Self-Supervised Contact Geometry Learning by GelStereo Visuotactile Sensing

Shaowei Cui, Rui Wang, Jingyi Hu, Chaofan Zhang, Lipeng Chen, and Shuo Wang, *Member, IEEE*

***Abstract*— Vision-based tactile sensors have recently shown promising contact information sensing capabilities in various fields, especially for dexterous robotic manipulation. However, dense contact geometry measurement is still a challenging problem. In this article, we update the design of our previous GelStereo tactile sensor and present a self-supervised contact geometry learning pipeline. Specifically, a self-supervised stereo-based depth estimation neural network (GS-DepthNet) is proposed to achieve real-time disparity estimation, and two specifically designed loss functions are proposed to accelerate the convergence of the network during the training process and improve the inference accuracy. Furthermore, extensive qualitative and quantitative experiments of perceived contact shape were performed on our GelStereo sensor. The experimental results verify the accuracy and robustness of the proposed contact geometry sensing pipeline. This updated GelStereo tactile sensor with dense contact geometric sensing capability has predictable application potential in the field of industrial and service robots.**

***Index Terms*— Depth estimation, robotic sensing systems, self-supervised learning, tactile sensors.**

I. Introduction

THE soft tissue with various tactile afferents attached to the hand deforms when the hand interacts with the object and thus provides information about the object's physical properties and the contact between the object and the hand [1]. This rich tactile information helps us complete daily grasping and dexterous manipulation tasks. Naturally, endowing robots with human-like tactile sensing systems will significantly improve the quality of industrial automation and home robot services [2], [3], which is also a long-standing research field in the robotics and automation communities [4]. Unfortunately, the most existing commercial tactile sensors for robot hands are matrix-distributed force sensors, far from the high-density tactile perception capabilities of human hands [5]. BioTac fingertip sensors use an incompressible liquid as an acoustic conductor to convey vibrations from the skin to a wide bandwidth pressure transducer, which can obtain force, vibration, and temperature information simultaneously [6]. However, the limited number of sensing elements of the BioTac sensor makes it unable to sense the shape of the contacting object, while humans can estimate it. The fingertip tactile system allows humans to roughly measure the geometry of objects in contact, which greatly enhances our perceptual capabilities in everyday grasping and manipulation tasks.

Recently, vision-based tactile sensors have shown promising contact information sensing capabilities, including force, geometry, and slip [7]. The vision retrographic sensing methods can achieve high-resolution tactile sensing even beyond human fingertips [8], which have been shown powerful performance in a variety of robotic manipulation tasks, such as in-hand manipulation [9], cable manipulation [10], and swing-up manipulation [11]. The TacTip sensors start by imitating the structure of the human fingertip tactile afferent, creatively placing pins into the gel layer so that the contact information can be converted into the displacement of the embedded pins [12]. This bionic scheme achieves satisfactory, stable, and efficient tactile information sensing, but it cannot measure contact geometry directly. Yamaguchi and Atkeson [13] develop FingerVision sensors, which adopt a transparent coating design, which gives the sensors proximity sensing ability. Furthermore, Du *et al.* [14] update the FingerVision sensor and propose a high-resolution 3-D deformation tracking method for contact geometry measurement. However, monocular sensing with a dense random color pattern may not meet the requirements of high-precision contact geometry reconstruction.

GelSight tactile sensors benefit from the creative use of the photometric stereo algorithm to reconstruct dense 3-D contact geometry directly [15]. The photometric stereo algorithms are further used in many GelSight-type sensors, such as

Manuscript received September 13, 2021; revised November 18, 2021; accepted November 30, 2021. Date of publication December 16, 2021; date of current version March 2, 2022. This work was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103003, in part by the National Natural Science Foundation of China under Grant U1913201, in part by the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) under Grant XDB32050100, in part by the CIE-Tencent Robotics X Rhino-Bird Focused Research Program, and in part by the Youth Innovation Promotion Association CAS under Grant 2020137. The Associate Editor coordinating the review process was Dr. Anirban Mukherjee. *(Corresponding author: Shuo Wang.)*

Shaowei Cui and Jingyi Hu are with the School of Future Technology, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: cuishaowei2017@ia.ac.cn).

Rui Wang is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Chaofan Zhang is with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Lipeng Chen is with the Tencent Robotics X Laboratory, Shenzhen 518054, China.

Shuo Wang is with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China, and also with the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China (e-mail: shuo.wang@ia.ac.cn).

Digital Object Identifier 10.1109/TIM.2021.3136181

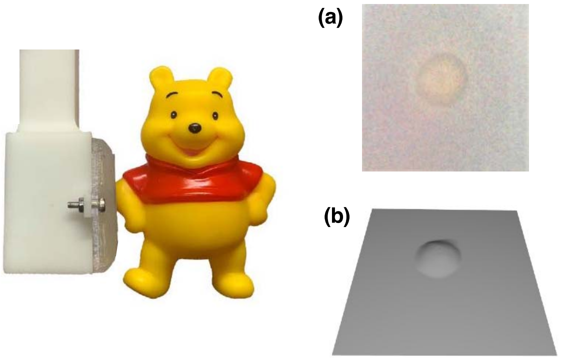

Fig. 1. GelStereo tactile sensor with 3-D high-resolution contact geometry measurement capability. (a) Tactile image. (b) Perceived 3-D contact geometry.

GelSlim [16], DIGIT [17], and GelTip [18]. However, these photometric stereo algorithms have high requirements for the design and fabrication of the light source. It is necessary to ensure that the different colors of light are evenly distributed on the internal contact surface and use a specific calibration board for color-normal registration, which poses a considerable obstacle to the fabrication of the GelSight sensors and ensures the accuracy of contact geometry sensing.

To address the above problems, we present a novel visuotactile sensor named GelStereo [3], which adopts a stereo vision system to capture contact depth information. Different from GelSight sensors using the photometric stereo algorithm for color-normal calibration, our GelStereo sensor uses a binocular stereo matching algorithm for depth estimation, which simplifies the design, fabrication, and calibration cost of the mechanical structure and 3-D reconstruction system. In our previous study [3], GelStereo sensors realize high-resolution ($<$1 mm) 3-D tactile point cloud sensing by a specific stereo matching algorithm.

