2 - u^2 - v^2}}\right) \tag{7a}$$

$$(G_u, G_v) = \nabla z \tag{7b}$$

$$u^2 + v^2 \leq r^{*2}. \tag{8}$$

A total of 30 images with a resolution of 1280 × 960 pixels were collected. The circular indentation pattern was annotated manually, generating binary masks with a value of 1 inside the spherical region and 0 elsewhere. Only the pixels within the masked region, along with their associated feature values $(u, v, I_R, I_G, I_B, G_u, G_v)$, were retained as valid training samples, yielding a total of 842 400 entries. Therefore, the gradient map $\vec{g}$ for each test image will be calculated, then the depth map ϕ of the contact surface can be reconstructed by solving a 2-D Poisson equation

$$\nabla^2 \phi = \nabla \cdot \vec{g} \tag{9}$$

where $\nabla^2\phi$ is the Laplacian of the depth map, and $\nabla \cdot \vec{g}$ represents the divergence of the gradient map.

Given the high computational speed requirements for image reconstruction in the target application, the Fourier transform method was selected to solve the Poisson equation. This method is well suited for global solutions and demonstrates greater

robustness when applied to data with conventional image characteristics. The computational procedure is as follows.

First, perform fast Fourier transform operation on $\nabla \cdot \vec{g}$,

$$F = \mathcal{F}(\nabla \cdot \vec{g}). \tag{10}$$

Second, solve the Poisson equation based on frequency domain division

$$\mathcal{F}(\phi) = \frac{\mathcal{F}(\nabla \cdot \vec{g})}{-k_x^2 - k_y^2} \tag{11}$$

where k_x and k_y are frequency variables.

Finally, perform the inverse Fourier transform on the solution, then the depth map ϕ can be reconstructed

$$\phi = \mathcal{F}^{-1}(\mathcal{F}(\phi)). \tag{12}$$

C. Switching Mechanism Between Two Sensing Modes

Mode switching is achieved via friction between the silicone film and the PET belt, ensuring stable film movement (see Supplementary Video). A 360-degree continuous rotation servo motor is controlled by an Arduino-based system using pulsewidth modulation (PWM). The motor operates at predefined speeds and directions based on serial commands. The control algorithm dynamically adjusts the pulse width of the control signal to achieve the desired rotational direction and speed.

Through systematic experiments, we identified the pulse width of 100 μs to initiate counterclockwise rotation, while 2500 μs to trigger clockwise motion, at which the transmission efficiency is maximized. When the target approaches the predefined distance threshold $D_t = 10$ cm, the mode-switching mechanism is activated, triggering the rotation of the servo motor and the illumination of the LED strip, thereby transitioning to the tactile measurement mode. To ensure precise control, the motor's rotation duration is fine-tuned using timed delays, allowing it to complete predefined movement before stopping.

IV. Performance Characterization

A. Distance Measurement Accuracy of Proximity Perception

Given the design of the proposed sensing system, a working range of 10–30 cm was considered suitable for most approaching scenarios. Accordingly, experiments were conducted to quantitatively assess the reliability of the proposed algorithm. The experimental setup is detailed in Fig. 5. We fabricated the test block and slide rail by 3-D printing technology (X1Carbon, Bambu Lab). To ensure controlled movement of the block in the desired direction, we designed the connection between the block and the slide rail as a mortise-and-tenon structure. The object block itself was designed with a "concave shape" to appear as a standard square from the perspective of the sensor. Inside the cavity of the block, a circular ring with an inner diameter of 2 mm and an outer diameter of 4 mm was incorporated. A transparent fishing line with a diameter of 1 mm was fixed to this ring, with its other end secured at the far end of the experimental platform. A 360° servo motor was placed at the far end of the platform, with its rotation speed controlled via a PWM signal emitted by an Arduino Mega 2560 board. To achieve the desired speed, we designed disc-shaped servo motor couplings of different sizes. In addition, a 2-mm-deep groove was created along the side of the disc to collect the retracted fishing line.

The sensor recorded videos of a block approaching at multiple speeds ($V_i \in \{2, 4, 10, 12.5, 17.5\}$ cm/s), with a frame rate of 30 fps and a resolution of 1280 × 960 pixels. Five images were extracted from each video, corresponding to real-world object distances of approximately $D_i \in \{30, 25, 20, 15, 10\}$ cm. As illustrated in Fig. 7(a), each raw image is processed through the DepthAnythingV2 model for depth estimation and the U-Net algorithm for segmentation. The nonlinear calibration function is then applied to the network outputs to obtain the predicted distances in world coordinates. As shown in Fig. 7(b), the predicted results across different approaching speeds maintain a consistent and reasonable alignment with the ground truth. Among them, the model achieves the highest accuracy at $V = 12.5$ cm/s, with a coefficient of determination of $R^2 = 0.9551$ and an MSE of 2.2470.

