Design and Benchmarking of a Multimodality Sensor for Robotic Manipulation With GAN-Based Cross-Modality Interpretation

Dandan Zhang , *Member, IEEE*, Wen Fan , *Graduate Student Member, IEEE*, Jialin Lin ,
Haoran Li , *Member, IEEE*, Qingzheng Cong, Weiru Liu , Nathan F. Lepora , *Member, IEEE*,
and Shan Luo , *Senior Member, IEEE*

Abstract—In this article, we present the design and benchmark of an innovative sensor, ViTacTip, which fulfills the demand for advanced multimodal sensing in a compact design. A notable feature of ViTacTip is its transparent skin, which incorporates a "see-through-skin" mechanism. This mechanism aims at capturing detailed object features upon contact, significantly improving both vision-based and proximity perception capabilities. In parallel, the biomimetic tips embedded in the sensor's skin are designed to amplify contact details, thus substantially augmenting tactile and derived force perception abilities. To demonstrate the multimodal capabilities of ViTacTip, we developed a multitask learning model that enables simultaneous recognition of hardness, material, and textures. To assess the functionality and validate the versatility of ViTacTip, we conducted extensive benchmarking experiments, including object recognition, contact point detection, pose regression, and grating identification. To facilitate seamless switching between various sensing modalities, we employed a generative adversarial network (GAN)-based approach. This method enhances the applicability of the ViTacTip sensor across diverse environments by enabling cross-modality interpretation.

Index Terms—Cross-modality interpretation, generative adversarial network (GAN), multimodality sensing, vision-based tactile sensor (VBTS).

I. INTRODUCTION

GIVEN the complexity of the real-world environment, a single sensing modality or method for data acquisition may prove inadequate for achieving a comprehensive perception. This has led to a notable demand in research aiming at integrating multimodality sensing in robotic systems for advanced manipulation tasks. These approaches encompass many strategies, such as combining features acquired from visual and tactile sensors within a shared latent space to augment robot perception [1]. Alternatively, some methods focus on mapping representations from one sensory modality to another [2] or leveraging information from one modality to guide exploration in another [3]. In these studies, visual cameras are employed to capture visual data, while tactile sensors are utilized to gather contact information. Most prior research has focused on integrating various sensors into robotic systems and fusing the resulting data for analysis. In this work, we aim to develop a unified multimodality sensor that endows robotic systems with versatile perception capabilities, including vision, tactile, proximity, and force sensing. The proposed sensing system significantly reduces hardware and computing requirements in compact robotic systems while significantly enhancing integration capabilities.

One of the critical challenges in multimodal sensor fusion is minimizing cross-sensitivity, which occurs when the input from one modality interferes with the output of another. This cross-sensitivity complicates the decoupling of data from different modalities, making it difficult to extract information specific to a particular modality for downstream tasks. To address these challenges, recent research has focused on strategies such as employing distinct signal patterns, implementing control mechanisms in composite materials, and optimizing the morphology of active materials [4]. Developing sensors that utilize materials with frequency-dependent characteristics, such as ionic/proton-based materials, has been explored for their potential to decouple stimuli [4]. Combining innovative structures with functional materials like piezoresistive electrodes in capacitive sensors has enabled multimodal sensing from a single unit [5]. However, the aforementioned approaches necessitate specialized hardware development, potentially complicating the manufacturing process and increasing hardware expenses significantly.

The recent development of generative adversarial networks (GANs)-based method [6], [7], [8] can be used to transfer sensing modality from one domain to another. Therefore, we aim to explore GANs-based approaches to enable seamless transitions among different data visualization modes, allowing the sensor

Fig. 1. Overview of the four perception capabilities of ViTacTip, including two principle modalities: visual and tactile sensing, as well as two derived modalities: proximity and force sensing.

to dynamically adapt to varied perceptual demands in diverse operational contexts.

In this article, we introduce a novel perception system that combines multiple sensory modes (visual, tactile, proximity, and force) within a compact design, as shown in Fig. 1. Its unique combination of the see-through-skin mechanism, biomimetic pin design, and GANs-based modality transition establishes ViTacTip as an advantageous multimodality sensor. We aim to expand the field of multimodality sensing by exemplifying its significant role in robotic perception technologies that can be used for complicated tasks in diverse environments. The sensor is designed as an integrated unit, capable of interpreting, and responding to various types of stimuli.

The *main contributions* are listed as follows.
1) The innovative development of the ViTacTip sensor, a multimodal fusion device capable of capturing both tactile information and visual details.
2) The implementation of a GAN-based methodology to enable seamless switching between visual and tactile sensing modalities.
3) The comparative studies of ViTacTip against the open-source TacTip sensor [9] and our in-house developed ViTac sensor.

In this work, we further elaborate on the multimodality sensing capabilities of ViTacTip by providing a detailed analysis of its visual and tactile sensing, along with two derived modalities (proximity and force perception). In addition, we significantly expand the experimental evaluation of the overall methodology. The additional contributions presented in this article are as follows.

1) We detail the multimodal perception capabilities of Vi-TacTip and present hardware benchmarking experiments covering contact point detection, pose regression, and grating identification.
2) We introduce a hierarchical multitask learning framework to demonstrate ViTacTip's advanced multimodal perception capabilities through hardness, material, and texture recognition.
3) We present a comprehensive analysis of the GAN-based cross-modality interpretation framework, including both qualitative and quantitative evaluations of the usability and robustness of the light removal GAN (LR-GAN) and marker removal GAN (MR-GAN) across diverse tasks and complex environments.

The rest of this article is organized as follows. Section II presents an overview of the related work in this field. Section III details the design and fabrication processes of ViTacTip. The multimodality sensing capabilities of ViTacTip are thoroughly explained in Section IV. Section V focuses on hardware benchmarking, including comparative studies with baseline sensors. Subsequently, Section VI introduces the development and application of LR-GAN and MR-GAN for effective sensing modality transitions. Discussions on the findings and potential avenues for future research are presented in Section VII. Finally, Section VIII concludes this article.

II. RELATED WORK

A. Vision-Based Tactile Sensors (VBTSs)

Most of the existing VBTSs are primarily categorized into several distinct types. The first type encompasses coated-type sensors, which are characterized by their reflective layers, with the GelSight family being a prominent example. GelSight [10] features a unique design with a thin reflective layer atop a clear elastomer layer, which is further supported by a flat acrylic plate. This configuration, illuminated parallel to the surface, facilitates the detection of deformations using photometric stereo methods. Subsequent advancements have further refined this technology [11]. For instance, GelTip [12] has enabled curved sensing surfaces with a single camera. In addition, OmniTact [13] employs five cameras to augment multidirectional sensing capabilities.

The second category encompasses purely marker-based sensors, which are available in various forms. For example, the GelForce sensor [14] incorporates markers suspended within an elastomer. Such movement of the pins introduces an amplification effect in the shear motion of these markers, significantly improving the tactile sensor's responsiveness to both shear and normal deformations on its detection surface. Moreover, the TacTip skin is made by three-dimensional (3-D) printing, which leads to a wide range of different customized designs [9]. In addition, DTac-type sensors [15], such as 9DTac [16] and C-Sight [17], employ a combination of translucent gel and an opaque layer. Variations in pixel darkness within the images allow for intensity-to-depth regression and can even facilitate force and torque estimation.

A fundamental principle of VBTSs is the transduction of tactile information into visual data, using the sensor's skin as the medium [18], [19], [20]. Tactile information is transformed into visual features, such as gel deformation, shadows, or marker movements, which are captured by internal cameras. This conversion enables the processing of tactile features leveraging the strengths of state-of-the-art deep learning-based vision techniques [21], [22], [23].

However, as single-modality tactile sensors, both GelSight and TacTip have a significant limitation: their inability to incorporate visual perception. In GelSight, the reflective coating inhibits light transmission, while in TacTip, the black skin blocks external light. As a result, these designs exclude critical information about the touched object, such as fine surface textures and precontact distance perception. This exclusion of visual data can hinder a comprehensive understanding of tactile perception processes, especially in tasks involving interactions with real-world physical environments [2]. These limitations motivate the development of a new sensor that integrates both visual and tactile modalities to enhance perception capabilities for robotic applications.

B. Multimodality Sensors

Multimodality data, which combines both visual and tactile information from the same object, has been shown to significantly improve performance in texture recognition tasks [24]. Therefore, the development of a well-designed multimodality sensor capable of simultaneously capturing visual and tactile data during interactions is of considerable importance.

FingerVision [25], a representative multimodality sensor, is fabricated entirely from transparent silicone. It can acquire visual information directly through the sensor's skin. However, this design is limited by the absence of an internal light source, making its reliability highly susceptible to variations in environmental lighting conditions. Moreover, the integration of visual and tactile perception modalities in this sensor is tightly coupled, potentially affecting its robustness and adaptability to diverse tasks.

SpecTac [26] introduces an innovative approach to multimodal sensing by integrating ultraviolet (UV) LEDs and randomly distributed UV fluorescent markers. When illuminated by the UV LEDs, these markers become visible, allowing for clear differentiation and tracking against the background. SpecTac enables seamless switching between visual and tactile sensing modes by controlling the activation of the UV LEDs, which toggles the visibility of the markers. However, its performance is constrained under low-light conditions.

TIRgel [27] employs a lens-shaped elastomer and an adjustable-focus camera to enable switching between visual and tactile sensing modes. It leverages total internal reflection (TIR) within the elastomer for tactile imaging. However, TIRgel features a relatively flat surface and lacks force-sensing capabilities, which limits its suitability for dynamic manipulation tasks.

Most of the aforementioned work utilizes hardware approaches to switch between different modalities. To the best of authors' knowledge, the software-based approach (such as the adoption of GANs) for modality switching in multimodality sensors remains rarely explored. Our primary objective is to create a multimodal sensor that includes vision, tactile, proximity, and force sensing functions. In addition, we plan to gather extensive information and improve the sensor's robustness across different environmental settings. To accomplish this target, we plan to utilize GAN-based models. These models are intended to ease the transition between sensing modes and boost the sensor's overall performance.

III. DESIGN AND FABRICATION

A. Design Considerations

Our goal in designing the ViTacTip sensor is to combine multimodal sensing's versatility with the design inspired by the human fingertip [28]. We aim to replicate the human skin's layered functionality, including the dermal and subcutaneous layers [9]. The sensor features a slim, flexible, and transparent rubber-like surface, designed with pin-like formations to resemble the touch-sensitive properties of human skin. Inside, it houses a transparent, highly flexible polymer that simulates the mechanical characteristics of the skin's deeper layers, which enhances the sensor's tactile response. An internal camera tracks markers on the internal structure of the rubber skin, allowing for precise, remote, and sensitive detection of surface deformations [28]. A ring-shaped light source is used to project uniform illumination downward, highlighting black markers on the tips of internal pins within the camera's view. The overview of the sensor's key components with an exploded view is shown in Fig. 2(a).

The transparent outer layer of the ViTacTip sensor serves a dual purpose. First, it facilitates proximity sensing, enabling the sensor to estimate the distance to objects approaching its surface without requiring direct contact, as illustrated in Fig. 2(b). Second, it allows for the visual observation of object features. Upon contact, the soft, flexible skin conforms to the object's shape, delivering detailed tactile information and supporting force measurements. This functionality is further enhanced by an internal supporting gel, which adjusts the orientation of embedded pins to amplify the richness of touch information.

The sketch of ViTacTip's working principles can be found in Fig. 2(c), including the illustrations of the following:
1) the projections of tactile deformations mapped by the pin-like black markers;
2) visual feature projections captured through the sensor's transparent skin;
3) internal lighting passes through the skin.

The sketch also compares the sensor's performance with and without the pin-like marker design. It demonstrates that the displacement of the markers, denoted as Δx, is significantly greater than that of the traditional dot-like marker when the sensor interacts with the same indenter.

