IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

StereoTac: a Novel Visuotactile Sensor that Combines Tactile Sensing with 3D Vision

Etienne Roberge[1], Guillaume Fornes[2], Jean-Philippe Roberge[1]

***Abstract*—Combining 3D vision with tactile sensing could unlock a greater level of dexterity for robots and improve several manipulation tasks. However, obtaining a close-up 3D view of the location where manipulation contacts occur can be challenging, particularly in confined spaces, cluttered environments, or without installing more sensors on the end effector. In this context, this paper presents *StereoTac*, a novel vision-based sensor that combines tactile sensing with 3D vision. The proposed sensor relies on stereoscopic vision to capture a 3D representation of the environment before contact and uses photometric stereo to reconstruct the tactile imprint generated by an object during contact. To this end, two cameras were integrated in a single sensor, whose interface is made of a transparent elastomer coated with a thin layer of paint with a level of transparency that can be adjusted by varying the sensor's internal lighting conditions. We describe the sensor's fabrication and evaluate its performance for both tactile perception and 3D vision. Our results show that the proposed sensor can reconstruct a 3D view of a scene just before grasping and perceive the tactile imprint after grasping, allowing for monitoring of the contact during manipulation.**

***Index Terms*—Tactile Sensing, Vision-Based Sensors, 3D vision, Perception for Manipulation and Grasping.**

I. Introduction

TACTILE sensing is often an important capability for robots that interact with their environment and that perform tasks such as grasping, manipulating and assembling objects. In recent years, significant progress has been made in developing tactile sensors that can assist robots with determining a myriad of physical attributes related to the objects to manipulate such as their pose, shape and even their texture at a very fine scale. These sensors can help robots to handle a wide range of objects—from small and delicate to large and irregular—which, by extension, contribute to increase their overall dexterity. When complementing vision, tactile sensing can also unlock the ability of achieving new, complex tasks that require additional information not provided by either sensing modality alone.

However, a common challenge in tactile sensing and visual perception during robotic manipulation is the reduced object visibility that will typically occur when a robot reaches for an object and/or when the gripper encompasses it as it performs its grasp. This is particularly true when the manipulator is used for reaching objects in confined spaces, such as cabinets or boxes, or when manipulation occurs in cluttered environments. In these scenarios, the robot will often create obstructions to cameras that are statically affixed to the robotic cell. This can limit the accuracy and reliability of the grasp, as the vision system cannot continuously and unobstructedly track the object, which may be prone to moving as the robot approaches. Alternatively, vision system(s) can be mounted directly to the robot wrist to get a closer perspective to where manipulation contacts will be generated. However, this generally increases the size of the tool at the end effector, which can in turn lead to additional constraints on movement / reduced motion and an overall augmented bulkiness of the system. This is particularly true when 3D vision is needed—even though small 3D vision systems exist, time-of-flight, structured light and stereoscopic vision to name a few, will generally require more internal components and more space than what is typically needed by 2D cameras. Combining 3D vision with tactile sensing could be an advantage in several manipulation tasks, but acquiring complete, unobstructed tridimensional view of the object close to where manipulation actually happens is often still considered a challenge.

Fig. 1. A photo showing *StereoTac*, a visuotactile sensor combining photometric stereo with stereoscopic vision.

To address this issue, we propose a novel vision-based tactile sensor that combines stereoscopic vision (3D vision) with photometric stereo (tactile), as displayed in Fig. 1. The proposed sensor uses a semi-transparent soft skin and two cameras to enable high-resolution, multi-modal tactile sensing for robotic manipulation. The sensor is inspired by the idea

Manuscript received: March 11, 2023; Revised June 18, 2023; Accepted July 30, 2023. This paper was recommended for publication by Editor Ashis Banerjee upon evaluation of the Associate Editor and Reviewers' comments. This work is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC, grant RGPIN 2022-04884) and the Fonds de recherche du Québec - Nature et technologies (FRQNT, grant 327363).

[1]Etienne Roberge and Jean-Philippe Roberge are with the Command and Robotics Laboratory, École de technologie supérieure, Montreal, Quebec, H3C1K3, Canada (e-mail: etienne.roberge.1@ens.etsmtl.ca; jean-philippe.roberge@etsmtl.ca).

[2]Guillaume Fornes is with the Bordeaux Institute of Technology, ENSEIRB-MATMECA, 1 avenue du Dr Albert Schweitzer B.P. 99 33402 Talence, France (e-mail: gfornes@enseirb-matmeca.fr).

Digital Object Identifier (DOI): see top of this page.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

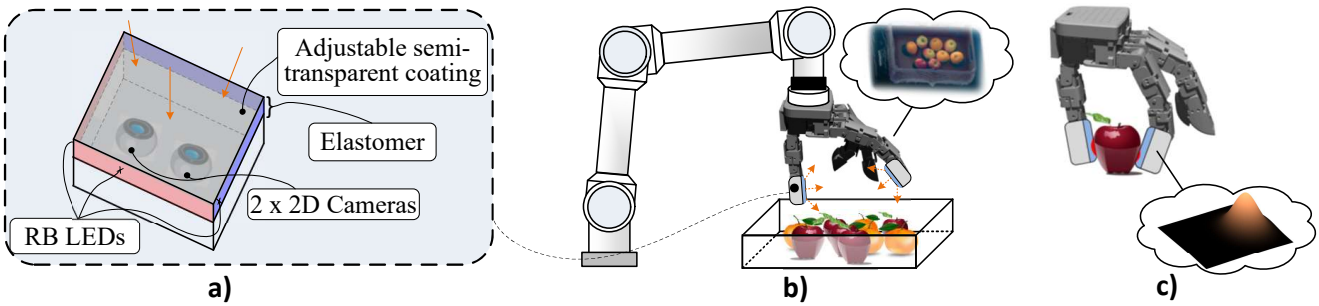

Fig. 2. An example of a use case scenario with StereoTac. a) StereoTac's main components; b) Here, two sensors are integrated to robotic hand and provide a stitched 3D close-up view for grasp planning; c) After an objet is picked up, the sensor allows the reconstruction of the tactile imprint.

of *whole-body vision* [1] which consists of having the ability to perceive tactile stimuli as well as being able to see through the skin. In our proposed implementation, the contact interface is made of a transparent elastomer covered with a thin layer of reflective paint, which allows two cameras to capture the deformation of the surface induced by the contact while still being able to see through the skin by varying internal lighting conditions. This allows the sensor to acquire 3D images of the scene, even when the manipulator is close (from 5 to 60 cm) to the grasping site, and to get tactile imprints after contacting the object, as illustrated in Fig. 2.