B. Fine Texture Capture and Morphological Reconstruction

Two characterization experiments are conducted to evaluate the sensor's tactile sensing ability: one quantifying the surface roughness detection threshold, and the other examining contact geometry reconstruction accuracy. The experimental platform is shown in Fig. 8. The tactile sensing unit of the sensor is mounted on a 3-D-printed immovable supporter. Within the sensor, the RGB LED strip is powered by a dc power supply (Maisheng, China). A laptop is used to acquire high-resolution images from the built-in OV5640 camera modules. A 3-D-printed planar pressing plate is mounted on a three-axis ball screw displacement platform, moving vertically along the screw slider. This design ensures that the resultant normal force applied to all measured objects remains consistent to the sensing plane, thereby minimizing measurement errors. All original data for the tactile modality characterization experiment were acquired using the experimental platform developed above.

1) Identification of Contact Surface Roughness: In this experiment, sandpaper with varying roughness (grit sizes of 150, 280, and 500 mesh) was sequentially applied to the sensor. Its ability to perceive fine textures was assessed by analyzing texture features extracted from the acquired images.

In this process, two kinds of tactile data are collected separately, one is a tactile image without any pressing trace, the other is a tactile image obtained by pressing sandpaper with different roughness on the measurement surface. The difference images were obtained by subtracting the tactile images without contact from those acquired under sandpaper compression with varying roughness. A logarithmic scale was employed to improve the visibility of the amplitude spectrum, followed by zero-frequency shifting to the center of the spectrum. Fig. 9(a) presents the grayscale representation of the original tactile image, while Fig. 9(b) illustrates the processed results obtained using the

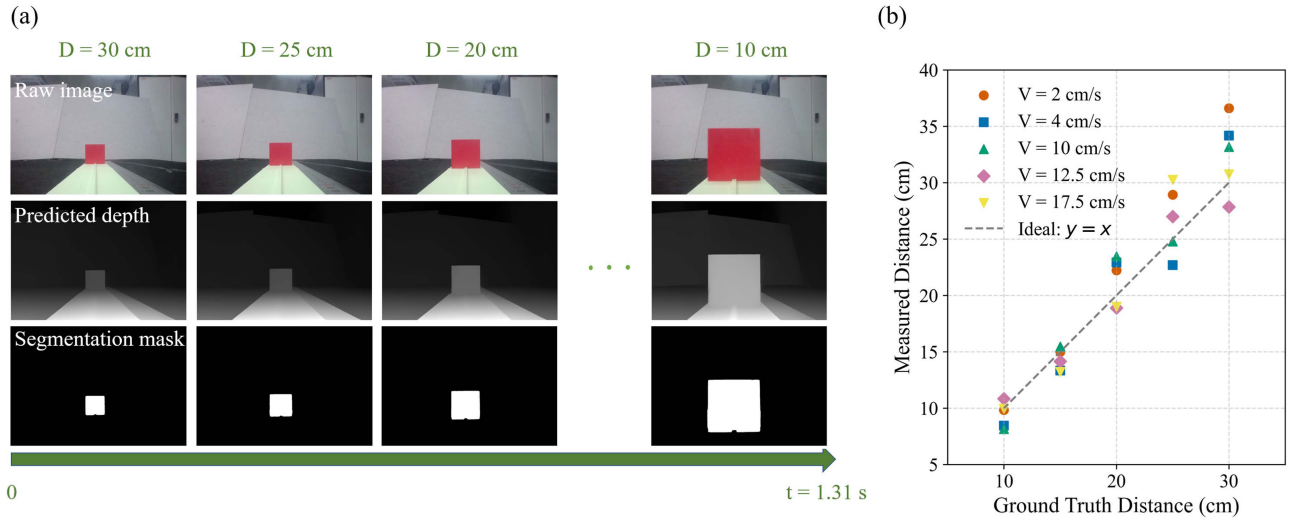

Fig. 7. Schematic and evaluation of the distance measurement algorithm. (a) Representative results at a speed of 12.5 cm/s, showing the original image, depth map, and segmentation mask from top to bottom in temporal sequence. (b) Experimental results of the distance measurement tests under various speeds with the nonlinear dynamic mapping model of proximity sensing.