B. Fabrication of ViTacTip

The detailed manufacturing process of the ViTacTip sensor is comprehensively illustrated in Fig. 3. The fabrication process
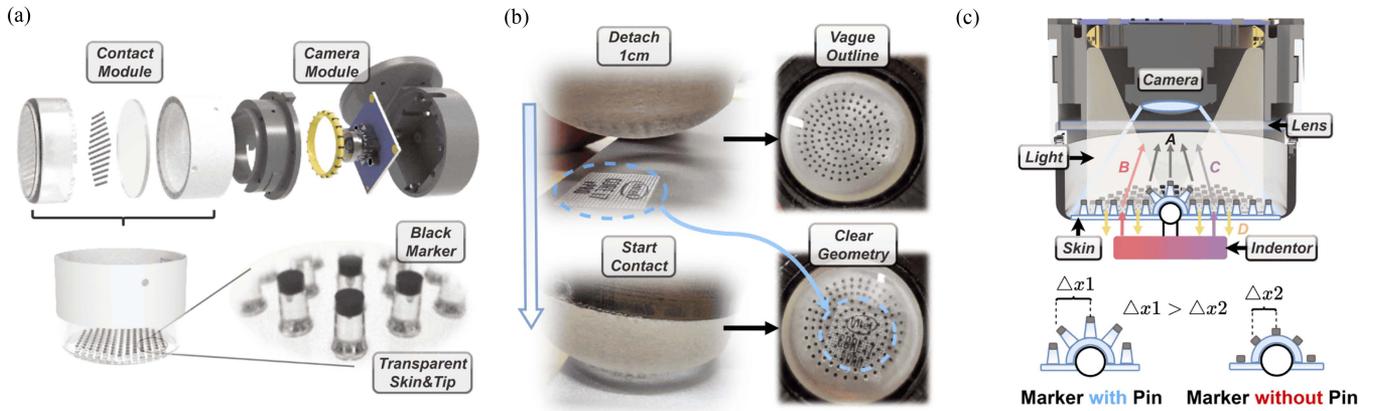
Fig. 2. (a) Architecture of the ViTacTip Sensor: An exploded view illustrating the sub-components, a side view of the assembled ViTacTip, and a detailed illustration of the sensor's tips and markers. (b) Demonstration of ViTacTip's proximity and vision perception capabilities. (c) Sketch of ViTacTip's working principles: A: Projections of tactile deformations mapped by the displacement ($\triangle x$) of black markers on the sensor tips. B/C: Visual feature projections captured through the sensor's transparent skin. D: Internal lighting passes through the skin, illuminating nearby objects. The sketch also compares the sensor's performance with and without a pin-like marker design.

Fig. 3. Design and fabrication schematic of the ViTacTip. (a) Entire Model Design: Exploded view of subcomponents. (b) Fabrication Process: The mounting bases and outer skin are 3-D printed. An acrylic lens is glued to the base, and the gel is prepared and injected [29], [30]. (c) Assembly Process: After fabricating the ViTacTip outer skin, it is assembled with the camera unit, illumination unit, and mounting base to construct the complete system.

is adapted from the open-source designs for the TacTip sensor.[1] However, the materials used for the fabrication of TacTip and ViTacTip are different. The black skin of the TacTip blocks the external light source for visual perception. TacTip primarily focuses on tactile sensing without integrated vision capabilities. In contrast, ViTacTip integrates vision modalities seamlessly with tactile sensing via adopting a transparent skin, which enhances the sensor's ability to capture and interpret detailed multimodal data. More specifically, ViTacTip uses materials with high light transmission properties that enhance the perception of visual features. As shown in Fig. 3(a), the overall ViTacTip sensor consists of eight subassemblies, which can be categorized into three major modules: structure, contact, and perception.

1) Structure Module: The structure module of the sensor system, which includes the camera base and housing part, serves as both a foundational and connective component in the sensor assembly. Fabricated using the Stratasys F370 printer, known for its precision and quality, the module is made from ABS (Acrylonitrile Butadiene Styrene), a thermoplastic polymer valued for its strength and durability. The high tensile strength of ABS ensures the module's resilience under physical stress, while its excellent impact resistance and toughness safeguard sensitive internal components, such as the cameras. Furthermore, ABS represents an economical manufacturing option, since it offers a balance between cost-effectiveness and high-quality production.

2) Contact Module: The fabrication of the contact module involves a three-step process. As depicted in Fig. 3(b), the assembly of the transparent skin, featuring pin-shaped markers on its inner surface, alongside a rigid base, is accomplished using the multimaterial Stratasys J826 3-D printer. Specifically, the skin and pins are constructed from Agilus30 Clear, a material known for its flexibility and high light transmission properties. It has translucent properties, which makes it suitable for applications where the perception of visual features of external objects is important. Agilus30 Clear has a Shore A hardness of 30–35,[2] which offers a blend of flexibility and firmness ideal for conforming to various shapes while maintaining structural integrity. Its polymerized density ranges from 1.14 to 1.15 g/cm^3, indicating high durability and stability. With a tensile strength of 2.4–3.1 MPa, it resists stretching and pulling forces, which is crucial for enduring mechanical stresses in tactile sensing. These properties make Agilus30 Clear an ideal material for equipping ViTacTip with the ability to accurately detect and adapt to object contours.

The ViTacTip markers and base are fabricated using VeroBlack, which is selected for its rigidity and opacity.[3] The pins are initially printed with Agilus30 Clear to ensure flexibility and transparency. Using the multimaterial capability of the Stratasys J826 3-D printer, the tips of the pins are then overprinted with VeroBlack in a seamless single print cycle. This process ensures precise placement of the black markers at the pin ends, resulting

in robust, well-defined markers that contrast effectively with the transparent portions.

After the 3-D printing process, a postprocessing step is necessary to remove the support material. The ViTacTip markers and base incorporate support structures crucial for preserving their integrity and geometry during the printing process. Support removal can be achieved either through chemical dissolution, where a chemical bath selectively dissolves the support material, or via a high-pressure water jet, which removes the support physically without chemical exposure. We primarily use the water-jet method in this article.

After cleaning up, a 1.5 mm thick acrylic board is laser-cut into a suitable shape and fixed on the print part through glue. Then, in the final step, a gel-like soft material is prepared by mixing two solutions, TECHSTL RTA27905 A/B, in a 1:1 ratio. Similar to the preparation of silicone, bubble removal through a vacuum machine is required to realize pure transparency. The prepared gel is carefully injected via a syringe into the space between the skin and the lens, with special attention to avoid the introduction of air bubbles. After 12 h of standing in the heating chamber, the internal gel gradually stabilized, and the entire contact module is ready to use. As shown in Fig. 3(c), the contact module will be assembled with the structure and perception modules to build an entire ViTacTip sensor.

3) Perception Module: The perception module in the sensor system consists of an ELP USBFHD06H-L180 camera[4] and ring-shaped LED lighting, both of which are essential for image data capture. For fair comparative studies during hardware benchmarking experiments, we use the same camera as TacTip [31]. The camera, equipped with a wide-angle lens, provides a broad field of view for thorough image capture. The flexible ring-shaped LED is selected for housing it to the limited space within the ViTacTip. We use a gelatinous diffusion material to envelop the LED ring. It effectively softens and evens out the light, which helps reduce glare and shadows. Such uniform light distribution is essential for multimodal sensing, as it ensures consistent illumination necessary for capturing high-quality images, thereby supporting the accurate and reliable perception.

IV. MULTIMODALITY SENSING

A. Principles of Multimodality Fusion

The ViTacTip sensor is designed with a transparent skin, allowing LED light to penetrate and illuminate an area that extends more than 20 mm beyond the sensor's surface. This transparent skin enables the sensor to capture visual information about nearby objects, including their contour, color, and position information. The sensor effectively discerns most features of an object when it is within a 5–10 mm range. Generally, as the ViTacTip approaches an object, proximity perception becomes increasingly significant. When the sensor is within 4 mm of an object, its visual perception capabilities are significantly

---

[1][Online]. Available: https://softroboticstoolkit.com/tactip

[2][Online]. Available: https://www.javelin-tech.com/3d/stratasys-materials/agilus-30/

[3][Online]. Available: https://www.dinsmoreinc.com/material/verowhite-veroblue-and-veroblack/

[4][Online]. Available: https://www.elpcctv.com/elp-2mp-full-hd-1080p-uvc-free-driver-low-light-imx322-usb-camera-wide-angle-for-android-p-333.html
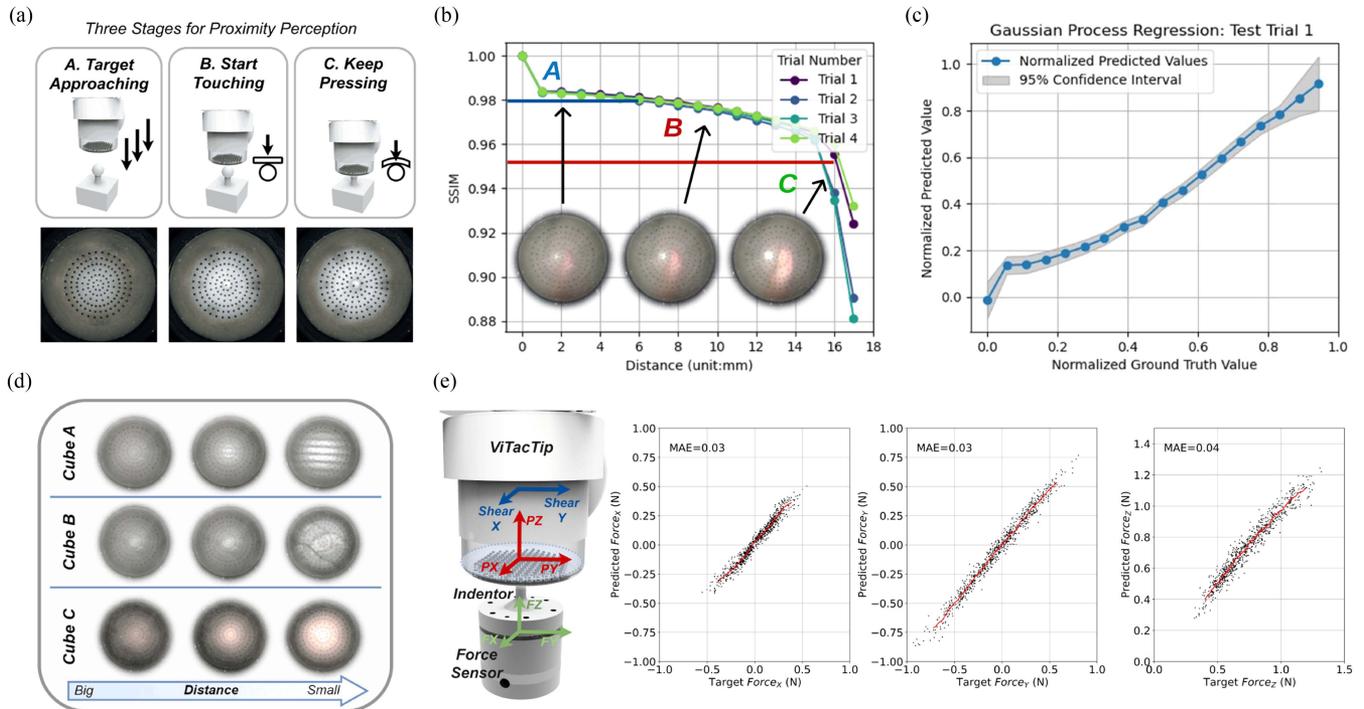
Fig. 4. (a) Three stages of the ViTacTip sensing process and the illustration of the proximity perception mechanism. (b) Curve showing the relationship between distance (0–18 mm) and SSIM values of images obtained from ViTacTip, using a human finger database as an example. Points "A," "B," and "C" correspond to the three stages illustrated in (a). The blue and red lines indicate the thresholds for segmenting the stages between A and B, and B and C, respectively. (c) GPR-based distance estimation using a database involves approaching human finger with ViTactip. (d) Examples of images captured during proximity perception between ViTacTip and three cubes with different textures. (e) Mean average errors (MAE) of ViTacTip in force estimation (F_x, F_y, F_z). Black dots represent the results of the trained model on the test dataset, while the red line represents the smoothed predictions.

enhanced. Upon direct contact, the sensor's functionality expands to include force detection and tactile sensing. This contact activates the sensor's camera, located at its base, to generate a detailed multimodal fusion image. This image combines visual and tactile data, providing a more comprehensive understanding of the object than single-modal sensors can achieve. In addition, changes in marker distribution upon contact indicate the sensor's force and tactile perception features.

