M^{3}Tac: A Multispectral Multimodal Visuotactile Sensor with Beyond-Human Sensory Capabilities

Shoujie Li, *Student Member, IEEE*, Haixin Yu, Guoping Pan, Huaze Tang, *Student Member, IEEE*, Jiawei Zhang, Linqi Ye, *Member, IEEE*, Xiao-Ping Zhang, *Fellow, IEEE*, Wenbo Ding, *Member, IEEE*

***Abstract*—To realize the exquisite interaction and precise manipulation for the robot, in this article, we propose a multispectral multimodal visuotactile sensor named M^{3}Tac, which combines visible, near-infrared, and mid-infrared imaging technologies for the first time and can exceed the sensing ability of human skin in terms of resolution (719 pixels/cm^2), temperature sensing range (-20∼130 ^{o}C), etc. The M^{3}Tac can not only realize high-quality sensing of deformation, texture, force, stickiness, and temperature comparable to human skin but also can realize proximity sensing that is lacking for human skin. To achieve this, we not only design a multispectral imaging system with an elastic film whose light penetrability can be regulated by the brightness of the light, but also develop corresponding algorithms, including the pixel-level force sensing with finite element method (accuracy: ±0.023 N), the proximity perception (accuracy: ±3.8 mm), the 3D reconstruction (accuracy: 0.33 mm), the super-resolution temperature sensing (accuracy: ±0.3 ^{o}C), the multimodal fusion classification (accuracy: 98%), and the stickiness recognition (accuracy: 98%). Finally, we conduct experiments to verify the effectiveness and application potential of our research. This paper has supplementary material available at https://sites.google.com/view/MTac-sensor.**

***Index Terms*—Visuotactile sensing, Multimodal sensing, Multispectral imaging**

I. Introduction

The exquisite interaction of humans with the world heavily relies on the multimodal tactile sensation of the skin. Thanks to the dense array of tactile receptors in the skin (as high as 90 units/cm^2 [1], [2] on the hand), humans are able to perceive various attributes like temperature, texture, shape, and stickiness through touch [3]. Similarly, for robots to realize complex and delicate manipulation like humans, they call for tactile sensors with multiple modalities and high performance. However, due to constraints of e-skin solutions [4] in the manufacturing process and costs, designing a tactile sensor with the desired attributes of high resolution, broad coverage, and multifunctionality still remains a formidable challenge.

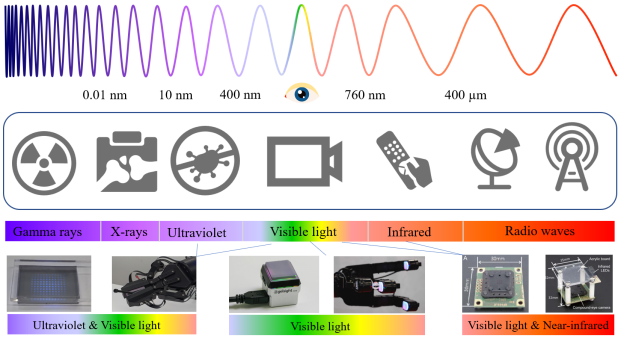

Fig. 1. Applications of different wavelengths of light. Visuotactile sensors combine UV and visible light: UVtac [5]; Visuotactile sensors in the visible light band: GelSight [6] & DIGIT [7]; Visuotactile sensors combine visible light and near-infrared: Tac [8].

With advancements in optical imaging and computer vision technologies, a novel field known as visuotactile perception technology has emerged in robotics [6], [9], [10], which offers high resolution, extensive coverage and stability at low costs. Consequently, visuotactile sensors have shown huge potential in areas such as force perception [11] and texture detection [12]. However, existing visuotactile sensors predominantly concentrate on capturing information within the visible light spectrum, which limits the perceptible dimensions and modalities of information, since light encompasses a vast spectrum of physical information beyond visible colors [13]. As shown in Fig. 1, the wavelength can be divided into X-rays, ultraviolet (UV) light, visible light, infrared light, etc. [14]. Different wavelengths of light have different physical properties, which can be applied in different fields. For instance, X-rays have strong penetrating properties and are used in industrial flaw detection and medical testing [15], UV can be used for sterilization [16], and infrared is used in night-vision monitoring equipment [17] or temperature detection [18]. Researchers have also tried to combine UV and visible light to reduce the impact of the marker on the detection of the object's contour [5], which makes up for some of the shortcomings of visible light imaging. Hence, if more wavelengths of light can be fused, it may greatly extend the sensing capability of

This work was supported by Shenzhen Key Laboratory of Ubiquitous Data Enabling (No. ZDSYS20220527171406015), Shenzhen Science and Technology Program (JCYJ20220530143013030), Guangdong Innovative and Entrepreneurial Research Team Program (2021ZT09L197), National Natural Science Foundation of China (62104125, 62003188), Shenzhen Higher Education Stable Support Program (WDZC20231129093657002), Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation (No. SZPR2023005). (Corresponding author: Linqi Ye & Wenbo Ding, yelinqi@shu.edu.cn & ding.wenbo@sz.tsinghua.edu.cn)

Shoujie Li, Haixin Yu, Guoping Pan, Huaze Tang, Jiawei Zhang, Xiao-Ping Zhang, Wenbo Ding are with Shenzhen Key Laboratory of Ubiquitous Data Enabling, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China.

Xiao-Ping Zhang and Wenbo Ding are also with the RISC-V International Open Source Laboratory, Tsinghua-Berkeley Shenzhen Institute, Shenzhen 518055, China.

Linqi Ye is with the Institute of Artificial Intelligence, Collaborative Innovation Center for the Marine Artificial Intelligence, Shanghai University, Shanghai 200444, China.

This paper has supplementary downloadable material available at https://sites.google.com/view/MTac-sensor.

visuotactile sensors.

In this paper, we propose a multispectral multimodal visuotactile sensor named M^{3}Tac, which combines visible light, near-infrared, and mid-infrared imaging technologies for the first time and can exceed the sensing ability of human skin in terms of resolution (719 pixels/cm^2), temperature sensing range (-20∼130 ^{o}C), functionality, etc. The sensor can not only realize high-quality sensing of deformation, texture, force, stickiness, and temperature comparable to human skin but also can realize proximity sensing that is lacking for human skin. The contributions of this paper are as follows:

- In the hardware aspect, we propose a high-resolution multispectral sensing system with a unidirectional perspective latex film, which can simultaneously realize the optical imaging in three different bands of 400∼700 nm (visible light), 930∼950 nm (near-infrared), and 5.5∼14 μm (mid-infrared).
- In the algorithm aspect, we propose a sensing library that includes the pixel-level force sensing and 3D reconstruction algorithms based on the near-infrared image, the proximity sensing algorithm based on the visible-light image, the super-resolution temperature sensing algorithm based on the mid-infrared image, the stickiness sensing algorithm based on the contact separation video, and the multimodal information fusion-based classification algorithm, respectively. The system can achieve a contact force sensing accuracy of 0.023 N, a 3D reconstruction accuracy of 0.33 mm, a proximity sensing accuracy of 3.8 mm, and a temperature sensing accuracy of 0.3 ^{o}C in the range of -20∼130 ^{o}C.
- To achieve pixel-level force sensing, we develop an automated acquisition & annotation system and propose a force-sensing network named FSwint-MAP based on the swin transformer, which is combined with finite element analysis to estimate the force at each pixel. Besides, we utilize the thermal residual effect to build a super-resolution temperature data acquisition system and propose a lightweight super-resolution network, which can achieve 8× temperature data super-resolution.
- Finally, we test the performance of M^{3}Tac through several experiments, including fragile and lightweight objects grasping, circuit board heat position detection, classification of liquid with different temperatures, and underwater pipeline anomalous hot spot detection experiment, which show that M^{3}Tac has a wide application in real-world scenarios.

The rest of this paper is organized as follows. The related work is reviewed in Section II. The hardware design is detailed in Section III. Section IV presents the contact force sensing algorithm, proximity perception algorithm, temperature sensing algorithm, stickiness classification algorithm, and multimodal fusion classification algorithm. Furthermore, experimental validations are provided in Section V. Finally, Section VI concludes this paper.

II. Related Work

The application of tactile sensors can help robots acquire information such as contact force [27], temperature [28], material [29], etc, which is of great significance in improving the perception and operation of robots. Current tactile sensors have high detection accuracy, but it is still difficult to obtain high-resolution tactile information due to the limitations of the manufacturing process and cost. To tackle this, visuotactile sensing technology [30] has been proposed and got widely recognized in academia and industry due to its high resolution and stability. Several visuotactile sensors, such as GelSight [6], Insight [31], have been developed. These sensors empower robots to perform complex tasks like cable manipulation [32], fabric recognition [33], and grasping of underwater objects [34]. Thus, the utilization of visuotactile sensors holds great promise in enhancing the overall performance of robotic operations. Next, we will review the current sensing technologies from four aspects: contact force sensing, proximity sensing, temperature sensing, and multimodal sensing.

A. Force Sensing

Force sensing is one of the most important functions of tactile sensors, which allows robots to sense the contact force to facilitate manipulation. After years of development, visuotactile sensors can obtain high-resolution tactile images along with high-accuracy force perception. For example, GelSight can achieve a force detection accuracy of 0.67 N by using a convolutional neural network [6]. GelSlim can achieve a detection accuracy of 0.32 N by using the finite element method (FEM) in combination with a neural network [35].

Force sensing for visuotactile sensors can be divided into data level and pixel level. Data-level force sensing can obtain the position and force of the entire contact area [36], [37]. Pixel-level force sensing can not only segment the position of the contact area but also acquire the tactile force data at each pixel location. [38]. Although pixel-level force sensing provides better detection performance, it requires a huge amount of data for sensors without markers, e.g., Insight collected and labeled 187,358 samples at 3,800 randomly selected initial contact locations to implement force sensing [31]. To solve the difficulty of data annotation, we propose a fully automated system for acquiring and annotating contact force at the pixel level, which can simultaneously obtain high-precision data on both the contact force and the contact position for each pixel, leveraging finite element analysis to ensure accuracy.

B. Promixity Sensing

In addition to contact force sensing, proximity sensing is also a hot issue in current tactile sensor research [39]. Proximity sensing enables the acquisition of the distance between an object and the sensor prior to physical contact, thereby affording the robot with increased planning time for its operations. There are many ways to realize proximity sensing. For example, Gilbert *et al.* designed an e-skin based on a magnetosensitive sensor, which uses non-contact perception to realize tasks such as virtual keyboards and dimming [40]. Ruth *et al.* proposed a bimodal tactile sensor that enables proximity and contact sensing based on capacitive sensors, which can significantly improve robotic grasping capabilities [41]. Liu *et al.* combined a triboelectric nanogenerator with a flexible

TABLE I
HARDWARE COMPARISON OF M^{3}TAC AND OTHER VISUOTACTILE SENSORS

Ref	Sensing skin			Lighting	Imaging system	Modality		Modality switch
	Material	Marker	Shape			Vision	Touch	
GelSight [6]	The coating gel	Black points	2D	RGB	RGB	✘	✔	✘
TacTip [19]	The black gel	White pins	2D	White	RGB	✘	✔	✘
Zhang *et al.* [20]	Thermochromic material	\	2D	White	RGB	✘	✔	✘
Yu *et al.* [21]	Thermochromic material	Black grates	2D	White	RGB	✘	✔	✘
HaptiTemp [22]	Thermochromic material	\	2D	White	RGB	✘	✔	✘
STS [23]	Unidirectional perspective gel	\	2D	RGB	RGB	✔	✔	Switching lights
SpecTac [24]	Transparent gel	Fluorescent marker	2D	UV	RGB	✔	✔	Switching lights
Tac [8]	Acrylic board	\	2D	Near-infrared	RGB + Near-infrared	✔	✔	Sync
FingerVision [25]	Transparent gel	Black points	2D	\	RGB	✔	✔	Sync
TIRgel [26]	Transparent gel	\	2D	White	RGB	✔	✔	Adjusting focus
M^{3}Tac (Ours)	Unidirectional perspective latex	\	2.5D	Near-infrared + White	RGB + Near-infrared + Mid-infrared	✔	✔	Sync

TABLE II
FUNCTIONAL COMPARISON OF M^{3}TAC AND OTHER VISUOTACTILE SENSORS

Ref	Force sensing	Reconstruction	Texture	Shape	Proximity	Temperature	Stickiness	Application environment	
								Light	Dark
GelSight [6]	✔	✔	✔	✔	✘	✘	✘	✔	✔
TacTip [19]	✔	✘	✘	✘	✘	✘	✘	✔	✔
Zhang *et al.* [20]	✘	✘	✔	✔	✘	5∼45 °C	✘	✔	✔
Yu *et al.* [21]	✔	✔	✔	✔	✘	25∼31 °C	✘	✔	✔
HaptiTemp [22]	✘	✘	✔	✔	✘	31∼50 °C	✘	✔	✔
STS [23]	✘	✘	✘	✔	✔	✘	✘	✔	✘
SpecTac [24]	✔	✘	✘	✔	✔	✘	✘	✔	✘
Tac [8]	✘	✔	✘	✔	✔	✘	✘	✔	✘
FingerVision [25]	✔	✘	✘	✔	✔	✘	✘	✔	✘
TIRgel [26]	✘	✘	✔	✔	✔	✘	✘	✔	✔
M^{3}Tac (Ours)	✔	✔	✔	✔	✔	-20∼130 °C	✔	✔	✔

robotic arm to propose a touchless robot control framework that allows interactive control without contact with the robotic arm [42]. Saloutos *et al.* proposed a tactile gripper with high sampling frequency by combining pressure sensors with time-of-flight proximity sensors, which achieved a grasp success rate above 90% while clearing over 100 items in an autonomous clutter-clearing task [43], [44].

As shown in Table I, to attain proximity perception utilizing visuotactile sensors, researchers have embarked on numerous studies focusing on sensing skins, primarily comprising transparent gel and unidirectional perspective gel technologies. Transparent gel can achieve proximity perception but it is difficult to achieve force perception or reconstruction because the visual camera will pass directly through the film to detect the objects outside. For example, Yamaguchi *et al.* designed FingerVision [25], a fully transparent tactile sensor for sensing skin, which is capable of directly observing the external environment but cannot discern the texture of the object being touched. To adapt to dark conditions, Zhang *et al.* proposed a new sensor, TIRgel [26], which has LEDs mounted around the sensing skin. This design ensures that the external illumination surpasses the internal sensor brightness even in dimly lit environments. An alternative approach involves employing a unidirectional perspective gel as the sensing skin. The transmittance of this gel is influenced by the differing brightness levels on its two sides, enabling the transition between proximity and contact sensing modes through the manipulation of internal brightness within the sensor. For example, Hogan *et al.* proposed the STS sensor [23], which integrates a special translucent film that facilitates the transition between vision and touch modes by adjusting the internal light brightness, but it is difficult to work in dark conditions.