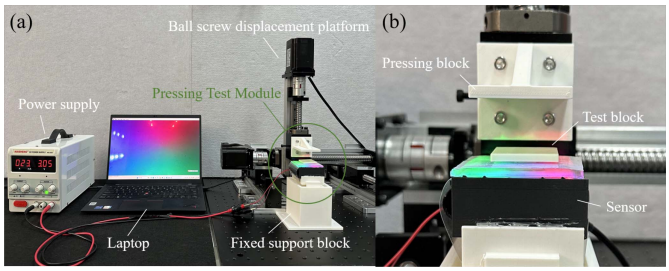

Fig. 8. Experimental platform for tactile sensing. (a) illustrates the overall layout of the tactile perception experimental platform; (b) ensures the resultant normal force remains perpendicular to the sensing plane to minimize measurement errors.

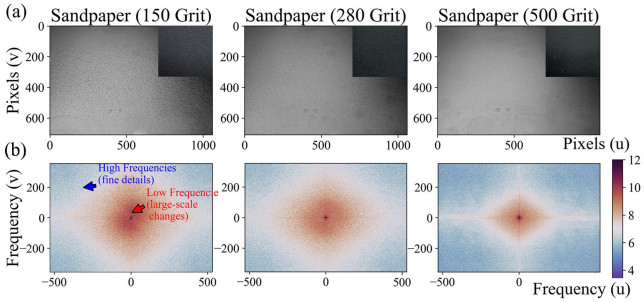

Fig. 9. Frequency domain analysis of tactile images. (a) Grayscale-processed tactile data of sandpapers with different grit sizes, with the real photo of the corresponding sandpaper shown in the upper right corner. (b) Frequency amplitude spectrum visualized on a logarithmic scale.

proposed method. The red region at the center corresponds to the low-frequency components, which represent large-scale texture features, whereas the blue regions at the image periphery indicate fine texture details. As the grit size of the sandpaper increases, the low-frequency region shrinks significantly, while the high-frequency region expands accordingly. These results demonstrate the sensor's ability to effectively distinguish surface roughness, achieving a fine-texture sensing resolution approximately 6 to 8 times higher than that of the human tactile perception system.

2) Reconstruction of Contact Surface Geometry: In this section, the geometry of the contact sensing surface is reconstructed using the previously introduced PS method. As shown in Fig. 10, we implement a fully connected (FC) neural network (FCNN) in PyTorch for tactile gradient image computation. The network consists of an input layer, four hidden layers, and an output layer. The input layer receives 5-D feature vectors [as shown in (4)], and each hidden layer comprises a FC layer followed by an ReLU activation function. The numbers of neurons in each hidden layer are 16, 64, 32, and 8, respectively. A dropout layer with a probability of 0.3 is applied before the final output layer (G_u, G_v) to mitigate overfitting. After obtaining the gradient map in (u, v) directions, the complete 3-D morphology of the tactile measurement surface can be obtained through Poisson reconstruction. The model is trained using the Adam optimizer with a learning rate of 3×10^{-5} to minimize the L1 loss function. An early stop mechanism is implemented to terminate the training process when the loss does not decrease for ten consecutive epochs. Furthermore, the training process is performed for 120 epochs, resulting in a regression model with a stable test MSE of approximately 0.04.

To quantitatively evaluate the performance of the geometric depth reconstruction method, a standard 8-mm-diameter calibration sphere was randomly pressed onto the sensing surface. By combining geometric information with image processing technologies (detailed in Fig. S6), we calculated the depth error between the reconstructed surface and the theoretical sphere

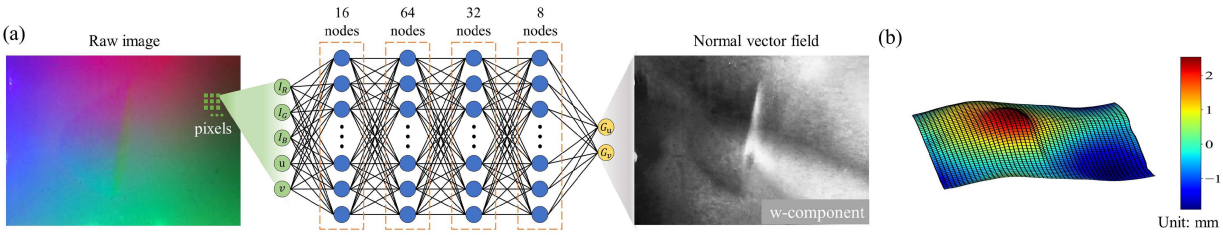

Fig. 10. Schematic of the tactile data surface reconstruction process. (a) Network architecture: A mapping from local pixel light intensities to geometric gradients. (b) Surface geometry obtained through Poisson reconstruction.