In this article, we further address the high-resolution 3-D contact geometry reconstruction problem, as shown in Fig. 1. We first consider using the mainstream stereo matching deep learning network [19] for disparity estimation and then reconstruct the depth information according to the triangulation principle. Unfortunately, the disparity labels used for supervised training are hard to obtain due to the limited imaging space of the GelStereo sensor. Inspired by the rapid progress in the field of unsupervised depth estimation [20]–[22], we address the GelStereo depth estimation problem by self-supervised disparity estimation using binocular stereo correspondences.

Similar to [14], we first update our GelStereo sensor by replacing the markers pattern with a semitransparent color pattern, which provides more texture for disparity estimation. Specifically, a self-supervised depth estimation neural network (GS-DepthNet) is proposed to inference contact depth information, and stereo correspondences between the binocular views of our updated GelStereo sensor are used to guide the network optimization. Furthermore, we observe that effective depth information is concentrated in the contact area, and depth calculations in other areas will result in a lot of redundant calculations. To this end, a contact semantic loss function is proposed to accelerate the convergence of the network and improve depth estimation accuracy. We also present a weighted left–right disparity consistency loss function to reduce the uncertainty during network optimization. Finally, extensive experiments are performed to verify the accuracy and generalization of the proposed self-supervised contact geometry sensing pipeline. The contributions of this study are summarized as follows.

1) A 3-D high-resolution contact geometry learning pipeline is proposed based on our updated GelStereo sensor. Specifically, a self-supervised stereo-based depth estimation neural network (GS-DepthNet) is presented using stereo correspondence guidance for training.
2) Since the contact area has greater significance in depth information, we propose a novel training loss that enforces consistency between the estimated disparities and contacts' semantic information, leading to improved performance and better network convergence than existing approaches. Moreover, we also present a weighted left–right disparity consistency loss function to reduce the uncertainty during network optimization.
3) Extensive qualitative and quantitative analyses of the perceived contact geometry are performed on the proposed GelStereo sensor. The experimental results verify the accuracy and generalization of the proposed high-resolution 3-D contact geometry measurement pipeline.

The rest of this article is organized as follows. In Section II, a brief introduction of the updated GelStereo sensor is presented. The self-supervised depth estimation neural network (GS-DepthNet) that uses binocular stereo correspondences is provided in Section III. The experimental setup and results are introduced in Sections IV and V, respectively. Finally, we conclude the proposed 3-D contact geometry measurement method in Section VI.

II. GelStereo Sensor

The main principle of the GelStereo sensor is to convert the tactile information into the geometric deformation of the silicon gel layer and adopt visual sensing methods to capture various tactile information contained in the 3-D contact geometry [3]. To improve the accuracy and resolution of the contact geometry, we update our previous design by replacing the dot pattern with a semitransparent dense color pattern. This colored pattern helps add texture to the contact imaging surface, making the disparity estimation more effortless and accurate.

An exploded view of the updated GelStereo sensor is shown in Fig. 2(a). The proposed sensor consists of three parts: a binocular camera module, a supporting tray, and a contact silicon gel module. Specifically, the binocular camera module we selected is small and can obtain stereo frames at over 30 FPS with just one USB port connection. The supporting tray is fabricated using white photosensitive resin, and the attached flexible LED strip provides a stable and uniform light field, thanks to the diffuse reflection propriety of the photosensitive resin. The contact silicon gel module is fabricated by an acrylic plate, a colored water-transfer pattern, and silicon gel with a shore hardness of 18 A (close to human skin). This compact

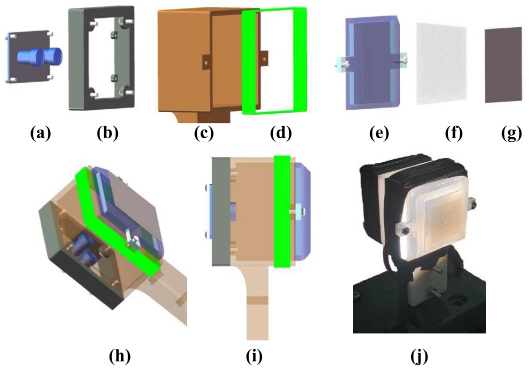

Fig. 2. Design of the proposed updated GelStereo tactile sensor. (a) Binocular camera module. (b) Camera support frame. (c) GelStereo tray. (d) LED strip. (e) Gel layer. (f) Water-transfer paper with a color pattern. (g) Painting layer of the GelStereo Sensor. (h) and (i) Side views of an assembled sensor. (j) Shot of the finished sensor.

design guarantees that the deformation of the elastomer will be represented by the changes of the pattern with high resolution and durability. Furthermore, these three modules are fabricated separately and connected by screws, and this separation design can ensure that a module can be easily replaced when it is broken.

III. Method

This section describes our self-supervised contact geometry learning pipeline using stereo correspondences. We first describe how we transform the GelStereo geometry measurement problem into an image reconstruction task and then the detailed architecture of the proposed GS-DepthNet is then introduced. Furthermore, besides the popular loss function used for the image reconstruction task, we also present two novel loss functions for the network optimization, which are specifically for the contact geometry sensing of the GelStereo sensor.

A. Geometry Measurement as Image Reconstruction

Our goal is to learn a function f that can predict the 3-D high resolution contact geometry, that is, the per-pixel scene depth, $d = f(P)$, given a rectified image pair $P = \{I^l, I^r\}$. Most existing learning-based methods treat this as a supervised problem, such as stereo matching or single image depth estimation, where they have color input images and their corresponding per-pixel target disparity (depth) for training. Unfortunately, it is not practical to acquire such ground-truth (GT) disparity (depth) data for the GelStereo sensor. On the one hand, the popular depth information acquisition devices, such as Intel real-sense cameras, cannot obtain reliable and precise depth in the limited imaging space of the sensor. On the other hand, the calculation methods [23], [24] of stereo disparity based on structured light are also difficult to perform in such a limited space. The intuition behind the proposed method is that, given a calibrated pair of GelStereo binocular cameras, if we can learn a function that is able to reconstruct one image from the other, then the transformation information about the 3-D contact geometry can be learned.

