GelRoller: A Rolling Vision-based Tactile Sensor for Large Surface Reconstruction Using Self-Supervised Photometric Stereo Method

Zhiyuan Zhang, Huan Ma, Yulin Zhou, Jingjing Ji, and Hua Yang*

Abstract— **Accurate perception of the surrounding environment stands as a primary objective for robots. Through tactile interaction, vision-based tactile sensors provide the capability to capture high-resolution and multi-modal surface information of objects, thereby facilitating robots in achieving more dexterous manipulations. However, the prevailing GelSight sensors entail intricate calibration procedures, posing challenges in their application on curved surfaces and requiring the maintenance of stable lighting conditions throughout experimentation. Additionally, constrained by shape and structure, current vision-based tactile sensors are predominantly applied to measurements within a limited area. In this study, we design a novel cylindrical vision-based tactile sensor that enables continuous and swift perception of large-scale object surfaces through rolling. To tackle the challenges posed by laborious calibration processes, we propose a self-supervised photometric stereo method based on deep learning, which eliminates pre-calibration requirements and enables the derivation of surface normals from a single image without relying on stable lighting conditions. Finally, we perform surface reconstruction from normal and point cloud registration on the multiple frames of images obtained by rolling the cylindrical sensor, resulting in large surface reconstruction. We compare our method with the representative lookup table method in the GelSight sensors. The results show that the proposed method enhances both reconstruction accuracy and robustness, thereby demonstrating the potential of the proposed sensor in large-scale surface reconstruction. Codes and mechanical structures are available at: https://github.com/ZhangZhiyuanZhang/GelRoller**

I. INTRODUCTION

Visual and tactile modalities serve as vital means for perceiving and comprehending the intricacies of our surroundings [1]. With computer vision and robotics advancement, the emergence of vision-based tactile sensors effectively combine the strengths of both modalities, enabling intelligent robots to engage in more dexterous and sophisticated manipulations [2]. Based on distinct measurement principles, vision-based tactile sensors can be classified into three types: the first, exemplified by GelForce [3], Soft-bubble [4], and FingerVision [5], estimates forces and torques by gauging the displacement of markers embedded within a gel layer; the second type, represented by GelStereo [6], [7], and [8], captures three-dimensional (3D) information about the contact area through the principles of stereo imaging using a binocular camera setup; the third type, demonstrated by GelSight [9], GelSlim [10], and DIGIT [11], employs RGB tricolored light sources for illumination and adopts photometric stereo algorithms to achieve high-precision, high-resolution reconstruction results of contact object surfaces. Regardless of the specific type of vision-based tactile sensors, their architecture comprises three primary components [12]: the contact module, the light source module, and the camera module. In conjunction with elaborately designed algorithms, these three modules empower vision-based tactile sensors to accomplish intricate measurement and perception tasks, including slip detection [13], object recognition [14], and object pose estimation [15].

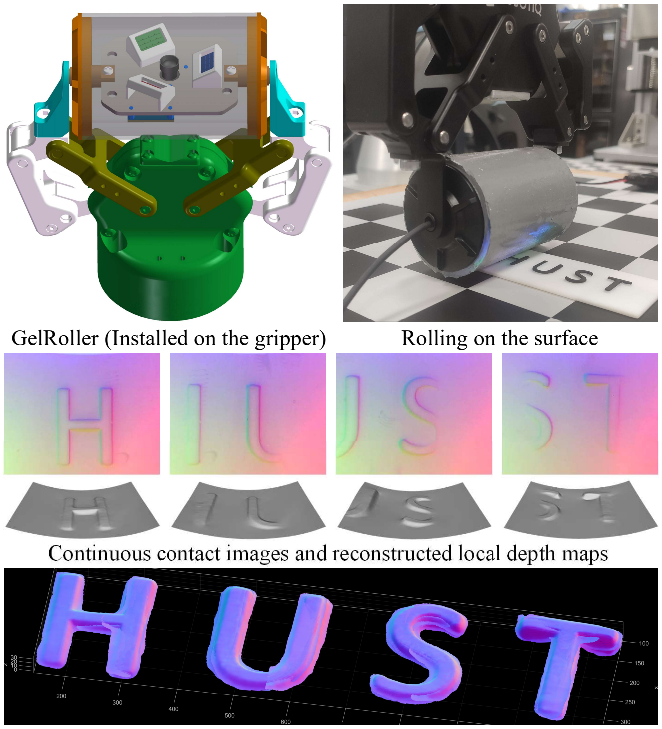
GelRoller (Installed on the gripper)
Rolling on the surface
Continuous contact images and reconstructed local depth maps
Large surface reconstruction result

Fig. 1. The designed GelRoller sensor and the reconstruction result using the proposed self-supervised learning method. This work aims at large surface reconstruction by harnessing the continuously captured images obtained by rolling the sensor. The proposed method can reconstruct the local surface with only a single input image and exhibits robustness under varying lighting conditions.