The main contributions of this paper are:

- The design of *StereoTac*, a novel visuotactile sensor that combines tactile sensing with 3D vision. To the best of our knowledge, it is the first time a vision-based tactile sensor that uses photometric stereo with stereoscopic vision was designed to capture both visual (3D) and tactile data from the same location;
- An analysis of the impact the level of transparency has on both tactile imprint reconstruction and 3D perception;
- Experimental results that demonstrate the sensor's ability to capture both tactile and 3D vision. The stereoscopic camera is compared to a similar off-the-shelf product and the whole sensor is integrated to a robotic arm for 3D object reconstruction (shown in the accompanying video).

In the next sections, we first provide background information about vision-based sensors, with a focus on sensors that have the ability to acquire external vision as well as tactile stimuli. In section III, we discuss StereoTac's operating principles as well as its fabrication process. In sections III-B and III-C, we present experimental results related to the performance of both tactile and visual data, and use of the sensor in the real world. Finally, we discuss the main results this work generated as well as future work opportunities.

II. Related Work

A. Vision-Based Tactile Sensors

Vision-based tactile sensors have brought significant interest in the last years, partly due to the fact that their data processing can leverage well-known computer vision techniques. Furthermore, these sensors benefit from recent progress in the field of cameras and are generally considered as high resolution, affordable and reliable sensors, when broadly compared to other transduction techniques [2]. Several approaches exist for the fabrication of vision-based tactile sensors. However, in most occurrences and although these sensing devices are equipped with camera(s), the visual perception is normally limited to the sensor's internal chamber. Some researchers have used frustrated total internal reflection (FTIR) by using an elastomer sitting on a light-conductive material [3], [4]. However, in this setting, the elastomer has to be non-transparent to generate a perceptible reflection to the camera, therefore limiting the camera's view to the sensor's internal chamber. Other approaches involve tracking pins [5], scattered particles [6] or markers [1], [7], which can be used to measure normal force, shear and torque. Nevertheless, the presence of such particles in the elastomer generally hinders efforts for external perception. Another popular principle is photometric stereo [8], which consists of generating a depth map by illuminating the elastomer coated with reflective pigments from different directions. Following this method, Gelsight sensors [9], Digit [10] and others [11] use distinct red, green and blue lighting from different directions to illuminate an elastomer that allows the generation of a depth map from a single image. Using photometric stereo with markers have also been proposed [11], [12], which allow geometric reconstruction along with the determination of force distribution on the contact medium. Classical stereophotometric approaches usually involve the application of a non-transparent specular or matte coating on the contact medium, which limits the camera's perception to the inside of the sensor. Other types of vision-based tactile sensors have been proposed, such as stereoscopic-based sensors that get 3D images of a membrane deformation without constrained illumination [13], [14], sensors using a mix of photometric stereo and structured light [15], and sensors based on illumination contrast and difference images [16]. However, all of the aforementioned sensors also suffer from the same limitation of using camera(s) that are able to capture only the inside portion of the sensor.

B. Vision-Based Sensors with Proximity/3D Geometry Sensing

Researchers have sought ways to integrate both tactile and visual perception in a compact device. One well-known example of such sensor is Fingervision [17], which combines tactile sensing through marker tracking with 2D vision. By analyzing multiple 2D images with optical flow, the authors were able to estimate the proximity of objects. Fingervision have been successfully used for cutting vegetables [18], however, the markers printed on the external layer create noise in the images and their use of a single camera limited perception to 2D. Yin et al. [7] integrate an air pressure sensor and an RGB camera for tactile sensing via marker tracking, along with a solid-state lidar for depth sensing, into a single sensor system. As noted by the authors, the use of markers results in visible artifacts

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

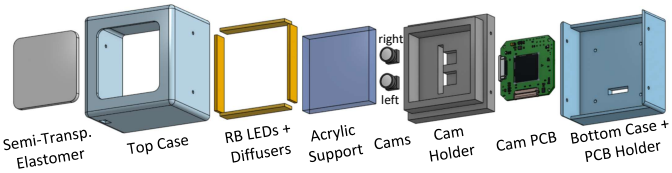

Fig. 3. Exploded view of the proposed sensor.

once the depth surpasses 40 mm, which only allowed reliably capturing depth for distances under 10 cm. Shimonomura et al. [19] combines FTIR sensing with stereoscopic vision. However, the sensor used FTIR technology with a rigid acrylic surface as the contact medium, which means it only provided information on the position and area of contact without capturing grasping forces or enabling contact imprint reconstruction. On the other hand, Hogan et al. [20] showed that it was possible to see through a semi-reflective coating by varying the sensor's internal lighting conditions. The authors have demonstrated the feasibility of estimating object proximity using an approach that required artificial vision techniques and moving the robot on a predetermined linear path to acquire images [21]. However, this method used a single camera, which prevented the sensor from perceiving the 3D geometries. Recently, Luu et al. [22] introduced a sensor with a contact medium made of polymer dispersed liquid crystal film with markers, which allowed the active control of the membrane's transparency. However, the markers generated noise in the 2D images and the two cameras integrated by the authors were facing each others in the current implementation, which would not be well-suited for acquiring 3D features.