Specifically, the GelStereo sensor addresses two images, I^l and I^r, corresponding to the left and right color images from a calibrated stereo pair. Instead of attempting to predict the per-pixel depth directly, we attempt to learn the dense correspondence map d^l that, when applied to the left image, allows us to reconstruct the right image, also known as $\tilde{I}^r$. Similarly, we can estimate the left image given the right one, $\tilde{I}^l = f(I^r, d^r)$, and $\text{Diff}(\tilde{I}^l, I^l)$ and $\text{Diff}(\tilde{I}^r, I^r)$ both can provide supervision information about the pixel-level correspondences learning. Furthermore, assuming that the images are rectified, d^l and d^r correspond to the image disparity, a scalar value per pixel can be used to reconstruct the geometry. Given the baseline distance b between the cameras and the camera focal length f, we can easily recover the dense depth D from the predicted disparity, $D = bf/d$, and the 3-D high-resolution geometry can be reconstructed by estimating normal on the obtained D.

In this way, the 3-D contact geometry measurement task can be defined as a self-supervised disparity estimation task using rectified stereo pairs, which is easy to capture and does not need additional manual calibration.

B. GS-DepthNet

Given $P = \{I^l, I^r\}$, the proposed GS-DepthNet estimates contact per-pixel depth by inferring left and right disparities (d^l and d^r), which can be enforced to be consistent with each other. Furthermore, the learned disparities generate color images, $\tilde{I}^l$ and $\tilde{I}^r$, by dense sampling from the input color images. Instead of only using the left image [20], we input the network with stereo pairs simultaneously.

Fig. 3 illustrates the architecture of the proposed GS-DepthNet network. Inspired by Mayer *et al.* [24], we adopt a similar encoder–decoder architecture in this image reconstruction task. The decoder uses skip connections [25] from the encoder's activation blocks, enabling it to learn at a higher resolution. Given predicted disparities, d^l and d^r, the GS-DepthNet network generates the predicted image using a bilinear sampler, resulting in a fully differentiable image formation model. At each inference, the network gradient can be optimized by left–right color image reconstruction losses and other disparity constraint losses.

C. Image Reconstruction Loss Function

Similar to other depth estimation networks [26], we train the proposed network with image appearance loss and some specific disparity constraint losses. Note that we output disparity predictions at four different scales (see Fig. 3), which double in spatial resolution at each of the subsequent scales. We define a loss $\mathcal{L}_s$ at each output scale s, forming the total loss at the sum $\mathcal{L} = \sum_{s=1}^{4} \mathcal{L}_s$. In this article, we formulate the loss module $\mathcal{L}_s$ as a combination of four terms

$$\mathcal{L}_s = w_{\text{ap}}\left(\mathcal{L}^l_{\text{ap}} + \mathcal{L}^r_{\text{ap}}\right) + w_{\text{ds}}\left(\mathcal{L}^l_{\text{ds}} + \mathcal{L}^r_{\text{ds}}\right) + w_{\text{lr}}(\mathcal{L}_{\text{lr}}) + w_{\text{se}}\left(\mathcal{L}^l_{\text{se}} + \mathcal{L}^r_{\text{se}}\right) \quad (1)$$

where $\mathcal{L}_{\text{ap}}$ encourages the reconstructed image to look similar to the corresponding input one, $\mathcal{L}_{\text{ds}}$ prefers the smoother disparities, $\mathcal{L}_{\text{lr}}$ enforces the predicted left and right disparities to be consistent, and $\mathcal{L}_{\text{se}}$ helps the network optimization only focusing on the contact area.

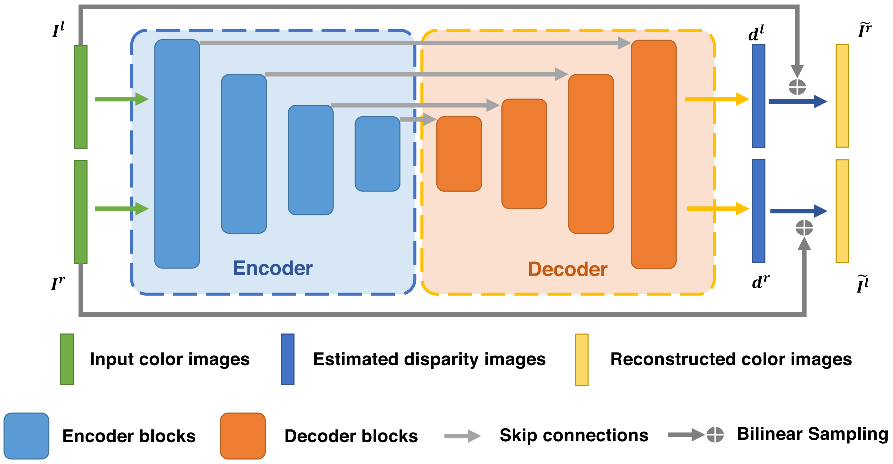

Fig. 3. Architecture of the proposed GS-DepthNet network. Given a rectified GelStereo color image pair, I^l, I^r, the proposed GS-DepthNet aims to learn the corresponding disparities, d^l, d^r, using an encoder–decoder architecture. The estimated left (right) disparity is then applied to the input color images to generate reconstructed right (left) image, and the image reconstruction losses and other disparities constraint losses are used to supervise the network optimization.

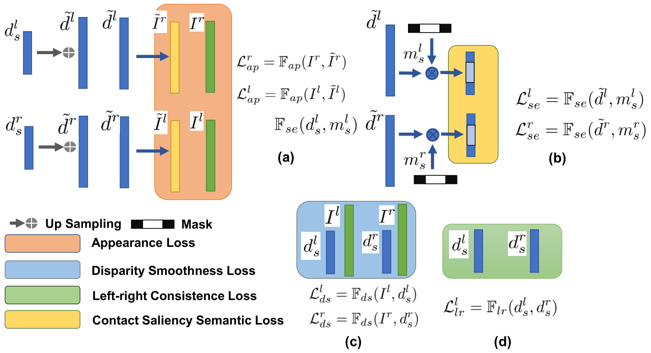
d_s^l
$\tilde{d}^l$
$\tilde{d}^l$
$\tilde{I}^r$
I^r
d_s^r
$\tilde{d}^r$
$\tilde{d}^r$
$\tilde{I}^l$
I^l
$\mathcal{L}_{ap}^r = \mathbb{F}_{ap}(I^r, \tilde{I}^r)$
$\mathcal{L}_{ap}^l = \mathbb{F}_{ap}(I^l, \tilde{I}^l)$
$\mathbb{F}_{se}(d_s^l, m_s^l)$
(a)
$\tilde{d}^l$
m_s^l
$\tilde{d}^r$
m_s^r
$\mathcal{L}_{se}^l = \mathbb{F}_{se}(\tilde{d}^l, m_s^l)$
$\mathcal{L}_{se}^r = \mathbb{F}_{se}(\tilde{d}^r, m_s^r)$
(b)
Up Sampling
Mask
Appearance Loss
Disparity Smoothness Loss
Left-right Consistence Loss
Contact Saliency Semantic Loss
I^l
I^r
d_s^l
d_s^r
$\mathcal{L}_{ds}^l = \mathbb{F}_{ds}(I^l, d_s^l)$
$\mathcal{L}_{ds}^r = \mathbb{F}_{ds}(I^r, d_s^r)$
(c)
d_s^l
d_s^r
$\mathcal{L}_{lr}^l = \mathbb{F}_{lr}(d_s^l, d_s^r)$
(d)