The ability to perceive and model the three-dimensional geometry of contact surfaces stands as a defining characteristic that sets GelSight-like vision-based tactile sensors apart from conventional tactile sensors. However, current GelSight sensors primarily perceive local regions of objects, lacking effective perception of overall information. [16]. Further-

* indicates corresponding author

All the authors are with State Key Laboratory of Intelligent Manufacturing Equipment and Technology (SKL-IMET), School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan, China huayang@hust.edu.cn

more, these sensors employ photometric stereo algorithms, establishing color-normal gradient calibration through geometric information from calibration balls and obtaining contact surface normals using a lookup table or neural network method [9]. The precision of their 3D reconstruction critically hinges on the calibration accuracy. Additionally, recalibration becomes necessary if lighting conditions change.

In this study, we introduce a novel cylindrical vision-based tactile sensor called GelRoller, designed to perceive large-scale object surfaces continuously and swiftly through rolling, as illustrated in Fig. 1. To overcome the limitations associated with the photometric stereo algorithm mentioned above, we propose a self-supervised learning method based on deep learning. The core of the proposed method is to use the background portion of a single image to compute the light sources' information and the foreground portion (the contact area between the gel and the object) to compute the desired normals. Furthermore, it replaces the conventional assumption of infinity point lighting model with near-point lighting model [17] that that better aligns with real-world scenarios. Surface reconstruction can be accomplished through a fast Poisson solver from surface normals. For large-scale surface reconstruction, we use the iterative closest point (ICP) method [18], which registers multiple single-frame reconstruction results into an integrated one.

To the best of our knowledge, this work is the first to reconstruct the surface of objects from a single image without calibration. The proposed method enables photometric stereo techniques in GelSight sensors even in changing lighting conditions and when calibration data is unavailable, significantly enhancing GelSight sensors' practicality.

II. Related Works

A. GelSight Sensors

A typical GelSight sensor employs a circularly symmetric arrangement of RGB tricolor light sources to illuminate the contact layer coated with a reflective material, while employing a monocular camera to capture images [9]. It assumes that the reflective layer's surface adheres to Lambertian surface properties, utilizing photometric stereo algorithms to reconstruct the surface normals of the contact area, subsequently yielding the surface shape. However, such sensor requires rigorous structural design and light source calibration [19]. GelSlim [10] enhances the optical path system of GelSight by implementing a mirror-reflective structure, eliminating the need for the camera to face the contact body. This innovation results in a more compact sensor design. DIGIT [11] offers a cost-effective, compact solution that delivers high-resolution tactile perception, greatly enhancing its suitability for robotic finger manipulation. GelSight Wedge [20] is designed to meet the mechanical specifications of compact robot fingers while preserving high-resolution 3D reconstruction capabilities. In contrast to the previous flat-structured contact layer, GelTip [21] proposes a finger-like shape sensor that can sense contacts on any location of its surface. Similarly, DenseTact [22] uses a spherical contact layer, making it exceptionally well-suited for sensing complex object surfaces.

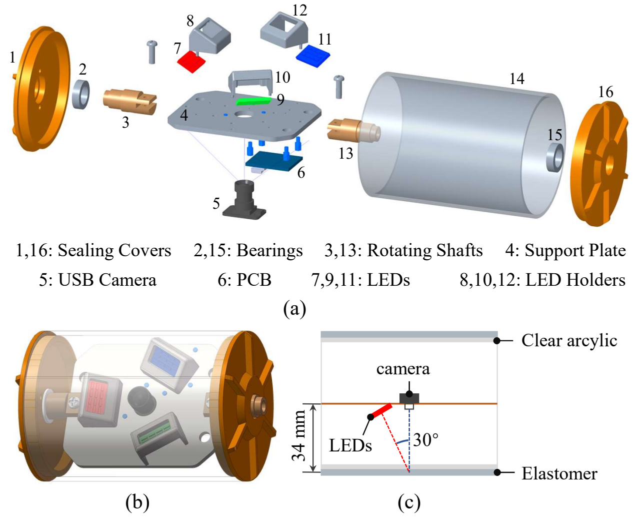
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
1,16: Sealing Covers 2,15: Bearings 3,13: Rotating Shafts 4: Support Plate
5: USB Camera 6: PCB 7,9,11: LEDs 8,10,12: LED Holders
(a)
Clear arcylic
camera
34 mm
LEDs
30°
Elastomer
(b)
(c)

Fig. 2. Design of GelRoller: (a) An exploded view that reveals its internal components; (b) The assembled CAD model. (c) A structural sketch.

B. Cylindrical Sensor and Large Surface Reconstruction

TouchRoller [16] is the first cylindrical vision-based tactile sensor to address the inefficiencies and time-consuming character associated with existing tactile sensors that require pressing multiple times on large surfaces. It allows quick assessment of large surfaces with high-resolution tactile sensing and efficient tactile images collection. For large-scale 3D reconstruction of object surfaces, [23] employs a concept similar to SLAM (Simultaneous Localization and Mapping), harnessing GelSight sensors to collect multiple frames of tactile information generated through repeated contact with objects. They incorporate point cloud registration algorithms, loop closure detection algorithms, and pose graph optimization algorithms to achieve highly precise surface reconstruction results. However, TouchRoller does not achieve the reconstruction of large-scale scenes, while [23] requires multiple presses on the objects for large-scale reconstruction, which is less efficient than rolling sensing.