However, both the transparent gel and unidirectional perspective gel methods rely on time-division multiplexing of

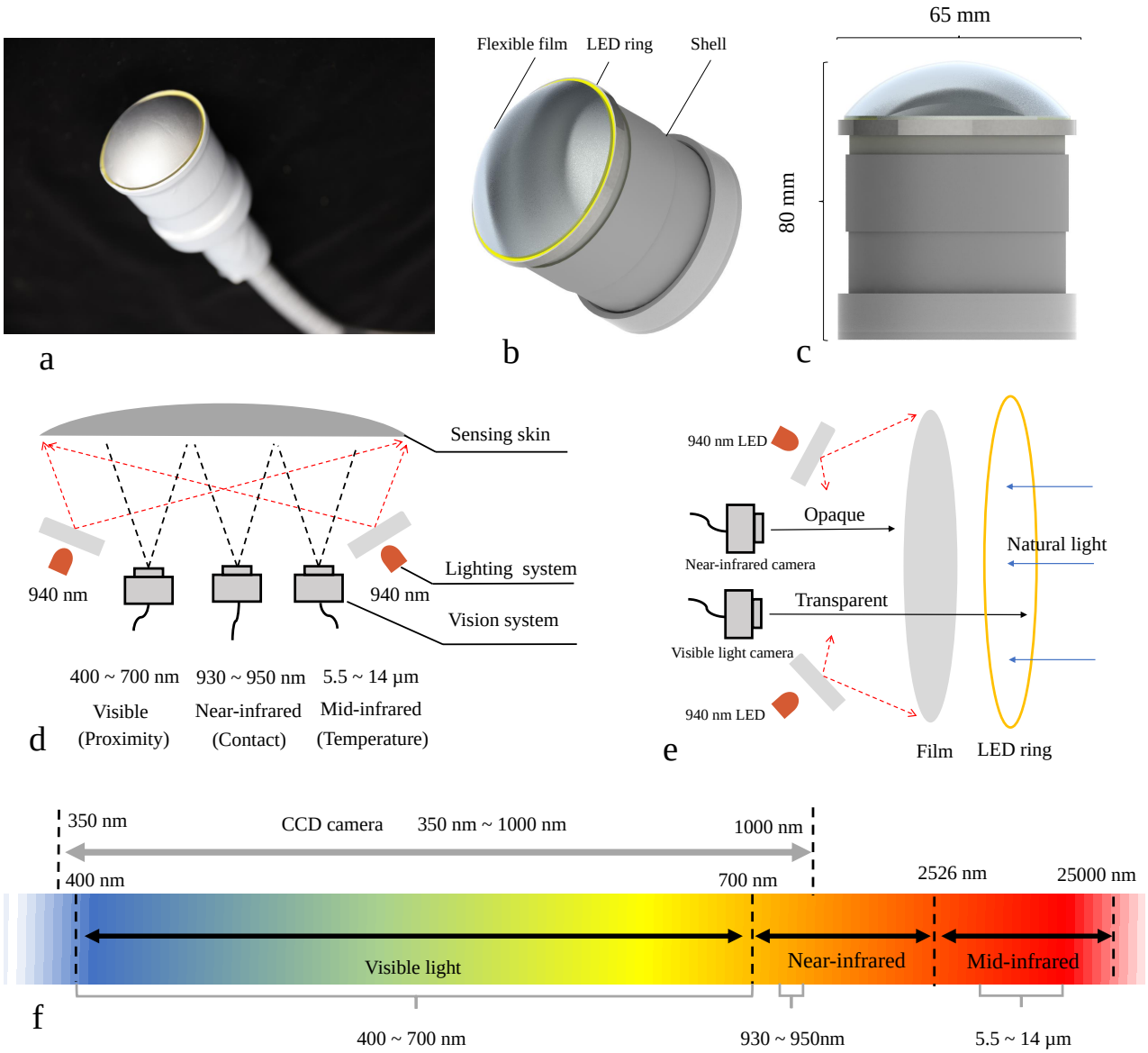

Fig. 2. Sensor overall design structure diagram. (a) The physical picture of M^{3}Tac; (b) Overall structure of M^{3}Tac; (c) Dimensions of M^{3}Tac; (d) M^{3}Tac optical imaging schematic; (e) Unidirectional perspective latex film transmission schematic; (f) Optical bands used in M^{3}Tac.

LEDs or adjustments to camera focus to alternate between proximity and contact sensing. This, to a certain extent, complicates the data acquisition process and necessitates precise control over the exposure time of the lighting elements. Furthermore, these techniques face difficulties in accurately estimating the distance of an object from the sensor when operating in the visual mode.

C. Temperature Sensing

Temperature is a crucial physical attribute of any object, and tactile sensors equipped with temperature-sensing abilities empower robots to detect and respond to variations in temperature within the objects they interact with. Common methods of temperature sensing include thermistors, thermocouples, etc. Rao *et al.* proposed an electronic skin that can differentiate between pressure and temperature at the same time, which combines triboelectric nanogenerator (TENG) with a thermosensitive electrode combining BiTO and rGO [45]. Husain *et al.* combined a fabric with a metal wire and achieved temperature measurement by measuring the resistance of the wire. While these techniques ensure the acquisition of precise temperature data, they are constrained by relatively low resolution [46].

In pursuit of achieving high-resolution temperature data, researchers have explored the integration of visuotactile sensing technology with thermochromic materials. Zhang *et al.* used thermochromic materials to design a visuotactile sensor that can sense temperatures from 5 to 45 ^{o}C, but this skin only has four states: black, pink, blue, and white [20]. To obtain a higher temperature detection accuracy, Chen *et al.* used the hue model to analyze the color change of the temperature sensing material and achieved a resolution of 0.4 ^{o}C in the temperature range of 25∼31 ^{o}C [21]. Abad *et al.* proposed a

thermochromic thin film with a fast temperature response and used an LAB model for color-to-temperature mapping [22].

Although these methods can realize temperature sensing, there are still two problems to be resolved:

- Limited temperature detection range. The diverse temperature ranges encountered by robots in home service scenarios pose a significant challenge, as they often exceed the detection capabilities of current sensors, thereby hindering their practical applicability in such environments.
- Repeated calibration for each sensor. While temperature sensing through the analysis of thermochromic material color changes is feasible, the intricate relationship between the film's thickness, material composition, and the resulting color variation complicates the direct establishment of a uniform, linear correlation between temperature and color.

D. Multimodal Sensing

With the extension of robotic applications, a single functional tactile sensor can no longer meet the task requirements of robots in variable scenarios. In this case, the study of multimodal tactile sensors becomes necessary. Most multimodal tactile sensors only combine two functions, such as pressure stimuli and temperature variations [47], tactile and proximity sensing [48], etc. Furthermore, the low resolution of these sensors makes it difficult to acquire information such as the contours and textures of the objects in contact.

As shown in Table I and Table II, to realize high-resolution tactile perception, researchers have made many improvements on the basis of classic visuotactile sensors such as GelSight [6], TacTip [19], etc. For example, to realize high-resolution temperature perception and texture perception, researchers combined thermochromic materials with visuotactile perception [20]–[22]. High-resolution proximity sensing was also realized by designing special optical films or adjusting the focal length [8], [23]–[26]. However, designing a tactile sensor that surpasses human skin's sensing capabilities, capable of concurrently achieving high-resolution texture, deformation, force, temperature, and proximity perception, remains a formidable yet crucial research endeavor. Such a sensor would represent a significant advancement in the field.

III. Hardware Design

A visuotactile sensor usually consists of a sensing skin, an imaging system, and a lighting system, which acquires tactile information by analyzing the deformation, texture, and color changes of the sensing skin through the imaging system. In this paper, we design a high-resolution visuotactile sensor that can acquire contact force, texture, proximity, and temperature information simultaneously, as shown in Fig. 2(a)(b)(c). To achieve this, we optimize the design of the sensor's imaging system, lighting system, and sensing skin.

The internal structure is shown in Fig. 2(d), to realize the detection of light with a broad spectrum of wavelengths, we design an imaging system that can detect visible, near-infrared, and mid-infrared light simultaneously. The visible-light camera is used for proximity sensing, the near-infrared camera is used for shape and texture detection, and the mid-infrared imaging system is used for temperature detection. To realize proximity sensing, we design a film with special optical properties, which looks opaque on the brighter side and transparent on the darker side, as shown in Fig. 2(e). By adjusting the brightness of different wavelengths of light on both sides of the elastic film, we can realize selective transmission of light. To control the light inside the sensor, we design a fully enclosed sensor housing, in which the visible light from the outside is brighter than that from the inside, so the elastomer film is transparent to visible light. To realize the detection of the deformation of the sensing skin, we construct a near-infrared light field inside, in which case the infrared light inside is stronger than that outside, and thus the film becomes opaque to infrared light. Therefore, proximity sensing is achieved through visible light where the sensing skin looks transparent to visible light from inside, and the contact information of the sensing skin is achieved through near-infrared light where the sensing skin looks opaque to near-infrared light from inside. In this way, our sensor can work on two sensing modes of proximity and contact simultaneously, without the need to switch lights to change the sensing mode.

A. Sensing Skin

The sensing skin is the core of the visuotactile sensors. Most of the sensing skins [6], [38] are made of acrylic as the support layer and elastic silicone as the deformation layer. Although this structure has a large force-sensing range, it is difficult to obtain contour information on objects with large surface gradients. In this paper, we utilize an elastic film as the sensing layer and replace the support layer by inflating the film. To guarantee that the internal air pressure of the sensor is greater than the external, we accomplished two parts of the work: Firstly, we sealed the sensor, which not only reduces the leakage of internal gas but also realizes a good waterproof effect. Secondly, we also inflate the sensor through the pressure stabilizing valve to ensure the stability of the internal air pressure. This scheme has three advantages:

- Better perception of the contour information of the object can be achieved. In addition to directly obtaining the texture and shape of the object in contact, the stickiness information of the object can also be analyzed through the process of separating from the object in contact.
- No additional transparent support layer is required. The wavelength detection range is 5.5$\sim$14 μm for the mid-infrared temperature sensor, but both acrylic and ordinary glass can not pass through light with a wavelength above 5 μm, and special optics are not only expensive but also less durable, which is not suitable for large-scale commercialization. The structure of this elastic inflatable film can skillfully avoid this problem.
- Faster thermal response is achieved. Since there is no need for an acrylic support, the film structure has a faster thermal response. As we tested, the temperature of the film can vary from 30 ^{o}C to 75 ^{o}C in 820 ms, which indicates a change rate of 54 ^{o}C/s.

To achieve the switch between proximity and contact sensing, we would like to have an elastic film that has a band-pass filter-like effect. However, designing a highly elastic, abrasion-resistant optical narrow-band filter film is generally very difficult because expansion, stretching, and abrasion of the elastomer can change the optical properties of the film. To achieve stable and reliable optical filtering, we create a silver film with high elasticity, durability, and stability. The light transmittance of this film is related to the difference in light intensity between the two sides of the film, with the brighter side of the film showing stronger reflective properties and the darker side of the film showing stronger transmissive properties. Based on this physical phenomenon, we only need to control the light intensity of different wavelengths of light on the inside and outside of the film to realize the filtering effect of different wavelengths of light.

Most previous studies [23], [49] on unidirectional perspective coatings used silicone materials. To ensure the film's elasticity, toughness, and wear-resistant performance under the premise of being as thin as possible, we used latex as the main material. This material can achieve the same thickness as silica gel and has better elasticity and toughness, but the fabrication process is also relatively complex. We cooperate with a latex production factory to produce this special film. We control the transmittance of the film by adjusting the proportion of silver powder added and the thickness of the elastic film. When the light intensity on both sides of the film reaches a certain gap, the effect of unidirectional visibility can be realized. With the help of this film, we can realize reversed unidirectional properties for visible and infrared light by controlling the brightness of light of different wavelengths.

B. Imaging System

To realize multifunctional tactile sensing, we design a sensing system which simultaneously acquires optical information in different wavelength bands.

To reduce the cost of the sensor, we use a charge-coupled device (CCD) as the main imaging device. The wavelength imaging range of the CCD is 350∼1000 nm, which includes the UV band below 400 nm, the visible band of 400∼700 nm, and the near-infrared band of 700∼1000 nm. However, prolonged exposure to UV light not only harms the human body, but also limits its large-scale popularization due to the high cost of UV narrow-band filters, so we decided not to apply UV light in M^{3}Tac.

Besides visible light information, light also contains temperature information. Due to the existence of internal thermal motion, an object will constantly radiate electromagnetic waves in all directions, which are the infrared waves with a band located between 0.75 and 100 μm [50]. The temperature range of most objects in life is mostly in the -5∼110 °C range [51], and the MLX90640 can detect temperatures in the range of -40 to 300 °C, which can meet the demands in most scenarios of life. Therefore, we add mid-infrared light into the imaging module to realize temperature sensing. Since longer wavelength light will require more specialized instruments for detection, which is not only costly but also large and not easy to install and deploy on tactile sensors. Therefore, we finally adopt the light of three bands in M^{3}Tac: visible light, near-infrared light, and mid-infrared light.

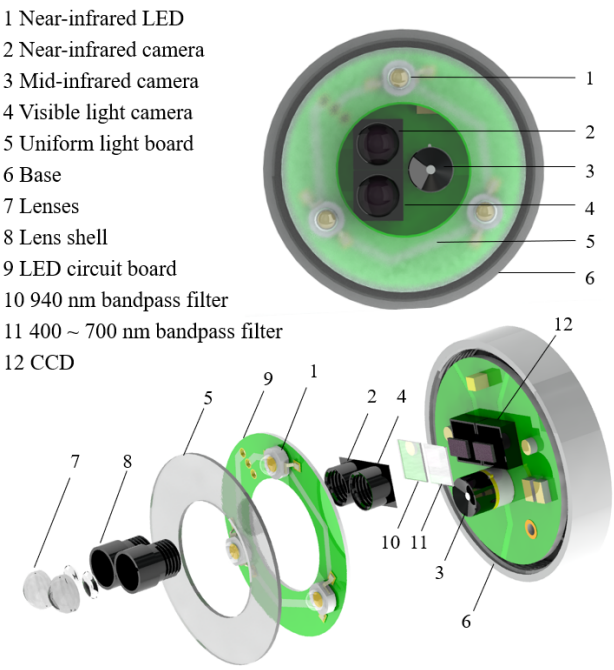

Fig. 3. Internal structure of the imaging system. Top: Structure diagram of the imaging system; Bottom: Exploded view of the imaging system.