To address the challenges posed by marker-induced artifact creation, the limitations of using a single camera, and the difficulties associated with FTIR perception that typically requires a rigid contact medium, we introduce StereoTac. This sensor employs stereoscopic vision for 3D imaging and photometric stereo for tactile sensing, and benefits from the use of an elastomer with adjustable transparency through lighting.

III. Design and Sensing Principles

A. Fabrication

The design objective is to create a visuotactile sensor that integrates stereoscopic vision to capture a 3D representation of the nearby environment before contact and uses photometric stereo to acquire a 3D depth map (tactile imprint) of the contact. This capability could complement existing tactile-based manipulation approaches in cluttered environments (e.g.: [23]), where the ability to accurately perceive the surrounding environment and the object being manipulated is crucial. Fig. 3 provides an exploded view of the elements used in the proposed design, which include two 2D cameras, the contact elastomer and the semi-transparent layer. In the following subsections, we describe how these elements were fabricated and assembled, along with specific design considerations.

1) Elastomer and Coatings: As per prior studies [20], P-595 silicone elastomer from Silicones Inc. was employed as the base material for all considered membranes, as depicted in Fig. 4. The elastomers were produced using a mold made of three 1/8 inch acrylic sheets, with one sheet cut to the desired membrane shape and the remaining two employed to compress the gel and flatten the surfaces. The transparent and

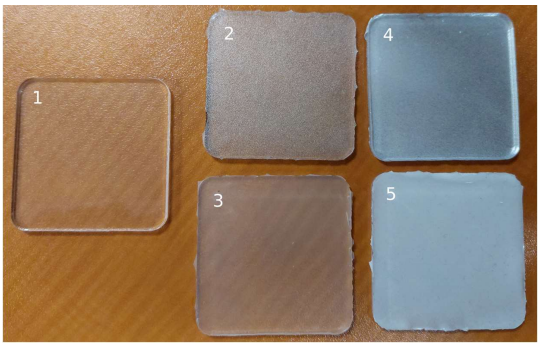

Fig. 4. The elastomers used in the comparative study: 1) Completely transparent, 2) Semi-transparent reflective, 3) Semi-transparent matte, 4) Opaque reflective, and 5) Opaque matte. It is important to note that membranes 1-3 are the focus of the study, while membranes 4-5 are not see-through and only used for comparison purposes in section III-C.

semi-transparent reflective membranes studied in this paper, which correspond to Fig. 4-#1 and Fig. 4-#2, were designed in accordance with the specifications outlined in [20]. To achieve the desired transparency, multiple layers of mirror-type spray paint (Rust-Oleum 267727) were applied onto the elastomer. A silicone protective layer was then sprayed over the paint layer to safeguard it. To create the protective layer, the same gel used for the membrane was mixed with a silicone thinner (Smooth-On NOVOCS Gloss) in a 2:1 ratio, resulting in a highly fluid gel that could be easily spread over the paint layer.

The semi-transparent matte membrane (Fig. 4-#3) was generated by incorporating 0.5% white silicone dye (Smooth-On Silc Pig White) by weight into the protective layer. Subsequently, the opaque membranes (Fig. 4-#4 and Fig. 4-#5) were produced by adding 3% white dye by weight to the protective layer. To create the reflective opaque membrane, a mirror paint layer was applied, similar to that of the semi-transparent reflective membrane, prior to coating with the opaque protective layer. The opacity of each membrane depicted in Fig. 4 was approximated using a stable light source and a lux meter (AP-881D from AOPUTTRIVER) positioned at a fixed distance. The sensing portion of the lux meter was covered by the membrane placed on an adapter, which only allowed light to pass through the membrane to reach the lux meter. The lux readings obtained in this manner and estimated opacity are reported in Table I.

TABLE I
Approximated Opacity for Each Membrane

Membrane Type	Total Light (Lux)	Approx. Opacity (%)
No membrane	466	n/a
Transparent	442	5.15
Semi-Trans. Reflective	352	24.46
Semi-Trans. Matte	363	22.10
Opaque Reflective	93	80.04
Opaque Matte	141	69.74

2) Lighting and acrylic support: For the internal illumination of the sensor, NeoPixel LED strips from Adafruit were positioned on the walls of the sensor enclosure at the same level as the acrylic membrane support. This configuration provides parallel illumination of the membrane surface, effectively reducing light leakage from the sensor. In particular, when using transparent or semi-transparent membranes, we aim to minimize light escaping from the sensor to prevent premature object illumination before contact. Further discussion on this

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

topic can be found in section IV-E. These lights are operated using an Arduino Micro communicating with a ROS Noetic package that we developed for StereoTac.

To minimize internal reflections of the lights on the acrylic support block, a gray filter (VViViD Smoke Black Gloss Vinyl) was applied to the sides, as demonstrated in [12]. Additionally, a 1624K57 LED diffuser from McMaster-Carr is employed to promote light diffusion instead of specularity.

3) Cameras: The prototype employs Odseven 160° variable focus cameras, which provide a broad field of view to encompass the entire membrane. These cameras have a resolution of 5 MP, allowing for image capture at 2592x1944 pixels at a rate of 30 frames per second (FPS). The cameras are operated by an Arducam USB3.0 Camera Shield capable of simultaneously acquiring images from two cameras. However, the cameras' circuitry required a modification for stereoscopic usage. Specifically, to synchronize the camera images, the clock component was removed from one camera's electrical circuit and that camera was connected to the clock of the second camera. This ensured precise synchronization between the cameras, allowing simultaneous image acquisition by the acquisition card. The cameras are positioned 14 mm apart, providing better depth resolution for close-up shots near the sensor, but decreasing resolution for more distant objects.