III. The Principle of GelRoller Sensor

A. Overview

The top row of Fig. 1 displays the external appearance and operation mode of GelRoller. In order to achieve high-resolution 3D reconstruction results using photometric stereo algorithms, GelRoller employs a light sources design similar to GelSight sensors, which evenly distributes R, G, and B light sources around a circumference for omnidirectional and multi-angle illumination. The mechanical structure of GelRoller is illustrated in Fig. 2(a), comprising the following components: 1) High-transparency acrylic cylinder serves as the support structure during rolling; 2) The light source module consists of an LED circuit board with 12 LEDs arranged in a 3×4 array, an LED holder, and a diffuser film. The diffuser film blurs the light from LEDs, achieving

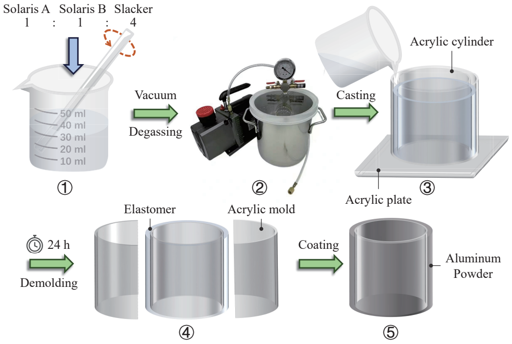

Fig. 3. Preparation of contact body: ① Multiple solutions are mixed according to the proportion and evenly stirred; ② Degassing the mixed solution in a vacuum chamber; ③ Slowly pour the mixed solution into the mold; ④ Demolding after waiting for 24 hours at room temperature; ⑤ Apply a thin layer of aluminum powder evenly onto the gel surface.

a uniform mixing of emitted light; 3) A stable USB camera is used for image capture; 4) A support plate secures the light source module and camera module, along with connected rotating shafts, which serve as the axis of rotation during GelRoller movement; 5) To ensure the relative stationarity of the support plate during rolling, bearings are employed to connect the rotating shafts and the rolling section; 6) Sealing covers reduce the impact of external light on the camera's imaging quality and act as connectors between the bearings and the acrylic cylinder.

B. Fabrication of GelRoller Sensor

Vision-based tactile sensors rely on the deformation of the gel to mirror the properties of the contacted objects. Hence, the production of the gel is a critical step. We have displayed the detailed production process of gel in Fig. 3, with specifics described in the caption of Fig. 3. Note that a thin layer of gel was applied over the aluminum powder layer. After the thin layer of gel has solidified, both the gel portion and the acrylic inner cylinder serve as the rolling component of GelRoller. The remaining components of GelRoller are as follows: The models of the red, green, and blue LEDs are NCD0603R1, NCD0603G1, and NCD0603B1, respectively; The model of the USB camera is HBVCAM-F2209HD V11 with a field of view of 100°; The bearing model is 6700; All other components shown in Fig. 2(a) are manufactured using 3D printing.

C. Algorithm for Single Frame Surface Reconstruction

The rendering equation under the illumination of red, green, and blue light sources is modeled as a linear combination of each component and is presented as follows:

$$\mathbf{I}_i = R_i \sum_{k=1}^{3} \beta_k \mathbf{e}_k \cos\langle \mathbf{l}_k, \mathbf{n}_i \rangle, \tag{1}$$

where the subscript i represents each pixel; $\mathbf{I}_i = (I_i^r, I_i^g, I_i^b)$ represents the R, G, B intensities of the ith pixel; the value of β_k ranging from 0 to 1 indicates the contribution of the kth light source's intensity to the final light intensity. R_i represents the diffuse reflectivity at the ith pixel; $\mathbf{e}_k = (e_k^r, e_k^g, e_k^b), k = 1,2,3$ are the intensities of the red, green, and blue light sources, respectively; Assuming the illuminated surface has Lambertian properties, $\mathbf{I}_i$ is directly proportional to the cosine of the angle (represented by $\langle \cdot, \cdot \rangle$) between the light direction $\mathbf{l}_k = (l_k^x, l_k^y, l_k^z)$ and the surface normal $\mathbf{n_i} = (n_i^x, n_i^y, n_i^z)$. Equation (1) is based on the infinity point lighting model, which is not realistic. Fig. 4 shows the non-uniformity of incident light intensity, i.e., the intensity is higher closer to the light source. We modify the assumption to use a near-point lighting model [17]: 1) the incident light direction at each pixel is the direction from the light source to the pixel; 2) the incident light intensity decreases quadratically with distance. Based on the modified lighting model, the rendering equation becomes as follows:

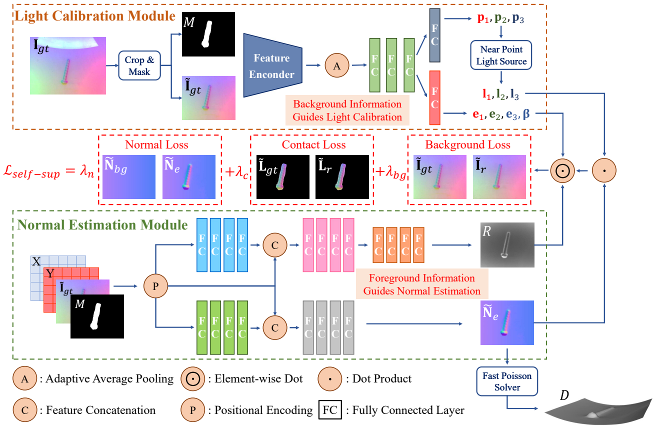
Light Calibration Module
$\mathbf{I}_{gt}$
Crop & Mask
M
$\tilde{\mathbf{I}}_{gt}$
Feature Enconder
A
FC
$\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3$
Near Point Light Source
$\mathbf{l}_1, \mathbf{l}_2, \mathbf{l}_3$
$\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \beta$
Background Information Guides Light Calibration
$\mathcal{L}_{self-sup} = \lambda_n$
Normal Loss
$\tilde{\mathbf{N}}_{bg}$
$\tilde{\mathbf{N}}_e$
$+\lambda_c$
Contact Loss
$\tilde{\mathbf{L}}_{gt}$
$\tilde{\mathbf{L}}_r$
$+\lambda_{bg}$
Background Loss
$\tilde{\mathbf{I}}_{gt}$
$\tilde{\mathbf{I}}_r$
Normal Estimation Module
X
Y
P
C
R
Foreground Information Guides Normal Estimation
$\tilde{\mathbf{N}}_e$
Fast Poisson Solver
D
A : Adaptive Average Pooling
⊙ : Element-wise Dot
· : Dot Product
C : Feature Concatenation
P : Positional Encoding
FC : Fully Connected Layer

Fig. 4. The proposed self-supervised photometric stereo method for single frame surface reconstruction using near-point lighting model, normal loss, contact loss, and background loss. The core of the proposed method is to use the background of a single image to compute the light sources' information and the foreground to compute the desired normals.

$$\mathbf{I}_i = R_i \sum_{k=1}^{3} \beta_k \frac{\mathbf{e}_k}{r_{k,i}^3} \cos\left\langle \mathbf{l}_{k,i}, \mathbf{n}_i \right\rangle, \tag{2}$$

where $r_{k,i}$ and $\mathbf{l}_{k,i}$ are the distance and direction between the kth light source and the ith pixel, respectively, and represented by:

$$\mathbf{l}_{k,i} = \mathbf{p}_k - \mathbf{x}_i, r_{k,i} = \|\mathbf{p}_k - \mathbf{x}_i\|_2, k = 1,2,3. \tag{3}$$

$\mathbf{p}_k = (p_k^x, p_k^y, p_k^z)$ and $\mathbf{x}_i = (x_i, y_i, z_i)$ are the positions of the kth light source and the ith pixel, respectively. Note that the above equations treat the LED array as equivalent to a single light source after passing through a diffuser. From (2), it can be seen that only $\mathbf{I}_i$ is known for each pixel, while all remaining variables are unknown, which results in a highly underdetermined equation. Therefore, reconstructing surface normals from a single image is extremely challenging. GelSight sensors use a pre-calibrated method to establish the mapping relationship between pixel intensities and surface normals. However, this method is unsuitable for varying lighting conditions and cannot be applied to curved surfaces. We adopt a self-supervised strategy to solve (2), fully leveraging the information in a single image to compute light sources' positions and intensities. Then, the contact

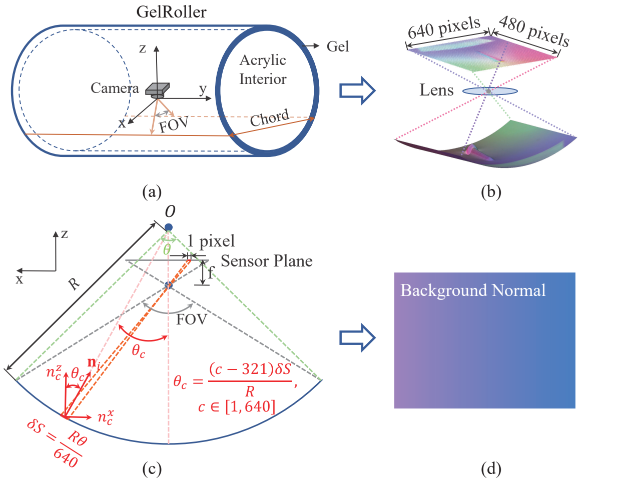

Fig. 5. Schematic for background normals computation. (a) GelRoller sketch; (b) Projection of the cylindrical surface onto the image plane; (c) Pixel positions and their corresponding surface positions; (d) Obtained background normals.

object's surface normals can be inferred by (2) using the obtained light sources information. The network architecture is shown in Fig. 4. We drew inspiration for network design from [24] and expanded it from a single light source to multiple light sources.

In order to obtain the parameters to be estimated in (2), we divide the network into two parts: In the Light Calibration Module, we perform cropping on the input image $\mathbf{I}_{gt}$ to retain well-illuminated portions $\tilde{\mathbf{I}}_{gt}$ and manually extract the mask M containing the contact region. These inputs are then fed into the neural network for feature extraction. Ultimately, fully connected (FC) layers predict the positions $\mathbf{p}$, intensities $\mathbf{e}$, and contribution coefficients β for the R, G, B light sources. It is noteworthy that, since the light source information is estimated adaptively for different input images, the proposed network is capable of addressing variations in illumination. In the Normal Estimation Module, we integrate position encoding, the cropped image $\tilde{\mathbf{I}}_{gt}$, and the mask M, utilizing fully connected layers to predict diffuse reflectivity R and surface normals $\tilde{\mathbf{N}}_e$. Finally, we incorporate elaborately designed normal loss, background loss, and contact loss as self-supervision signals to guide the optimization of network parameters, which will be discussed in the following paragraphs.