As shown in Fig. 3, the imaging module consists of a visible-light imaging unit, a near-infrared imaging unit, and a mid-infrared imaging unit. First, the visible-light imaging unit consists of a CCD sensor, a 400∼700 nm bandpass filter, and a lens. The near-infrared detection unit consists of a CCD sensor, a 930∼950 nm bandpass filter, and a lens. Band-pass filters can eliminate the effect of other bands of light on detection. MLX90640 is used as the mid-infrared imaging unit, which can achieve wavelength detection between 5.5∼14 μm. It has characteristics of low cost, small size, and a wide temperature sensing range. The resolution of MLX90640 is 24×32, and the maximum sampling rate can be 1 Mhz, but a high sampling frequency will cause a decrease in its detection accuracy. To achieve a higher detection accuracy, we use the sampling frequency of 4 Hz. In addition, the MLX90640 sensor includes two models, D55 and D110, of which the D110 has a field of view of 110^o, and the D55 has a field of view of 55^o. To obtain a larger detection area, we chose the MLX90640-D110 as our mid-infrared imaging unit.

C. Lighting System

The transmittance of the elastic film we designed is affected by the light intensity, so we can adjust the intensity of different light to control the film's transmittance under different wavelength bands. Firstly, to realize the effect of opacity under near-infrared light, we need to make the intensity of infrared light inside the sensor greater than that outside. Therefore, we installed an array of 940 nm near-infrared LEDs inside the sensor and used a uniform light board to ensure the

uniformity of the internal light. Secondly, to realize the effect of transparency of the elastic film under visible light, we need to ensure the intensity of visible light inside the sensor is weaker than that outside the sensor. Since the entire sensor is a sealed structure, there is no other light source inside the sensor except for the 940 nm light. So in a bright environment, the intensity of visible light outside of the sensor is greater than that inside. To further guarantee that the sensor works even in dark conditions, we install an LED ring on the outside edge of the sensor to ensure that the intensity of visible light on the outside of the sensor is always greater than that on the inside, as shown in Fig. 2(b).

Through continuous optimization and improvement of the sensing skin, imaging system, and lighting system, the sensors can work indoors, outdoors, and in dark conditions. To better demonstrate this, we adopt a spectrometer to detect the light intensity in sunlight, indoors, darkness, and the inside of the sensor, the results are shown in Fig. 4. From the results, we can see that the light intensity inside the sensor peaks at 940 nm, with an intensity of 4800 dn[1], and the intensity of the other bands is almost 0. The light intensity in sunlight is mainly concentrated in the range of 450∼750 nm, with an intensity of 1000∼1800 dn, the light in indoor environments is mainly concentrated in the range of 450∼650 nm, with an intensity of 20∼60 dn, and the intensity of each band under dark conditions is almost 0. Therefore, both in sunlight and indoor conditions, the sensor can meet the visible light intensity inside is less than outside, and the infrared light intensity inside is stronger than outside. To adapt the sensor to dark conditions, we also mounted LEDs on the edge of the sensor's sensing skin, which allows the sensor to work in dark conditions.

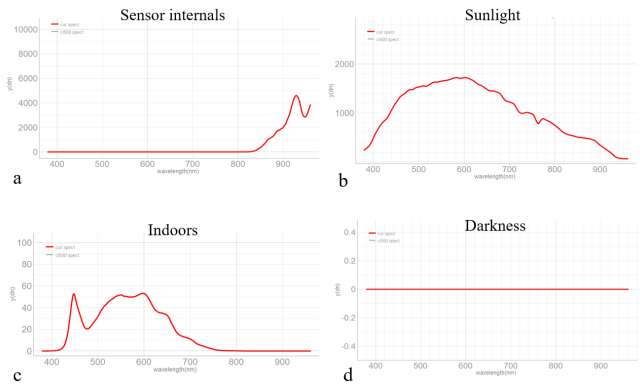

Fig. 4. Spectrograms of different scenes. (a) Sensor internal; (b) In sunlight; (c) Indoor; (d) In the dark.

IV. Algorithm Design

M^{3}Tac can not only realize the functions of contact force sensing, texture sensing, and 3D reconstruction that traditional visuotactile sensors have but also realize proximity sensing and high-resolution, wide-range temperature sensing. Here, we will introduce the sensing algorithms one by one.

[1]Digital Number (dn) is a raw, uncalibrated digital value that represents the brightness of a pixel in a remote sensing image.

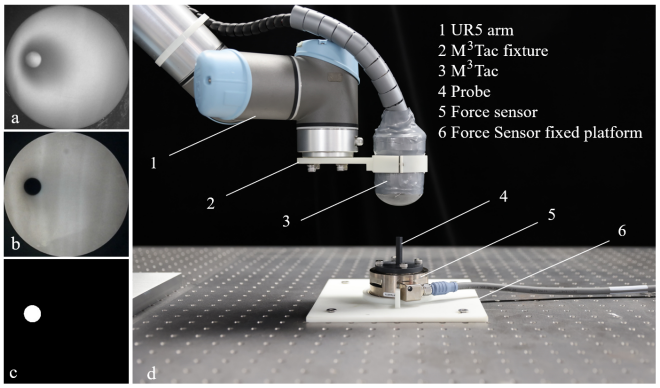

Fig. 5. Pixel-level automated acquisition & annotation platform. (a) Images captured by the near-infrared camera; (b) Images captured by the visible-light camera; (c) The mask of the contact position is obtained from threshold processing of the visible-light image.

A. Contact Force Sensing (calibration: near-infrared & visible; sensing: near-infrared)

Compared to data-level force sensing, pixel-level force sensing has a higher resolution but also requires more data. To tackle this, we propose a pixel-level automated acquisition & annotation platform, which can realize contact force sensing using industrial force sensors (ATI Gamma force sensor) as well as pixel-level contact area segmentation with the help of multispectral imaging. To obtain the force distribution of each pixel at the contact area, we establish a finite element model of the inflatable elastomer and propose a pixel-level force sensing network based on a swin transformer [52].

1) Pixel-level automated acquisition & annotation platform: M^{3}Tac can acquire both visible-light and near-infrared images. As shown in Fig. 5(a), the near-infrared image is related to the deformation of the elastic film only, which can be adopted for accurate and stable contact force sensing. The visible-light image is related to the contact position information. The elastic film is transparent in visible light, and we can see from Fig. 5(b) that the contact position between the film and the probe will be obviously darkened in the visible-light image. So we can quickly and efficiently get the mask information of the contact position through the color threshold processing algorithm, as shown in Fig. 5(c). Considering that the mask will have an offset error when the distance of the probe from the lens changes, we apply displacement compensation to get accurate annotation information.

The purpose of displacement compensation is to minimize the deviation between visible and near-infrared images, which mainly consists of four parts: intrinsic parameters calibration, image cropping, perspective transformation, and sampling fine-tuning, as shown in Fig. 6. First, we perform the intrinsic parameters calibration using the method described in [53], which can reduce the effect of camera lens aberrations on imaging. In the second part, we capture two sample images after calibration. Next, we crop down the circular area on the sensor surface, and we can see from Fig. 6 that the circle is still distorted. To obtain a regular circle, we next correct the circle using the perspective transformation method [54]. To further minimize the error, we will take several samples with different

positions and fine-tune the image by rotating and translating them. Finally, we will get the image after alignment. Keeping the parameters obtained from the above algorithm, batch data acquisition can be performed.

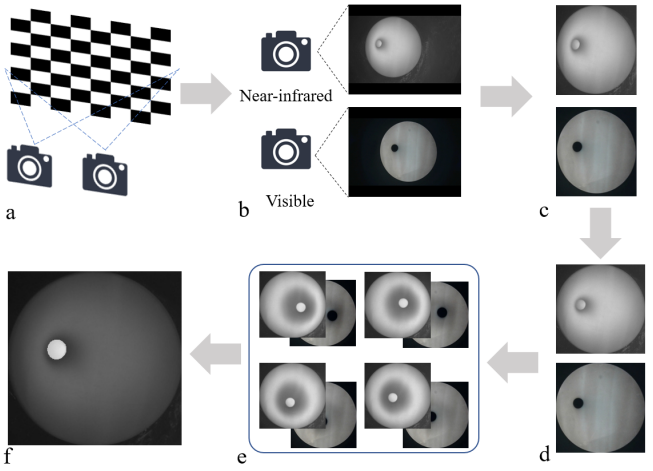

Fig. 6. The displacement compensation algorithm. (a) Intrinsic parameters calibration; (b) Image acquisition; (c) Image cropping; (d) Perspective transformation; (e) Sampling & fine tuning; (f) Recording parameters and testing.

Based on this principle, we propose an automatic pixel-level force sensing data acquisition & annotation system, as shown in Fig. 5(d). Compared with other annotation systems, the proposed annotation system can not only automatically obtain the contact force information, but also obtain the contact position information, which greatly reduces the cost of acquiring the dataset. In the system, black probes are used. This is because compared to the white color, the black probe will be more visible in the visible light imaging, as shown in Fig. 7, and near-infrared imaging is independent of the color of the contact object. Since the force sensing algorithm we proposed only needs to acquire near-infrared images (in the experimental stage), the force sensing performance of this sensor is not affected by the color of the object. Based on this annotation system, we can obtain more than 50,000 pixel-level annotation data in one day, which plays a significant role in promoting the industrialization of visuotactile sensors.

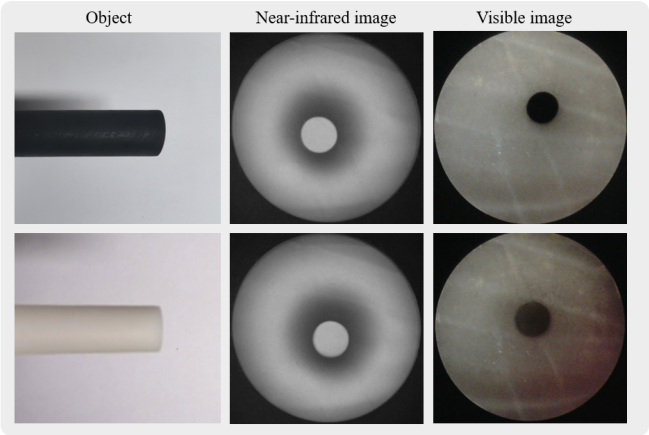

Fig. 7. Detection results of cameras of different bands.

2) Finite element analysis: Although calibration using industrial force sensors can obtain an accurate total force, it is not possible to obtain the force at each pixel of the contact area. To estimate the contact force at each pixel as accurately as possible, we model the elastic inflatable film using finite element analysis to simulate the non-linear elastic behavior of interaction between the membrane and the force sensor. We apply COMSOL multi-physics simulator [55] to conduct finite element analysis, which is a powerful tool to simulate the behavior of various physical phenomena. Specifically, we utilize the Ogden material model to replicate the real-world, non-linear elastic behavior of the membrane in contact with the sensor. The Ogden material model [56] is a common method for large-strain contact condition analysis of biomaterials, soft tissues, and rubbers, which shows excellent performance for large deformations compared to traditional models such as Mooney-Rivlin or Neo-Hookean [57].

The strain energy density function in the Ogden model is defined as

$$\omega = \sum_{i=1}^{3} \frac{\mu_i}{\alpha_i}(\lambda_i^{\alpha_i} - 1), \tag{1}$$

where ω stands for the strain energy, $\{\mu_i\}_{i=1}^3$ are the material parameters, $\{\alpha_i\}_{i=1}^3$ are dimensionless material parameters, and $\{\lambda_i\}_{i=1}^3$ denote the principal stretches. In COMSOL, the strain energy function of the Ogden model can be directly incorporated into the Material node. Regarding the establishment of material constants, we referred to classical Ogden parameters for natural rubber, and the detailed calculation process is outlined in in [58], [59]. Once the model is set up, COMSOL utilizes the FEM to numerically solve the underlying equations, thereby allowing us to observe the stress-strain behavior under different scenarios. The parameter of Ogden is set as $\mu_1 = 0.63$ MPa, $\mu_2 = 1.2$ kPa, $\mu_3 = -0.01$ MPa, $\alpha_1 = 1.3$, $\alpha_2 = 5$, and $\alpha_3 = -2$. The membranes are almost incompressible, so the bulk modulus [60] is set to $\kappa = 10$ kPa. Further, to simulate the pressure from air compressed the membranes, the pressure-density relation is leveraged, assuming that the confined air is adiabatic. Therefore, the pressure of confined air under deformation can be written as

$$P = (\rho/\rho_0)^\gamma P_0 = (V_0/V)^\gamma P_0, \tag{2}$$

where P, P_0, ρ, ρ_0, V, and V_0 denote air pressure, density, and volume after and before deformation, respectively. The constant γ is the heat capacity ratio set as $\gamma = 1.4$. Given the calculated P, the load on the inner side of membranes will be

$$\Delta P = P - P_0 = P_0\left((V_0/V)^\gamma - 1\right). \tag{3}$$

Measured by professional instruments, we know that the air pressure inside the M^{3}Tac is 5 kPa above standard air pressure, and the thickness of the elastic film is 0.2 mm. In the simulation, we set the original air pressure inside membranes as $P_0 = 0.106$ MPa, which denotes extra pressure of 5 kPa added to standard air pressure. To facilitate the analysis and calculations, we make two assumptions here: 1) The boundary effect of the inflatable elastomer is ignored. 2) Like previous work [61]–[63], we ignore the effect of gravity on the perception algorithm in our experiments.

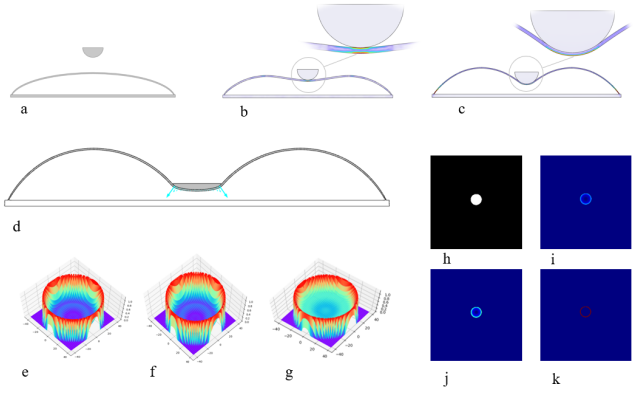

Fig. 8. Finite element analysis. (a) The initial state; (b) When contact occurs without sealing; (c) When contact occurs with sealing; (d) A cylinder of 7.5 mm diameter and 1.5 mm height in contact with an inflatable film; (e) (f) (g) Distribution of forces in the X, Y, and Z directions; (h) The mask of the contact position; (i)(j)(k) Distribution of forces in the X, Y, and Z directions after combination with the mask.