Fig. 4. Diagram of our proposed loss module for training GS-DepthNet. At each disparity scale s, the decoder block outputs the left and right disparities d_s^l and d_s^r. We up-sample depth predictions at each scale to input resolution and compute all losses. (a) Appearance losses are obtained by computing the differences between the reconstructed color image and the inputs. (b) Predicted disparities are masked by the contact masks, which are output by a pretrained semantic network, and the contact semantic losses are computed by the masked disparities. (c) Disparity smoothness losses are computed by the estimated disparities and input images. (d) Weighted left–right disparity consistency loss is computed by d_s^l and d_s^r directly.

Fig. 4 illustrates how the loss module works in each disparity scale. Inspired by the improvement of multi-scale depth estimation in [26], we first up-sample depth predictions at each intermediate layer to the input resolution ($\tilde{d}^l$ and $\tilde{d}^r$) for appearance and contact semantic losses, reducing texture-copy artifacts. Next, we take the left image as an example to introduce our loss items successively.

1) Appearance Loss: During training, the most critical loss for supervising the network optimization is the image reconstruction loss. Given the left color image I^l, our network learns to estimate a right image by sampling pixels from it. In this article, we adopt the bilinear sampler to complete this sampling operation [25]. The bilinear sampler is locally fully differentiable and integrates seamlessly into our fully convolution architecture, which means that we do not require any simplification or approximation of our cost function.

Referring to the existing monocular depth estimation image reconstruction loss functions, we define the appearance loss function $\mathcal{L}_{\text{ap}}^l$ as a combination of an $L1$ and a single scale SSIM [27] term

$$\begin{aligned}\mathcal{L}_{\text{ap}}^l &= \mathbb{F}_{\text{ap}}\left(I^l, \tilde{I}^l\right)\\ &= \frac{1}{N}\sum_{i,j}\alpha\frac{1-\text{SSIM}\left(I_{ij}^l, \tilde{I}_{ij}^l\right)}{2} + (1-\alpha)L1\left(I_{ij}^l, \tilde{I}_{ij}^l\right)\end{aligned} \quad (2)$$

where N is the number of pixels in images, and I_{ij}^l and $\tilde{I}_{ij}^l$ indicate the $\{i, j\}$th pixel of the input image I^l and the corresponding reconstructed image $\tilde{I}^l$, respectively. α is a balance coefficient between SSIM and $L1$ loss, which we set it as 0.85 in this article.

2) Disparity Smoothness Loss: The disparity smoothness loss is used to enforce that the estimated disparities are locally smooth, which is a commonly used loss term in depth estimation problems. We formulate this loss term as an $L1$ penalty on the disparity gradients ∂d and weight it with the different directions of the gradients of image ∂I

$$\begin{aligned}\mathcal{L}_{\text{ds}}^r &= \mathbb{F}_{\text{ds}}\left(I^r, d_s^r\right)\\ &= \frac{1}{N}\sum_{i,j}\left|\partial_x d_{i,j}^l\right| e^{-L1\left(\partial_x I_{i,j}^l\right)} + \left|\partial_y d_{i,j}^l\right| e^{-L1\left(\partial_y I_{i,j}^l\right)}\end{aligned} \quad (3)$$

where $\partial_x d_{i,j}$ and $\partial_x I_{i,j}^l$ indicate the gradients of the disparity and image in the x-direction, respectively.

D. GelStereo Loss Function

Besides the appearance and disparity smoothness loss that are usually used for image reconstruction tasks, we also present two novel loss functions that specially designed for GelStereo sensing.

1) Weighted Left–Right Disparity Consistency Loss: In this stereo-based disparity estimation task, the proposed GS-DepthNet network predicts both the left and right disparities d^l and d^r. However, the two disparities are not theoretically, d^l and d^r should be consistent with each other at pixel-level. To bring this coherence to the network training process, we introduce an $L1$ left–right disparity consistency penalty as part of our loss module. Moreover, due to the problem of multi-medium refraction, these two disparities are not completely the same in our sensing scene. According to our observations, the central area is basically the same, but the farther away from the central area, the consistency decreases. As a result, we present a decay weight for this loss term, w_i, which is a linear function of the abscissa (i) of each disparity value $(d_{i,j})$

$$\begin{aligned}\mathcal{L}_{\mathrm{lr}} &= \mathbb{F}_{\mathrm{lr}}\left(d_s^l, d_s^r\right)\\ &= w_i * \left(\frac{1}{N}\sum_{i,j}\left|d_{ij}^l + d_{ij+d_{ij}^l}^r\right| + \left|d_{ij}^r + d_{ij+d_{ij}^r}^l\right|\right) \quad (4)\\ w_i &= 2/W \times (W/2 - \mathrm{abs}(i - W/2)) \quad (5)\end{aligned}$$

where $d_{ij+d_{ij}^l}^r$ means the estimated disparity value corresponding to the translation of $d_{i,j}^l$ pixels in d_{ij}^r, which should be the opposite number of d_{ij}^l. In other words, the obtained left–right disparities d_{ij}^l and d_{ij}^r should have the opposite value in each pixel. W denotes the width of the image in pixels.

2) Contact Semantic Loss: This semantic loss is the critical insight of the designed loss module and is designed specifically for our GelStereo situation. We observe that effective depth information is concentrated in the contact area, and depth calculations in other areas will bring a lot of redundant calculations. To this end, we present this semantic loss to penalize the valid depth in the non-contact area of the GelStereo Sensor. Similarly, we formulate this loss using $L1$ penalty

$$\mathcal{L}_{\mathrm{se}}^l = \mathbb{F}_{\mathrm{se}}\left(\tilde{d}^l, m_s^l\right) = \frac{1}{M}\sum_{i,j}\left|\tilde{d}_{ij}^l\right| \times m_{ij}^l \quad (6)$$

where m_{ij}^l is a mask indicator scalar. When $\{i, j\}$ belongs to the contact area, we set its value as 0, otherwise it is 1. M is the number of pixels that are not in the contact area, and the contact area is determined by the contact semantic mask output of a pretrained semantic segmentation network.