Based on the geometric characteristics of the contact area on GelRoller's curved surface, we can calculate the normals of the curved background captured by the camera. As depicted in Fig. 5(b), the cylindrical surface is projected onto an image with a resolution of 640 × 480 pixels. Through manual measurements, we can determine the chord length shown in Fig. 5(a) and further ascertain the size of the central angle θ in Fig. 5(c) based on the radius of the circle R. Therefore, the arc length of the curved surface that the camera can capture is $R\theta$. The arc length corresponding to one pixel in the image's width direction projected onto the surface can be approximated as $\delta S = R\theta/640$, as illustrated in Fig. 5(c). The angle between this projection point and the z-axis can be expressed as:

$$\theta_c = \frac{(c-321)\,\delta S}{R}, c \in [1,640], \tag{4}$$

where, c represents the column index. Based on the geometric relationships in Fig. 5(c), the normal vector at this point can be represented as:

$$n_c^x = \sin\theta_c, \quad n_c^y = 0, \quad n_c^z = \cos\theta_c. \tag{5}$$

Ultimately, the visualized normals of the curved surface background are presented as shown in Fig. 5(d), where the three-channel colors respectively denote the magnitudes of the normal vector components in the x, y, and z directions.

The influence of geometric errors caused by gel refraction is neglected due to the thinness of the gel. Therefore, We treat the calculated background normals $\tilde{\mathbf{N}}_{bg}$ as approximated ground truth labels to constrain the network's estimated background normals $\tilde{\mathbf{N}}_e$:

$$L_n = \frac{\sum_i (1-M_i)\rho_n\left(\tilde{\mathbf{N}}_{bg}, \tilde{\mathbf{N}}_e\right)}{\sum_i (1-M_i)}, \tag{6}$$

where $\tilde{\cdot}$ represents extracting a local data region from the original resolution; M represents a mask that can be obtained through manual drawing. $M_i = 1$ indicates that the ith pixel is in the foreground, while $M_i = 0$ indicates that the ith pixel is in the background; $\rho_n(\cdot,\cdot)$ denotes the similarity metric formula, and we employ the L1 norm in our implementation. Combined with the photometric consistency between the input image $\tilde{\mathbf{I}}_{gt}$ and the output rendering image $\tilde{\mathbf{I}}_r$ in the background portion:

$$L_{bg} = \frac{\sum_i (1-M_i)\rho_{bg}\left(\tilde{\mathbf{I}}_{gt}, \tilde{\mathbf{I}}_r\right)}{\sum_i (1-M_i)}, \tag{7}$$

where $\rho_{bg}(\cdot,\cdot)$ denotes the similarity metric formula, and we adopt the L1 norm in our implementation. By minimizing (6) and (7), we can obtain accurate light source information. This is because only when the correct light source positions and intensities are estimated can we obtain background normals and pixel intensities that match the ground truth.

The contact region often exhibits rich textural features; thus, both L1 loss and SSIM (Structural Similarity Index) loss are used to constrain the similarity between the rendered image and the input image in terms of pixel intensities and structural information over the contact region:

$$L_c = \frac{\sum_i M_i \left(\alpha \frac{1-\mathrm{SSIM}\left(\tilde{\mathbf{I}}_{gt}, \tilde{\mathbf{I}}_r\right)}{2} + (1-\alpha)\left\|\tilde{\mathbf{I}}_{gt} - \tilde{\mathbf{I}}_r\right\|_1\right)}{\sum_i M_i}, \tag{8}$$

where α is the parameter that controls the balance between SSIM and L1, often set to 0.85 in practical applications [25]. Finally, the self-supervised loss function used in the proposed method is:

$$L_{self-sup} = \lambda_n L_n + \lambda_c L_c + \lambda_{bg} L_{bg}, \tag{9}$$

where λ_n, λ_c, and λ_{bg} are coefficients that control the balance between each loss. By harnessing neural networks to execute

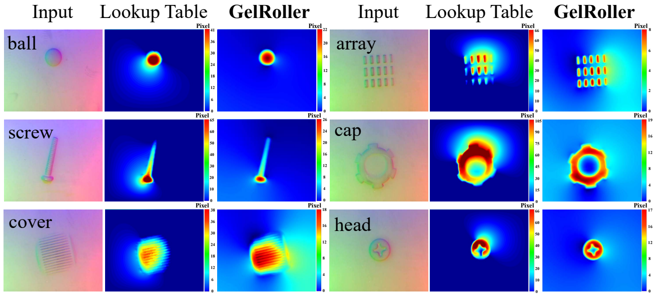

Fig. 6. Comparison of the proposed method with the lookup table method in the reconstruction of object surfaces. Reconstruction results are represented by depth maps. For ease of presentation, we extracted the normals of the contact region and set the normals of the background region to (0, 0, 1), before proceeding with Poisson reconstruction.