B. Stereoscopic Vision

Although the 14 mm spacing between the cameras limits the depth perception resolution at longer distances, it allows the sensor to be more compact and to perceive objects in a closer range, which is helpful for fine manipulation and more suited to confined and/or cluttered spaces (e.g.: [23]). To calibrate both the intrinsic and extrinsic parameters of the cameras, the stereo camera calibration utility available in the camera_calibration package [24] on ROS was used. The calibration was performed using an 8x6 checkerboard with 17mm tiles. The information obtained from calibration enables image rectification. Indeed, due to the cameras having wide-angle lenses, the initial images are distorted and require rectification prior to utilization. Using the Q Matrix from calibration, the Stereo Block Matching utility provided by OpenCV is utilized to generate the disparity map and 3D projection of the scene. Finally, the points are filtered by statistically removing outliers using the Open3D [25] library. The range of 3D vision extends from 5 to 60 cm. Below 5 cm, the blind spot between the two cameras prevents the acquisition of relevant data, and above 60 cm, the 3D data becomes incoherent due to low disparity. The 3D point cloud from stereoscopy is 320 x 280 (L x W) voxels, acquired at 2.1 FPS. The stereoscopic field of view is 67.6° x 60.0°, while for 2D vision only, it's 102.3° x 87.2°.

C. Tactile Sensing

To ensure that only the deformation of the membrane is captured while StereoTac is operated in tactile mode, the exposure of the cameras is reduced and the LED arrays inside the sensor are illuminated to make the luminosity high enough to capture reflections. Moreover, to eliminate any potential contamination from external lighting in tactile mode, an HSV filter is applied to isolate only the red and blue tones

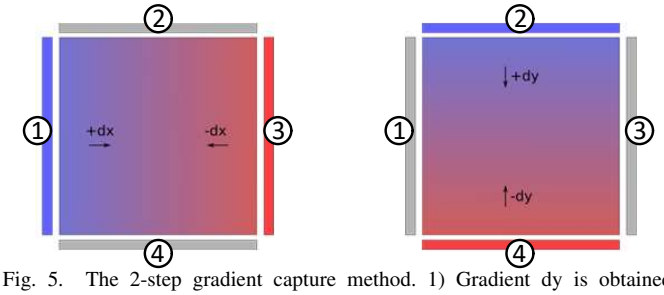

Fig. 5. The 2-step gradient capture method. 1) Gradient dy is obtained by illuminating the membrane with LED array 2 and 4 with blue and red respectively. 2) Gradient dx is obtained by illuminating the membrane with LED array 1 and 3 with blue and red respectively.

corresponding to the LEDs' illumination. The LED arrays are placed on the four sides of the square-shaped lens, providing a range of angles to illuminate the membrane and enabling the detection of fine surface details. When the goal is only to obtain the shape of the contacts on the membrane, all the LED arrays can be illuminated to observe the reflection on the membrane's surface with the camera. However, when the goal is to perform 3D reconstruction of the contact on the membrane, directed light can be used to capture the reflection gradient and create a 3D representation of the tactile imprint.

*1) **Capturing contact gradients from membrane illumination**:* By utilizing the reflectance map of the elastomer, it is possible to determine the gradient at each pixel based on the intensity of the measured colors. Regarding the StereoTac sensor, the membrane is illuminated by placing the sensor lights at the elastomer's perimeter in a square configuration, as displayed in Fig. 3. This setup provides four feasible illumination angles. However, a three-angle approach using RGB colors on three distinct axes, as is frequently done with photometric stereo-based sensors, would not be sufficient to observe the complete contact gradients. Specifically, the utilization of three colors simultaneously on three different, orthogonal angles would only provide a fraction of the information in the fourth direction where no directed light is employed. To address this limitation, we utilized the placement of the illumination axes to illuminate the elastomer in two steps, as shown in Fig. 5. Given the parallel arrangement of the LED rows, the x-gradient (dx) information is obtained by illuminating rows 1 and 3 with distinct colors. Similarly, the y-gradient (dy) information is obtained from rows 2 and 4. Since illumination is done from only two directions at the same time, we employed blue-red pairs to facilitate color segmentation during image processing. The dx and dy gradients are captured sequentially by alternating the illuminated LED pairs, with dx and dy LEDs alternating at a rate of 4 Hz. An example of the resulting images for the semi-transparent reflective membrane (Fig. 4-#2) can be observed in Fig. 6.

After acquiring the dx and dy gradient images, an HSV filter is applied to retain only the pixels corresponding to the red and blue tones of the LEDs. This step eliminates potential contamination captured by the cameras originating from external lighting. As different membranes are investigated during this study, and their exact reflectance maps are unknown, a simple, 3-hidden fully-connected layer neural network is employed, as proposed by Wang et al. [12]. Calibration is performed for each membrane by capturing 30 images, with a calibration

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

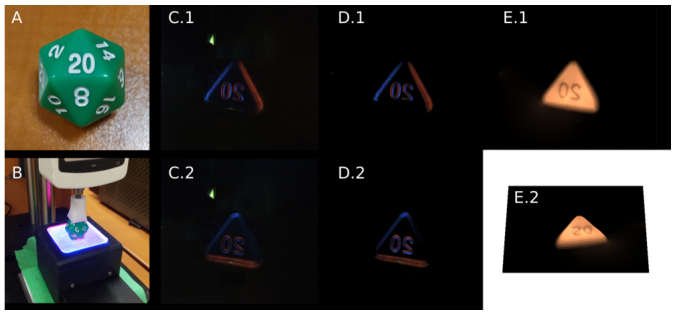

Fig. 6. Process for the 3D reconstruction of contacts. A) The example-object (D20). B) D20 contacting the reflective semi-transparent membrane. C.1-2) Images obtained during the two-step illumination dx-dy: note that the white/green spot is a ceiling light in the room where the image was taken (339 Lux). D.1-2) Images obtained after filtering C.1-2 using a color filter. E1-2) 3D reconstruction obtained with gradients calculated using images D.1-2.

ball of known diameter pressed at various positions on the membrane. After using the HSV mask to eliminate non-contact information from the image, and by knowing the diameter of the ball and the exact center position of the ball in the image, the dx and dy values of each pixel in contact with the ball can be easily determined using a simple distance relationship from its center. Indeed: $d_x = \sin^{-1}((p_x - c_x)/r)$, where p is the position of the pixel, c is the position of the center of the ball in pixels and r is the radius of the ball in pixels.