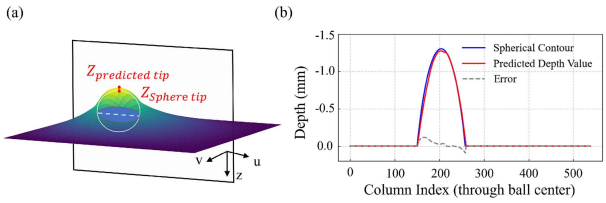

Fig. 11. Characterization of the geometric reconstruction method. (a) Reconstructed surface of the sensor under indentation by a standard 8 mm-diameter sphere. (b) Comparison between the extracted depth profile along the sphere's central column and the theoretical spherical surface.

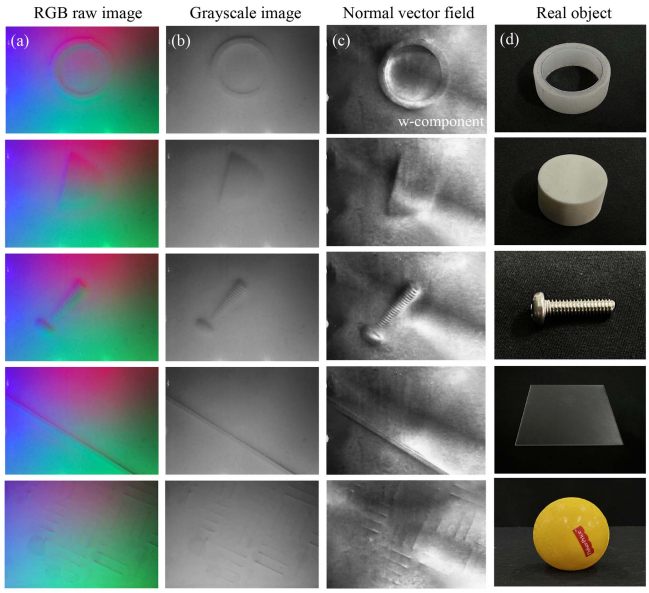

Fig. 12. Reconstruction performance of geometric texture. (a) Raw image data captured by the palm sensor. (b) Grayscale images of raw data, illustrating geometric information before processing. (c) Normal surface in *w*-axis calculated by the proposed FCNN algorithm, which indicates the depth information and surface orientation. (d) Real objects.

surface. The column of reconstructed depth values aligned with the centerline of the standard sphere was extracted and compared with the theoretical height profile of a perfect spherical surface, as shown in Fig. 11(b), the absolute mean error is approximately 0.0239 mm. Furthermore, several objects commonly operated by robotic grippers were selected to evaluate the texture reconstruction performance of the V-T PALM, with the results presented in Fig. 12.

V. Applications

A. Preplan Grasping and Subtlety Identification

Grasping and object recognition are fundamental in industrial robotics, while robust pregrasp planning and distinguishing visually similar objects with various subtly textures remain challenging. Therefore, we integrated the sensor into a gripper and evaluated the proposed system's perceptual advantages in approach-to-grasp scenarios, combining proximity sensing, external vision, and tactile sensing. Specifically, our design enhanced grasping efficiency by enabling autoswitch modalities and optimizing grasp execution timing, while also allowing for precise differentiation of visually similar objects.

1) Robust Distance Measurement Across Different Targets: The soft gripper offers high dexterity and adaptability while maintaining a simple structure and control scheme, making it suitable for rapidly building a testing platform for grasping objects with diverse shapes, textures, and hardness [11], [13], [30]. As shown in Fig. 13(a), the V-T PALM was integrated with four pneumatic soft fingers and was mounted on a 6-DOF robotic arm (Jaka Zu7, JAKA). The target object was placed stationary on a horizontal tabletop, with the measuring surface of the palm oriented vertically. Under the proximity mode, the palm was mechanically coupled with the end of the robotic arm and approached the target at a constant speed of 80 mm/s. For different target objects, the segmentation masks used in the ranging algorithm were obtained via HSV color space thresholding. When the measured distance reached 10 cm, the modality switching procedure was triggered. Upon detecting definitive contact, the grasping procedure was then initiated.