To simplify the calculation, we study a slice of an original 3D model and ensure the equivalence of the study through geometric symmetry, the results obtained are shown in Fig. 8. From Fig. 8(a)(b)(c)(d), we can see that the contact force occurs at the center when the interior is not inflated, while in the inflated state, the contact force is mainly concentrated at the edges. By decomposing the contact force, we can get the contact force information of the contact position in X, Y, and Z directions, and ensure the consistency of each dimension by normalization, the results obtained are shown in Fig. 8(e)(f)(g)(i)(j)(k).

3) Contact force sensing algorithm: In recent years, a large number of transformer structures [64]–[66] have been used for tasks like image segmentation and depth estimation due to their excellent performance. Among them, the swin transformer [52] stands out due to its local self-attention layers, hierarchical feature mapping, and lower computational complexity, making it a universal backbone in the field of computer vision. In the task of force estimation, where the prediction of dense force distribution is required, the hierarchical structure of the swin transformer is particularly advantageous.

Additionally, the U-Net structure [67] is commonly employed in various medical imaging tasks. Its skip connections effectively facilitate local-global semantic feature learning. For force distribution maps represented after finite element analysis, local features aid in learning the surface variations of the force map, while global features help learn the magnitude of the force values. To combine both strengths, we employ the transformer-based U-shaped encoder-decoder architecture proposed in [68]. This architecture has been verified to yield the best results in our experiments.

To obtain a more accurate distribution of contact force, we combine the results of finite element analysis with the contact force and propose a new force-sensing network, FSwint-MAP, as shown in Fig. 9. The input of the network is a 224×224 near-infrared image.

The force distribution map effectively represents the forces at different positions in the contact area between the probe and the film. The peak force map not only indicates the magnitude of the total contact force, but also directly outputs the forces in three separate X, Y, and Z directions, representing the magnitude of the force and the contour of the force-receiving area. We use the smooth L1 loss for these two different outputs, with the loss function $\mathscr{L}$ being:

$$\mathscr{L} = \beta_1 \mathscr{L}_{\mathrm{PFM}} + \beta_2 \mathscr{L}_{\mathrm{FDM}}, \tag{4}$$

where $\mathscr{L}_{\mathrm{PFM}}$ denotes the loss of the peak force map, and $\mathscr{L}_{\mathrm{FDM}}$ denotes the loss of the force distribution map. β_1 and β_2 are adjustable parameters. During the training process, we set β_1 and β_2 to 1. Without using a pre-trained model, we input the near-infrared images directly into the network for training and use the peak force map and force distribution map as labels. FSwint-MAP is trained using an Nvidia 3090 GPU and an AMD 3990X CPU, with a batch size set to 72. The model is trained for 35 epochs with a learning rate of 0.0001.

After training, we will post-processed the output. The output of the neural network is the contact area *Seg* between the probe and the sensed skin, the total force F_k, and the distribution of the force M_k, $k \in \{x, y, z\}$, k denotes the component of the force in the x,y,z directions. After obtaining Seg, M_k, F_k through neural networks, we first compute the total force T_k, for all pixels (i, j) in M_k:

$$T_k = \sum_{i,j} M_k(i, j), (i, j) \in Seg. \tag{5}$$

Finally, we divide the total force by the sum of the individual pixel points by multiplying by the value of each pixel point to get the distribution of the force at each pixel $Res_k(i, j)$:

$$Res_k(i, j) = F_k / T_k \times M_k(i, j). \tag{6}$$

B. Proximity Perception (visible light)

Although robots can obtain stable and reliable contact force information through contact sensing, proximity sensing is also important when performing the grasping of some fragile objects. To obtain the distance between the object and the sensor, we propose a proximity perception method.

As shown in Fig. 10(a), since the film looks transparent for visible light, the visible-light camera can see a shadow of the object on the film. As the distance from the object to the sensor changes, the sharpness of the image detected by the visible-light camera through the film changes. The closer the object is to the sensor, the higher the sharpness it produces, and when in complete contact the shadow is almost purely black. Based on this principle, we propose a proximity perception algorithm, as shown in Fig. 10(b). We first remove the influence of background on proximity perception by the background difference method and then reduce the influence of noise by the filtering algorithm. Since we use the background difference method, the influence of light will produce some noise. To better extract effective information, we introduce a noise threshold and invert these invalid regions.

For M^3Tac, it is difficult for us to obtain information about the depth of an object in space as in a depth camera, so we use the closest distance between the object and the tactile sensor as the detection value. We take the minimum

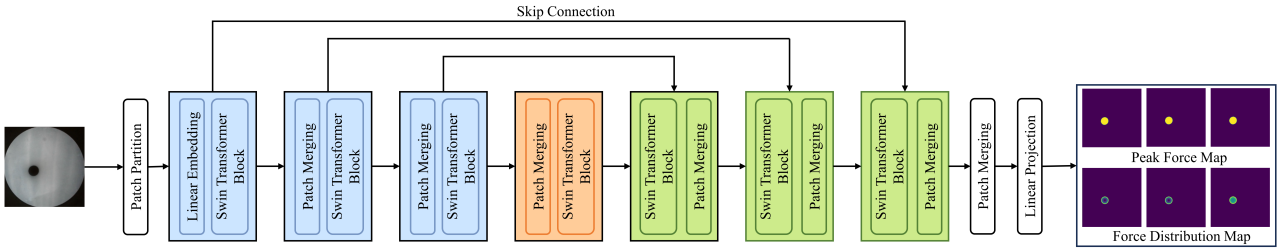

Fig. 9. Contact force sensing algorithm: FSwint-MAP.

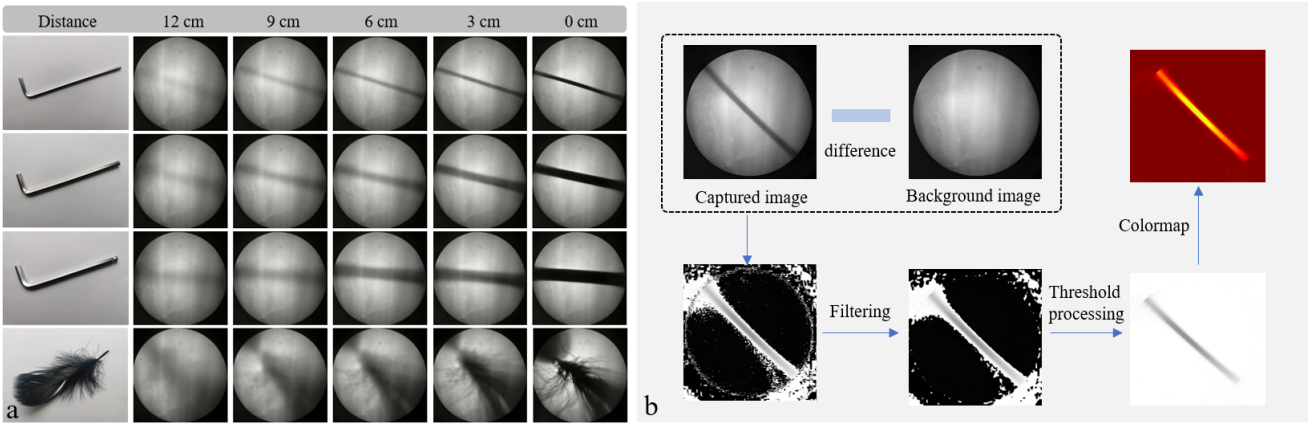

Fig. 10. Proximity perception. (a) Proximity test of objects at different distances; (b) Proximity sensing algorithm.

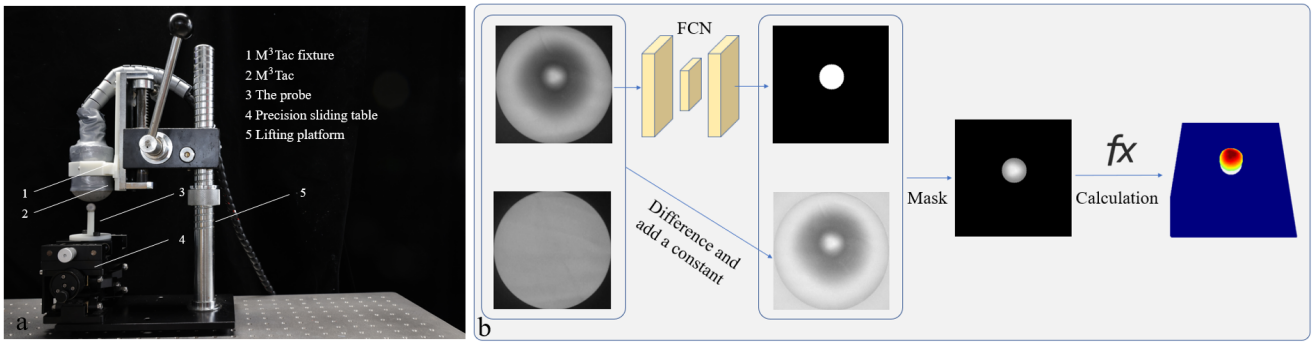

Fig. 11. 3D reconstruction. (a) 3D calibration platform; (b) 3D reconstruction algorithm.

intensity value of the pixel after processing and establish an equation between this value and the distance, through which the distance between the object and the sensor can be estimated. Finally, for better visualization, we map the obtained distance to a colormap.

C. 3D Reconstruction (near-infrared)

3D reconstruction is one of the most important functions of visuotactile sensors. The current 3D reconstruction technology mainly uses the photometric stereo method, which calculates the normals based on the brightness of different colored rays on the sensor surface. Although the photometric stereo method has a high degree of versatility, its accuracy is greatly affected by the position and uniformity of the light. Generally speaking, the photometric stereo method requires the use of three colors of lights to ensure the existence of a certain angle difference between the lights of different colors, which puts forward high requirements on the design of the sensor. In particular, the proposed sensor M^{3}Tac employs an inflatable elastic film, on whose surface it is very difficult to realize the illumination of multiple lights at a certain angle.

Thanks to the sealed structure of the sensor and the mounting of the uniform light board, uniform illumination of near-infrared light can be realized on the surface of the sensor, and the lighting intensity is related to the distance of the sensing skin from the uniform light board. When the sensor is in contact with an object, the area closer to the light is more luminous, and using this luminance information the reconstruction of luminance information on the sensor surface can be realized. However, in addition to the contact area, its boundary area will appear as a shadow due to the occlusion of the contact area. Since we only use one type of light,

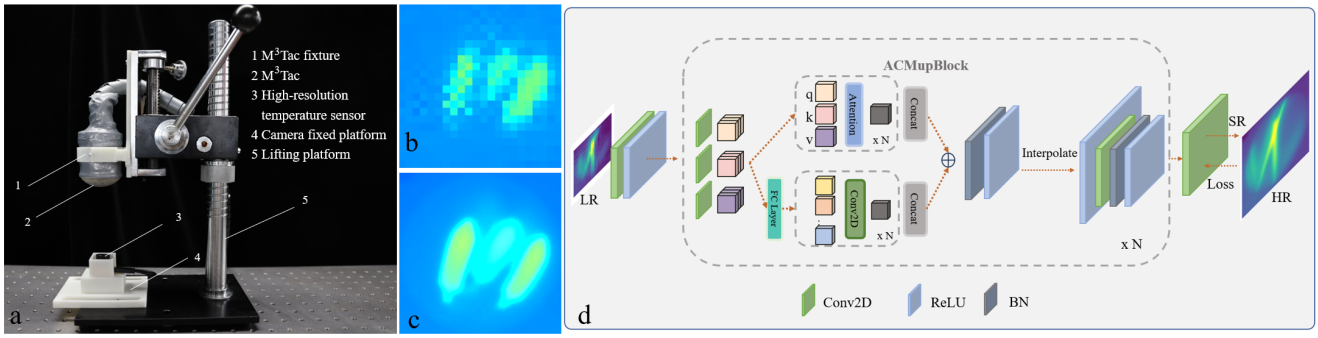

Fig. 12. Super-resolution temperature perception. (a) Temperature data acquisition platform; (b) Low-resolution temperature image; (c) High-resolution temperature image; (d) Temperature super-resolution algorithm, which employs the ACmix organically fusing convolutional neural network and self-attention mechanism and progressively realizes super-resolution.

it is difficult to separate the shadow part. To exclude the interference of the shadow part on the reconstruction, we propose a two-step reconstruction scheme, which first obtains the information of the contact region through the contact area segmentation network, and then realizes the reconstruction of the contact region based on the brightness. This approach not only avoids the influence of shadows on the 3D reconstruction, but also accurately obtains the contact information, and its realization process is shown in Fig. 11(b).

To obtain the relationship between luminance and depth, we design a calibration system, as shown in Fig. 11(a). Unlike the photometric stereo method, we do not need to solve for the normal directions corresponding to different pixels, we only need to establish an equation for luminance versus depth by fitting the equation, i.e.,

$$D(i,j) = \mathfrak{F}\left(B(i,j)\right), \tag{7}$$

where $B(i,j)$ and $D(i,j)$ denote the brightness and depth of the (i,j) pixel in the image, and $\mathfrak{F}(\cdot)$ is the fitting function obtained by calibrating the image.

D. Temperature Sensing (mid-infrared)

High-resolution temperature information acquisition is very challenging. The traditional method of using thermocouple arrays makes it difficult to achieve high resolution, while infrared thermometry provides a new way to achieve high-resolution temperature detection.

The high-resolution mid-infrared temperature measurement device is costly. For a device with a resolution of 100×100 above, the price is often more than 200 US dollars, which limits the mass production of the sensor. Besides, the higher the resolution, the larger its volume tends to be, which also prevents the deployment of tactile sensors with limited space inside. To achieve low cost and miniaturization, we used MLX90640 as the temperature sensing unit, which can achieve a resolution of 24×32 and a field of view of 155°. In addition, to achieve higher-resolution temperature sensing, we propose a temperature data calibration platform along with lightweight super-resolution algorithms for mid-infrared temperature images.

1) Temperature data acquisition platform: Super-resolution algorithms can be categorized into two types: traditional methods and deep learning methods [69], [70]. Traditional methods can improve the resolution of the image without using a referenced high-resolution image, but the accuracy is poor. Deep learning methods require high-resolution reference data but usually achieve better performance. We choose to apply the deep learning method. To obtain high-resolution reference images, we build a temperature data acquisition platform as shown in Fig. 12(a)(b)(c), consisting of M^{3}Tac, a holder, and a high-resolution temperature sensor (InfiRay T2S+, resolution: 192×256). Since M^{3}Tac utilizes an ultra-thin elastic film as the sensing skin, the inner and outer surfaces of the sensing skin almost have the same temperature. After contacting the sensor surface with objects of different temperatures, we remove the object and the temperature of the film will be maintained for a short time. At this moment, we obtain paired temperature images, one externally with the high-resolution temperature sensor and one internally with the low-resolution sensor. We collected a total of 800 groups of data and each group contains one high-resolution image and one low-resolution image.