All non-zero pixels obtained from this acquisition are utilized to train the MLP neural network. The 1-by-4 inputs of the network consist of the R-B values and the x-y position of each pixel, the 1-by-2 outputs are the gradients. x-y values are included to counteract light attenuation, where pixels distant from the light source appear dimmer. This addresses the disparity where pixels closer to a light source appear brighter, despite the same contact gradient, than those further away.

*2) **Reconstructing the 3D Tactile Imprint using dx-dy**:* As demonstrated by Wang et al. [12], it is possible to use a 2D Fast-Poisson solver to compute the depth (z) of each pixel given the dx and dy values of each pixel from an image capture. This approach is particularly advantageous when handling noisy or incomplete gradient data, as it typically yields smoother results during 3D reconstruction. In our experiments, we found that this reconstruction method provided a more accurate tactile imprint with finer details. Finally, by knowing the number of pixels per mm in the obtained images (in our case, 15 pixels/mm), we can determine the measured depth in millimeters. We used the open-source Python code from Doerner [26] to compute the Fast Poisson algorithm. Fig. 7 provides an overview of the reconstruction outcome for several objects with different shapes.

IV. Experiments

A. Evaluation of Visual Depth through different membranes

The performance of the sensor in vision mode is evaluated using *Z-accuracy*, *RMS error* (spatial noise) and *temporal noise*. These metrics are commonly used in the evaluation of depth cameras and correspond to the recommended metrics in the "Camera Depth Testing Methodology" from Intel [27].

To perform empirical experiments on the membranes (#1, #2 and #3 in Fig. 4), we positioned the sensor at different distances (10 cm, 15 cm, 20 cm, 25 cm, and 30 cm) away

Fig. 7. Examples of 3D contact reconstruction made with the semi-transparent reflective membrane. Left, the real objects (M5 screw, wood screw, D12 and D0). Center, top view of the 3D imprints. Right, isometric view of the imprints.

from a completely flat surface with a checkered pattern, adding texture to the surface. At each position, ten RGB-D images were captured with each of the membranes. Additionally, the same experiments were repeated using the Intel RealSense D405 camera, which is an off-the-shelf short-range stereoscopic camera, for comparison purposes.

*1) **Z-accuracy**:* Z-accuracy (or absolute error) evaluates depth accuracy by using a pre-measured ground truth as a reference. This metric reflects the precision of depth data on a per-pixel basis, comparing it with the ground truth for each captured depth image at a fixed distance. To isolate camera positioning errors, the depth is computed relative to a best-fit plane within the point cloud. The Z-accuracy is obtained by calculating

$$Z_{acc} = \frac{\text{med}(D(x,y) - GT)}{GT} \times 100\%, \tag{1}$$

where med is the median function, $D(x,y)$ is the calculated depth at pixel positions (x,y) and GT is the ground-truth depth value. The experiment results can be observed in Fig. 8.

The precision of depth measurement was found to be affected by the use of a semi-transparent membrane. As expected from the presence of distortions created by using a semi-transparent interface, the presence of matte or reflective finishes created more variations in depth perception and resulted in a generally less accurate median values ($\sim 0.5-9\%$) in the distribution, when compared to the Intel camera or StereoTac equipped with a transparent membrane.

*2) **RMS error**:* The Root Mean Square (RMS) error (or spatial noise) of StereoTac's depth measurements are evaluated. While the variation in depth values of each pixel within a ROI does not directly measure accuracy, it is an

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

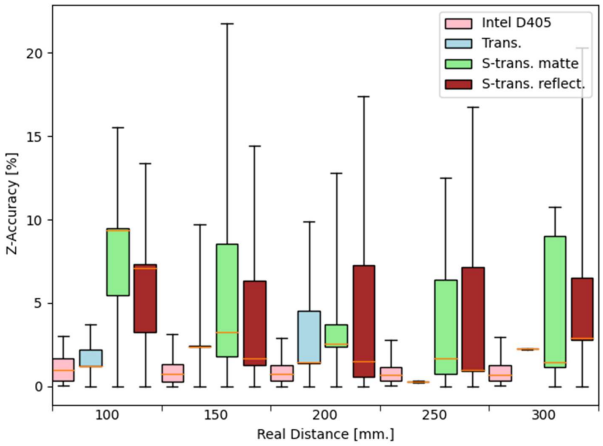

Fig. 8. Z-Accuracy measured on a flat surface using different membranes at different distances.

important metric as it provides information on the consistency and repeatability of the measurements within a specific area by evaluating the spatial depth uniformity. The RMS error is obtained by calculating

$$\text{RMS}_{err} = 100\% \times \sqrt{\frac{\sum (D(x,y) - F(x,y))^2}{n}} / GT, \quad (2)$$

where $F(x,y)$ are the depths value of the fitted plane at pixel positions (x,y), GT is the ground-truth depth value and n is the number of pixels. The results from these experiments can be observed in table II, which reveals that the reflective semi-transparent membrane exhibits a higher RMS error.

TABLE II
DISTRIBUTION OF 10 RMSE VALUES OBTAINED ON A FLAT SURFACE BY DIFFERENT MEMBRANES AT DIFFERENT DISTANCES, (%) [$\mu : \sigma$]

Dist.	Intel D405	Transparent	Semi-trans. matte	Semi-trans. reflective
10 cm.	0.43 : 0.01	4.1 : 1.66	2.18 : 0.1	8.06 : 1.43
15 cm.	0.6 : 0.02	2.3 : 0.7	3.52 : 0.59	4.70 : 0.79
20 cm.	0.86 : 0.03	2.23 : 0.06	5.91 : 3.19	6.65 : 0.48
25 cm.	1.01 : 0.04	1.92 : 0.52	3.08 : 0.06	4.76 : 0.43
30 cm.	1.15 : 0.02	1.89 : 0.18	3.74 : 0.13	6.15 $\pm$ 1.35

*3) **Temporal Noise**:* Temporal Noise assesses uniformity across sequential frames. The fluctuation in depth values for each pixel was monitored over ten consecutive images at each distance. Temporal noise was quantified by determining the standard deviation of depth values across these ten images of a flat surface. The median of these standard deviations within the ROI served as the temporal noise measure. The findings from this experiment are presented in Table III.