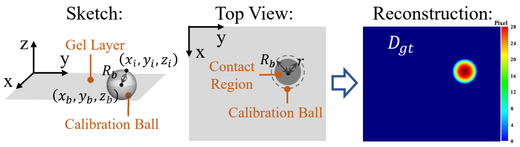

Fig. 7. Reconstruct the ground truth depth map using the geometry of the calibration ball.

the abovementioned strategies, we can derive surface normals for the contact area using only one input image. The surface shape is computed through the use of a fast Poisson solver.

D. Large Surface Reconstruction Using Multiple Frames

Once we have obtained the local surface reconstruction results, we use the ICP method to align the results from multiple frames to achieve large-scale reconstruction. Since ICP is sensitive to the initial pose for iterations, we can employ features like FPFH (Fast Point Feature Histogram) [26] for a coarse matching step to provide an initial pose. However, FPFH feature matching can often result in misalignment in cases where features are not prominent. Thanks to GelRoller's characteristic of moving in a single direction during rolling, we can use translation along the rolling direction as the initial pose when applying ICP.

IV. Experimental Analysis

A. Experimental Setup

The dimensions of the acrylic cylinder used for GelRoller are as follows: inner diameter 61 mm, outer diameter 65 mm, height 90 mm. The average thickness of the gel is 3 mm. The resolution used for the USB camera is 640×480 pixels. We crop the original image to a well-illuminated region of 449×373 pixels for reconstruction, as shown in Fig. 4. The coefficients in the self-supervised loss function, namely λ_n, λ_c, and λ_{bg} are set to 0.4, 3.0, and 2.0, respectively. The network is trained for 30,000 iterations on the first image, and subsequent inputs are fine-tuned for 1,000 to achieve good estimation results. Training for 1,000 iterations using a single NVIDIA Geforce RTX 3090 takes about one and a half minutes.

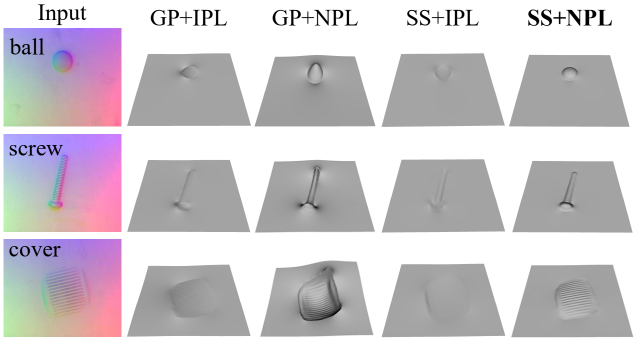

Fig. 8. Using 3D depth maps to present the results of various component ablation experiments. GP: global photometric loss; IPL: infinity point lighting model; NPL: near-point lighting model; SS: the proposed self-supervised loss.

B. Comparison with Lookup Table Method

Fig. 6 compares the performance of the proposed method with the lookup table method [27] in object reconstruction. We pressed a calibration sphere with a diameter of 4 mm at different positions on the gel and captured 38 images to construct the lookup table. On the other hand, all the proposed method needs is just a single captured image. As seen in Fig. 6, the lookup table method relies on pre-calibrated data, and the reconstruction requires identical lighting conditions as during calibration; otherwise, the reconstruction results are unsatisfactory. In contrast, the proposed self-supervised method is not dependent on lighting conditions and can effectively reconstruct the object's surface.

To further quantify the accuracy of the self-supervised method employed by GelRoller in its reconstructions, we collected five sets of tactile images generated by pressing calibration balls with a diameter of 6 mm. Due to the well-defined geometric properties of the spheres, we can approximate the actual size corresponding to each pixel in the contact region based on the radius of the pressed circle in the tactile images. As illustrated in Fig. 7, the calculation process is as follows: assuming the obtained pressed circle radius is r, the calibration sphere ball is R_b, and the coordinates of the calibration ball's center are (x_b, y_b, z_b), then the coordinates of the pixel with index i (represents by (x_i, y_i, z_i)) in the contact region satisfy:

$$(x_i - x_b)^2 + (y_i - y_b)^2 + (z_i - z_b)^2 = R_b^2, \tag{10}$$

where,

$$z_b = -\sqrt{R_b^2 - r^2}, \tag{11}$$

and x_b, y_b can be obtained from the input image. The values of z (unit: pixel) obtained from the solution of (10) can be converted to a reconstructed depth map D_{gt} (unit: mm) using the ratio, f (0.0658 mm/pixel in our experiments). We define the reconstruction accuracy as the mean absolute error (MAE) between the estimated relative depth map $\hat{D}_e$ in the contact region (represents by M) and D_{gt}:

$$\mathrm{MAE} = \frac{\sum_i M_i \left\| \hat{D}_e - D_{gt} \right\|_1}{\sum_i M_i}, \tag{12}$$

TABLE I

THE MAE FOR BOTH METHODS ON THE TACTILE IMAGES (UNIT: MM).

Image Index	#1	#2	#3	#4	#5	Average
Lookup Table	0.320	0.317	0.353	0.317	0.380	0.337
GelRoller	**0.134**	**0.182**	**0.215**	**0.218**	**0.125**	**0.175**

TABLE II

THE AAE FOR BOTH METHODS ON THE TACTILE IMAGES (UNIT: °).