In this section, we evaluate the force sensing, proximity sensing, and vision-tactile fusion-based comprehensive recognition capabilities of the ViTacTip. We begin by illustrating the sensor's proximity and force sensing capabilities (see Sections IV-B and IV-C, respectively). Next, we explore and validate the sensor's advanced capabilities in capturing a broader range of visual and tactile information through two tasks: i) shape (object) recognition (see Section IV-D); and ii) hardness, material and texture sensing (see Section IV-E).

B. Proximity Sensing

1) Task Description: Proximity sensing is crucial for human–robot interaction and robotic manipulation tasks. Traditional coated-type sensors [10] and marker-type sensors [28] (with black skin) lack this capability. In contrast, the ViTacTip sensor, with its transparent skin, offers a unique proximity-sensing capability. To evaluate this capability, a detailed experiment was conducted by mounting the ViTacTip sensor on the

end-effector of a Ufactory 850 robotic arm.[5] The experiment setup is illustrated in Fig. 4(a).

To evaluate the reliability and variability of distance measurements, we collected data across various scenarios involving interactions with a sharp object of complex shape, a flat desktop, a human finger, and three cubes covered with fabrics featuring different texture patterns, referred to as Cube A, Cube B, and Cube C, respectively. The resulting datasets are referred to as the "sharp object dataset," "human finger dataset," "flat desktop dataset," "Cube A/B/C Dataset," respectively.

During experiments, we controlled the ViTacTip to gradually approach the target object and acquire the corresponding perception image in real time. The sensor was directed toward the object gradually until direct contact was established, which marked a critical point in the experiment. Notably, the robot's movement did not stop at the initial contact; it continued an additional 4 mm downward. Each trial of the experiment was divided into distinct phases based on the sensor's distance from the object. In Stage A, when the sensor was positioned within 20–40 mm of the object, the initial activation of its proximity detection feature was observed. As the sensor approached the target, entering Stage B, the sensor started to reflect a more precise visualization of the object. Upon making contact with the object, as depicted in Stage C, a notable change was observed in the sensor's output. This stage was characterized by changes

[5][Online]. Available: https://www.ufactory.cc/ufactory-850/

TABLE I
SUMMARY FOR DISTANCE PREDICTION TESTS ACROSS DATASETS

Dataset	Trial 1	Trial 2	Trial 3	Trial 4	Mean
Sharp object	0.0009	0.0009	0.0009	0.0012	0.0010
Human finger	0.0045	0.0007	0.0026	0.0039	0.0029
Flat desktop	0.0074	0.0054	0.0067	0.0045	0.0060
Dataset	**Trial 1**	**Trial 2**	**Trial 3**	**Trial 4**	**Mean**
Cube A	0.0016	0.0016	0.0017	0.0015	0.0016
Cube B	0.0015	0.0011	0.0021	0.0012	0.0016
Cube C	0.0017	0.0020	0.0022	0.0027	0.0022

in the distribution of markers, signaling the commencement of tactile and force perception.

We use the structural similarity index measure (SSIM) [32] to estimate object distances for proximity perception, a choice inspired by previous studies [33]. An image captured when the ViTacTip is far from the target object serves as the reference image. The SSIM is then calculated between the real-time image and this reference image. Fig. 4(b) illustrates the relationship between distance and SSIM values of images obtained from ViTacTip, using the human finger database as examples, where points "A," "B," and "C" correspond to the three stages shown in Fig. 4(a). The blue and red lines represent the thresholds used to segment the stages between i) Stage A and B, ii) Stage B and C, respectively.

However, the effectiveness of SSIM in proximity sensing is influenced by object characteristics, such as texture, color, and reflectivity [34]. For instance, smooth surfaces, such as desks or nontextured walls, often exhibit limited textural variability. This makes distance estimation more challenging compared to textured surfaces, such as a human hand or fabric-covered objects. Therefore, we explore a machine learning-based approach to enhance the robustness for proximity perception capability. More specifically, SSIM is used as a feature for the Gaussian process regression (GPR) model-based distance estimation, equipped with a radial basis function kernel and supplemented by a constant kernel to account for data variability.

2) Results Analysis: The mean squared error (MSE) between two images is calculated to quantify the errors between ground truth distance estimation values and model predictions for proximity perception. As shown in Table I, in the sharp object dataset, MSE values were highly consistent, ranging from 0.0009 to 0.0012, with an average of 0.0010, indicating reliable predictions across trials. In contrast, the flat desktop dataset showed greater variation in MSE values, ranging from 0.0045 to 0.0074, with an average of 0.0060. For the human finger dataset, the MSE values remained relatively stable. The average MSE for this dataset was 0.0029, which is slightly higher than that of the sharp object dataset but significantly lower than flat desktop dataset. This reflects the sensor's improved ability to capture fine details, compared to simpler desktop scenarios. An example of the comparison between distance estimation using the GPR model and the ground truth data is shown in Fig. 4(c).

The GPR model delivered reliable proximity estimation across the Cube datasets. Examples of images captured from proximity perception between ViTacTip and the three different

cubes with various textures are shown in Fig. 4(d). For Cube A, MSE values ranged from 0.0015 to 0.0017, with an average of 0.0016, reflecting stable predictions throughout the trials. Cube B demonstrated similar stability, with MSE values between 0.0011 and 0.0021, also averaging at 0.0016. Cube C exhibited slightly greater variation, with MSE values between 0.0017 and 0.0027 and an average of 0.0022. These results confirm the model's robustness in estimating proximity for various objects, with minor differences likely due to texture variations.

The overall performance highlights the ViTacTip sensor's reliability across various proximity scenarios. It performs particularly well in detecting sharp objects, human fingers, and cubes with complex textures (Cube A and B). However, there is some variability in proximity detection for the flat desktop and Cube C, which have simpler textures. This may be due to the lack of distinct features, slightly reducing the sensor's accuracy in SSIM-based proximity perception. To further enhance the sensor's proximity sensing capabilities, we will incorporate advanced deep learning models capable of adapting to environmental changes and varying object properties in the future.

C. Force Sensing

1) Task Description: Force sensors are crucial for robotics. However, the high cost of these sensors can often be a barrier for many businesses and individuals who require them. For example, standard force sensors like the ATI (Axia80-M20)[6] or NBIT[7] are priced between $3000 and $5000. Given this high cost, there is an increasing interest in the robotics community to explore tactile sensors as a low-cost alternative to these traditional high-end force sensors.

In our research, we explore the potential of ViTacTip for 3-axis force estimation, while the same method can be extended for 3-axis torque estimation. We conducted an experiment to explore the usage of ViTacTip as a force sensor. We positioned a commercial NBIT force sensor beneath the ViTacTip,[8] which was mounted on the wrist of a low-cost robotic arm (MG400, Dobot). The robotic arm can be used to control the ViTacTip to approach the targeted force sensor. This setup allowed us to simultaneously gather data from the force sensor and the ViTacTip during their interaction. A dataset of 3000 images was collected to train a deep neural network model. We ensured that the pose values $[X(mm), Z(mm), \theta(°)]$ varied within $[-5, 5]$ for X, $[-1, 1]$ for Z, and $[-45, 45]$ for θ, respectively [29].

The ViTacTip sensor was designed to prioritize high sensitivity in detecting subtle force variations, a critical requirement in robotic applications for precise control and preventing damage during tasks such as grasping, assembly, and handling. Its small force measuring range ensures operation within its most sensitive regime, enabling accurate detection and measurement of subtle forces. This capability is particularly valuable for safely and efficiently interacting with fragile materials or soft objects [35].

---

[6][Online]. Available: https://www.ati-ia.com/index.aspx
[7][Online]. Available: http://www.iibtcn.com
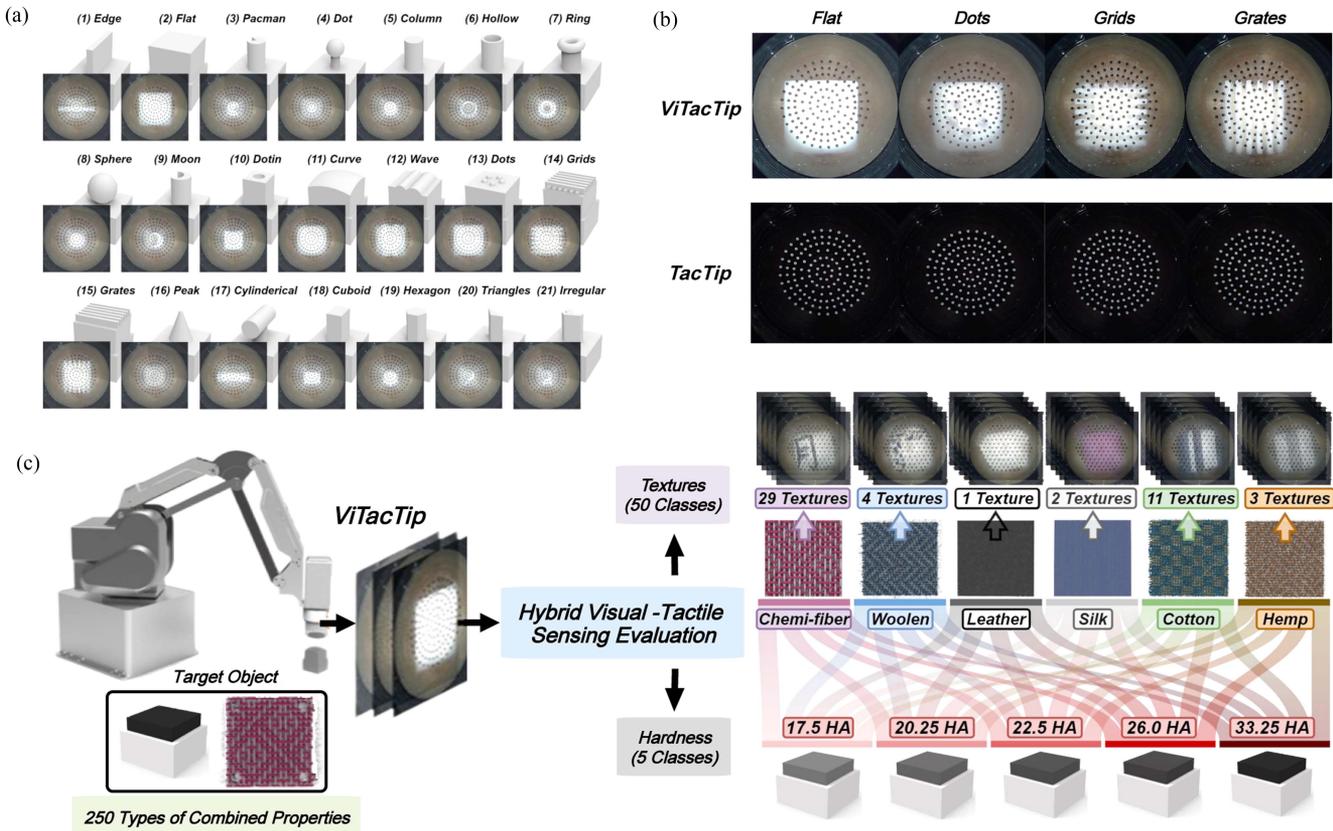[8][Online]. Available: https://www.sensor360.org/brand/7473

Fig. 5. (a) Samples for the object recognition task: 21 objects with different shapes, as detailed in [32]. Real imaging from ViTacTip allows recognition of object localization and contour, with marker distribution adapting to the contact shape. (b) High-resolution perception images from ViTacTip and TacTip obtained by interacting with typical objects, illustrating challenges in shape differentiation using tactile information alone without visual support. (c) Experimental setup: a desktop Dobot robotic arm (MG400) for data collection [29], examples of texture samples used in experiments, and a schematic of hybrid visual-tactile sensing evaluation. Covering the elastomer with fabric requires ViTacTip to perceive both the elastomer's hardness and the fabric's texture upon contact.