The results show that the use of a matte semi-transparent membrane does not significantly affect the noise level over time. However, the use of the reflective semi-transparent membrane adds observable noise. For example, at 10cm, the temporal noise would be around $\pm$ 3.72 mm. This increase may be due to the clarity of the membrane, where the reflective semi-transparent membrane has tiny visible paint spots when viewed up close by the cameras.

TABLE III
TEMPORAL ERRORS OBTAINED ON A FLAT SURFACE BY DIFFERENT MEMBRANES AT DIFFERENT DISTANCES. [%]

Dist.	Intel D405	Trans.	Semi-trans. matte	Semi-trans. reflective
10 cm.	0.29	1.25	0.75	3.72
15 cm.	0.46	0.81	1.42	3.48
20 cm.	0.69	0.72	1.84	3.12
25 cm.	0.81	1.17	1.02	2.86
30 cm.	0.96	1.64	0.89	4.57

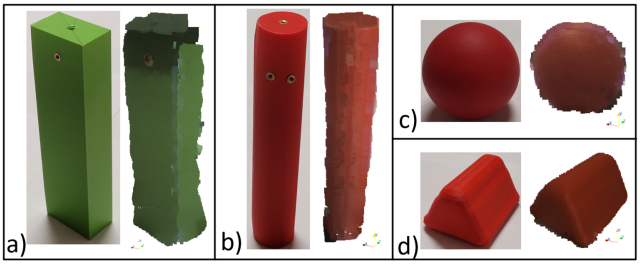

Fig. 9. Comparison between real (left) and reconstructed (right) prototypical shapes: a) a rectangular prism, b) a cylinder, c) a ball and d) a tringular prism.

B. 3D Object Reconstruction

Mounted on a manipulator arm, StereoTac facilitates object reconstruction, as evidenced in the supplementary video. The process involves acquiring four point clouds from orthogonal viewpoints, merged using the ICP algorithm with robot poses as initial estimates. Artifact reduction utilizes statistical and radius outlier filters from Open3d, with voxel downsampling ensuring uniform density. Fig. 9 and Table IV compare reconstructed (acquired with a semi-matte membrane) and reference geometries, including volume and RMSEs. Object volumes are estimated using the point cloud's convex hull, while RMSE calculation employs eq. 2, referencing the prototypical shape CAD as ground truth. While the reconstructed volume tends to be underestimated, these results are consequent with results from Table II.

TABLE IV
VOLUMES COMPARISON AND RMSE VALUES FOR PROTOTYPICAL SHAPES

	Rect. Prism	Cyl.	Ball	Triang. Prism
Real vol. (cm^3)	449.4	255.64	143.79	199.8
Est. vol. (cm^3)	430.23	245.88	138.14	193.25
Est. error. (%)	4.27	3.82	3.93	3.28
RMSE (%)	5.97	3.92	4.19	3.39

C. Tactile Imprint Experiments

To evaluate the accuracy and stability of depth measurements in tactile mode, we conducted empirical experiments using a flat object of a known size, pressing it onto the sensor at a consistent depth. For this, we used a 13mm diameter disk mounted on a Mark-10 Manual Lever Operated Test Stand, ensuring perpendicular pressure on the membrane and the capture of a truly flat image. A Mitutoyo Absolute 543-693 vertical vernier, mounted on the test stand, was utilized to confirm the disk's penetration depth into the membrane.

To obtain a reliable estimation of the depth reconstruction error for each membrane, pressure disks were consecutively pressed onto the sensor 30 times at random locations to a depth of 1 mm. The methods discussed in section III-C1 were then used to reconstruct the disk. Depth measurements for each trial were obtained by taking the average of the depth values of the flat surface of the disk in the image. Table V provides the mean and standard deviation values for each membrane.

The mean values are generally below 1mm for all membranes except for those obtained using the Opaque-Reflective membrane. This may be due to several factors. For example, since a neural network was employed to obtain depth measurements, the calibration process of the membranes plays a

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

TABLE V
MEAN AND STANDARD DEVIATION OF DEPTH MEASUREMENTS—1MM INDENTATION WITH A DISK ONTO EACH MEMBRANE 30 TIMES. [MM.]

Membrane type	Mean	Std
Clear	0.915	0.179
Semi-Reflective	0.837	**0.085**
Semi-Matte	0.6143	0.121
Opaque-Reflective	1.094	**0.071**
Opaque-Matte	0.853	0.091

crucial role in determining the accuracy of the depth estimates. The calibration was performed using a sphere with a textured surface, and it is possible that the estimation of depth on the smooth surface of the disk altered the perspective of the measurements. However, it is worth noting that the two reflective membranes yielded the smallest standard deviations in measurement, indicating that depth measurement using them resulted in more consistent outcomes.

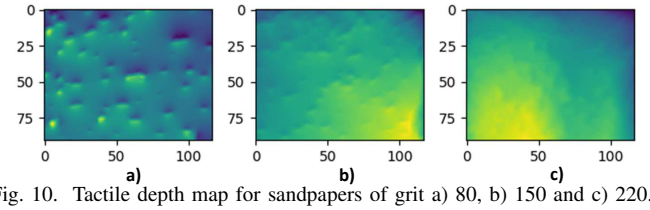

Fig. 10. Tactile depth map for sandpapers of grit a) 80, b) 150 and c) 220.