The input pressures of each soft actuator are regulated by an electronic regulator (QB1XANEEN100P400KPG, Proportion-Air), which is connected to a control module (MicroLabBox-DS1202, dSPACE). The sensor transmits image data to the PC via a serial port, while a control module (Arduino Mega 2560, Arduino SRL) handles communication between the sensor and the PC. The sensor's image signals were continuously monitored in real time to support logical decision-making. Upon initiation of a grasp command, MATLAB generates and transmits control signals at a frequency of $F_s = 50$ Hz [shown in Fig. 13(b)]. In this experiment, the target object's position and the robot arm's motion trajectory were fixed. We evaluated the success rate of the bimodal hand in autonomously sensing modalities switching and executing grasps for targets with different sizes, shapes and colors, achieving a 100% grasping success rate with a 78.6%

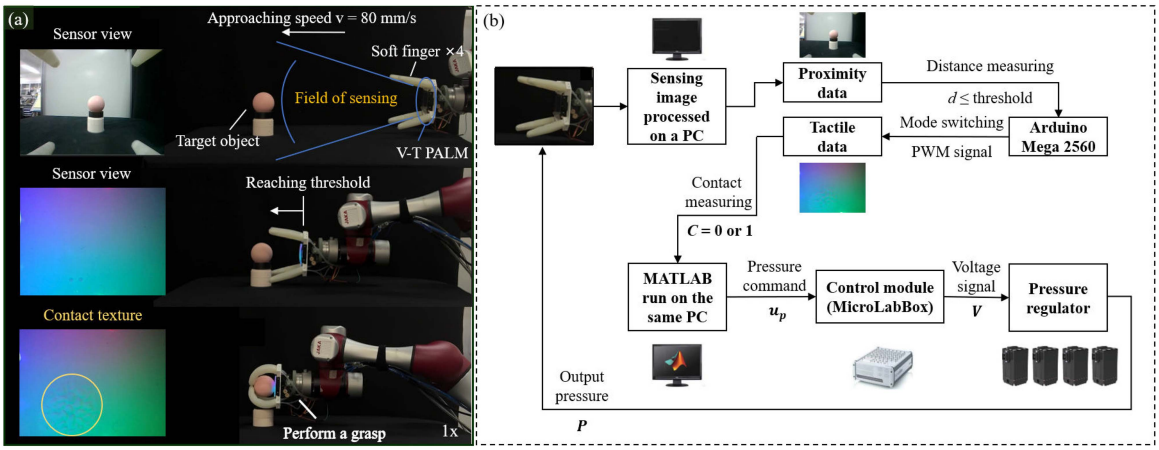

Fig. 13. Schematic illustration of the operational workflow and experimental platform. (a) Steps of the complete experimental process and the images from sensor perspective. (b) Control logic of the platform.

Fig. 14. Experimental results of preplanned grasping and subtle identification tests. (a) Grasping performance for various objects. (b)–(d) Reconstruction results and roughness analysis of the tactile palm.

ranging accuracy, as presented in the Supplementary Video and Fig. 14(a).

2) Distinguishing Visually Similar Objects: In high-speed sorting applications, distinguishing target objects with highly similar visual characteristics (e.g., color and shape) poses a significant challenge, particularly when relying solely on vision sensors. To demonstrate the effectiveness of our approach, we conducted controlled experiments with two test groups: 1) pink yoga balls featuring different surface textures; and 2) oranges at various stages of ripeness. As shown in Fig. 14(b)–(d), for each obtained result, we applied the wavelet transform to compute the energy of wavelet decomposition coefficients across different scales. In addition, the gray-level co-occurrence matrix was utilized to quantify grayscale contrast variations. The experimental results demonstrate that our sensing system can effectively differentiate target objects, even when their visual attributes are highly similar, highlighting its potential for improving high-speed sorting performance.

B. Card Insertion Experiment

The conveyor-driven transmission mechanism naturally endows the haptic palm with one degree of freedom in movement, enabling fine-tuning of the target's orientation without altering the finger grasping state. As shown in Fig. 15(a), when the robotic arm's end effector is aligned with the upper edge of the "transfer gap," simply releasing the hand grip to perform a card throwing action does not guarantee that the successfully pass through the narrow gap. By introducing an automatic perception–action control loop based on tactile sensing for grasp refinement, the card delivery performance is significantly improved, as illustrated in Fig. 15(b). The host computer continuously monitors the average indentation depth at the card–palm contact region from the tactile image. After the gripper reaches the target place above the box, a "down-and-up-and-finetune" or "down-and-release" control logic will be cyclic triggered. During the execution of "down" command, if the real-time measured depth exceeds a predefined threshold (0.15 mm), the system determines that the insertion is blocked, then initiates the "down-and-up-and-finetune" loop. Otherwise, the gripper releases the card to complete the throw. At the beginning of the experiment, the card was manually handed to the gripper, which naturally introduced a slight and random tilt to its initial pose, ensuring the fairness of the controlled trials. The successful execution of the throwing task offers a new possibility for soft robotic hand dexterous manipulation.