2) Super-resolution algorithms: To achieve higher-resolution temperature sensing using low-resolution sensors, we design a lightweight super-resolution algorithm for thermal images, named PACmixSR (Progressive-ACmix Super Resolution), as shown in Fig. 12(d). The algorithm employs the ACmix [71] organically fusing convolutional neural network and self-attention mechanism and progressively realizes super-resolution. Temperature sensing is one of the basic functions of M^{3}Tac, based on which we can do a lot of extended tasks, and a lightweight super-resolution network can save more computational resources for subsequent tasks. The overall model parameters only count 0.71 MB, achieving a balance between lightweight and performance.

The self-attention mechanism is decomposed into the generation of query, keys, values, and feature aggregation phases. Let $\mathbf{F}\in\mathbb{R}^{C_{\text{in}}\times H\times M}$, $\mathbf{G}\in\mathbb{R}^{C_{\text{out}}\times H\times M}$ denote the input and output features, and H, M represent the height and width of the image, respectively. $\mathbf{f}_{i,j}\in\mathbb{R}^{C_{\text{in}}}$, $\mathbf{g}_{i,j}\in\mathbb{R}^{C_{\text{out}}}$ represent the corresponding tensors of pixel (i,j), the attention module computation can be defined as:

$$\mathbf{q}_{i,j}^{(n)} = \mathbf{W}_{\text{q}}^{n}\mathbf{f}_{i,j}, \quad \mathbf{k}_{i,j}^{(n)} = \mathbf{W}_{\text{k}}^{(n)}\mathbf{f}_{i,j}, \quad \mathbf{v}_{i,j}^{(n)} = \mathbf{W}_{\text{v}}^{(n)}\mathbf{f}_{i,j}, \tag{8}$$

$$\mathbf{g}^{\mathrm{s}}_{i,j} = \bigvee_{n=1}^{N} \left(\sum_{(a,b)\in N_c(i,j)} \mathfrak{A}\left(\mathbf{q}^{(n)}_{i,j}, \mathbf{k}^{(n)}_{a,b}\right) \mathbf{v}^{(n)}_{a,b} \right), \tag{9}$$

where $\bigvee_{n=1}^{N}$ represents the concatenation of the outputs of N self-attention heads, and $\mathbf{g}^{\mathrm{s}}_{i,j}$ is the self-attention part of outputs. $\mathfrak{A}(\cdot,\cdot)$ represents the attention weight function. $\mathbf{W}^{(n)}_{\mathrm{q}}, \mathbf{W}^{(n)}_{\mathrm{k}}, \mathbf{W}^{(n)}_{\mathrm{v}}$ are the weight matrices of query (q), keys (k), and values (v). Let R represent kernel size, $N_R(i,j)$ is a local region centered around (i,j) with margin width of R.

The convolutional layer can be similarly decomposed into two phases for generating feature maps of the corresponding pixels of the convolutional kernel and mixing them up by using the Shift operator. However, the first phase occupies the major amount of computation for both modules. Therefore, the same first stage can be fused to simplify the computation. The computation of the convolutional layer can be defined as:

$$\mathbf{p}^{(n)}_{i,j} = \bigvee \left(\mathbf{q}^{(n)}_{i,j}, \mathbf{k}^{(n)}_{i,j}, \mathbf{v}^{(n)}_{i,j} \right), \tag{10}$$

$$\mathbf{K}^{R^2}_{i,j} = \mathfrak{FC}\left(\bigvee_{n=1}^{N} \mathbf{p}^{(n)}_{i,j} \right), \tag{11}$$

$$\mathbf{g}^{\mathrm{c}}_{i,j} = \sum_{y=0}^{R-1} \sum_{x=0}^{R-1} \mathfrak{Shift}\left(\mathbf{K}^{R\times y+x+1}_{i,j}, y-\lfloor R/2 \rfloor, x-\lfloor R/2 \rfloor \right), \tag{12}$$

$$\mathfrak{Shift}(\mathbf{f}_{i,j}, \Delta x, \Delta y) = \mathbf{f}_{i-\Delta x, j-\Delta y}, \tag{13}$$

where $\mathbf{p}^{(n)}_{i,j}$ is the concatenation of $\mathbf{q}^{(n)}_{i,j}, \mathbf{k}^{(n)}_{i,j}, \mathbf{v}^{(n)}_{i,j}$ in channel dimension, and $\mathfrak{FC}(\cdot)$ denotes kernel feature extractor composed of fully connected neural network. The final output is the combination of the features of self-attention layer $\mathbf{g}^{\mathrm{s}}_{i,j}$ and convolutional layer $\mathbf{g}^{\mathrm{c}}_{i,j}$ with weights η_1, η_2:

$$\mathbf{g}_{i,j} = \eta_1 \mathbf{g}^{\mathrm{s}}_{i,j} + \eta_2 \mathbf{g}^{\mathrm{c}}_{i,j}. \tag{14}$$

Comprising $\log_2(t) - 1$ ACmixResidualBlocks and an Upsample module, where t donates the scale factor, the overall network architecture of PACmixSR makes a great process in being lightweight. Additionally, by adopting a progressive super-resolution strategy, it achieves more stable performance. The combination of ACmixResidualBlock and the progressive strategy enables PACmixSR to strike a balance between lightweight design and performance efficiency.

E. Multimodal Classification (near-infrared + mid-infrared)

In a home or laboratory scenario, we often need to manipulate objects with different textures, shapes, and temperatures, and sometimes the object's excessively high temperature or cold temperature can easily damage human skin. And for some transparent containers in life, such as cups and bottles, it is difficult to detect and classify them using only a camera [72]. The use of M^{3}Tac not only allows us to obtain information about the temperature of the object at the time of contact but also allows us to realize the classification of objects during the grasping process. To realize this function, we design a multimodal classification network, where temperature and texture images are concatenated together using GoogleNet [73]. The architecture of this network is shown in Fig. 13.

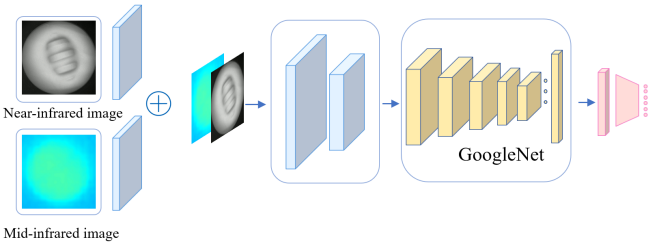

Fig. 13. Multimodal classification network.

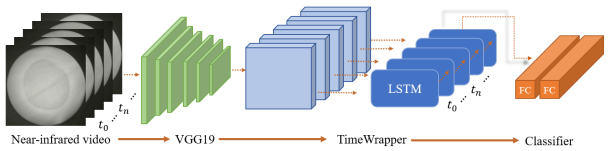

Fig. 14. Stickiness recognition network.

F. Stickiness Recognition (near-infrared)

In addition to temperature, stickiness is also a common property of objects. Human beings [74] can also determine the stickiness of an object by sensing the deformation of the skin during the contact-separation process which [75] is one of the ways to measure the stickiness of adhesive tapes. Inspired by this process, we record the contact and separation process of M^{3}Tac with the object, which can realize the stickiness classification of the object. We use VGGNet as a feature encoder and two-layer LSTM to capture temporal relationships. The network framework is shown in Fig. 14. Considering the limit of computing resources, TimeWrapper [76] is added to balance the usage of GPU memory and inference time.

V. Experiments

To test the performance of M^{3}Tac, we conduct contact force sensing experiments (**Exp. 1**), super-resolution temperature sensing experiments (**Exp. 2**), 3D reconstruction experiments (**Exp. 3**), stickiness classification experiments (**Exp. 4**), liquid temperature classification experiments (**Exp. 5**), and the proximity sensing experiment (**Exp. 6**), which verified that the sensor has excellent performance in contact force sensing, temperature sensing, deformation detection, multimodal classification and so on. In addition, to test the application potential of the sensor in real scenarios, we also conducted fragile object grasping experiments (**Exp. 7**), circuit board heating location detection experiments (**Exp. 8**), underwater pipeline abnormal hot spot detection experiments (**Exp. 9**). These experiments illustrate the high application value of M^{3}Tac.

A. Exp. 1: Contact Force Sensing

To test the precision of force perception, we employ an automatic acquisition & annotation platform as shown in Fig. 5, and collect data using six different sizes of probes depicted in Fig. 15. We gathered a total of 48,000 data, with 8,000 data collected for each type of probe, and 80% of these data are utilized as the training dataset. We use near-infrared

TABLE III
COMPARISON OF FORCE PREDICTION ACCURACY

Method	X-peak↓	Y-peak↓	Z-peak↓	Mean↓	X-map			Y-map			Z-map		
					1.05↑	1.10↑	1.25↑	1.05↑	1.10↑	1.25↑	1.05↑	1.10↑	1.25↑
ResNet-FCN [31]	0.031	0.039	0.300	0.123	0.245	0.424	0.685	0.187	0.342	0.609	0.297	0.484	0.741
FSwint-XYZ (Ours)	0.023	**0.015**	0.082	0.040	0.745	0.871	0.946	0.473	0.765	0.931	**0.571**	0.744	0.895
FSwint-MAP (Ours)	**0.017**	**0.015**	**0.035**	**0.023**	**0.754**	**0.879**	**0.954**	**0.553**	**0.792**	**0.939**	0.548	**0.745**	**0.899**

images as input and use the processed visible-light images as the mask. The proposed FSwint-MAP network employs a transformer-based U-shaped encoder-decoder architecture and utilizes skip connections for local-global semantic feature learning, which can effectively extract the force characteristics within the images.

To validate the effectiveness of our network and output scheme, we compare it with two classic schemes as baselines. One is the FSwint-XYZ, which utilizes the transformer architecture, and the peak force map is no longer outputted. Instead, it directly outputs the individual force values in the X, Y, and Z directions, which are adopted by the latest force-sensing algorithm framework [36]. Due to the lack of open-source code, we reproduce and modify the network based on the one they provided. The other baseline is to change the Swin-Unet [68] structure to the ResNet-FCN structure.

In the experiments, we adopt the δ_τ metric for the outputted force distribution map, a common measure in the field of depth estimation, representing the percentage of pixels within the error range of τ. The error range τ is defined as follows:

$$\max\left(f/\hat{f}, \hat{f}/f\right) < \tau, \tag{15}$$

where f and $\hat{f}$ represent the force distribution map for the ground truth and prediction, respectively. δ_τ denotes the percentage of pixels within the error range τ. We set the values of τ to 1.05, 1.10, and 1.25, respectively. According to the results in Table III, it can be seen that ResNet-FCN [31] has a significant gap compared to FSwint-XYZ and our proposed FSwint-MAP. This is primarily because the transformer architecture has an advantage over the ResNet architecture [77] in dense prediction. FSwint-XYZ, having the same main network structure as our FSwint-MAP scheme, therefore achieves similar results in the evaluation of the force distribution map as our proposed FSwint-MAP network. To evaluate the performance of our network and FSwint-XYZ in predicting peak force, we convert the peak force map into forces in the X, Y, and Z directions. The conversion process involves taking the average force within all mask areas and calculating the average force prediction error in the test set.

As can be seen from Table III, FSwint-MAP performs better compared to FSwint-XYZ. This is because we perform dense force estimation in FSwint-MAP, ensuring that even if the prediction is not good at certain points, the error will be averaged out by other points, so FSwint-MAP converges more easily. Besides, by predicting the peak force map, the force's area of effect and center can be more easily discerned. It also allows for the prediction of multiple contact points, an aspect that is lacking when predicting individual forces in the X, Y, and Z directions. The performance of ResNet-FCN is relatively poor, indicating that the transformer architecture has an advantage over the ResNet architecture in dense prediction tasks.Besides, we have open-sourced our code, data and training weights[2] & https://cloud.tsinghua.edu.cn/d/ with a detailed description of the force capability.

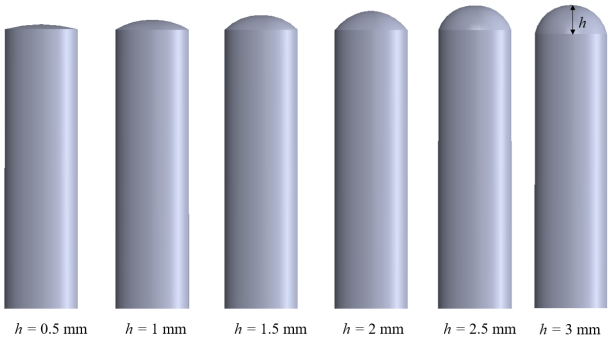

Fig. 15. Different sizes of probes used for force testing.

B. Exp. 2: Temperature Sensing

To test the temperature detection accuracy of the sensors, we build an experimental platform as shown in Fig. 16(a), which contains the M^{3}Tac sensor, the heating platform and the thermocouple sensor. We adjusted the temperature of the heating platform and compared the error between the thermocouple sensor and M^{3}Tac. Considering the temperature range of the equipment, we take 40 detection points, as shown in Fig. 16(b), and adopt Mean Absolute Error (MAE) as an evaluation metric:

$$MAE = \frac{1}{m}\sum_{i=1}^{m}\left|y_i - \widehat{y_i}\right|, \tag{16}$$

where y_i, $\tilde{y}_i$ and m denote the true value, the predicted value and the number of sample sets.

After testing, the temperature detection accuracy of M^{3}Tac can reach 0.3 ^{o}C. In the temperature range test, the sensor can achieve temperature detection between -20 and 130 ^{o}, the extreme temperature is 140 ^{o}C, and the temperature response speed can reach 54 ^{o}C/s.