D. Tactile Perception of Sandpaper Grit Grade

The ability to estimate surface roughness can enhance object manipulation and discrimination. We qualitatively assessed the StereoTac sensor's ability to perceive roughness using sandpaper of various grit grades. As shown in Fig. 10, the sensor's tactile depth map reveals that grit60 sandpaper (with grain sizes around 336-425 micrometers) produces significant asperities in the depth map. Grit150 creates less pronounced protrusions, while Grit220 (grains around 190-265 micrometers) results in irregularities that are less visible. These outcomes suggest the sensor may serve to gauge roughness to a certain degree.

E. Discussion

Fig. 11 showcases a qualitative comparison of StereoTac's 3D vision capabilities using the three see-through membranes. Stereoscopic assessments have noted that semi-reflective membranes generally undermine depth measurement reliability. The errors detected are typically limited to a maximum of 2 centimeters over a distance of 30 cm. While these errors bear significance in the context of precise robotic grasping, real-time depth readings are still feasible to estimate the actual depth required to reach a target. Moreover, the utilization of stereoscopy is influenced by ambient brightness. Integrating external LEDs to the sensor in the future could help adjusting the ambient brightness for uniform readings. Additionally, the membrane's cleanliness can impact the depth reading's precision. For example, handling oily or dirty objects may leave residue on the membrane, which could potentially hinder reliable readings. A viable solution is washing the membranes with soap and water, as the silicone layer on the contact interface enables cleaning without degradation. Alternatively, since they are cost-effective and simple to manufacture, the membranes can be replaced periodically.

Regarding the tactile properties of the membranes, Fig. 7 qualitatively demonstrates that semi-transparent membranes

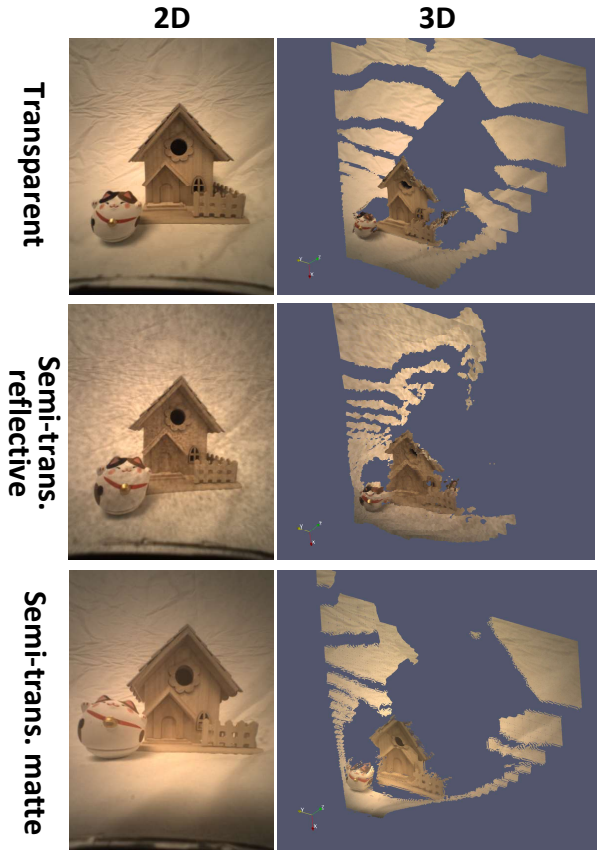

Fig. 11. Comparison of 2D and 3D perception through different membranes. For a more comprehensive understanding of the sensor's 3D vision capabilities, the reader is invited to refer to the accompanying video.

yield 3D reconstructions suitable for object recognition during grasping. The empirical results of experiments, presented in Table V, suggest that using reflective semi-transparent membranes yields the most reliable reconstruction for our sensor type. The choice of membrane type introduces a trade-off, dependent on whether the priority is optimizing 3D vision or tactile map reconstruction. For tasks demanding fine tactile details, like texture recognition or delicate manipulation, a semi-reflective membrane would be more appropriate due to its enhanced tactile sensing performance. However, for tasks necessitating accurate 3D object recognition and positioning, a semi-matte membrane might be preferred.

However, it is important to note that the use of transparent or semi-transparent membranes comes with noise effects that

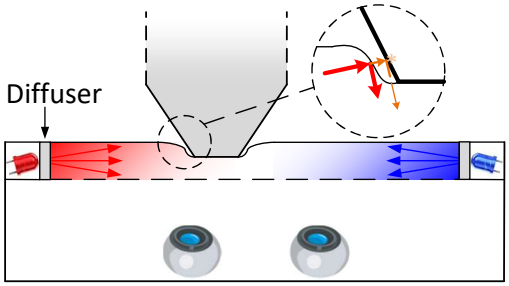

Fig. 12. Illustration of the effect of semi-transparent/transparent coating on a soft elastomer. As shown in the top-right corner: with an elastomer coated in opaque reflective paint, only the red light ray is reflected to the camera. In contrast, only the orange light ray is visible when a completely transparent membrane is used. When employing a semi-transparent coating, the camera perceives both the red and orange rays.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.

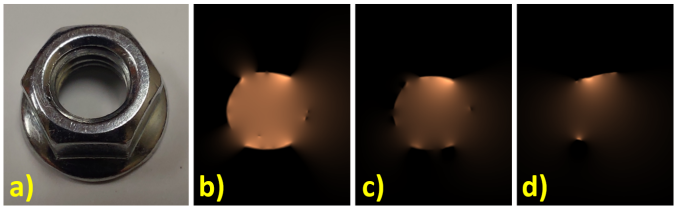

Fig. 13. Example of contact measuring error with a reflective object held at 1mm. from the membrane. 1) Reflective bolt. 2) Transparent membrane. 3) Semi-transparent opaque membrane. 4) Semi-transparent reflective membrane.

would not be present with opaque membranes. A fraction of the sensor's internal illumination escapes through the semi-transparency of the membrane, and depending on the object's color, incident angle or reflectivity, depth readings may be altered, as shown in Fig. 12.