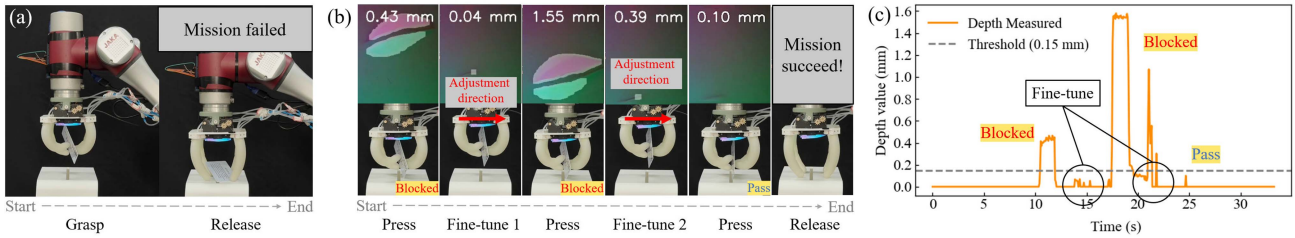

Fig. 15. Display the use of V-T PALM's DOF in card delivery application. (a) A failure case of direct throwing without adjustment. (b) Successful fine-tuning process guided by real-time tactile feedback; the top row shows the average indentation depth signal used for decision-making. (c) Tactile signals captured during the successful trial in (b).

VI. Conclusion

In this article, we proposed a visual–tactile dual-mode sensor with a sliding sensing window. By introducing a mechanical transmission design, it can seamlessly switch between proximity and tactile sensing modes, enabling accurate long-range proximity perception while simultaneously maintaining ultra-high-resolution texture sensing and reconstruction capabilities. Experimental results demonstrate that the proposed proximity perception system accurately measures external distances across various target velocities, achieving the highest accuracy at a speed of 12.5 cm/s, with a coefficient of determination of $R^2 = 0.9551$ and an MSE of 2.2470. The tactile mode of the proposed sensor enables high-precision 3-D reconstruction and accurate distinction of surface roughness, achieving a sensing resolution approximately 6–8 times higher than that of the human fingertip. By integrating the dual-mode palm sensor with soft robotic fingers, the system demonstrates strong potential for adaptive grasping, object classification, and in-hand manipulation. Experiments show a 100% success rate in reliably grasping objects with various shapes, as well as accurate classification of visually similar objects through surface texture reconstruction. This enables the differentiation of material properties, such as surface roughness and compliance (e.g., identifying the ripeness of fruits). Moreover, the "card throwing" experiment validates the effectiveness of the palm's intrinsic DOF in supporting fine-tuned in-hand adjustments.

The integration of both modalities improves the robot's efficiency in executing various operation tasks. Future work will focus on enhancing the perception framework to support proximity sensing of transparent objects, increasing the upper limit of tactile sensing resolution, and incorporating calibration data from objects of different shapes and sizes into the training of the depth regression head, with an aim to further improve ranging robustness and expand the applicability of the proposed sensor in more challenging scenarios.

Acknowledgment

The authors would like to thank Ms. Jiawen Yu for her suggestions on the experimental design of this article and to Mr. Xinyu Yang for his support in data processing.