To test the performance of the proposed super-resolution algorithm, we acquired 800 groups of data in the range of

[2]The supplementary material links are https://github.com/T-DaVinci/M^{3}Tac/tree/main/M^{3}Tac

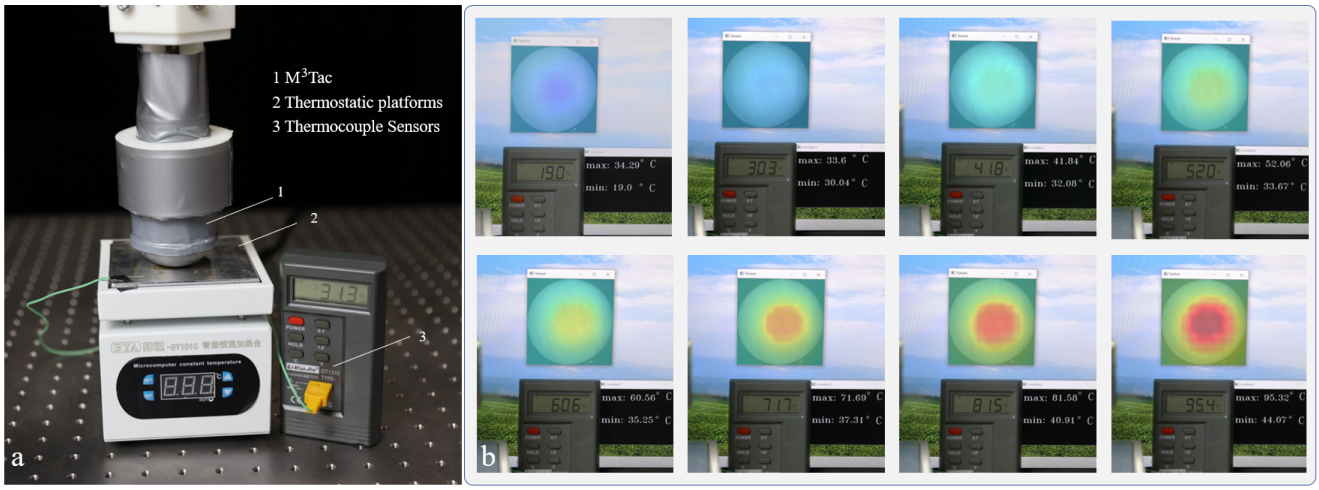

Fig. 16. Temperature sensing accuracy test experiment. (a) Experimental platform; (b) Test results at different temperatures.

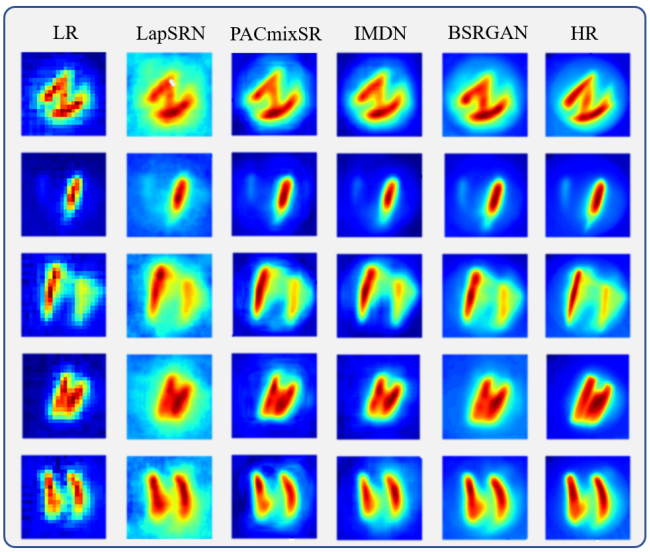

Fig. 17. Comparison of the results of different super-resolution models (LR: low-resolution data; HR: high-resolution data).

TABLE IV
SUPER-RESOLUTION ALGORITHM PERFORMANCE COMPARISON

Method	PSNR↑	SSIM↑	#Params↓	FPS↑
LapSRN [78]	30.82	0.9682	4.98M	416.16
PACmixSR (Ours)	31.50	0.9705	**0.71M**	**445.77**
IMDN [79]	32.13	0.9733	3.44M	337.58
BSRGAN [80]	**36.90**	**0.9846**	63.69M	32.84

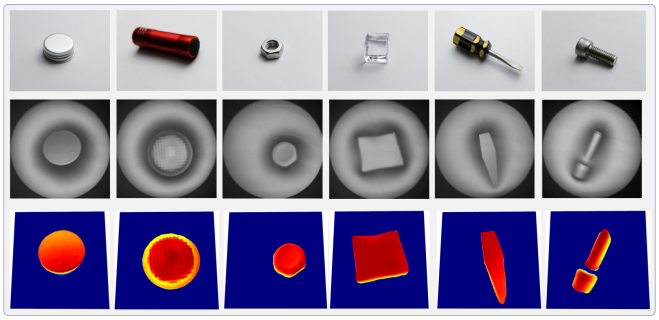

Fig. 18. 3D reconstruction results for different objects. From left to right: Bottle cap, flashlight, nut, cube, screwdriver, screw.

0∼100 °C, and each group contains a high-resolution image and a low-resolution image. To demonstrate the performance of PACmixSR under the guarantee of model lightweight, we compare it with the classical super-resolution models LapSRN [78], IMDN [79], and BSRGAN [80]. All models use the Adam optimizer with a learning rate of 0.004 and MultiStepLR to adjust the learning rate, with a milestone set every 5,000 steps. After cropping and aligning the thermal sensor images, we obtain a low-resolution image size of 22×22 and a high-resolution image size of 176×176. Table IV shows the quantitative comparisons, where PACmixSR achieves the minimum model parameters count of 0.71 MB and the highest inference speed of 445.77 frames per second (FPS). In comparison with LapSRN, it achieves a better super-resolution performance with a smaller memory occupation while using a progressive super-resolution network structure. Although BSRGAN achieves the highest peak signal-to-noise ratio (PSNR) of 36.90 and structural similarity index (SSIM) of 0.9846, the model parameters count is nearly increased to 90 times. Comparison with the lightweight model IMDN reveals that our model parameters are reduced to 21% and the FPS is improved by 32% at a negligible performance sacrifice. Fig. 17 visualizes the results of the model testing. Besides, the function of the camera in the mid-infrared band is to obtain the temperature information on the surface of the sensing skin, the mid-infrared light cannot transmit through the sensing skin, so it is also not affected by the color of the object.

C. Exp. 3: 3D Reconstruction

To test the detection accuracy of the reconstruction algorithm, we utilize a 14 mm diameter sphere for calibration. After calibration, we captured 10 images and obtained the MAE of 0.33 mm, the calculation equation is as follows:

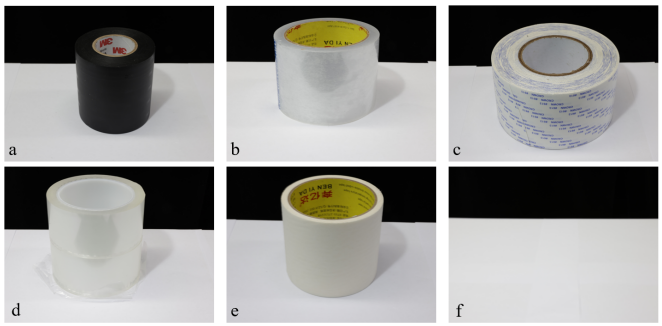

Fig. 19. Tapes used for stickiness testing. (a) Electrical insulating tape; (b) Transparent plastic tape; (c) Nano double-sided tape; (d) Normal double-sided tape; (e) Paper tape; (f) Paper.

$$MAE = \frac{1}{m \times n} \sum_{i=1}^{m} \sum_{j=1}^{n} \left| Seg(x_i, y_j) - \widehat{Seg}(x_i, y_j) \right|, \tag{17}$$

where $Seg(x_i, y_j)$, $\widehat{Seg}(x_i, y_j)$, Seg and $m \times n$ denote the true value, the predicted value, the contact area and the number of sample sets, and $(x_i, y_j) \in Seg$.

In addition, for the selection of fitting algorithms, we compare the least squares and random forest algorithms. The least squares method has a faster reconstruction speed compared to the random forest algorithm.

We perform the reconstruction of objects with different shapes, such as bottle caps, flashlights, nuts, glass cubes, screwdrivers, screws, etc. When performing 3D reconstruction, we first segment the contact area using a neural network and later reconstruct the segmented area using the reconstruction algorithm, so the final display is planar, and the results obtained are shown in Fig. 18.

D. Exp. 4: Stickiness Recognition

To test the effectiveness of the stickiness classification network, we chose five tapes with different stickiness as well as a tabletop for testing, as shown in Fig. 19. We record video data for classification by near-infrared camera when the sensor contacts and separates from the tape. In our experiments, we collected 100 sets of data for each type of tape and 600 arrays in total. To test the detection effectiveness, we divided the training and testing sets by 8:2. After 60 epochs of training, the classification accuracy can reach 98%, which validates the feasibility of M^{3}Tac for object stickiness classification.

E. Exp. 5: Liquid Bottle Classification

Based on the multimodal sensing ability of M^{3}Tac, we can realize not only the texture classification of grasping objects but also the temperature detection of objects. To test the application potential of our proposed sensor for grasping and classification, we design grasping experiments for liquid bottles with different temperatures. The grasping platform is shown in Fig. 20(a)(b)(c), a parallel gripper is designed for the gripping experiment, which is mounted on a Franka arm. One side of the gripper is equipped with an M^{3}Tac sensor and the other side is equipped with an ATI mini45 force sensor.

We choose five bottles with similar diameters and sizes but different textures and fill them with water of different temperatures, as shown in Fig. 20(d)(e). To better quantify the temperature index, we classify the solution into three categories: normal temperature (20∼30 ^{o}C), cold temperature(0 $\sim$ 20 ^{o}C), and hot temperature (30∼100 ^{o}C), 150 sets of data are collected for each bottle in each temperature range, and each group of data contain a near-infrared image and a mid-infrared image, a total of 150×3×5 sets of data are collected. As shown in Fig. 22, we adopt accuracy as the main evaluation metric, the calculation equation is as follows:

$$Accuracy = n_{\text{correct}} / n_{\text{total}}, \tag{18}$$

where n_{correct} and n_{total} denote the number of successfully predicted samples and the total number of samples.

After 50 training rounds, the network achieves a classification accuracy of 98%. These experimental results underscore the sensor's significant practical utility, particularly in the realms of home service and chemical laboratories.

F. Exp. 6: Proximity Sensing

To test the contact force sensing performance of M^{3}Tac, we built an experimental platform as shown in Fig. 23(Left). The platform consists of a distance sensor, a liftable platform, etc. The distance sensor can obtain the true value of the object's distance from the sensor surface, and the liftable platform can ensure the stability of the object during data collection. In the experiment, we tested four kinds of objects with different sizes and colors, and each kind of object collected 50 sets of data in the range of 0∼12 cm, as shown in Fig. 23(Right)[3]. The proximity algorithm was used to establish the fitting equation, and the MAE between the true value and the predicted value directly was calculated. Finally, we obtained the perception accuracy of the four objects as 0.222, 0.255, 0.371, and 0.286 cm, and the perception accuracy is below 3.8 mm.

G. Exp. 7: Fragile and Light Objects Grasping

To verify the value of our proposed proximity perception in practical applications, we conduct grasping experiments on fragile and light objects, as shown in Fig. 24. In the experiment, we used feathers and pencil lead as the grasping objects. Feathers are light and soft, which makes grasping very difficult. The pencil lead is brittle and thin, which makes it very challenging to detect and grasp.

As shown in Fig. 20(a), to show the advantages of the M^{3}Tac in gripping fragile and light objects, a force sensor ATI min45 is mounted on the other side of the parallel gripper, which has a force resolution of 1/16 N. To better detect the proximity information, we also installed an LED ring on the other side, to ensure a stable brightness. During the gripping of a feather and a pencil lead, M^{3}Tac is able to detect the imperceptible contact information despite the absence of any valid signal from the ATI force sensor, which shows the feasibility of the M^{3}Tac to achieve stable gripping of fragile and light objects.

[3]P: Predicted value, M: Measured value by distance sensors.

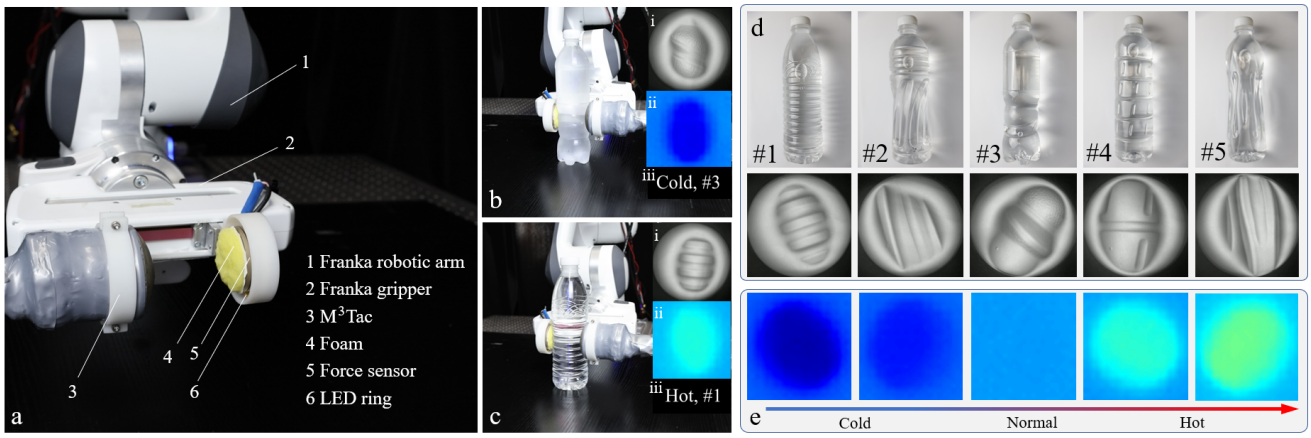

Fig. 20. Liquid bottle classification experiment. (a) Grasping platform; (b) Experiment on cold object grasping and classification: (i) Texture and contour information of the object; (ii) Temperature information; (iii) Recognition results; (c) Experiment on hot object grasping and classification; (d) Objects ready to be grasped; (e) Temperature classification.

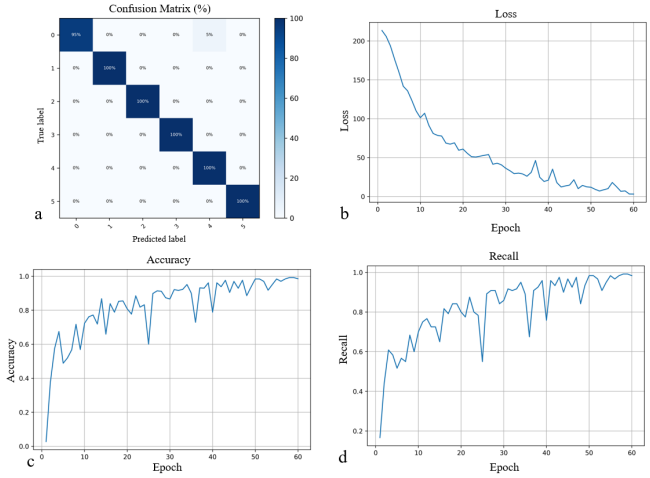

Fig. 21. The result of stickiness recognition experiment. (a) Confusion matrix; (b) Loss curve; (c) Accuracy curve; (d) Recall curve.