2) Results Analysis: The experiment began with a standard press on the sensor and introduced a shear displacement along the horizontal plane. This action was designed to exert a 3-D force on the sensor's surface, simulating real-world scenarios where forces are not limited to a single direction. In addition, we varied the intensity of external light sources during the experiment to assess the impact of changing ambient light conditions on the sensor's performance. Our primary objective was to accurately predict the forces (Fx, Fy, Fz) resulting from normal pressing and the horizontal shearing motions.

As shown in Fig. 4(e), our results demonstrated a remarkable ability to predict 3-D forces accurately. Specifically, we observed a horizontal force estimation error between -0.5 and 0.5 N and a normal pressure prediction error of 0.04 N. These results suggest that the ViTacTip not only serves as a feasible alternative to traditional force sensors in terms of cost but also provides reliable and precise force perception. The estimated cost for each ViTacTip sensor is about $50, including the costs for 3-D printing and camera.

D. Multimodality Perception for Object Recognition

1) Task Description: The ability to capture comprehensive object shape information is essential for robots performing a

wide range of tasks [36]. Luo et al. [32] established this task as a benchmark for evaluating a tactile sensor's capability to capture shape information, using objects with distinct geometric configurations, such as 3D-printed spheres, crescents, triangles, and others.

The ViTacTip sensor is well-suited for object recognition tasks. We hypothesize that its accuracy in object recognition can be improved due to the integration of both vision and tactile perception. We conducted experiments involving interactions between the sensor and 21 indentors of diverse 3-D shapes to evaluate the ViTacTip's effectiveness in capturing detailed information for object recognition [37]. Examples of these interactions are illustrated in Fig. 5(a). Through ViTacTip's real-time imaging, the localization and contour of each object are discernible. Concurrently, the distribution of markers on the sensor's skin changes, reflecting the shape of the contact.

Objects with shapes like "Grids," "Grates," "Curves," and "Dots" [see Fig. 5(b)] present significant challenges to pure tactile perception. These shapes appear indistinguishable when relying solely on the deformation detected by the sensor's skin, due to the limited resolution of pin-like markers. This is where the ViTacTip's visual capabilities become critical. By providing a detailed visual representation of these complex shapes, the ViTacTip effectively overcomes the limitations of tactile-only

perception, highlighting its multimodal advantage in distinguishing objects that are difficult to differentiate through touch alone.

We initiated the experiments by collecting training data. This involved securely positioning the target objects, referred to as stimuli, on a stable surface. The ViTacTip sensor was then applied to each stimulus at predetermined contact poses to ensure a standardized data collection process. The center point position of each object's surface was defined as [0, 0, 0, 0], in terms of [X(mm), Y(mm), Z(mm), θ(°)]. For each object, 500 images were collected, resulting in a total of 10 500 images for 21 objects. The range of contact poses between the ViTacTip sensor and the object was defined as [5, 5, 1, 90] to [−5, −5, −1, −90]. To ensure precise deformation for accurate sensing, the sensor was progressively pressed down by 5 mm at each contact point. Contact positions were systematically determined by evenly sampling within these predefined ranges, providing a uniform dataset representing the sensor's response to various points of contact. Following data collection, we fine-tuned DenseNet121 [38] to process the images captured by ViTacTip. The embeddings extracted during this process were subsequently passed through a softmax function for classifying different objects.

2) Results Analysis: In the object recognition task, ViTacTip achieved a nearly perfect accuracy at 99.91%, which far exceeds the 88.03% achieved by TacTip. As can be seen from Fig. 5(a), the visual information within the multimodality feature plays a significant role, as it clearly shows the contours of the object below, as well as the fine details of the contact surface. The spatial resolution and feature sensitivity of this pixel-level information surpass that of TacTip, which relies solely on sparse marker patterns. This advantage is particularly evident for objects with similar general shapes but differing fine details, such as the Cuboid (18) and the Hexagon (19). By integrating visual and tactile data, ViTacTip achieved superior recognition and differentiation of complex shapes, highlighting its potential as a powerful tool for automated object recognition tasks.

E. Multimodality Perception for Hardness, Material, and Texture Recognition

1) Task Description: The unique strength of the ViTacTip sensor lies in its ability to fuse multiple modalities. To thoroughly evaluate the effectiveness of this design, we conducted a comprehensive assessment of its multimodality sensing capabilities.

For this evaluation, we fabricated five distinct elastomers using 3-D printing technology. These elastomers were designed with varying hardness levels, measured on the Shore A scale, ranging from 17.5 to 33.25 HA. Each elastomer measured 30 mm × 30 mm × 10 mm and was securely affixed to a rigid base to ensure stability during testing. In addition, a diverse set of fabric samples was used to evaluate the sensor's texture recognition capabilities, comprising 50 distinct textures.

During the data collection and evaluation process, each fabric sample was placed over the elastomers. This setup enabled the ViTacTip sensor to visually capture the fabric's surface patterns while discerning the hardness of the underlying elastomer through tactile sensing. By isolating these two modalities, the experimental design effectively assessed the sensor's visual and tactile performance independently. To ensure comprehensive testing, all possible combinations of fabrics and elastomers were evaluated, resulting in a total of 250 combinations. This approach enabled the generation of a dataset that thoroughly assessed the ViTacTip sensor's ability to visually distinguish different textures and recognize various hardness levels via tactile perception.

The experimental setup and examples of the collected data are illustrated in Fig. 5(c). These results demonstrate that the ViTacTip sensor captures not only tactile information but also visual features, such as the color and design of contact objects. The captured images facilitate a detailed analysis of the spatial arrangement and deformation of tactile elements, which is essential for precise tactile perception. This approach significantly enhances the sensor's ability to interpret complex tactile data by providing visual context, thereby improving its utility in applications requiring accurate object identification and precise manipulation capabilities.

In this task, the pose range was defined within bounds of [3, 3, 0, 90] and [−3, −3, 0, −90] for [X(mm), Y(mm), Z(mm), θ(°)]. For each combination within these bounds, 100 images were collected, resulting in a total of 25 000 images across 250 combinations. The dataset was subsequently divided into training, validation, and test sets with a 5:3:2 ratio. These datasets were used to train and evaluate a multitask learning model, as detailed in the following.

Our network architecture employs a hierarchical multihead structure, where each head specializes in classifying one specific property: hardness, material, or texture. The network leverages the feature vector output from DenseNet [39], which is distributed among the three subnetwork heads. It is important to recognize the hierarchical relationship between material and texture. While different fabrics may be woven from the same material, they can exhibit a wide range of textures. Material properties often require a combination of tactile and visual perception to infer general features, such as the fibrous texture of wool, the smoothness of silk, and the roughness of hemp. In contrast, specific textures rely more heavily on visual information, including pattern geometry and color distribution, which are unique to each fiber type. This hierarchical relationship between material and texture properties is reflected in the design of the network. Details of the network structure and its implementation are illustrated in Fig. 6.

2) Results Analysis: After training, the network achieved recognition rates of 98.81% for material identification and 97.78% for texture identification. These high accuracy rates are particularly noteworthy considering the task's complexity: the sensor was required to differentiate among 50 distinct fabric patterns while these fabrics were placed over elastomers with varying hardness properties. The presence of the elastomers beneath the fabrics added an additional layer of complexity, potentially hindering the sensor's ability to accurately identify
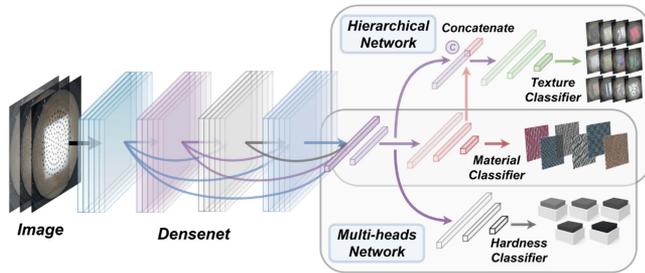
Fig. 6. Neural network architecture of hierarchical multitask learning for material, texture, and hardness recognition.

TABLE II
HARDWARE BENCHMARKING RESULT

Sensors	Grating ↑	Pose Err. (mm)↓				Contact Pt. Err. (mm)↓			
	Acc	X	Z	θ	Ave	P_x	P_y	P_z	Ave
Tactip	94.60%	0.16	0.11	0.47	0.25	0.56	0.57	0.10	0.41
ViTac	99.72%	0.39	0.12	0.34	0.28	0.43	0.43	0.07	0.31
ViTacTip	**99.72%**	**0.14**	**0.08**	**0.24**	**0.15**	**0.34**	**0.34**	**0.08**	**0.25**

The bold values are used to emphasize key results or significant findings.

material and texture. Despite these challenges, the network demonstrated remarkable robustness and efficacy.

Moreover, the network successfully recognizes varying levels of hardness between elastomers, achieving an accuracy of 97.47%. This result highlights the sensor's ability to discern subtle hardness variations through contact. The sensor's internal camera efficiently captures contact deformation, using markers to translate tactile information into measurable data via image processing. These results highlight the ViTacTip sensor's potential in applications such as cloth manipulation, where accurate identification of materials and textures is essential.

V. HARDWARE BENCHMARKING

A. Overview

In this section, we present hardware benchmarking experiments to showcase ViTacTip's properties across the following three tasks:

1) contact point detection;
2) pose regression;
3) grating identification.

For comparative analysis, we included two additional tactile sensors. The first sensor, referred to as *TacTip* [9], features pin-shaped markers embedded in an opaque (black) skin. To ensure a fair comparison, the TacTip was configured with the same marker distribution as the ViTacTip, consisting of seven concentric circles of markers (137 in total) and a skin thickness of 0.8 mm. This configuration balances contact sensitivity and wear resistance. The opaque skin of the TacTip effectively blocks out external light sources. The second sensor, referred to as *ViTac*, uses only transparent skin and differs from the ViTacTip by excluding internal markers while maintaining the same skin setup, including thickness and material composition.

To ensure fair and consistent comparisons, we used DenseNet121 as the backbone architecture for all tasks and sensors. The comparative results provide valuable insights into the advantages of the ViTacTip and identify potential areas for improvement in practical applications. Samples collected from the TacTip, ViTac, and ViTacTip sensors during the hardware benchmarking experiments are shown in Fig. 7(a).

B. Contact Point Detection

This experiment focused on accurately predicting the relative spatial location of the stimulus' contact point on the sensor's skin, represented by the coordinates (Px, Py, Pz). In a comparative analysis of performance metrics among TacTip, ViTac, and ViTacTip, ViTacTip demonstrates superior accuracy in detecting the contact point on the x and y axes, achieving the smallest mean errors of 0.34 mm in both dimensions. In contrast, TacTip exhibits the largest contact point estimation errors across all three dimensions, with values of 0.56 mm, 0.57 mm, and 0.10 mm, respectively.

The comparative results for contact point localization using the three sensors (TacTip, ViTac, and ViTacTip) are shown in Table II and Fig. 7(b). ViTacTip demonstrated exceptional precision in localizing contact points, even under challenging conditions such as varying shear forces and changes in lighting. This high performance is primarily attributed to the incorporation of additional visual information, which effectively mitigates positioning errors caused by shear movements and ensures reliable and consistent results.

C. Pose Regression

To effectively evaluate the sensor's proprioception capabilities, pose regression tests were conducted [31]. Previous studies have shown that the TacTip sensor is capable of pose regression [31], [40], [41]. Building on this foundation, we conducted a regression analysis using deep learning methodologies, focusing on predicting the sensor's pose relative to the boundary of a square stimulus.

The measured parameters included the horizontal distance (X) from the sensor's center to the boundary, the depth of press (Z) along the Z-axis, and the angle of rotation (θ) along the Z-axis. We collected 3000 images each for ViTac, TacTip, and ViTacTip, resulting in a total of 9000 images. The dataset was randomly divided into training, validation, and test sets in an 8:1:1 ratio. To ensure a diverse range of pose values, X(mm), Z(mm), and θ(°), were varied within the ranges of $[-5, 5]$, $[-1, 1]$, and $[-45, 45]$, respectively.