Image Index	#1	#2	#3	#4	#5	Average
Lookup Table	24.68	23.40	24.24	23.54	23.86	23.95
GelRoller	**15.18**	**17.59**	**17.71**	**15.43**	**14.92**	**16.17**

where,

$$\hat{D}_e = D_e - \frac{\sum_i (1 - M_i) D_e}{\sum_i (1 - M_i)}, \tag{13}$$

and D_e is the estimated depth map. Similarly, based on the geometric properties of the spheres, we can obtain the ground truth normal $\mathbf{N}_{gt}$ for each pixel. After obtaining the estimated normal $\mathbf{N}_e$, we can calculate the average angular error (AAE) in the contact region:

$$\mathrm{AAE} = \frac{\sum_i M_i \left(\arccos\left(\mathbf{N}_e \cdot \mathbf{N}_{gt}\right)\right)}{\sum_i M_i}. \tag{14}$$

The MAE and AAE for five sets of images using both the lookup table and the proposed method are shown in Table I and Table II, respectively. It can be observed that the accuracy of the proposed method is significantly higher than that of the lookup table method, with MAE and AAE approximately 50% and 70% of the lookup table method, respectively. This clearly demonstrates that the proposed method is capable of reconstructing surface information of objects with greater accuracy.

C. Ablation Study

To investigate the effectiveness of the proposed self-supervised loss function and the improved near-point lighting (NPL) model, we compared four different setups: using global photometric (GP) loss + infinity point lighting (IPL) model, GP + NPL, self-supervised (SS) loss function + IPL, and SS + NPL (i.e., the proposed method). The global photometric loss function is defined as follows:

$$L_{global} = \left\| \tilde{\mathbf{I}}_{gt} - \tilde{\mathbf{I}}_r \right\|_1. \tag{15}$$

The comparison results are shown in Fig. 8. It can be observed that the effective reconstruction of single-frame images is achieved only when the proposed self-supervised loss function is combined with the near-point lighting model.

D. Large Surface Reconstruction Results

Finally, we mounted GelRoller onto a mechanical gripper and collected data using it as depicted in the first row of Fig. 1. Subsequently, we applied the proposed self-supervised reconstruction algorithm to reconstruct the surface of a large-scale object: a text logo with dimensions of 13.5 mm in length, 2.5 mm in width, and 1 mm in height, as shown in Fig. 9(a). Thanks to the continuous tactile sensing mode of GelRoller, we recorded a video at 30 frames per second while rolling GelRoller along the object's length. Then, 25 frames are selected from the video and fed into the network. Using the point cloud registration methods described in Section III-D, we merged the 25 reconstructed point clouds into a contact one, which is shown in Fig. 9(b). Fig. 9(c) indicates that the reconstructed object is mostly on the same plane, which aligns with the actual scenario. The reconstruction result demonstrates the feasibility of large-scale surface reconstruction using GelRoller.

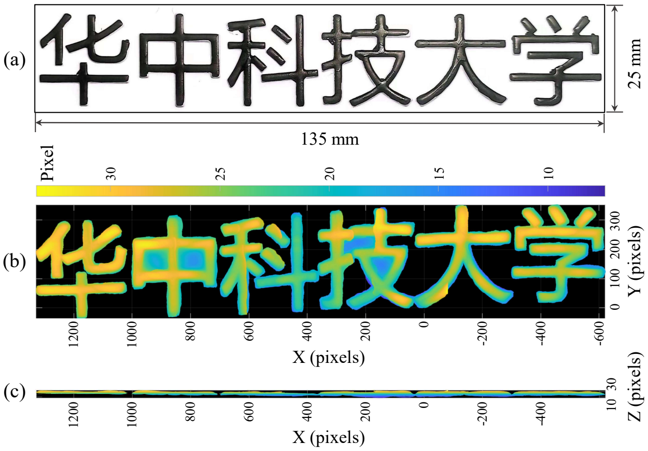

Fig. 9. Large-scale surface reconstruction result. (a) Reconstructed object and its dimensions; (b) Reconstruction results with colors representing height values; (c) Display of height consistency in the reconstruction results.

V. CONCLUSIONS

In this study, we introduced GelRoller sensor based on rolling perception and a self-supervised photometric stereo reconstruction method to achieve high-precision reconstruction of large-scale surfaces. GelRoller enables continuous and swift tactile perception of object surfaces. The self-supervised algorithm can reconstruct the surface with only a single image, eliminating the need for pre-calibration procedures. In the experimental section, we compared our approach with lookup table methods and demonstrated the effectiveness of each module in our proposed method through ablation experiments. The experimental results demonstrate the practicality and potential of GelRoller sensor and self-supervised method in large-scale surface reconstruction.

ACKNOWLEDGMENT

This work was supported by the Joint Funds of National Natural Science Foundation of China (Grant No. U222A20208), the Natural Science Foundation Innovation Group Project of Hubei Province (Grant No. 2022CFA018), and the Key Research and Development Program of Guangdong Province (Grant No. 2022B0202010001-2).

REFERENCES

[1] Alexander C Abad and Anuradha Ranasinghe. Visuotactile sensors with emphasis on gelsight sensor: A review. *IEEE Sensors Journal*, 20(14):7628–7638, 2020.