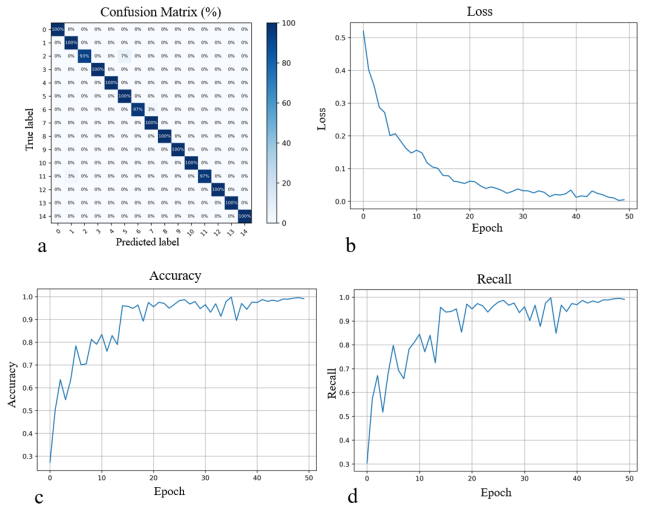

Fig. 22. The result of liquid bottle classification experiment. (a) Confusion matrix; (b) Loss curve; (c) Accuracy curve; (d) Recall curve.

H. Exp. 8: Circuit Board Heat Detection

M^{3}Tac combines multi-spectral information, which can not only obtain the temperature information of the contacting object but also the contour and depth of the object. Using this property, we can locate the heat generation position of the circuit board through contact and reconstruct the shape of the circuit board surface, as shown in Fig. 25. Compared with the traditional infrared camera, M^{3}Tac can detect the installation problems of the components on the surface of the circuit while acquiring the location of heat generation, and finding the problems of the circuit more accurately. This fully utilizes the advantages of M^{3}Tac's large detection area, high deformation range, and wide temperature detection range. In addition, the sensor can detect temperatures up to 130 ^{o}C, which meets the need to detect the thermal distribution of circuit boards.

I. Exp. 9: Underwater Pipeline Anomalous Hot Spot Detection

Thanks to the excellent leakproofness and anti-interference capabilities, M^{3}Tac can also be applied to underwater scenarios. To illustrate this, we design an underwater pipeline heating position detection experiment, as shown in Fig. 26. We put heat-generating electrodes inside the pipe to simulate the heat-generating line and use the tactile tracing algorithm to guide the movement of the robotic arm. When touching the heat-generating position, M^{3}Tac can detect the contour of the pipeline while acquiring the location of anomalous heat generation, which has important applications for realizing underwater detection.

VI. Conclusion & Discussion

In this paper, we combine multispectral imaging with unidirectional perspective latex film to propose a multispectral multifunctional visuotactile sensor (M^{3}Tac) that can realize force, deformation, texture, proximity, temperature, and stickiness sensing. Firstly, to obtain pixel-level force sensing information, we build an automated pixel-level data acquisition & annotation system and utilize finite element analysis and FSwint-MAP network to estimate the contact force information of

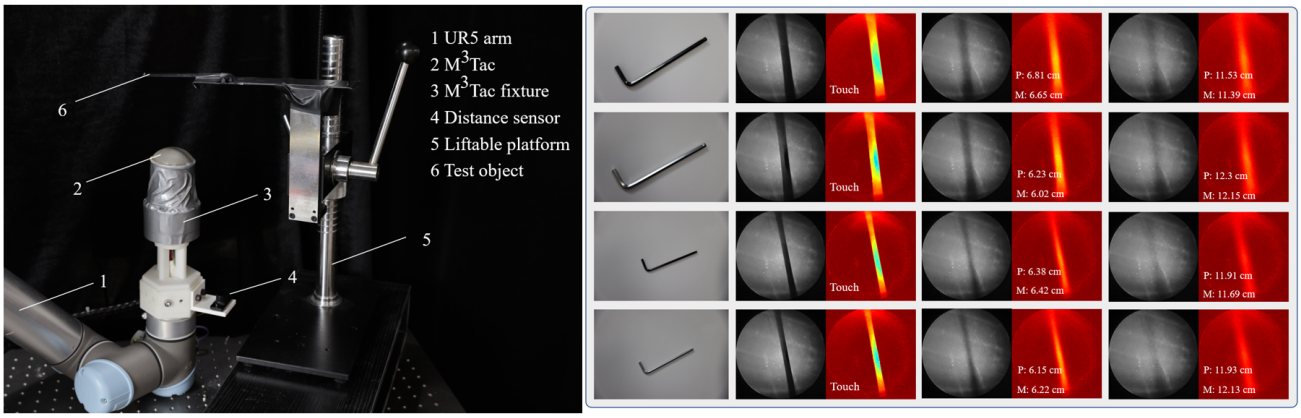

Fig. 23. The proximity sensing experiment. Left: proximity sensing test platform. Right: test results for different objects.

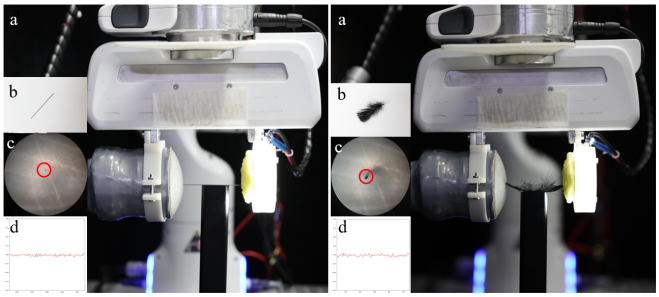

Fig. 24. Fragile and light objects grasping experiment. Left: (a) Pencil lead grasping; (b) Pencil lead; (c) Proximity image (Red circle indicates contact position); (d) Force detected by industrial force sensors. Right: (a) Feather grasping; (b) Feather; (c) Proximity image (Red circle indicates contact position); (d) Force detected by industrial force sensors.

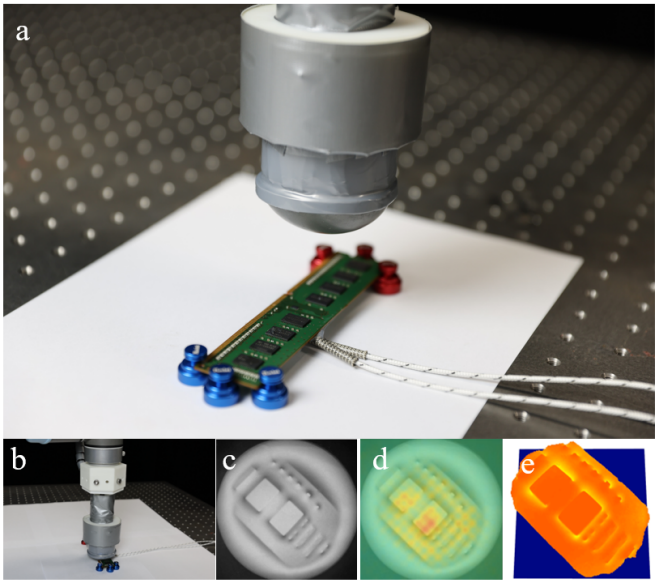

Fig. 25. Circuit board heat position detection experiment. (a) Experimental scenario; (b) The sensor makes contact with the circuit board; (c) Near-infrared image; (d) Temperature and near-infrared fusion information; (e) Depth image.

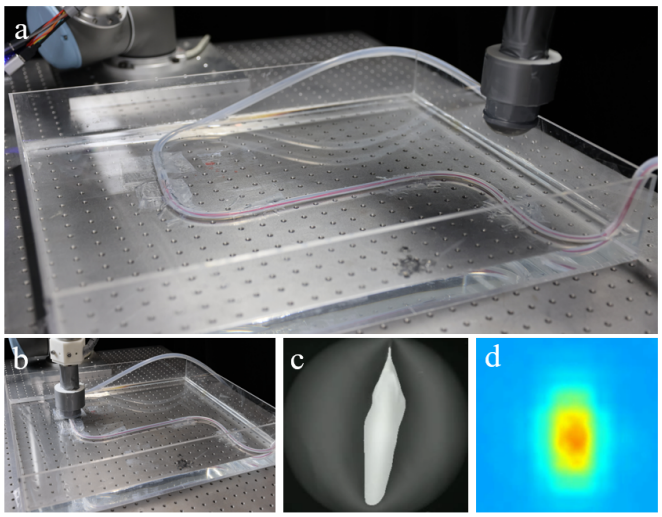

Fig. 26. Underwater pipeline anomalous hot spot detection experiment. (a) Experimental scenario; (b) Finding hot spots; (c) Pipeline information obtained from the segmentation network; (d) Temperature information.

each pixel, which achieves an accuracy of 0.023 N. Secondly, we propose a 3D reconstruction method based on luminance information, which not only realizes depth reconstruction but also information extraction from the contact area. Next, to realize high-resolution temperature sensing, we propose a lightweight super-resolution network, which can get 172×172 high-resolution temperature information, the sensing accuracy can reach 0.3 °C, and the sensing range can be up to -20∼130 °C. The temperature response speed can reach 54 °C/s. In addition, we also propose a multimodal classification algorithm and a stickiness classification method. Finally, to verify the application value of the sensors, we propose a fragile object grasping experiment, a circuit board heating position detection experiment, and an underwater pipeline heating position detection experiment. These tests serve to underscore the sensor's real-world applicability across various domains, including home service, industrial inspection, and underwater operations.

M^{3}Tac is the first sensor to simultaneously realize high-resolution proximity, force, deformation, temperature, sticki-

ness, and texture sensing. In some aspects, it can even surpass the sensing ability of human skin, which is of great significance in advancing the development of tactile sensors. M^{3}Tac adopts a multispectral imaging technique, which provides new ideas for developing visuotactile sensors. On the other hand, based on the M^{3}Tac sensor, we propose a complete algorithmic framework and data acquisition system, which not only reduces the workload during sensor calibration but also plays a role in promoting the industrialization of visuotactile sensors. In the contact force sensing method, we explore the finite element analysis of the elastic inflatable film when contact occurs, which provides theoretical support for sensors adopting elastic inflatable films. In addition, the elastic film used in our sensing skin not only avoids the influence of acrylic glass on mid-infrared temperature detection but also realizes stickiness classification based on the contact-separation process. However, the sensor also has some drawbacks, such as pixel-level force sensing that needs to obtain the finite element model of the contacting object, proximity sensing that is related to the size of the object, and excessive sensor size.

REFERENCES

[1] G. Corniani and H. P. Saal, "Tactile innervation densities across the whole body," *Journal of Neurophysiology*, vol. 124, no. 4, pp. 1229–1240, 2020.

[2] A. B. Vallbo, R. S. Johansson, *et al.*, "Properties of cutaneous mechanoreceptors in the human hand related to touch sensation," *Human Neurobiology*, vol. 3, no. 1, pp. 3–14, 1984.

[3] A. Handler and D. D. Ginty, "The mechanosensory neurons of touch and their mechanisms of activation," *Nature Reviews Neuroscience*, vol. 22, no. 9, pp. 521–537, 2021.

[4] X. Wang, L. Dong, H. Zhang, R. Yu, C. Pan, and Z. L. Wang, "Recent progress in electronic skin," *Advanced Science*, vol. 2, no. 10, p. 1500169, 2015.

[5] 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 Robotics and Automation Letters*, vol. 7, no. 3, pp. 6036–6043, 2022.

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

[7] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, *et al.*, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[8] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 138–143, 2016.

[9] 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 Transactions on Industrial Electronics*, vol. 69, no. 6, pp. 6015–6025, 2021.

[10] C. Wu, A. C. Wang, W. Ding, H. Guo, and Z. L. Wang, "Triboelectric nanogenerator: a foundation of the energy for the new era," *Advanced Energy Materials*, vol. 9, no. 1, p. 1802906, 2019.

[11] X. Lin and M. Wiertlewski, "Sensing the frictional state of a robotic skin via subtractive color mixing," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2386–2392, 2019.

[12] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4842–4849, 2018.

[13] D. L. Pavia, G. M. Lampman, G. S. Kriz, and J. A. Vyvyan, *Introduction to spectroscopy*. USA: Cengage Learning, 2014.

[14] J. Zwinkels, "Light, electromagnetic spectrum," *Encyclopedia of Color Science and Technology*, vol. 8071, pp. 1–8, 2015.

[15] W. Khan, N. Zaki, and L. Ali, "Intelligent pneumonia identification from chest X-rays: A systematic literature review," *IEEE Access*, vol. 9, pp. 51747–51771, 2021.

[16] P. Chanprakon, T. Sae-Oung, T. Treebupachatsakul, P. Hannanta-Anan, and W. Piyawattanametha, "An ultra-violet sterilization robot for disinfection," in *5th International Conference on Engineering, Applied Sciences and Technology (ICEAST)*, pp. 1–4, 2019.

[17] A. Nowosielski, K. Małecki, P. Forczmański, A. Smoliński, and K. Krzywicki, "Embedded night-vision system for pedestrian detection," *IEEE Sensors Journal*, vol. 20, no. 16, pp. 9293–9304, 2020.

[18] H.-Y. Chen, A. Chen, and C. Chen, "Investigation of the impact of infrared sensors on core body temperature monitoring by comparing measurement sites," *Sensors*, vol. 20, no. 10, p. 2885, 2020.

[19] 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.

[20] S. Zhang, J. Shan, F. Sun, B. Fang, and Y. Yang, "Multimode fusion perception for transparent glass recognition," *Industrial Robot: the International Journal of Robotics Research and Application*, vol. 49, no. 4, pp. 625–633, 2022.

[21] C. Yu, L. Lindenroth, J. Hu, J. Back, G. Abrahams, and H. Liu, "A vision-based soft somatosensory system for distributed pressure and temperature sensing," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3323–3329, 2020.

[22] A. C. Abad, D. Reid, and A. Ranasinghe, "HaptiTemp: A next-generation thermosensitive GelSight-like visuotactile sensor," *IEEE Sensors Journal*, vol. 22, no. 3, pp. 2722–2734, 2021.

[23] 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 *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1218–1227, 2021.

[24] Q. Wang, Y. Du, and M. Y. Wang, "SpecTac: A visual-tactile dual-modality sensor using UV illumination," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10844–10850, 2022.