The comparative performance of the three sensors in pose regression is presented in Fig. 8(a), while examples of different pose values represented by various perception images are visualized in Fig. 8(b). Compared to TacTip, ViTacTip demonstrates substantial improvements in reducing pose regression errors for X (mm), Z (mm), and θ (°), achieving reductions from 0.16 mm/0.11 mm/0.47° to 0.14 mm/0.08 mm/0.24°. The most significant enhancement is observed in orientation estimation, with a 49% improvement.

Although ViTac surpasses TacTip in θ estimation, its lack of a pin-like structure leads to the largest errors in X and Z estimations among the three sensors, reaching 0.39 mm and
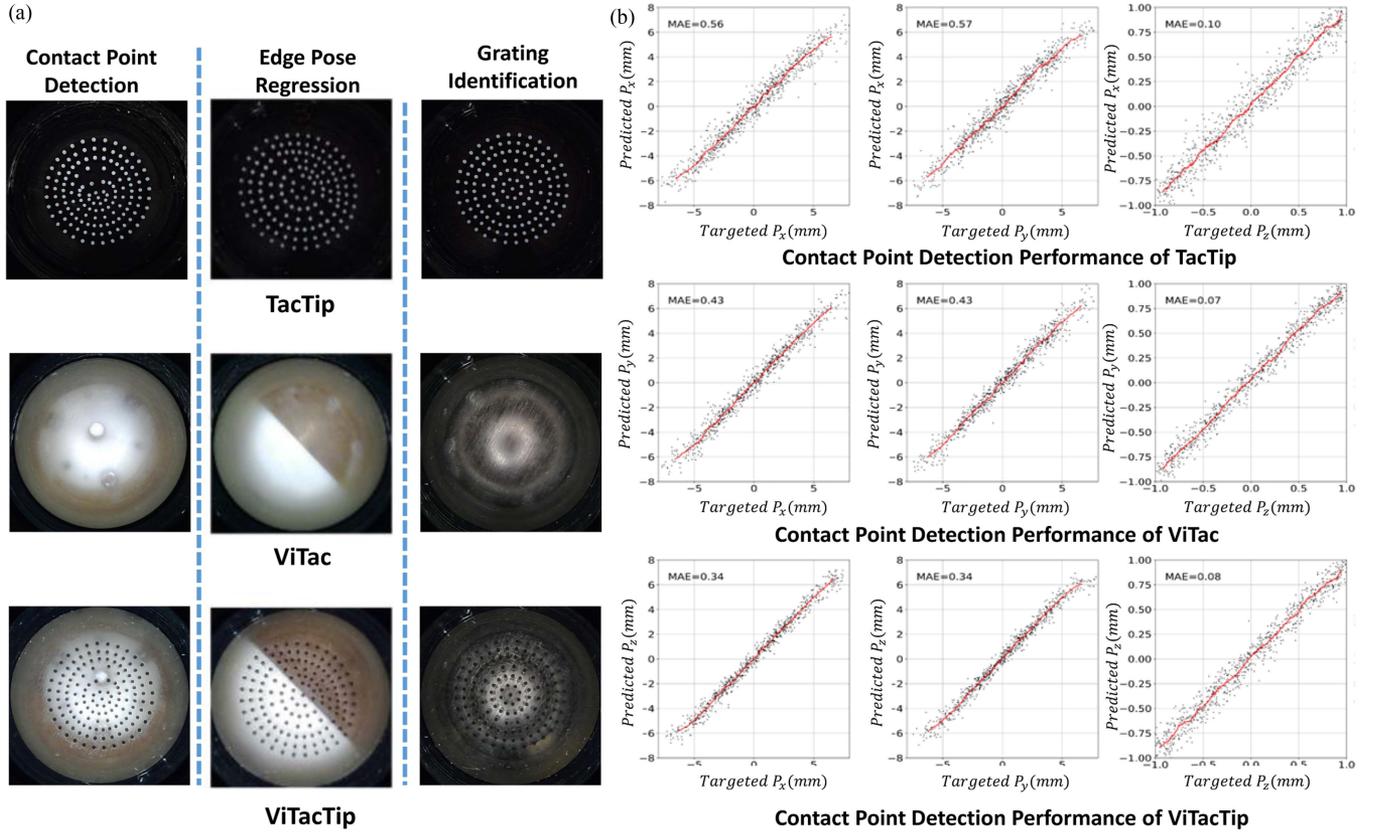
Fig. 7.   (a) Overview of the samples collected using TacTip, ViTac, and ViTacTip for hardware benchmarking experiments, including contact point detection, pose regression, and grating identification. (b) Comparative analysis of TacTip, ViTac, and ViTacTip based on MAE for the contact point detection task (P_x, P_y, P_z).

0.12 mm, respectively. These results indicate that TacTip and ViTacTip, both incorporating pins and markers, are more effective in amplifying deformations caused by skin pressing against edges, enhancing pose regression accuracy. Overall, ViTacTip achieves the best performance.

D. Grating Identification

To estimate the spatial resolution of the ViTacTip, our study employed objects with various levels of detail as outlined in [10]. A set of grating boards with different density specifications was applied to evaluate the spatial resolution of our proposed sensor. As shown in Fig. 9(a), we categorized the spatial resolution of these boards into three groups for tactile sensing: millimeters (0 mm, 1 mm, 2 mm), half-millimeters (0 mm, 0.5 mm, 1 mm, 1.5 mm, 2 mm), and quarter-millimeters (1 mm, 1.25 mm, 1.5 mm, 1.75 mm, 2 mm). In our analysis, we processed a substantial dataset of 3500 data points, divided across the 7 categories of grating board specifications (0 mm/Flat, 0.5 mm, 1 mm, 1.25 mm, 1.5 mm, 1.75 mm, 2 mm). The data were split in a 7:2:1 ratio for training, validation, and testing.

Our research revealed that the TacTip sensor exhibited lower precision in detecting quarter-millimeter measurements. It struggled to accurately distinguish spacings of 1.25 mm, 1.5 mm, and 1.75 mm, resulting in a reduced test accuracy of 94.60%. In contrast, both the ViTac and ViTacTip sensors demonstrated significantly better performance, achieving an impressive test accuracy of 99.72%. This comparison underscores the superior

sensitivity and precision of the ViTac and ViTacTip technologies in tasks requiring fine spatial resolution.

The improvement in spatial resolution achieved by ViTacTip, compared to TacTip, can be attributed to its integration of visual features. These visual capabilities enable ViTacTip to overcome certain limitations inherent to TacTip, such as constraints related to marker density and the conformability of the sensor's skin. By combining visual data with tactile sensing, ViTacTip significantly enhances its ability to discern fine spatial differences. In addition to the results presented in Table II, the confusion matrices for the grating identification task performed by the three sensors (TacTip, ViTac, and ViTacTip) are shown in Fig. 9(b).

VI. GAN-Based Multimodal Interpretation

A. Overview

The physical multimodality fusion approach used in ViTacTip offers both advantages and disadvantages. For instance, while the marker layout in ViTacTip may not be dense enough to discern fine tactile/force-relevant information effectively, the additional visual modality compensates by providing richer details, such as color and fine texture-relevant features. However, the presence of black markers can introduce noise into the visual data, potentially distorting visual perception. In addition, the sensor's transparent skin, even with its internal light source, is still vulnerable to interference from external ambient lighting. The colorful background further complicates conventional tactile processing techniques, such as marker detection
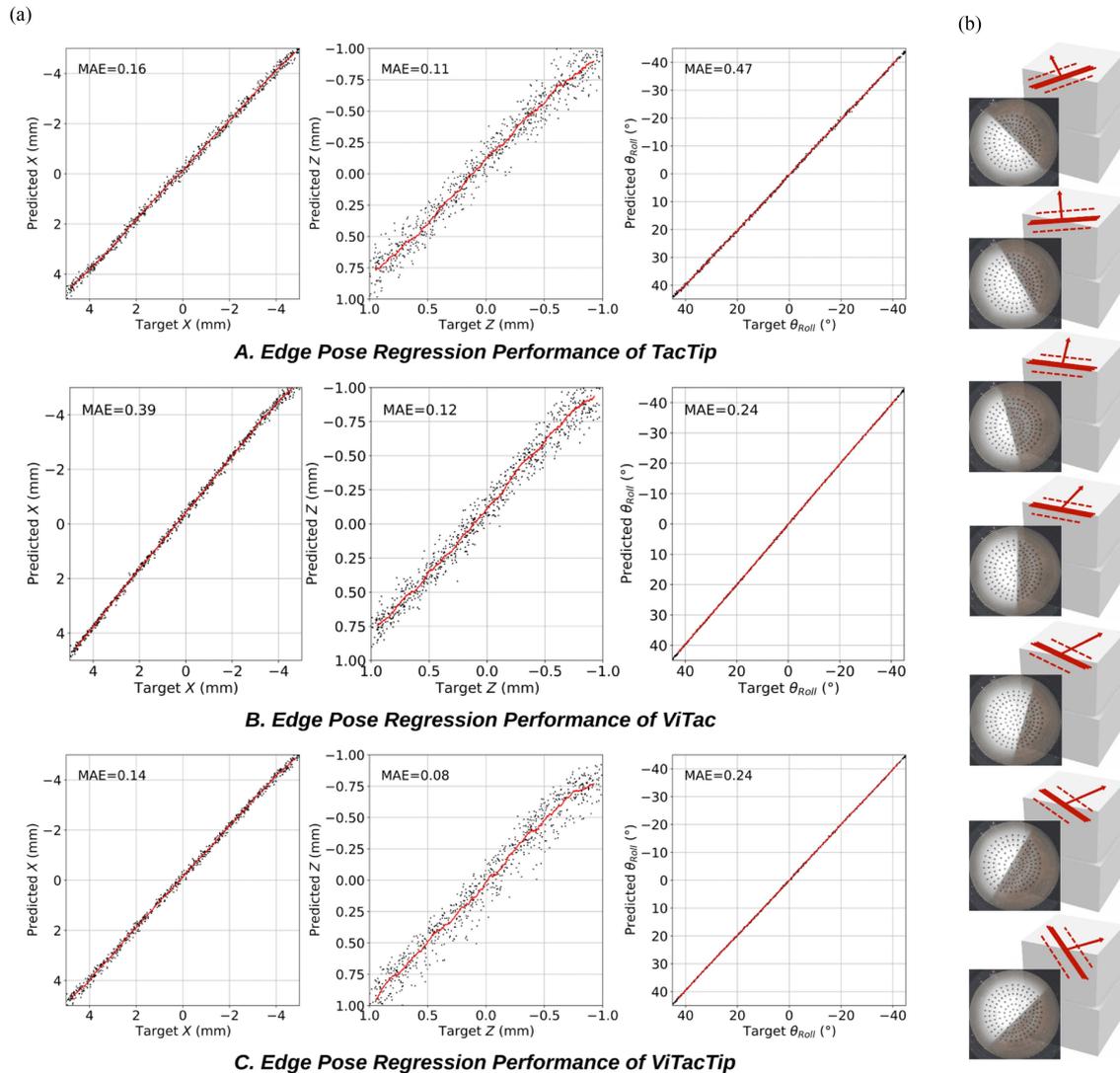
Fig. 8. (a) Evaluation result of TacTip, ViTac, and ViTacTip on pose regression for horizontal distance X, press depth Z, rotation angle θ. (b) Samples for the pose regression task using ViTacTip sensor.

and tracking. To address these challenges, we proposed two GAN-based solutions: one for marker removal and another for light interference mitigation.

B. Method Description

We collected three distinct datasets for our experiment using the ViTacTip, ViTac, and TacTip sensors. The TacTip sensor demonstrates minimal sensitivity to external light sources but lacks visual perception capabilities. In contrast, the ViTac sensor, which is designed without internal pins or markers and consists solely of transparent skin, exhibits higher sensitivity to ambient light due to its simplified design. However, it compensates for this limitation by offering enhanced visual perception capabilities, allowing for improved detection of the objects it interacts with.

We trained the MR-GAN to transform data from ViTacTip to ViTac. The output from MR-GAN replicates ViTac's data, effectively eliminating occlusions caused by markers. Similarly, the LR-GAN processes ViTacTip data to emulate TacTip,

mitigating the effects of ambient light interference. At the core of this approach is the separate extraction of visual and tactile information. This separation enhances the sensor's adaptability, enabling it to be tailored for a wide range of applications.