E. Disparity to Geometry

After the network inference, we obtain the left and right disparities at full resolution d^l and d^r. We define the left camera optical center coordinate system as the contact coordinate system in this article. The disparity d^l are first averaged by d^r

$$\hat{d}_{ij}^l = \frac{\left|d_{ij}^l - d_{ij+d_{ij}^l}^r\right|}{2}. \quad (7)$$

Next, the dense 3-D tactile point cloud can be computed by the triangulation principle of the binocular vision system. Given the disparity value at $\{i, j\}$th pixel, its corresponding 3-D tactile point P can be calculated by

$$P = Q \times p \quad (8)$$

$$p = \left[i, j, \hat{d}_{ij}^l, 1\right] \quad (9)$$

where i and j indicate the horizontal and vertical axis coordinates of I, respectively. The disparity of this point is denoted by $\hat{d}_{ij}^l$, and p is the homogeneous coordinate of the point in the image coordinate system. Q is a 4 × 4 perspective transformation matrix computed by stereo calibration. In this way, the estimated disparity can be converted into a dense 3-D point cloud in the contact coordinate system by triangulation and normal estimation.

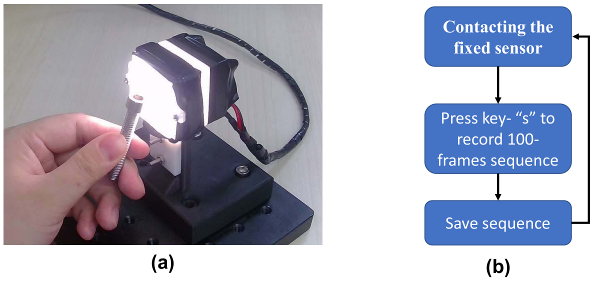
Contacting the fixed sensor
Press key- "s" to record 100-frames sequence
Save sequence
(a)
(b)

Fig. 5. (a) GelStereo data collection equipment. (b) Block diagram of the data collection process.

IV. Experiments: Design and Setup

The primary goal of our experiments is to examine the accuracy and generalization of the proposed 3-D high-resolution contact geometry measurement method sensed by the GelStereo sensor. In this section, we design a series of qualitative and quantitative experiments to answer the following two questions.

1) How are the measurement accuracy and generalization ability of the proposed self-supervised contact geometry sensing pipeline when dealing with different contact situations?
2) Does the proposed GS-DepthNet outperform the existing methods for this task?

To train the proposed self-supervised depth estimation network, we first introduce a GelStereo contact dataset built by capturing the GelStereo sensor's reading when contacting different objects with various contact configurations. Furthermore, to quantitatively evaluate the accuracy of the perceived disparities, we designed a GT disparity collection pipeline and generated a small batch of disparity data with GT labels.

A. GelStereo Contact Dataset

The GelStereo contact dataset is built to provide extensive rectified stereo image pairs used to train the proposed GS-DepthNet. Specifically, we design a data collection equipment to collect GelStereo images, which ensures that the data collection can be done by one person, as shown in Fig. 5(a). When collecting, we use one hand to contact the object with the collection setup in different contact configurations, and the other hand interacts with the keyboard to save the data. In this article, we set the key "s" as the trigger key of the save function, which can save 100 consecutive frames of binocular image sequences. A block diagram of this collection process is shown in Fig. 5(b).

Totally, we collected about 48k contact stereo image pairs, which are generated by contacting about 40 objects, and some of these objects are shown in Fig. 6.

Fig. 6. Some examples of objects in contact with the GelStereo sensor.

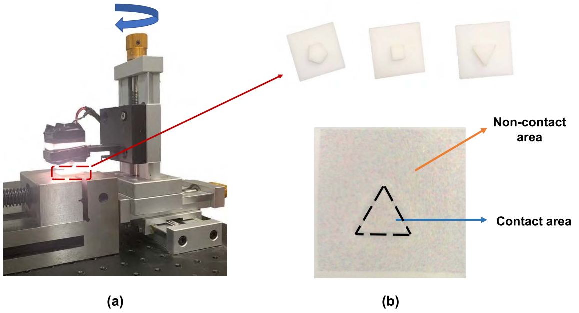

Fig. 7. GT disparity acquisition pipeline. (a) Data collection setup. (b) Schematic of manual labeling result.

B. GT Disparity Acquisition

Fig. 7(a) illustrates the GT disparity acquisition setup, which includes an optical board, a 3-D operation slide rail, a fixing tong, and a fixed GelStereo sensor. The GelStereo sensor and various workpieces with flat surfaces are fixed at the slide rail and tongs, respectively, during collection, and both the slider and tongs are fixed at the optical platform. In this way, we use mechanical constraints to ensure that the disparity map generated by the contact has a smooth and consistent value in the contact area, which can be treated as GT disparities.

The pipeline consists of three steps.

1) Fix a workpiece on the tongs and adjust the contact state by operating the slide rail.
2) Determine the contact area manually for each sample and compute the contact mask.
3) Compute the disparity value of contact and non-contact areas by the Blob Matching (BM) algorithm [28].

For the first step, we make contact between each workpiece and ten different sensor surface areas and sample stereo pairs from 0.5- to 1.5-mm contact depth at 0.5-mm intervals. For each workpiece, 30 stereo samples are collected, and a total of 90 samples for three workpieces are collected to extract GT disparities.

Fig. 7(b) shows an example of the contact area labeled manually. The surface of the sensor can be divided into two parts: contact area and non-contact area. Finally, the disparity value in contact and the non-contact area is determined semi-automatically, computed by an average disparity obtained by a robust stereo matching algorithm provided by OpenCV.

C. Implementation Details

The network, which is implemented by PyTorch [29], contains 16 million trainable parameters and takes about 6 h to train using four RTX GPU on our contact stereo dataset of 50k image pairs for 40 epochs. The proposed model is trained from scratch, with a batch size of 32 using Adam, where $\beta_1 = 0.9$, $\beta_2 = 0.900$, and $\epsilon = 10e^{-8}$. We use an initial learning rate of $\lambda = 1e^{-4}$ which we keep constant for the first 30 epochs before halving it every 10 epochs until the end. For a 256 × 256 image inference, RTX GPU takes less than 25 ms, and GTX 3070 Laptop GPU takes about 45 ms.