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

[26] S. Zhang, Y. Sun, J. Shan, Z. Chen, F. Sun, Y. Yang, and B. Fang, "TIRgel: A visuo-tactile sensor with total internal reflection mechanism for external observation and contact detection," *IEEE Robotics and Automation Letters*, 2023.

[27] P. Puangmali, K. Althoefer, L. D. Seneviratne, D. Murphy, and P. Dasgupta, "State-of-the-art in force and tactile sensing for minimally invasive surgery," *IEEE Sensors Journal*, vol. 8, no. 4, pp. 371–381, 2008.

[28] Y. Wang, Y. Lu, D. Mei, and L. Zhu, "Liquid metal-based wearable tactile sensor for both temperature and contact force sensing," *IEEE Sensors Journal*, vol. 21, no. 2, pp. 1694–1703, 2020.

[29] X. Qu, Z. Liu, P. Tan, C. Wang, Y. Liu, H. Feng, D. Luo, Z. Li, and Z. L. Wang, "Artificial tactile perception smart finger for material identification based on triboelectric sensing," *Science Advances*, vol. 8, no. 31, p. eabq2521, 2022.

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

[31] H. Sun, K. J. Kuchenbecker, and G. Martius, "A soft thumb-sized vision-based sensor with accurate all-round force perception," *Nature Machine Intelligence*, vol. 4, no. 2, pp. 135–145, 2022.

[32] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[33] 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 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2722–2727, 2018.

[34] S. Li, X. Yin, C. Xia, L. Ye, X. Wang, and B. Liang, "TaTa: A universal jamming gripper with high-quality tactile perception and its application to underwater manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6151–6157, 2022.

[35] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using GelSlim and inverse FEM," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5418–5424, 2019.

[36] W. K. Do, B. Jurewicz, and M. Kennedy, "DenseTact 2.0: Optical tactile sensor for shape and force reconstruction," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 12549–12555, 2023.

[37] C. Trueeb, C. Sferrazza, and R. D'Andrea, "Towards vision-based robotic skins: a data-driven, multi-camera tactile sensor," in *IEEE International Conference on Soft Robotics (RoboSoft)*, pp. 333–338, 2020.

[38] I. H. Taylor, S. Dong, and A. Rodriguez, "GelSlim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10781–10787, 2022.

[39] Y. Ye, C. Zhang, C. He, X. Wang, J. Huang, and J. Deng, "A review on applications of capacitive displacement sensing for capacitive proximity sensor," *IEEE Access*, vol. 8, pp. 45325–45342, 2020.

[40] G. S. Cañón Bermúdez, D. D. Karnaushenko, D. Karnaushenko, A. Lebanov, L. Bischoff, M. Kaltenbrunner, J. Fassbender, O. G. Schmidt, and D. Makarov, "Magnetosensitive e-skins with directional perception for augmented reality," *Science Advances*, vol. 4, no. 1, p. eaao2623, 2018.

[41] S. R. A. Ruth, V. R. Feig, M.-g. Kim, Y. Khan, J. K. Phong, and Z. Bao, "Flexible fringe effect capacitive sensors with simultaneous high-performance contact and non-contact sensing capabilities," *Small Structures*, vol. 2, no. 2, p. 2000079, 2021.

[42] W. Liu, Y. Duo, J. Liu, F. Yuan, L. Li, L. Li, G. Wang, B. Chen, S. Wang, H. Yang, *et al.*, "Touchless interactive teaching of soft robots through flexible bimodal sensory interfaces," *Nature Communications*, vol. 13, no. 1, p. 5030, 2022.

[43] A. SaLoutos, H. Kim, E. Stanger-Jones, M. Guo, and S. Kim, "Towards robust autonomous grasping with reflexes using high-bandwidth sensing and actuation," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10254–10260, 2023.

[44] A. SaLoutos, E. Stanger-Jones, M. Guo, H. Kim, and S. Kim, "Design of a multimodal fingertip sensor for dynamic manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8017–8024, 2023.

[45] J. Rao, Z. Chen, D. Zhao, R. Ma, W. Yi, C. Zhang, D. Liu, X. Chen, Y. Yang, X. Wang, *et al.*, "Tactile electronic skin to simultaneously detect and distinguish between temperature and pressure based on a triboelectric nanogenerator," *Nano Energy*, vol. 75, p. 105073, 2020.

[46] M. D. Husain, R. Kennon, and T. Dias, "Design and fabrication of temperature sensing fabric," *Journal of Industrial Textiles*, vol. 44, no. 3, pp. 398–417, 2014.

[47] F. A. Viola, A. Spanu, P. C. Ricci, A. Bonfiglio, and P. Cosseddu, "Ultrathin, flexible and multimodal tactile sensors based on organic field-effect transistors," *Scientific Reports*, vol. 8, no. 1, p. 8073, 2018.

[48] J. Ge, X. Wang, M. Drack, O. Volkov, M. Liang, G. S. Cañón Bermúdez, R. Illing, C. Wang, S. Zhou, J. Fassbender, *et al.*, "A bimodal soft electronic skin for tactile and touchless interaction in real time," *Nature Communications*, vol. 10, no. 1, p. 4405, 2019.

[49] 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 Robotics and Automation Letters*, vol. 7, no. 4, pp. 10865–10872, 2022.

[50] J. R. Howell, M. P. Mengüç, K. Daun, and R. Siegel, *Thermal radiation heat transfer*. Boca Raton: CRC Press, 2020.

[51] E. Stefanelli, F. Cordella, C. Gentile, and L. Zollo, "Hand prosthesis sensorimotor control inspired by the human somatosensory system," *Robotics*, vol. 12, no. 5, p. 136, 2023.

[52] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *International Conference on Computer Vision (ICCV)*, pp. 10012–10022, 2021.

[53] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[54] J. Mezirow, "Perspective transformation," *Adult education*, vol. 28, no. 2, pp. 100–110, 1978.

[55] "The official website of COMSOL." https://www.comsol.org/.

[56] Y. Başar and M. Itskov, "Finite element formulation of the ogden material model with application to rubber-like shells," *International Journal for Numerical Methods in Engineering*, vol. 42, no. 7, pp. 1279–1305, 1998.

[57] R. W. Ogden, "Large deformation isotropic elasticity–on the correlation of theory and experiment for incompressible rubberlike solids," *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, vol. 326, no. 1567, pp. 565–584, 1972.

[58] L. Gornet, G. Marckmann, R. Desmorat, and P. Charrier, "A new isotropic hyperelastic strain energy function in terms of invariants and its derivation into a pseudo-elastic model for Mullins effect," *Constitutive Models for Rubbers VII*, pp. 265–271, 2012.

[59] G. Greto, *JST-SPH: a total Lagrangian, stabilised meshless methodology for mixed systems of conservation laws in nonlinear solid dynamics*. PhD thesis, Cardiff University, 2018.

[60] D. Tabor, "The bulk modulus of rubber," *Polymer*, vol. 35, no. 13, pp. 2759–2763, 1994.

[61] Y. Zhang, Z. Kan, Y. Yang, Y. A. Tse, and M. Y. Wang, "Effective estimation of contact force and torque for vision-based tactile sensors with helmholtz–hodge decomposition," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4094–4101, 2019.

[62] V. Kakani, X. Cui, M. Ma, and H. Kim, "Vision-based tactile sensor mechanism for the estimation of contact position and force distribution using deep learning," *Sensors*, vol. 21, no. 5, p. 1920, 2021.

[63] K. Sato, K. Kamiyama, H. Nii, N. Kawakami, and S. Tachi, "Measurement of force vector field of robotic finger using vision-based haptic sensor," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 488–493, 2008.

[64] Y. Li, T. Yao, Y. Pan, and T. Mei, "Contextual transformer networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1489–1500, 2022.

[65] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," in *International Conference on Computer Vision (ICCV)*, pp. 3163–3172, 2021.

[66] S. Yun, M. Jeong, R. Kim, J. Kang, and H. J. Kim, "Graph transformer networks," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[67] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

[68] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, "Swin-Unet: Unet-like pure transformer for medical image segmentation," in *European Conference on Computer Vision (ECCV)*, pp. 205–218, 2022.

[69] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

[70] D. C. Lepcha, B. Goyal, A. Dogra, and V. Goyal, "Image super-resolution: A comprehensive review, recent trends, challenges and applications," *Information Fusion*, vol. 91, pp. 230–260, 2023.

[71] X. Pan, C. Ge, R. Lu, S. Song, G. Chen, Z. Huang, and G. Huang, "On the integration of self-attention and convolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 815–825, 2022.

[72] S. Li, H. Yu, W. Ding, H. Liu, L. Ye, C. Xia, X. Wang, and X.-P. Zhang, "Visual–tactile fusion for transparent object grasping in complex backgrounds," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3838–3856, 2023.

[73] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

[74] R. Jiang and Y. Wang, "Study of the human stickiness perception of wet fabric on the volar forearm via two contact modes: friction and adhesion-separation," *Perception*, vol. 49, no. 12, pp. 1311–1332, 2020.

[75] J. Johnston, "The physical testing of pressure-sensitve adhesive systems," *Handbook of Adhesive Technology. 2nd ed. New York: Marcel Dekker Inc*, pp. 253–72, 2003.

[76] K. Dan, Y. Liangzhe, L. Yandong, Z. Li, B. Matthew, and G. Boqing, "MoViNets: Mobile video networks for efficient video recognition," *arXiv preprint arXiv:2103.11511*, 2021.

[77] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

[78] W.-S. Lai, J.-B. Huang, N. Ahuja, and M.-H. Yang, "Deep laplacian pyramid networks for fast and accurate super-resolution," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 624–632, 2017.

[79] Z. Hui, X. Gao, Y. Yang, and X. Wang, "Lightweight image super-resolution with information multi-distillation network," in *27th ACM International Conference on Multimedia (ACM MM)*, pp. 2024–2032, 2019.

[80] K. Zhang, J. Liang, L. Van Gool, and R. Timofte, "Designing a practical degradation model for deep blind image super-resolution," in *International Conference on Computer Vision (ICCV)*, pp. 4791–4800, 2021.

This article has been accepted for publication in IEEE Transactions on Robotics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TRO.2024.3462931

Shoujie Li received the B.Eng. degree in electronic information engineering from the College of Oceanography and Space Informatics, China University of Petroleum, Tsingtao, China, in 2020. He is currently pursuing toward Ph.D. degree in Tsinghua-Berkeley Shenzhen Institute, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

His research interests include tactile perception, grasping, and machine learning. He received the Outstanding Mechanisms and Design Paper Finalists in ICRA 2022 and the Best Application Paper Finalists in IROS 2023. He won first place in the Robotic Grasping of Manipulation Competition-Picking in Clutter in ICRA 2024.

Haixin Yu received the B.S. degree in automation from Northeastern University, Shenyang, China, in 2021. He is currently working toward the M.S. degree in Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

His research interests include robotics, machine learning, and computer vision.

Guoping Pan received the B.S. degree in Intelligence Science and Technology from Xidian University, Shaanxi, China, in 2023. He is currently working toward the M.S. degree at Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. His research interests include reinforcement learning, robotics, and computer vision.

Huaze Tang Huaze Tang received the B.S. degree in communication engineering from Southeast University, Nanjing, China, in 2021. He is currently pursuing a Ph.D. degree in data science and information technology with the Smart Sensing and Robotics (SSR) Group at Tsinghua-Berkeley Shenzhen Institute, Tsinghua University, Shenzhen, China.

His research interests focus on signal processing with deep learning, reinforcement learning, and multi-agent systems.

Jiawei Zhang received the B.S. degree in Automation Engineering from the University of Electronic Science and Technology of China in 2018, and the M.S. degree in Electronic and Information Engineering from Tsinghua Shenzhen International Graduate School in 2023. He is currently working at a humanoid robot startup, AGIBOT.

His research focuses on object segmentation, pose estimation, photorealistic rendering, and robotic grasping based on sim2real.

Linqi Ye received bachelor's and Ph.D. degrees in control science and engineering from Tianjin University, Tianjin, China, in 2014 and 2019, respectively. From 2016 to 2017, he was a Visiting Scholar with the State University of New York at Buffalo, Buffalo, NY, USA. From 2017 to 2018, he was a Visiting Scientist at Cornell University, Ithaca, NY, USA. From 2019 to 2022, he was a Postdoc with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. He is currently an Associate Professor at Shanghai University, Shanghai, China.

His research interests include humanoid robots and reinforcement learning.

Xiao-Ping Zhang received B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, both in Electronic Engineering. He holds an MBA in Finance, Economics and Entrepreneurship with Honors from the University of Chicago Booth School of Business, Chicago, IL. He is the founding Dean of Institute of Data and Information (iDI) at Tsinghua Shenzhen International Graduate School (SIGS), Chair Professor at Tsinghua SIGS and Tsinghua-Berkeley Shenzhen Institute (TBSI), Tsinghua University. He had been with the Department of Electrical, Computer and Biomedical Engineering, Toronto Metropolitan University (Formerly Ryerson University), Toronto, ON, Canada, as a Professor and the Director of the Communication and Signal Processing Applications Laboratory (CASPAL), and has served as the Program Director of Graduate Studies. He was cross-appointed to the Finance Department at the Ted Rogers School of Management, Toronto Metropolitan University. His research interests include image and multimedia content analysis, sensor networks and IoT, machine learning/AI, statistical signal processing, and applications in big data, finance, and marketing.

Dr. Zhang is Fellow of the Canadian Academy of Engineering, Fellow of the Engineering Institute of Canada, Fellow of the IEEE, a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is the general co-chair for 2021 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2021). He is the general co-chair for 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the general co-chair for 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He was an elected member of IEEE International Conference on Multimedia and Expo (ICME) steering committee. He is the general chair for ICME2024. He is Editor-in-Chief for the IEEE Journal of Selected Topics in Signal Processing. He is Senior Area Editor for IEEE Transactions on Image Processing. He served as Senior Area Editor for IEEE Transactions on Signal Processing and Associate Editor for IEEE Transactions on Signal Processing, IEEE Transactions on Image Processing, IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Signal Processing Letters. He was selected as IEEE Distinguished Lecturer by Signal Processing Society and by Circuits and Systems Society.

Wenbo Ding received the BS and PhD degrees (Hons.) from Tsinghua University in 2011 and 2016, respectively. He worked as a postdoctoral research fellow at Georgia Tech under the supervision of Professor Z. L. Wang from 2016 to 2019. He is now an associate professor and PhD supervisor at Tsinghua-Berkeley Shenzhen Institute, Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, where he leads the Smart Sensing and Robotics (SSR) group. He has received many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva, the IEEE Scott Helt Memorial Award, and the IAS Residential Fellow.

His research interests are diverse and interdisciplinary, which include signal processing for robotics, embodied AI, human-robot interfaces and multi-agent systems.