References

[1] J. W. James, A. Church, L. Cramphorn, and N. F. Lepora, "Tactile model o: Fabrication and testing of a 3D-printed, three-fingered tactile robot hand," *Soft Robot.*, vol. 8, no. 5, pp. 594–610, 2021, pMID: 33337925.
[2] S. Suresh et al., "Neuralfeels with neural fields: Visuotactile perception for in-hand manipulation," *Sci. Robot.*, vol. 9, no. 96, 2024, Art. no. eadl0628.
[3] J. Xu, L. Wu, C. Lin, D. Zhao, and H. Xu, "Dtactive: A vision-based tactile sensor with active surface," 2024, *arXiv:2410.08337*.
[4] J. Zhang et al., "Dexterous hand towards intelligent manufacturing: A review of technologies, trends, and potential applications," *Robot. Comput.-Integr. Manuf.*, vol. 95, 2025, Art. no. 103021.
[5] A. B. Wan, H. Sanghyun, J. Sangyoon, B. Franklin, and P. Jang-Ung, "Transparent and flexible fingerprint sensor array with multiplexed detection of tactile pressure and skin temperature," *Nature Commun.*, vol. 9, no. 1, pp. 2458–, 2018.
[6] J.-Y. Yoo, M.-H. Seo, J.-S. Lee, K.-W. Choi, M.-S. Jo, and J.-B. Yoon, "Industrial grade, bending-insensitive, transparent nanoforce touch sensor via enhanced percolation effect in a hierarchical nanocomposite film," *Adv. Funct. Materials*, vol. 28, no. 42, 2018, Art. no. 1804721.
[7] M. L. Hammock, A. Chortos, B. C.-K. Tee, J. B.-H. Tok, and Z. Bao, "25th anniversary article: The evolution of electronic skin (e-skin): A brief history, design considerations, and recent progress," *Adv. Materials*, vol. 25, no. 42, pp. 5997–6038, 2013.
[8] Z. Pi, J. Zhang, C. Wen, Z.-b. Zhang, and D. Wu, "Flexible piezoelectric nanogenerator made of poly (vinylidenefluoride-co-trifluoroethylene)(PVDF-TrFE) thin film," *Nano Energy*, vol. 7, pp. 33–41, 2014.
[9] F. Huang, T. Chen, J. Si, X. Pham, and X. Hou, "Fiber laser based on a fiber Bragg grating and its application in high-temperature sensing," *Opt. Commun.*, vol. 452, pp. 233–237, 2019.
[10] G. Gu et al., "A soft neuroprosthetic hand providing simultaneous myoelectric control and tactile feedback," *Nature Biomed. Eng.*, vol. 7, no. 4, pp. 589–598, 2023.
[11] H. Zhao, K. O'Brien, S. Li, and R. F. Shepherd, "Optoelectronically innervated soft prosthetic hand via stretchable optical waveguides," *Sci. Robot.*, vol. 1, no. 1, 2016, Art. no. eaai7529.
[12] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017, Art. no. 2762.
[13] N. Zhang et al., "Soft robotic hand with tactile palm-finger coordination," *Nature Commun.*, vol. 16, no. 1, 2025, Art. no. 2395.
[14] S. Zhang, Y. Yang, J. Shan, F. Sun, H. Xue, and B. Fang, "PalmTac: A vision-based tactile sensor leveraging distributed-modality design and modal-matching recognition for soft hand perception," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 3, pp. 288–298, Apr. 2024.
[15] J. Di et al., "Using fiber optic bundles to miniaturize vision-based tactile sensors," *IEEE Trans. Robot.*, vol. 41, pp. 62–81, 2025.
[16] I. Taylor, S. Dong, and A. Rodriguez, "GelSlim3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *Proc. Int. Conf. Robot. Automat.*, 2021, pp. 10781–10787.
[17] S. E. Navarro et al., "Proximity perception in human-centered robotics: A survey on sensing systems and applications," *IEEE Trans. Robot.*, vol. 38, no. 3, pp. 1599–1620, Jun. 2022.
[18] B. Wu, T. Jiang, Z. Yu, Q. Zhou, J. Jiao, and M. L. Jin, "Proximity sensing electronic skin: Principles, characteristics, and applications," *Adv. Sci.*, vol. 11, no. 13, 2024, Art. no. 2308560.

[19] O. M. Andrychowicz et al., "Learning dexterous in-hand manipulation," *Int. J. Robot. Res.*, vol. 39, no. 1, pp. 3–20, 2020.
[20] W. Wei et al., "Learning human-like functional grasping for multi-finger hands from few demonstrations," *IEEE Trans. Robot.*, vol. 40, pp. 3897–3916, 2024.
[21] G. Jin, X. Yu, Y. Chen, and J. Li, "SCARA system: Bin picking system of revolution-symmetry objects," *IEEE Trans. Ind. Electron.*, vol. 71, no. 9, pp. 10976–10986, Sep. 2024.
[22] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2016, pp. 138–143.
[23] Q. Wang, Y. Du, and M. Y. Wang, "Spectac: A visual-tactile dual-modality sensor using uv illumination," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 10844–10850.
[24] F. R. Hogan, J.-F. Tremblay, B. H. Baghi, M. Jenkin, K. Siddiqi, and G. Dudek, "Finger-STS: Combined proximity and tactile sensing for robotic manipulation," *IEEE Robot. Automat. Lett.*, vol. 7, no. 4, pp. 10865–10872, Oct. 2022.
[25] S. Zhang et al., "TIRgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robot. Automat. Lett.*, vol. 8, no. 10, pp. 6307–6314, Oct. 2023.
[26] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1070–1077.
[27] L. Yang et al., "Depth anything V2," in *Proc. 38th Int. Conf. Neural Inf. Process. Syst.*, 2024, pp. 21875–21911.
[28] Y. Cai and S. Yuan, "In-hand manipulation in power grasp: Design of an adaptive robot hand with active surfaces," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 10296–10 302.
[29] S. Yuan et al., "Tactile-reactive roller grasper," *IEEE Trans. Robot.*, vol. 41, pp. 1938–1955, 2025.
[30] X. Wang et al., "Bionic soft fingers with hybrid variable stiffness mechanisms for multimode grasping," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 1957–1963.