Both MR-GAN and LR-GAN employ the Pix2Pix GAN architecture [42] for perception image translation tasks. Pix2Pix GAN conditions the generation of an output image on a corresponding input image, making it well-suited for modality switching among the ViTacTip, ViTac, and TacTip sensors. The MR-GAN and LR-GAN networks comprise two main components: a Generator (G) and a Discriminator (D). The generator adopts a U-Net-style architecture, while the discriminator is implemented as a PatchGAN [42]. The objective of the Pix2Pix GAN is formulated as a mini–max game between the generator and the discriminator.

C. Results Analysis

1) Qualitative Results Analysis: Fig. 10(a) demonstrates that MR-GAN effectively removes all markers from ViTacTip data
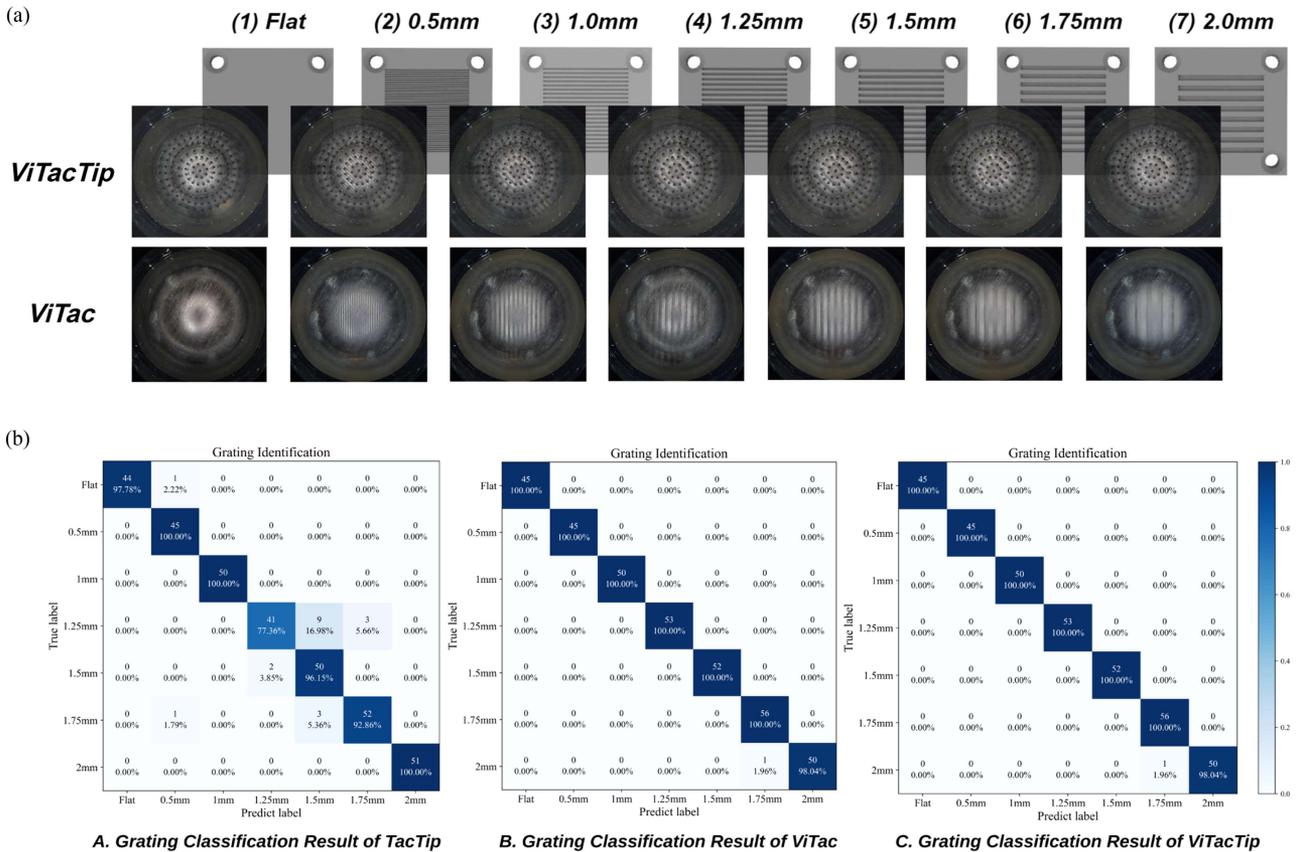
Fig. 9.   (a) Samples for the grating identification task and the perception images obtained using ViTacTip and ViTac. (b) Evaluation results of TacTip, ViTac, and ViTacTip on grating identification, presented as a confusion matrix.

collected during object recognition and grating identification tasks, while preserving fine perception image features through visual perception. The resulting images closely resemble those obtained directly from the ViTac sensor. These results indicate that MR-GAN successfully eliminates markers and retains intricate texture details, significantly improving the clarity of tactile information, which is essential for enhancing accuracy in contact visualization during object recognition. Additional examples of marker removal using MR-GAN for the object recognition task are presented in Fig. 10(b).

Similarly, Fig. 10(a) illustrates the processing of ViTacTip data for grating identification and pose regression tasks using LR-GAN. The transformed data successfully emulates the Tac-Tip style by replacing the background and external lighting with a black skin and white markers. This transformation enables the application of marker-based tactile algorithms to ViTacTip, enhancing its robustness against ambient light interference and improving its functionality in varied lighting conditions. More examples for the pose regression task, shown in Fig. 10(c), demonstrate the successful conversion of ViTacTip data into TacTip-style data, featuring pure markers. This highlights the versatility and effectiveness of LR-GAN in adapting ViTacTip data for marker-based tactile processing.

Additional results on modality switching and the generation of various data visualization formats are presented in Fig. 10(d). As shown in Fig. 10(e), the GAN-generated images maintain consistent quality even under varying ambient lighting

conditions. Fig. 11 provides examples of modality conversion using MR-GAN and LR-GAN on cube-like objects with irregular-shaped patterns and striped cloth-like patterns. These results demonstrate the proposed method's applicability to complex scenarios, highlighting its versatility and robustness.

2) Quantitative Results Analysis: To evaluate the quality of the data generated by the GAN models, 40 images were randomly selected from each task for quantitative analysis using three metrics: MSE, peak signal-to-noise ratio (PSNR), and SSIM. MSE quantifies the average squared difference between the pixel values of the ground-truth image in the target domain (referred to as the "target image") and the corresponding image generated by the GAN model. A lower MSE value indicates greater similarity and better alignment between the two images, as it reflects smaller pixel-wise differences. PSNR, expressed in decibels (dB), measures the peak error between the target and generated images. Higher PSNR values indicate better image quality and greater similarity. Specifically, a PSNR value exceeding 40 dB reflects extremely high similarity, while values below 20 dB denote low similarity. SSIM evaluates the structural similarity between the target and generated images by considering luminance, contrast, and structure. SSIM values range from 0 to 1, with values closer to 1 indicating higher similarity and better visual quality.

According to the results summarized in Table III, all PSNR values exceed 40 dB, and the average SSIM scores are greater than 0.7. These results confirm that the GAN-generated outputs
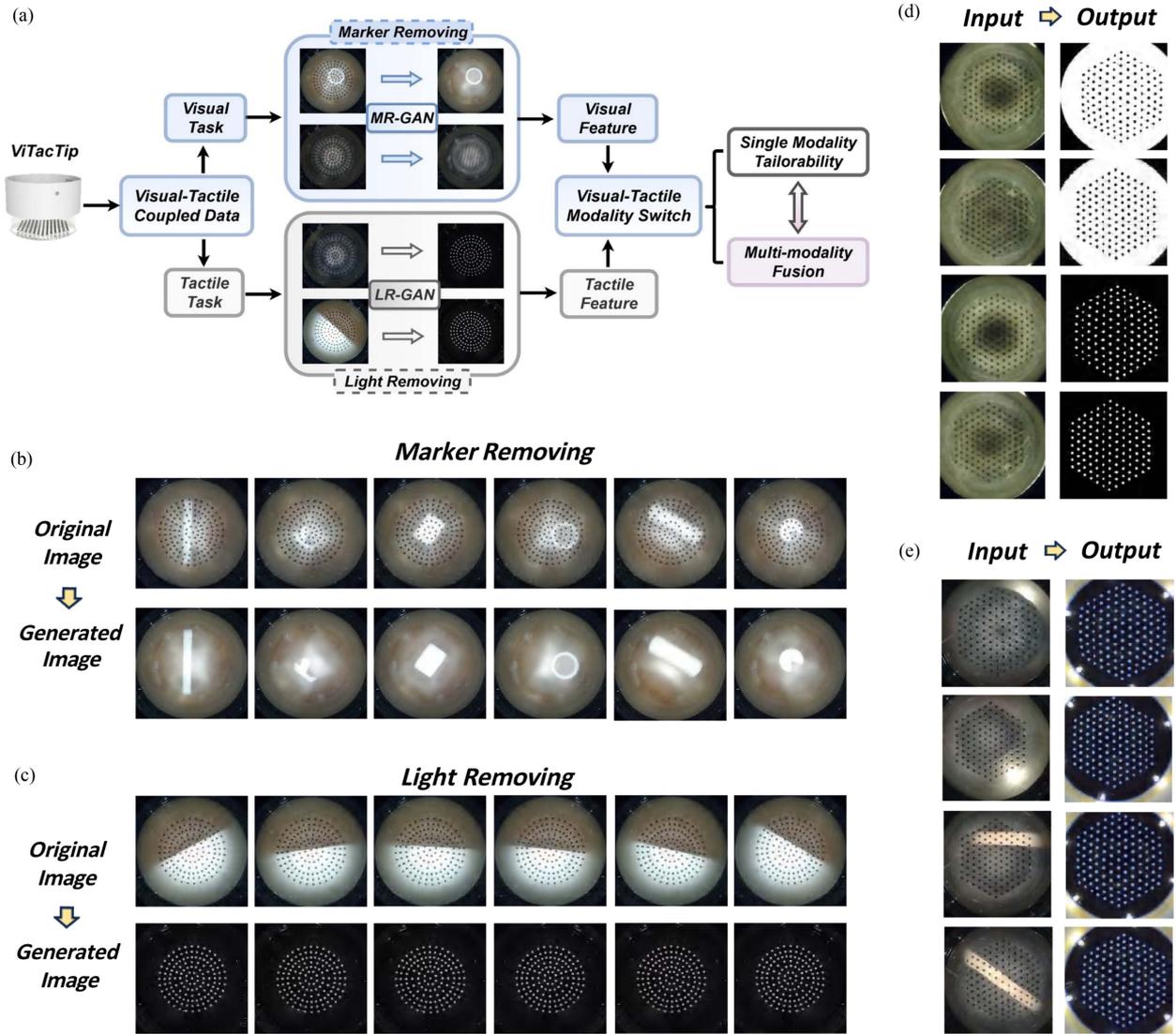
Fig. 10. (a) Overview of the modality-switch framework: Two independent GAN models are trained on datasets from ViTac, ViTacTip, and TacTip to perform marker removal and light removal. (b) Results of the MR-GAN modality switch for the object recognition task, converting ViTacTip-style images into ViTac-style images. (c) Results of the LR-GAN modality switch for the pose regression task, converting ViTacTip-style images into TacTip-style images. (d) Transformation of ViTacTip images into binary outputs isolating the tips, showcasing the versatility of the proposed GANs. (e) Demonstration of consistent image quality in GAN-generated outputs, regardless of variations in ambient lighting conditions.

TABLE III
QUANTITATIVE EVALUATION: COMPARISON OF GROUND-TRUTH DATA WITH LR-GAN AND MR-GAN-BASED MODALITY TRANSFER RESULTS

Scenarios	MSE↓	PSNR↑	SSIM↑
MR similarity on grating data	0.0405	62.078	0.805
MR similarity on object data	0.0645	60.125	0.837
MR Similarity on Texture Data	0.0105	70.534	0.932
LR Similarity on Grating Data	0.0104	70.063	0.892
LR Similarity on Pose Data	0.0429	62.663	0.708
LR Similarity on Texture Data	0.0191	71.498	0.727

"Texture Data" refers to data from cube-like objects with irregular-shaped patterns.

achieve a high degree of similarity to the ground truth images. In summary, the proposed GAN models successfully executed modality switching, performing both marker-removing and light-removing tasks across different scenarios, including grating boards, various objects, and random poses.