The architecture of our semantic segmentation model is FPN, and we select RedNeXt as its encoder in this article. For training this segmentation model, it takes about 1 h using one RTX GPU on 500 images, sampled and labeled manually uniformly from the contact stereo dataset, for 100 epochs. The proposed model is trained from scratch, with a batch size of 32 using Adam, and the learning rate is set as $1e^{-4}$.

For the loss module, we set the weights of appearance loss, left–right disparity consistency loss, and contact semantic term to $w_{\mathrm{ap}} = 1$, $w_{\mathrm{lr}} = 1$, and $w_{\mathrm{se}} = 1e^{-4}$, respectively. As a result of multi-scale output, the typical disparity of neighboring pixels will differ by a factor of 2 between each scale (as we are upsampling the output by a factor of 2). To correct this, we multiply the disparity smoothness term w_{ds} by r for each scale to obtain equivalent smoothing at each level. The weight of the disparity smoothness loss is set to $w_{\mathrm{ds}} = 0.01/r$, where r is the down-scaling factor of the corresponding layer in relation to the resolution of the input image. Using a scaled sigmoid nonlinearity, the possible output disparities are constrained to be between 0 and $d_{\max} = 12$ pixels at a given output scale (256 × 256).

D. Baselines Comparisons

We also compare the proposed GS-DepthNet with some existing methods. The first category is general stereo matching algorithms, including BM algorithms [28] and semi-global matching method (SGBM) [30]. The second category is self-supervised stereo matching, and the most representative of which is SsSMNet [31]. Besides, we also test monocular unsupervised depth estimation methods such as MonoDepth [20] and MonoDepth2 [26], in this article.

We evaluate each method using several evaluation methods form prior studies [32].

1) *Threshold:* % of y_i s.t. $\max(y/y^*, y^*/y) = \delta <$ thr.
2) *Absolute Relative difference (Abs Rel):* $(1/|T|)\sum_{y\in T}|y - y^*|/y^*$.
3) *Squared Relative difference (Sq Rel):* $(1/|T|)\sum_{y\in T}||y - y^*||^2/y^*$.
4) *RMSE:* $((1/|T|)\sum_{y\in T}||y - y^*||^2)^{1/2}$.
5) *RMSE (log):* $((1/|T|)\sum_{y\in T}||\log y - \log y^*||^2)^{1/2}$.

Here, T is the number of image pairs for evaluation, and y and y^* indicate the predicted and labeled disparity maps, respectively. Note that we perform the error evaluation on perceived disparities directly instead of the traditional depth value, which is hard to obtain.

V. Experiments: Results

A. Quantitative Results

Table I shows quantitative results with some examples of disparities shown in Fig. 8. The results show that our method outperforms all other reference methods. The traditional stereo matching methods perform much better than the existing unsupervised learning-based methods. This is because the matching scenarios contained in our test dataset are simple, and the

TABLE I

Disparity Error Analysis Results of Different Methods. S: Traditional Stereo Matching Methods. U+S: UnSupervised/Self-Supervised Stereo Matching Methods. U+M: Unsupervised Monocular Depth Estimation Methods. Our No LR: The Proposed Model Without the Left–Right Disparity Consistency Loss Supervision. Our No SE: Our Model Without the Contact Semantic Loss Supervision. Light Green Terms: Lower Is Better. Dark Green Terms: Higher Is Better

Method	Category	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
BM (block_size, 7) [28]	S	1.353	11.407	3.353	12.852	0.707	0.713	0.714
BM (block_size, 15)	S	0.456	1.815	1.329	10.366	0.796	0.801	0.802
BM (block_size, 21)	S	0.367	1.131	1.037	9.738	0.818	0.823	0.823
BM (block_size, 33)	S	0.296	0.935	0.929	8.651	0.857	0.861	0.861
SGBM (block_size, 7) [30]	S	2.710	24.058	4.859	17.178	0.488	0.495	0.496
SGBM (block_size, 15)	S	1.611	12.777	3.472	14.972	0.603	0.608	0.609
SGBM (block_size, 21)	S	1.312	9.571	2.974	14.247	0.637	0.642	0.643
SGBM (block_size, 33)	S	0.615	2.822	1.627	11.467	0.759	0.762	0.763
SsSMNet [31]	U+S	2.541	9.728	3.113	11.712	0.402	0.625	0.706
MonoDepth [20]	U+M	2.201	9.068	3.006	11.328	0.509	0.711	0.792
MonoDetph2 [26]	U+M	2.141	8.494	2.909	11.340	0.463	0.673	0.756
Ours no LR	U+S	0.275	0.975	0.956	3.098	0.896	0.918	0.928
Ours no SE	U+S	1.264	5.786	1.872	9.573	0.635	0.718	0.746
Ours	U+S	**0.243**	**0.769**	**0.815**	**2.157**	**0.926**	**0.937**	**0.942**

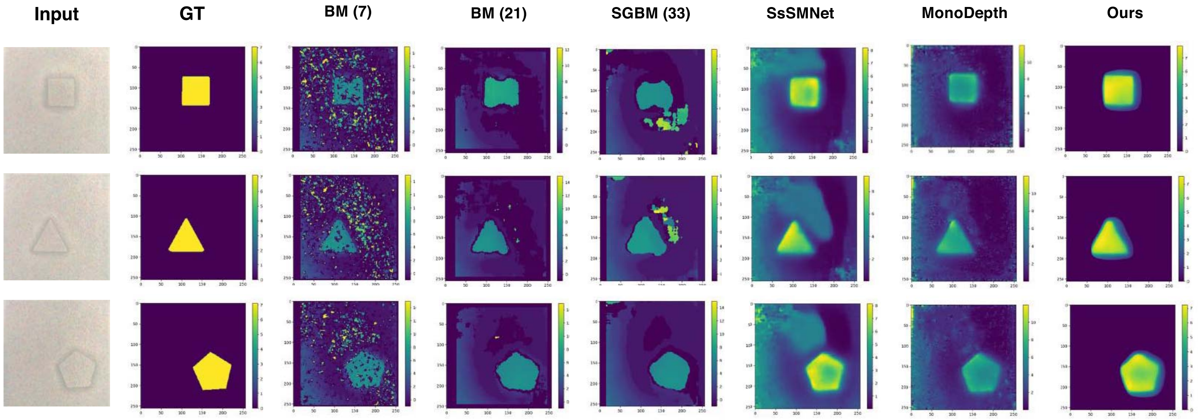

Fig. 8. Obtained disparity visualization of various methods, and the proposed GS-DepthNet achieves superior qualitative results on our test dataset.

learning-based method will lead to local convergence, which introduces much noise into the inference process. Furthermore, we also find that the learning-based methods achieve smoother and more accurate prediction than the traditional methods in the contact area, where is more important for contact geometry measurement.