[2] Umer Hameed Shah, Rajkumar Muthusamy, Dongming Gan, Yahya Zweiri, and Lakmal Seneviratne. On the design and development of vision-based tactile sensors. *Journal of Intelligent & Robotic Systems*, 102:1–27, 2021.

[3] Katsunari Sato, Kazuto Kamiyama, Naoki Kawakami, and Susumu Tachi. Finger-shaped gelforce: sensor for measuring surface traction fields for robotic hand. *IEEE Transactions on Haptics*, 3(1):37–47, 2009.

[4] Naveen Kuppuswamy, Alex Alspach, Avinash Uttamchandani, Sam Creasey, Takuya Ikeda, and Russ Tedrake. Soft-bubble grippers for robust and perceptive manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9917–9924. IEEE, 2020.

[5] Yazhan Zhang, Zicheng Kan, Yu Alexander Tse, Yang Yang, and Michael Yu Wang. Fingervision tactile sensor design and slip detection using convolutional lstm network. *arXiv preprint arXiv:1810.02653*, 2018.

[6] Shaowei Cui, Rui Wang, Jingyi Hu, Junhang Wei, Shuo Wang, and Zheng Lou. In-hand object localization using a novel high-resolution visuotactile sensor. *IEEE Transactions on Industrial Electronics*, 69(6):6015–6025, 2021.

[7] Vijay Kakani, Xuenan Cui, Mingjie Ma, and Hakil Kim. Vision-based tactile sensor mechanism for the estimation of contact position and force distribution using deep learning. *Sensors*, 21(5):1920, 2021.

[8] Huan Ma, Jingjing Ji, and Kok-Meng Lee. Effects of refraction model on binocular visuotactile sensing of 3-d deformation. *IEEE Sensors Journal*, 22(18):17727–17736, 2022.

[9] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[10] Elliott Donlon, Siyuan Dong, Melody Liu, Jianhua Li, Edward Adelson, and Alberto Rodriguez. Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1927–1934. IEEE, 2018.

[11] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg 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*, 5(3):3838–3845, 2020.

[12] Jingjing Ji, Yuting Liu, and Huan Ma. Model-based 3d contact geometry perception for visual tactile sensor. *Sensors*, 22(17):6470, 2022.

[13] Jasper Wollaston James and Nathan F Lepora. Slip detection for grasp stabilization with a multifingered tactile robot hand. *IEEE Transactions on Robotics*, 37(2):506–519, 2020.

[14] Pietro Falco, Shuang Lu, Andrea Cirillo, Ciro Natale, Salvatore Pirozzi, and Dongheui Lee. Cross-modal visuo-tactile object recognition using robotic active exploration. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5273–5280. IEEE, 2017.

[15] Maria Bauza, Antonia Bronars, and Alberto Rodriguez. Tac2pose: Tactile object pose estimation from the first touch. *arXiv preprint arXiv:2204.11701*, 2022.

[16] Guanqun Cao, Jiaqi Jiang, Chen Lu, Daniel Fernandes Gomes, and Shan Luo. Touchroller: A rolling optical tactile sensor for rapid assessment of textures for large surface areas. *Sensors*, 23(5):2661, 2023.

[17] Wuyuan Xie, Chengkai Dai, and Charlie CL Wang. Photometric stereo with near point lighting: A solution by mesh deformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4585–4593, 2015.

[18] Yang Chen and Gérard Medioni. Object modelling by registration of multiple range images. *Image and vision computing*, 10(3):145–155, 1992.

[19] Shixin Zhang, Zixi Chen, Yuan Gao, Weiwei Wan, Jianhua Shan, Hongxiang Xue, Fuchun Sun, Yiyong Yang, and Bin Fang. Hardware technology of vision-based tactile sensor: A review. *IEEE Sensors Journal*, 2022.

[20] Shaoxiong Wang, Yu She, Branden Romero, and Edward Adelson. Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6468–6475. IEEE, 2021.

[21] Daniel Fernandes Gomes, Zhonglin Lin, and Shan Luo. Geltip: A finger-shaped optical tactile sensor for robotic manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9903–9909. IEEE, 2020.

[22] Won Kyung Do and Monroe Kennedy. Densetact: Optical tactile sensor for dense shape reconstruction. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6188–6194. IEEE, 2022.

[23] Junyuan Lu, Zeyu Wan, and Yu Zhang. Tac2structure: Object surface reconstruction only through multi times touch. *IEEE Robotics and Automation Letters*, 8(3):1391–1398, 2023.

[24] Junxuan Li and Hongdong Li. Self-calibrating photometric stereo by neural inverse rendering. In *European Conference on Computer Vision*, pages 166–183. Springer, 2022.

[25] Liang Liu, Guangyao Zhai, Wenlong Ye, and Yong Liu. Unsupervised learning of scene flow estimation fusing with local rigidity. In *IJCAI*, pages 876–882, 2019.

[26] Radu Bogdan Rusu, Nico Blodow, and Michael Beetz. Fast point feature histograms (fpfh) for 3d registration. In *2009 IEEE international conference on robotics and automation*, pages 3212–3217. IEEE, 2009.

[27] Jianhua Li, Siyuan Dong, and Edward H Adelson. End-to-end pixel-wise surface normal estimation with convolutional neural networks and shape reconstruction using gelsight sensor. In *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1292–1297. IEEE, 2018.