D. Robustness Evaluation of Proposed Method

To further assess the robustness and capabilities of the GAN models in complex environments involving challenging tasks, their performance was evaluated under conditions with both fixed and random light sources. The LR-GAN model demonstrated remarkable consistency, maintaining high image quality across both lighting scenarios. It achieved an average PSNR of 65.57 and an SSIM of 0.954, with an MSE of approximately 0.0208. In contrast, the MR-GAN model exhibited some variability in performance. Under fixed lighting conditions,
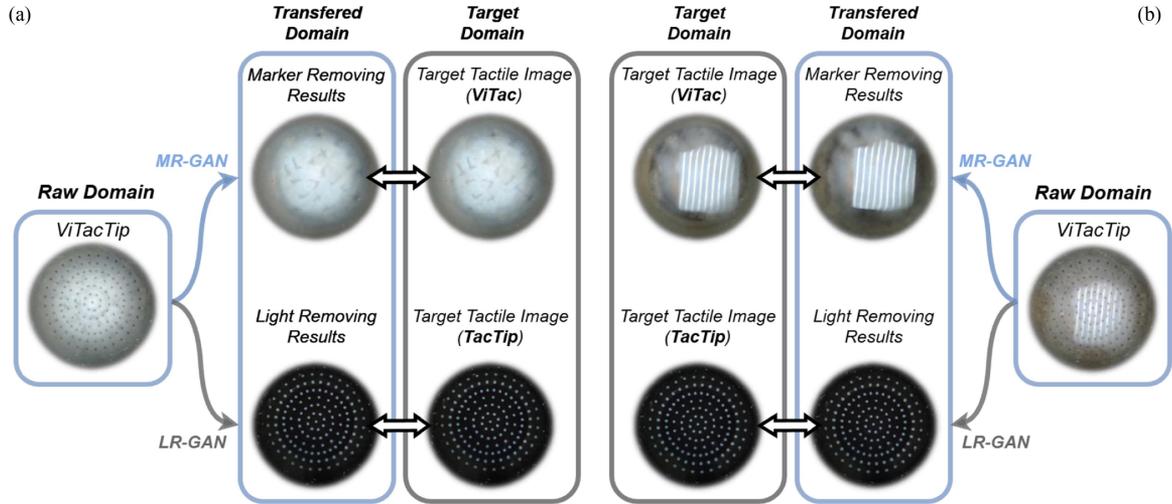
Fig. 11. Examples of modality conversion when applying MR-GAN and LR-GAN to cube-like objects with complex textures. (a) Irregularly shaped patterns. (b) Striped cloth.

TABLE IV
PERFORMANCE EVALUATION OF GAN MODELS UNDER DIFFERENT LIGHTING CONDITIONS WHEN APPLYING MR-GAN AND LR-GAN ON CUBE-LIKE OBJECTS WITH COMPLEX TEXTURES (STRIPED CLOTH-LIKE PATTERN)

Scenarios	MSE↓	PSNR↑	SSIM↑
MR, Fixed Light Source	0.0383	63.077	0.926
MR, Random Light Source	0.0799	59.217	0.837
LR, Fixed Light Source	0.0208	65.568	0.954
LR, Random Light Source	0.0208	65.569	0.954

it achieved moderately high image quality with a PSNR of 63.08 and an SSIM of 0.926. However, under random lighting conditions, its performance declined slightly, with the PSNR decreasing to 59.22 and the SSIM dropping to 0.837. The MSE also increased to 0.0799, indicating a higher error rate and greater sensitivity to lighting changes compared to the LR-GAN model. Table IV summarizes these findings.

Both GAN models significantly enhance the adaptability of the ViTacTip sensor. However, the LR-GAN model demonstrates superior robustness across varying environmental conditions. In contrast, the MR-GAN model performs effectively under consistent lighting but shows a decline in performance under variable lighting. This indicates the need for further refinements to enhance its robustness and achieve adaptability comparable to the LR-GAN model in dynamic environments.

VII. DISCUSSIONS AND FUTURE WORK

A. Discussions

1) Comparison With Single-Modality VBTS: GelSight-type sensors are renowned for their high spatial resolution in tactile perception. Unlike the mold-cast elastomer skin of GelSight sensors, the ViTacTip features a 3D-printed skin [29], enabling customizable shapes and dimensions. Its ultrasoft, gel-filled construction further enhances conformance to curved surfaces,

making it ideal for tasks demanding precise force sensing, a large dynamic range, and safe object contact.

Furthermore, the ViTacTip sensor also benefits from the easy modification of its pin density or shape through 3-D printing, which can improve tactile perception capability. However, increasing the pin density may impact visual perception capabilities, as the pins can obstruct the detection of fine geometric details on targeted objects. To address this, we need to find an optimal balance between tactile and visual perception capabilities through careful design of the markers and improvements in the generalizability of GAN-based modality switch approach.

In addition, while maintaining the same volume, weight, and robustness as TacTip, the integration of multimodal perception marks a significant advancement. This positions ViTacTip as the sensor of choice for applications demanding both detailed tactile information and visual perception of fine features. The incorporation of multimodal sensing elevates environmental interaction and object recognition by enabling the simultaneous detection of shape, material, and texture, rather than relying solely on touch-based data.

2) Comparison With Multimodality VBTS: Here, we conduct a comparative analysis of state-of-the-art multimodality sensors developed between 2016 and 2023, as outlined in Table V. This analysis focuses on critical aspects of tactile sensor design, including the sensing approach, spatial resolution, surface shape, modality coupling, and the capability to switch between modalities.

Upon examination, ViTacTip can capture fine details of both tactile and visual in the captured frames, delivering rich data for subsequent analysis. First, ViTacTip serves as a proficient proximity sensor, capable of perceiving precise visual information through its see-through-skin mechanism. This mechanism, coupled with high spatial resolution, is important for applications such as object and texture recognition. In addition, the pin-shaped markers in ViTacTip exhibit high sensitivity to dynamic tactile events, effectively monitoring touch deformations caused by normal or shear forces. This feature is

TABLE V
COMPARATIVE ANALYSIS WITH STATE-OF-THE-ART MULTIMODALITY SENSORS

Sensor	Sensing approach	Resolution	Shape	Modality coupling	Modality switch
FingerVision[25]	Vision and markers	-	2-D	High	None
STS [43]	Vision and coating	Micron	3-D	High	None
Finger-STS [44]	Vision and coating and markers	Micron	3-D	High	None
Kazuhiro's [45]	Vision and TIR	Micron	2-D	None	Compound-eye Camera (Visable/IR)
SpecTac [26]	Vision and UV markers	-	2-D	None	UV LEDs (ON/OFF)
UVtac [46]	Vision and coating and UV markers	-	2-D	None	UV LEDs (ON/OFF)
StereoTac [47]	Vision and coating	Sub-mm	3-D	None	LEDs (ON/OFF)
StereoTac [47]	Vision and coating	Sub-mm	3-D	None	LEDs (ON/OFF)
TIRgel [27]	Vision and TIR	Micron	2-D	None	LEDs (ON/OFF) & Focus (Near/Far)
VisTac [48]	Vision and coating	Micron	2-D	None	LEDs (ON/OFF)
ViTacTip (Ours)	**Vision and markers**	**Micron**	**3-D**	**Decoupled by GAN**	**MR-GAN & LR-GAN**

The bold values are used to emphasize key results or significant findings.

especially valuable when the sensor's skin conforms to or moves along a touched surface. In contrast, sensors like STS [43] and TIRgel [49] lack force-sensing capabilities. ViTacTip, however, offers a more comprehensive suite of functionalities, including force regression, shape recognition, and proximity sensing. In addition, unlike Kazuhiro's sensor [45], whose contact surface is flat, ViTacTip's pin-like markers can be easily printed on 3-D curved surfaces, which can enhance its sensing capability in complex scenarios.

Unlike other multimodal VBTSs [25], [43] that lack a modality-switching function, ViTacTip enables independent access to perception data from different modality domains. This design effectively mitigates the challenges associated with modality coupling, enabling seamless extraction and utilization of individual modalities as required. Through its GAN-based image modality translation, ViTacTip operates efficiently across multiple modes, including multimodal fusion, visual perception, and tactile perception.

Most other multimodal VBTSs, such as SpecTac [26], UVtac [46], and TIRgel [49], depend on state-switching capabilities of external hardware, such as UV LEDs or camera focus adjustments. Hardware-based multimodal systems often face the challenge of missing significant information during modality switching gaps, particularly in dynamic scenarios. ViTacTip addresses this limitation by adopting a software-based approach, eliminating the redundancy and complexity associated with hardware design for multimodal sensors. By bypassing hardware dependencies, ViTacTip minimizes delays, significantly enhancing its efficiency and robustness in environments where high response speed is critical. Notably, the GAN-based approach employed in ViTacTip not only optimizes its performance but also offers adaptability for integration with other types of multimodal sensors, thereby broadening its potential applications and versatility across diverse fields.

In summary, the ViTacTip sensor stands out with its unique blend of hardware functionalities and software innovations, surpassing traditional single-modality VBTSs and existing multimodal VBTSs through the pioneering use of GANs for the seamless modality switch. The integration of MR-GAN and LR-GAN ensures smooth transitions between modalities. In our proposed GAN-based modality switch framework, the MR-GAN enables ViTacTip to capture fine visual details without dot-like noise caused by the pin-shaped markers. Meanwhile, the LR-GAN improves ViTacTip's adaptability to diverse lighting conditions and environmental settings, demonstrating its performance in real-world applications where other multimodal VBTSs may struggle.

B. Future Work

We plan to evaluate the sensor's effectiveness in scenarios involving multiple contact points. In addition, we will conduct comparative analyses with various types of VBTSs, focusing on aspects such as resolution, sensitivity, and adaptability to diverse environmental conditions. Comparing ViTacTip with established GelSight-type sensors is expected to provide a more comprehensive understanding of its performance. However, it is important to note that existing GelSight-type sensors involve numerous parameters that may influence their softness. For example, while the commercial DIGIT sensor is a GelSight-type sensor, it is not suitable for a fair comparison in this study due to its significantly smaller sensing area. In future research, we aim to develop standardized approaches for benchmarking tactile sensor hardware.

Furthermore, we will explore the integration of ViTacTip into various types of robotic hands. Its shape can be easily customized for seamless incorporation into the fingertips or palms of robotic hands. The multimodal sensing capabilities of ViTacTip have the potential to significantly enhance robotic dexterity, making it a valuable tool for complex manipulation tasks. In addition, ViTacTip can drive advancements in multimodal robot learning. This integration is particularly advantageous for applications requiring delicate handling, such as electronics manufacturing assembly lines, as well as laboratory and domestic settings where contact-rich manipulation is crucial.

VIII. CONCLUSION

In this study, we introduce ViTacTip, a novel sensor with multimodal capabilities comprising two principal modalities (vision and tactile) and two derived modalities (proximity and force). Hardware benchmarking of ViTacTip across tasks such as

grating identification, pose regression, and contact point detection demonstrates its versatility in a wide range of applications.

The ViTacTip's performance was benchmarked against the TacTip and ViTac sensors. In these comparisons, specific features such as the see-through-skin mechanism and biomimetic pins were intentionally omitted to align its design with those of TacTip and ViTac. This approach serves as an ablation study, allowing for the assessment of the impact of these features on overall performance. Furthermore, a hierarchical multitask learning framework highlights ViTacTip's advanced capabilities, enabling it to simultaneously identify hardness, materials, and textures. These multimodal capabilities extend beyond the limitations of TacTip and ViTac, which are confined to single-modality perception.