Ablation Studies We also perform ablation studies on the proposed model to verify the effect of each term in the designed loss module. The appearance and disparity smoothness losses are essential and have been proved to be indispensable in the depth estimation problems [20]. Therefore, we only perform ablation studies on the weighted left–right disparity consistency loss and contact semantic loss, which are specifically designed for the GelStereo sensing.

From Table I, we can find that both LR and SE loss items are significant for this GelStereo depth estimation problem. Specifically, the SE loss is more critical than the LR item, consistent with our observation and intuition. As shown in Fig. 8, the obvious disadvantage of the existing methods is that they cannot ignore the depth information of the non-contact area, which brings great interference during the network optimization. The experimental results also show that the proposed GS-DepthNet, which incorporates contact semantic information, achieves satisfactory performance.

TABLE II

Measurement Error Mean/Std Analysis (mm). Tri: Triangle Workpiece. Quad: Quadrilateral Workpiece. Penta: Pentagon Workpiece

Work-piece		1.5-0.5	1.5-1.0	1.0-0.5	All
Tri	Mean	**-0.105**	-0.079	-0.035	-0.076
	Std	0.094	0.081	0.049	0.064
Quad	Mean	-0.099	-0.083	-0.441	-0.067
	Std	0.083	0.072	0.357	0.059
Penta	Mean	-0.045	**0.006**	0.027	-0.014
	Std	0.051	0.025	0.018	0.029
All	**Mean**	**-0.083**	**-0.052**	**-0.019**	**-0.051**
	Std	**0.064**	**0.039**	**0.042**	**0.047**

Furthermore, we also evaluate the accuracy of the measured depth obtained by the proposed 3-D contact geometry pipeline. In particular, we manually collect the measured depth of each workpiece in the contact area when measuring 0.5, 1.0, and 1.5 mm to determine three relative measurement differences of $1.5 - 0.5$ mm, $1.5 - 1.0$ mm, and $1.0 - 0.5$ mm, which are used for error analysis.

Table II shows that the measured geometric error is about 0.05 mm on average, and the maximum error is less than 0.15 mm, which is completely tolerable in the daily

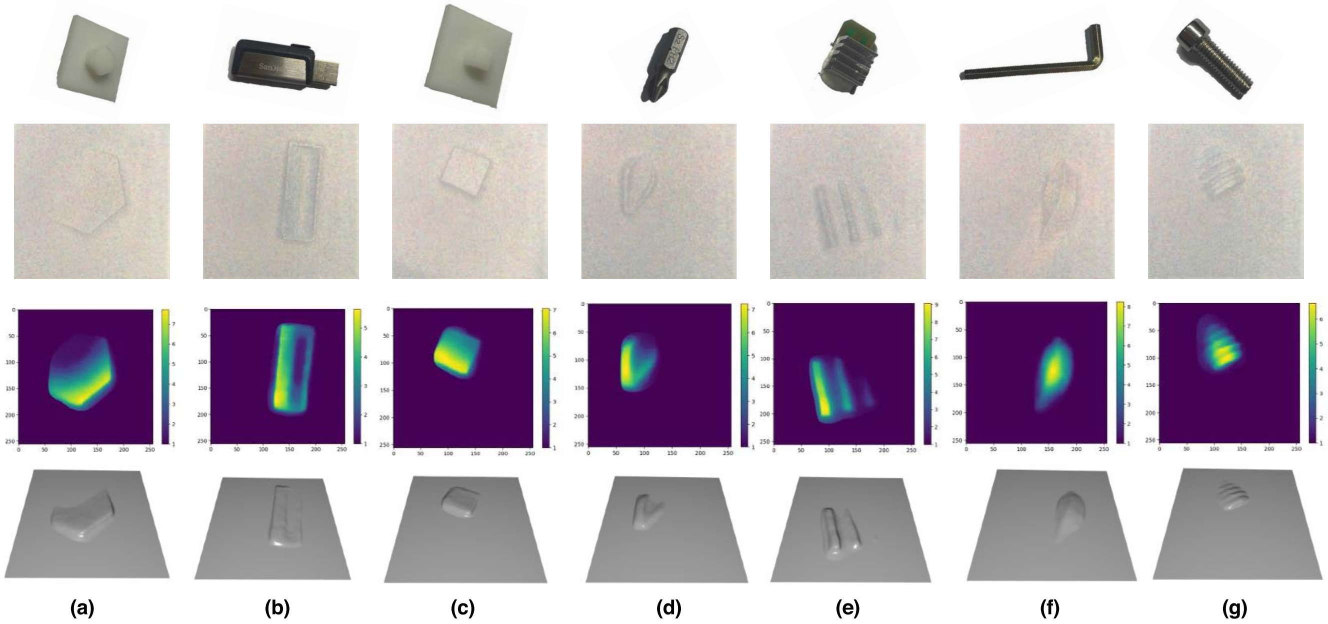

Fig. 9. 3-D contact geometry visualization of the GelStereo sensor contact with different objects. First row: objects. Second row: tactile images. Third row: perceived tactile disparities. Last row: the corresponding 3-D high-resolution contact geometry of each contact. (a) Hexagonal workpiece. (b) USB flash drive. (c) Quadrilateral workpiece. (d) Screwdriver head. (e) Heat sink. (f) L-shaped tool. (g) M3 screw.

manipulation tasks, verifying the accuracy of the proposed 3-D high-resolution contact geometry measurement method from a quantitative perspective.

B. Contact Geometry Visualization

Fig. 9 visualizes the qualitative results when the sensor contacts various objects with different contact configurations. The first row shows various contacted objects, such as screws, keys, and so on. Note that these objects are not used for training the network. When the GelStereo sensor contacts these objects, the corresponding perceived tactile images, disparity maps, and high-resolution geometry are shown in the second, third, and last row, respectively.