This phenomenon was created on purpose and is observable in Fig. 13. In this scenario, a reflective bolt was suspended 1 mm above the membrane, which corresponds to the longest distance away from the sensor at which gradients could be perceived. The red and blue lights emitted from the sensor hit the reflective surfaces of the bolt and returned to the sensor, causing an inaccurate measurement. The transparent membrane is obviously more susceptible to this issue, but semi-transparent membranes are also slightly affected.

V. Conclusion

This article presented the development of StereoTac, a novel tactile sensor with 3D vision capabilities. The feasibility of combining multiple modalities for robotic manipulation in a single sensor was demonstrated by incorporating a second camera inside a visuo-tactile sensor and using a semi-transparent contact membrane. Several areas of improvement will be explored as future work to improve different aspects of the sensor. For example, tactile imprints were reconstructed using only one camera in this work. However, using the two available cameras of the sensor for this task could potentially increase the accuracy of the reconstruction, as recently done in [13], [14]. Also, stereophotometry currently utilizes red and blue LEDs, primary colors that are 120 degrees apart on the hue axis. Yet, employing primary colors separated by 180 degrees, like yellow and blue, could potentially boost the hue filter's performance, thereby improving tactile reconstruction. Furthermore, it was observed that the 3D view of the sensor environment was less reliable when using reflective semi-transparent membranes. Depth post-processing techniques like temporal filtering or edge-preserving filtering [28] could be beneficial in reducing distortions generated by semi-transparent membranes. Lastly, work will aim to miniaturize StereoTac for effortless integration onto the fingertips of future robotic grippers. Simultaneously, informed by studies such as those by Lei et al. [29], investigations will pursue the incorporation of StereoTac's current version into the palm of a developing gripper, as showcased in the supplementary video.

References

[1] A. Yamaguchi and C. Atkeson, "Optical skin for robots: Tactile sensing and whole-body vision," in *Carnegie Mellon University*, 07 2017.

[2] A. Yamaguchi and C. G. Atkeson, "Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision?," *Advanced Robotics*, vol. 33, no. 14, 2019.

[3] S. Begej, "Planar and finger-shaped optical tactile sensors for robotic applications," *IEEE Journal on Robotics and Automation*, 1988.

[4] M. Ohka, H. Kobayashi, J. Takata, and Y. Mitsuya, "An experimental optical three-axis tactile sensor featured with hemispherical surface," *Journ. of Adv. Mech. Des. Syst. and Man.*, vol. 2, pp. 860–873, 2008.

[5] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 2018. PMID: 29297773.

[6] C. Sferrazza and R. D'Andrea, "Design, motivation and evaluation of a full-resolution optical tactile sensor," *Sensors*, vol. 19, no. 4, 2019.

[7] J. Yin, G. M. Campbell, J. Pikul, and M. Yim, "Multimodal proximity and visuotactile sensing with a selectively transmissive soft membrane," 2022.

[8] R. J. Woodham, "Photometric Method For Determining Surface Orientation From Multiple Images," *Opt. Eng.*, vol. 19, no. 1, 1980.

[9] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1070–1077, 2009.

[10] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, pp. 3838–3845, jul 2020.

[11] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[12] S. Wang, Y. She, B. Romero, and E. H. Adelson, "Gelsight wedge: Measuring high-resolution 3d contact geometry with a compact robot finger," *CoRR*, vol. abs/2106.08851, 2021.

[13] J. Hu, S. Cui, S. Wang, C. Zhang, R. Wang, L. Chen, and Y. Li, "Gelstereo palm: a novel curved visuotactile sensor for 3d geometry sensing," *IEEE Transactions on Industrial Informatics*, pp. 1–10, 2023.

[14] 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. on Industrial Electronics*, vol. 69, no. 6, pp. 6015–6025, 2022.

[15] H. Sun, K. J. Kuchenbecker, and G. Martius, "A soft thumb-sized vision-based sensor with accurate all-round force perception," *CoRR*, 2021.

[16] C. Lin, Z. Lin, S. Wang, and H. Xu, "Dtact: A vision-based tactile sensor that measures high-resolution 3d geometry from darkness," 2022.

[17] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pp. 241–248, 2017.

[18] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pp. 1045–1051, 2016.

[19] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," in *2016 IEEE Int. Conf. on Rob. and Aut. (ICRA)*, pp. 138–143, 2016.

[20] F. R. Hogan, M. Jenkin, S. Rezaei-Shoshtari, Y. Girdhar, D. Meger, and G. Dudek, "Seeing through your skin: Recognizing objects with a novel visuotactile sensor," in *Proc. of the IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*, pp. 1218–1227, January 2021.

[21] 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 R-A Letters*, pp. 1–8, 2022.

[22] Q. K. Luu, D. Q. Nguyen, N. H. Nguyen, and V. A. Ho, "Soft robotic link with controllable transparency for vision-based tactile and proximity sensing," 2022.

[23] R. Thomasson, E. Roberge, M. Cutkosky, and J. Roberge, "Going in blind: Object motion classification using distributed tactile sensing for safe reaching in clutter," in *2022 IEEE International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[24] "camera_calibration - package summary." http://wiki.ros.org/camera_calibration. Accessed: 2023-03-06.

[25] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," 2018.

[26] J. Doerner, "Fast poisson reconstruction in python." https://gist.github.com/jackdoerner/b9b5e62a4c3893c76e4c.

[27] Intel, "Camera depth testing methodology." https://dev.intelrealsense.com/docs/camera-depth-testing-methodology, Jan 2021.

[28] Intel, "Depth post-processing for intel realsense depth cameras." https://dev.intelrealsense.com/docs/depth-post-processing, Jan 2021.

[29] Z. Lei, X. Deng, Y. Wang, Z. Li, X. Xiao, D. Han, F. Chen, and M. Li, "A biomimetic tactile palm for robotic object manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11500–11507, 2022.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2024, Yokohama, Japan. Cite as RA-L paper.