The integration of MR-GAN and LR-GAN models with ViTacTip's sensor data represents a significant advancement. These GAN models empower the sensor to handle diverse tactile sensing tasks, including fine feature recognition and dynamic force estimation. By transforming ViTacTip's data to emulate the output styles of TacTip and ViTac using LR-GAN and MR-GAN, respectively, the sensor achieves enhanced versatility and operational effectiveness in various scenarios.

ACKNOWLEDGMENT

REFERENCES

[1] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 2722–2727.
[2] J.-T. Lee, D. Bollegala, and S. Luo, ""Touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception," in *Proc. Int. Conf. Robot. Autom.*, 2019, pp. 4276–4282.
[3] J. Jiang, G. Cao, A. Butterworth, T.-T. Do, and S. Luo, "Where shall i touch? Vision-guided tactile poking for transparent object grasping," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 1, pp. 233–244, Feb. 2023.
[4] R. Yang, W. Zhang, N. Tiwari, H. Yan, T. Li, and H. Cheng, "Multimodal sensors with decoupled sensing mechanisms," *Adv. Sci.*, vol. 9, no. 26, 2022, Art. no. 2202470.
[5] P. Szczuko, A. Harasimiuk, and A. Czyżewski, "Evaluation of decision fusion methods for multimodal biometrics in the banking application," *Sensors*, vol. 22, no. 6, 2022, Art. no. 2356.
[6] I. Goodfellow et al., "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.
[7] Y. Lin, M. Comi, A. Church, D. Zhang, and N. F. Lepora, "Attention for robot touch: Tactile saliency prediction for robust sim-to-real tactile control," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 10806–10812.
[8] D. Saxena and J. Cao, "Generative adversarial networks (GANs) challenges, solutions, and future directions," *ACM Comput. Surveys*, vol. 54, no. 3, pp. 1–42, 2021.
[9] N. F. Lepora, "Soft biomimetic optical tactile sensing with the TacTip: A review," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21131–21143, Oct. 2021.
[10] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017, Art. no. 2762.
[11] 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.*, 2018, pp. 1927–1934.
[12] D. F. Gomes, Z. Lin, and S. Luo, "GelTip: A finger-shaped optical tactile sensor for robotic manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9903–9909.
[13] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "OmniTact: A multi-directional high-resolution touch sensor," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 618–624.

[14] K. Kamiyama, K. Vlack, T. Mizota, H. Kajimoto, K. Kawakami, and S. Tachi, "Vision-based sensor for real-time measuring of surface traction fields," *IEEE Comput. Graph. Appl.*, vol. 25, no. 1, pp. 68–75, Jan./Feb. 2005.
[15] C. Lin, Z. Lin, S. Wang, and H. Xu, "Dtact: A vision-based tactile sensor that measures high-resolution 3D geometry directly from darkness," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2023, pp. 10359–10366.
[16] C. Lin, H. Zhang, J. Xu, L. Wu, and H. Xu, "9Dtact: A compact vision-based tactile sensor for accurate 3D shape reconstruction and generalizable 6D force estimation," *IEEE Robot. Automat. Lett.*, 2023.
[17] W. Fan, H. Li, Y. Xing, and D. Zhang, "Design and evaluation of a rapid monolithic manufacturing technique for a novel vision-based tactile sensor: C-sight," *Sensors*, vol. 24, no. 14, 2024, Art. no. 4603.
[18] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors*, vol. 19, no. 18, 2019, Art. no. 3933.
[19] W. Fan, H. Li, and D. Zhang, "Crystaltac: 3D-printed vision-based tactile sensor family through rapid monolithic manufacturing technique," 2024, *arXiv:2408.00638.*
[20] W. Fan, H. Li, and D. Zhang, "Magictac: A novel high-resolution 3D multi-layer grid-based tactile sensor," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 388–394.
[21] W. Fan et al., "Graph neural networks for interpretable tactile sensing," in *Proc. IEEE 27th Int. Conf. Automat. Comput.*, 2022, pp. 1–6.
[22] Z. Abderrahmane, G. Ganesh, A. Crosnier, and A. Cherubini, "A deep learning framework for tactile recognition of known as well as novel objects," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 423–432, Jan. 2020.
[23] W. Fan, M. Yang, Y. Xing, N. F. Lepora, and D. Zhang, "Tac-VGNN: A voronoi graph neural network for pose-based tactile servoing," in *Proc. 2023 IEEE Int. Conf. Robot. Automat.*, IEEE, 2023, pp. 10373–10379.
[24] G. Cao, J. Jiang, D. Bollegala, M. Li, and S. Luo, "Multimodal zero-shot learning for tactile texture recognition," *Robot. Auton. Syst.*, vol. 176, 2024, Art. no. 104688.
[25] A. Yamaguchi and C. G. Atkeson, "Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables," in *Proc. IEEE-RAS 16th Int. Conf. Humanoid Robots*, 2016, pp. 1045–1051.
[26] Q. Wang, Y. Du, and M. Y. Wang, "Spectac: A visual-tactile dual-modality sensor using uv illumination," in *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 10844–10850.
[27] S. Zhang et al., "TIRgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robot. Autom. Lett.*, vol. 8, no. 10, pp. 6307–6314, Oct. 2023.
[28] 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, 2018.
[29] N. F. Lepora, Y. Lin, B. Money-Coomes, and J. Lloyd, "DigiTac: A DIGIT-TacTip hybrid tactile sensor for comparing low-cost high-resolution robot touch," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9382–9388, Oct. 2022.
[30] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10754–10761, Oct. 2022.
[31] N. F. Lepora and J. Lloyd, "Pose-based tactile servoing: Controlled soft touch using deep learning," *IEEE Robot. Autom. Mag.*, vol. 28, no. 4, pp. 43–55, Dec. 2021.
[32] D. F. Gomes, P. Paoletti, and S. Luo, "Generation of GelSight tactile images for Sim2Real learning," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 4177–4184, Apr. 2021.
[33] Y. Cao, F. Luo, and Y. Li, "Toward better SSIM loss for unsupervised monocular depth estimation," in *Proc. Int. Conf. Image Graph.*, Springer, 2023, pp. 81–92.
[34] B. Rasti, Y. Chang, E. Dalsasso, L. Denis, and P. Ghamisi, "Image restoration for remote sensing: Overview and toolbox," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 201–230, Jun. 2022.
[35] L. He, Q. Lu, S.-A. Abad, N. Rojas, and T. Nanayakkara, "Soft fingertips with tactile sensing and active deformation for robust grasping of delicate objects," *IEEE Robot. Automat. Lett.*, vol. 5, no. 2, pp. 2714–2721, Apr. 2020.
[36] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.
[37] D. F. Gomes, Z. Lin, and S. Luo, "Geltip: A finger-shaped optical tactile sensor for robotic manipulation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9903–9909.
[38] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," 2014, *arXiv:1404.1869.*

[39] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[40] N. F. Lepora and J. Lloyd, "Optimal deep learning for robot touch: Training accurate pose models of 3D surfaces and edges," *IEEE Robot. Autom. Mag.*, vol. 27, no. 2, pp. 66–77, Jun. 2020.

[41] J. Lloyd and N. F. Lepora, "Pose and shear-based tactile servoing," *Int. J. Robot. Res.*, vol. 43, no. 7, pp. 1024–1055, 2024.

[42] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.

[43] 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. IEEE Winter Conf. Appl. Comput. Vis.*, 2021, pp. 1217–1226.

[44] F. R. Hogan, J.-F. Tremblay, B. H. Baghi, M. Jenkin, K. Siddiqi, and G. Dudek, "Finger-STS: Combined proximity and tactile sensing for robotic manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10865–10872, Oct. 2022.

[45] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," in *Proc. 2016 IEEE Int. Conf. Robot. Autom.*, 2016, pp. 138–143.

[46] W. Kim, W. D. Kim, J.-J. Kim, C.-H. Kim, and J. Kim, "UVtac: Switchable UV marker-based tactile sensing finger for effective force estimation and object localization," *IEEE Robot. Autom. Lett.*, vol. 7, no. 3, pp. 6036–6043, Jul. 2022.

[47] E. Roberge, G. Fornes, and J.-P. Roberge, "StereoTac: A novel visuotactile sensor that combines tactile sensing with 3D vision," *IEEE Robot. Autom. Lett.*, vol. 8, no. 10, pp. 6291–6298, Oct. 2023.

[48] S. Athar, G. Patel, Z. Xu, Q. Qiu, and Y. She, "VisTac toward a unified multimodal sensing finger for robotic manipulation," *IEEE Sensors J.*, vol. 23, no. 20, pp. 25440–25450, Oct. 2023.

[49] S. Zhang et al., "TIRgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robot. Autom. Lett.*, vol. 8, no. 10, pp. 6307–6314, Oct. 2023.

Dandan Zhang (Member, IEEE) received the Ph.D. degree in robotics from the Department of Computing, Imperial College London, London, U.K., in 2021.

She is currently a Lecturer (Assistant Professor) with the Department of Bioengineering. She leads the Multi-Scale Embodied Intelligence Lab, Imperial-X Initiative, Imperial College London. She was a Lecturer with the University of Bristol, Bristol, U.K., affiliarted with Bristol Robotics Lab (2021–2023). Her research interest include intersection of robotics and AI, with main applications in healthcare.

Wen Fan (Graduate Student Member, IEEE) received the M.Sc. degree in advanced control and systems engineering from the University of Manchester, Manchester, U.K., in 2021. He is currently working toward the Ph.D. degree in robotics with the Multi-Scale Embodied Intelligence Lab, Department of Bioengineering, Imperial-X Initiative, Imperial College London, London, U.K.

His research interests include tactile robotics.

Jialin Lin received the M.Sc. degree in robotics from the University of Bristol, Bristol, U.K., in 2022. He is currently working toward the first-year Ph.D. degree in robotics with the Imperial-X Initiative, Imperial College London, London, U.K.

His research focuses on magnetic microrobots.

Haoran Li (Member, IEEE) received the M.Sc. degree in robotics in 2020 from the University of Bristol, Bristol, U.K., where he is currently working toward the Ph.D. degree in engineering mathematics with the School of Engineering Mathematics and Technology, Bristol Robotics Laboratory.

His research interests include tactile sensors and robotic hands.

Qingzheng Cong received the M.Sc. degree in robotics from the University of Bristol, Bristol, U.K., in 2023.

He is currently a Research Assistant with the Imperial-X Initiative, Imperial College London, London, U.K. His research focuses on tactile sensors and robotic hands.

Weiru Liu received the Ph.D. in computer science from the University of Edinburgh, U.K., in 1995. She is currently the Chair of Artificial Intelligence (AI) with the University of Bristol, Bristol, U.K. She is currently a Co-Director on the EPSRC Centre for Doctoral Training in Future Autonomous and Robotic Systems: Towards Ubiquity (FARSCOPE-TU, 2019–2028). From 2023 to 2027, she is also a Co-Investigator on the ESRC funded Centre for Sociodigital Futures, where she leads research on examining social implications of AI and the importance of social science in designing future AI technologies. Her research interests include, explainable AI, data-driven intelligent autonomous systems; planning; machine learning; cyber-physical systems; large-scale sensor network data analytics, and event modeling, reasoning, and correlation in uncertain environments.

Nathan F. Lepora (Member, IEEE) received the B.A. degree in mathematics in 1992 and the Ph.D. degree in theoretical physics from the University of Cambridge, Cambridge, U.K., in 1999

He is currently a Professor of Robotics and AI with the University of Bristol, Bristol, U.K., leading the Dexterous Robotics Group. His research focuses on human-like dexterity using soft tactile robots and AI.

Dr. Lepora was the recipient of the Leverhulme Research Leadership Award and co-edited the award-winning book, *Living Machines*.

Shan Luo (Senior Member, IEEE) received the B.Eng. degree in automatic control from the China University of Petroleum, Dongying, China, in 2012 and the Ph.D. degree in robotics from King's College London, London, U.K., in 2016.

He is currently a Reader (Associate Professor) in Robotics and AI with King's College London. He was a Lecturer with the University of Liverpool, Liverpool, U.K., and a Research Fellow with Harvard University, Cambridge, MA, USA, and the University of Leeds, Leeds, U.K. His research interests include tactile sensing, robot learning, and visual-tactile perception.