These tactile images show that the proposed contact geometry sensing method can obtain high-resolution and contact-sensitive 3-D tactile information, particularly when contacting an M3 screw [see Fig. 9(d)], demonstrating the efficacy of the proposed method and the updated GelStereo sensor.

VI. Conclusion

This study presents a 3-D high-resolution contact geometry sensing pipeline using an updated GelStereo sensor. Specifically, a self-supervised stereo-based depth estimation network is proposed to address the GelStereo contact geometry measurement problem. Furthermore, we design two novel loss functions to accelerate the network convergence and improve the inference accuracy. Extensive quantitative and qualitative experiments are performed to verify the accuracy and generalization of the proposed method. The results show that the GS-DepthNet outperforms existing unsupervised stereo matching (depth estimation) methods in terms of qualitative performance, and the accuracy of the proposed geometry measurement pipeline is within 0.15 mm. Besides, the contact geometry visualization results indicate that the proposed pipeline can handle various contact situations, which verifies its generalization.

In the future, we will explore valuable industrial tasks using this contact geometric information provided by our GelStereo sensor, such as tactile 3-D object reconstruction, tactile SLAM, industrial defect detection, and so on.

References

[1] R. S. Johansson and J. R. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature Rev. Neurosci.*, vol. 10, no. 5, pp. 345–359, Apr. 2009.

[2] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Trans. Robot.*, vol. 36, no. 6, pp. 1619–1634, Dec. 2020.

[3] S. Cui, R. Wang, J. Hu, J. Wei, S. Wang, and Z. Lou, "In-hand object localization using a novel high-resolution visuotactile sensor," *IEEE Trans. Ind. Electron.*, early access, Jun. 24, 2021, doi: 10.1109/TIE.2021.3090697.

[4] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, Jun. 2019, Art. no. eaat8414.

[5] 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, Oct. 2021.

[6] Y. S. Narang, B. Sundaralingam, K. Van Wyk, A. Mousavian, and D. Fox, "Interpreting and predicting tactile signals for the SynTouch BioTac," 2021, *arXiv:2101.05452*.

[7] A. Rodriguez, "The unstable queen: Uncertainty, mechanics, and tactile feedback," *Sci. Robot.*, vol. 6, p. 54, May 2021.

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

[9] B. Ward-Cherrier, N. Rojas, and N. F. Lepora, "Model-free precise in-hand manipulation with a 3D-printed tactile gripper," *IEEE Robot. Autom. Lett.*, vol. 2, no. 4, pp. 2056–2063, Oct. 2017.

[10] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," 2019, *arXiv:1910.02860*.

[11] C. Wang, S. Wang, B. Romero, F. Veiga, and E. Adelson, "SwingBot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2020, pp. 5633–5640.

[12] B. Ward-Cherrier *et al.*, "The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies," *Soft Robot.*, vol. 5, no. 2, pp. 216–227, Apr. 2018.

[13] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using FingerVision," in *Proc. IEEE-RAS 17th Int. Conf. Hum. Robot.*, Nov. 2017, pp. 241–248.

[14] Y. Du, G. Zhang, Y. Zhang, and M. Y. Wang, "High-resolution 3-dimensional contact deformation tracking for FingerVision sensor with dense random color pattern," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2147–2154, Apr. 2021.

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

[16] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "GelSlim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Oct. 2018, pp. 1927–1934.

[17] B. Romero, F. Veiga, and E. Adelson, "Soft, round, high resolution tactile fingertip sensors for dexterous robotic manipulation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 4796–4802.

[18] D. Fernandes Gomes, Z. Lin, and S. Luo, "GelTip: A finger-shaped optical tactile sensor for robotic manipulation," 2020, *arXiv:2008.05404*.

[19] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2018, pp. 5410–5418.

[20] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, May 2017, pp. 270–279.

[21] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.

[22] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2485–2494.

[23] D. Scharstein *et al.*, "High-resolution stereo datasets with subpixel-accurate ground truth," in *Proc. German Conf. Pattern Recognit.* Cham, Switzerland: Springer, 2014, pp. 31–42.

[24] N. Mayer *et al.*, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4040–4048.

[25] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2015, pp. 3431–3440.

[26] C. Godard, O. M. Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 3828–3838.

[27] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[28] T. Tao, J. C. Koo, and H. R. Choi, "A fast block matching algorthim for stereo correspondence," in *Proc. IEEE Conf. Cybern. Intell. Syst.*, Oct. 2008, pp. 38–41.

[29] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*.

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

[31] Y. Zhong, Y. Dai, and H. Li, "Self-supervised learning for stereo matching with self-improving ability," 2017, *arXiv:1709.00930*.

[32] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*.

Shaowei Cui received the B.E. degree from the South China University of Technology, Guangzhou, China, in 2017. He is currently pursuing the Ph.D. degree in robotics with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, focusing on visuo-tactile perception and robotic dexterous manipulation.

Rui Wang received the B.E. degree in automation from the Beijing Institute of Technology, Beijing, China, in 2013, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, in 2018.

He is currently an Associate Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences. His current research interests include robot perception, control, and learning.

Jingyi Hu received the B.E. degree from the Beijing University of Posts and Telecommunications, Beijing, China, in 2019. She is currently pursuing the Ph.D. degree in control theory and control engineering with the Institute of Automation, Chinese Academy of Sciences, Beijing, focusing on vision-based tactile sensing and robotic dexterous manipulation.

Chaofan Zhang received the B.E. degree from Central South University, Changsha, China, in 2020. She is currently pursuing the Ph.D. degree in robotics with the Institute of Automation, Chinese Academy of Sciences, Beijing, China, focusing on visuo-tactile perception and robotic dexterous manipulation.

Lipeng Chen received the Ph.D. degree from the Robotic Manipulation Laboratory, University of Leeds, Leeds, U.K., in 2020.

He was a Post-Doctoral Researcher with The University of Edinburgh, Edinburgh, U.K. He is currently the Senior Research Scientist of the Tencent Robotics X Laboratory, Shenzhen, China. His research interests include autonomous robotic manipulation and physical human–robot collaboration.

Shuo Wang (Member, IEEE) received the B.E. degree in electrical engineering from the Shenyang Architecture and Civil Engineering Institute, Shenyang, China, in 1995, the M.E. degree in industrial automation from Northeastern University, Shenyang, in 1998, and the Ph.D. degree in control theory and control engineering from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2001.

He is currently a Professor with the State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences; the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing; and the Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai, China. His research interests include biomimetic robot, robot skill learning, and multirobot systems.