Yueshi Dong received the B.E. (Hons.) degree in mechanical engineering from Chongqing University, Chongqing, China, in 2022. She is currently working toward the Ph.D. degree in mechatronic engineering with Shanghai Jiao Tong University, Shanghai, China.

Her research interests include visual–tactile sensing and robotic manipulation, with an emphasis on integrating tactile perception into intelligent soft robotic systems.

Jieji Ren received the B.Sc. and M.Sc. degrees in optics from the Harbin Institute of Technology, Harbin, China, in 2013 and 2015, respectively, and the Ph.D. degree in mechatronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2022 .

Since 2022, he has been with Shanghai Jiao Tong University as an Assistant Researcher. His research interests include camera-based tactile sensing and its applications in soft robotics.

Zhenle Liu is currently working toward the B.E. degree in mechanical engineering with Shanghai Jiao Tong University, Shanghai, China.

His research interests include the design of soft robots and haptic feedback systems.

Zhanxuan Peng is currently working toward the B.E. degree in mechanical engineering with Shanghai Jiao Tong University, Shanghai, China.

His research interests include end-effector design and motion control of hyper-redundant manipulators.

Zihao Yuan received the B.E. degree in mechanical engineering, in 2022, from Shanghai Jiao Tong University, Shanghai, China, where he is currently working toward the Ph.D. degree in mechanical engineering.

His research interests include the design and control of pneumatic soft robots and bioinspired soft robotic systems.

Ningbin Zhang received the B.S. and M.S. degrees in mechanical engineering from Zhejiang Sci-Tech University, Hangzhou, China, in 2012 and 2016, respectively, and the Ph.D. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2022.

He is currently a Postdoctoral Research Fellow with the School of Mechanical Engineering, Shanghai Jiao Tong University. He has authored more than 30 academic papers published in multidisciplinary and robotics journals and conferences, including *Nature Biomedical Engineering*, *Nature Communications*, *Science Advances*, and the IEEE International Conference on Robotics and Automation. His research interests include robotic hands, soft–rigid mechanism design, and tactile sensing.

Guoying Gu (Senior Member, IEEE) received the B.E. degree (Hons.) in electronic science and technology and the Ph.D. degree (Hons.) in mechatronic engineering from Shanghai Jiao Tong University, Shanghai, China, in 2006 and 2012, respectively.

He was a Humboldt Fellow with the University of Oldenburg, Oldenburg, Germany, and was a Visiting Scholar with the Massachusetts Institute of Technology, Cambridge, MA, USA; the National University of Singapore, Singapore; and Concordia University, Montreal, QC, Canada. He is currently a Distinguished Professor with the School of Mechanical Engineering, Shanghai Jiao Tong University.

He has authored or coauthored of over 100 publications, including articles in *Science Robotics*, *Nature Biomedical Engineering*, *Nature Reviews Materials*, *Nature Materials*, *Nature Communications*, *Science Advances*, *Advanced Materials*, *Soft Robotics*, *Science China* series, the IEEE/ASME Transactions, book chapters, and international conference proceedings. His research interests include soft robotics, bioinspired and wearable robots, and smart materials for sensing, actuation, and motion control.

Dr. Gu was the recipient of the National Science Fund for Distinguished Young Scholars and the XPLORER Prize. He has also served for several international journals as an Editorial Board Member, Topic Editor, or Guest Editor, and for numerous international conferences and symposia as Chair, Co-Chair, Associate Editor, or Program Committee Member. He was an Associate Editor for *Soft Robotics*, IEEE TRANSACTIONS ON ROBOTICS, and IEEE ROBOTICS AND AUTOMATION LETTERS.