VTire: A Bimodal Visuotactile Tire With High-Resolution Sensing Capability

Shoujie Li, *Graduate Student Member, IEEE*, Jianle Xu, Tong Wu, Yang Yang, Yanbo Chen, Xueqian Wang, *Member, IEEE*, Wenbo Ding, *Member, IEEE*, and Xiao-Ping Zhang, *Fellow, IEEE*

***Abstract*—Developing smart tires with high sensing capability is significant for improving the moving stability and environmental adaptability of wheeled robots and vehicles. However, due to the classical manufacturing design, it is always challenging for tires to infer external information precisely. To this end, this article introduces a bimodal sensing tire, which can simultaneously capture tactile and visual data. By leveraging the emerging visuotactile techniques, the proposed smart tire can realize various functions, including terrain recognition, ground crack detection, load sensing, and tire damage detection. Besides, we optimize the material and structure of the tire to ensure its outstanding elasticity, toughness, hardness, and transparency. In terms of algorithms, a transformer-based multimodal classification algorithm, a load detection method based on finite element analysis, and a contact segmentation algorithm have been developed. Furthermore, we construct an intelligent mobile platform to validate the system's effectiveness and develop visual and tactile datasets in complex terrains. The experimental results show that our multimodal terrain sensing algorithm can achieve a classification accuracy of 99.2% , a tire damage detection accuracy of 97% , a 98% success rate in object search, and the ability to withstand tire loading weights exceeding 35 kg.**

***Index Terms*—Multimodal classification, smart tires, tactile sensor, visuotactile sensing.**

Received 3 December 2024; revised 10 February 2025 and 26 March 2025; accepted 27 April 2025. Recommended by Technical Editor E. Kayacan and Senior Editor J. Ueda. This work was supported in part by the National Key R&D Program of China under Grant 2024YFB3816000, in part by Shenzhen Key Laboratory of Ubiquitous Data Enabling under Grant ZDSYS20220527171406015, in part by Shenzhen Science and Technology Program under Grant JCYJ20220530143013030, in part by the Guangdong Innovative and Entrepreneurial Research Team Program under Grant 2021ZT09L197, in part by the National Natural Science Foundation of China under Grant 62104125 and 62003188, in part by the Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation under Grant SZPR2023005, and in part by Meituan. *(Shoujie Li and Jianle Xu contributed equally to this work.) (Corresponding authors: Xueqian Wang; Wenbo Ding.)*

Shoujie Li, Jianle Xu, Tong Wu, Yanbo Chen, Xueqian Wang, and Xiao-Ping Zhang are with Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China (e-mail: wang.xq@sz.tsinghua.edu.cn).

Wenbo Ding is with the Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China, and also with the RISC-V International Open Source Laboratory, Shenzhen 518055, China (e-mail: ding.wenbo@sz.tsinghua.edu.cn).

Yang Yang is with the Department of Mechanics Science and Engineering, Sichuan University, Chengdu 610065, China.

In addition, we open-source our algorithms, hardware, and datasets at https://sites.google.com/view/vtire.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TMECH.2025.3566394.

Digital Object Identifier 10.1109/TMECH.2025.3566394

I. Introduction

TIRE is a crucial actuating component, which has a wide range of applications in vehicles [1] and wheeled bipedal robots [2]. As the only part of the robot in contact with the ground, tires could enhance the driving stability as well as serve as the most dependable source of information regarding the ground. However, due to the classical manufacturing and sensing technology, the existing tires only obtain limited information, such as load, speed, acceleration, etc. [3], which cannot realize pixel-level texture sensing. Designing a smart tire with high-resolution tactile sensing ability can solve the problem of terrain sensing under complex scenarios and realize more functions, including ground crack detection, ground object search, and tire damage detection, dramatically improving the robot's or vehicle's sensing ability.

To enhance the sensing ability of tires, researchers usually adopt a variety of force [4], [5] and optical sensors [6], [7], [8] inside the tire, which could provide the tire with the ability to sense its states, such as load, tension, and so on. Nevertheless, due to the obstruction of the tire casing, such methods cannot obtain external information, such as the road texture, cracks, etc. With advancements in tactile perception and optical imaging technologies, a novel technique called visuotactile perception [9], [10], [11] has emerged. This technology detects tactile information by observing surface deformations through a camera, offering high resolution and extensive sensing coverage. However, tires' opacity, hardness, and toughness pose difficulties in integrating visuotactile perception technology.

In this article, by improving tires' materials, mechanical structure, and manufacturing process, we propose a bimodal smart tire named VTire based on visuotactile sensing techniques. As shown in Fig. 1, VTire overcomes the bottlenecks in the resolution of tactile sensing and sensing area of traditional smart tires and realizes the functions of sensing ground texture, cracks, and bumps, which are challenging for traditional smart tires. The contributions of this article are as follows.

1) *Innovative Manufacturing Process:* We propose a tire manufacturing method that capitalizes on visuotactile perception, ensuring high-resolution tactile data acquisition. Based on the unique structural design, our tires

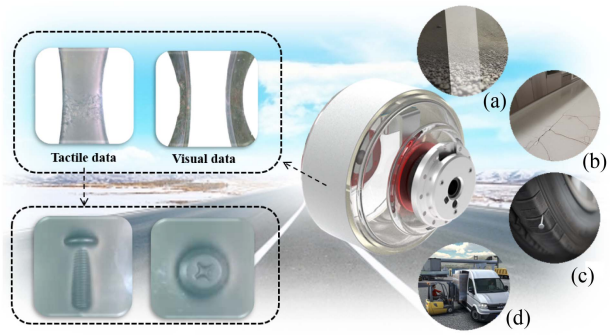

Fig. 1. Introduction to the bimodal Tire. Functions that can be achieved with tires. (a) Terrain classification in complex scenes. (b) Ground cracks and objects searching. (c) Tire damage detection. (d) Weight-loaded detection.

acquire both tactile and visual information, which enhances the perception of the external environment.

2) *Advanced Sensing Algorithms:* Our research introduces a series of sophisticated sensing algorithms tailored for smart tires. These algorithms include a transformer-based multimodal classification algorithm, a load detection algorithm using finite element analysis (FEA), and a contact segmentation algorithm.
3) *Comprehensive Dataset:* To evaluate the performance of our algorithms, we assembled a diverse dataset. This dataset encompasses various terrains, surface textures, tire damage scenarios, and ground cracks, providing a comprehensive evaluation framework.
4) *Numerous Validation Experiments:* To verify the performance of the tires, we not only built a smart mobile platform but also designed several real-world experiments. Results indicate remarkable accuracy rates, including 0.75-kg weight sensing accuracy, 98% crack segmentation accuracy, 98% object detection success rate, and 97% damage detection accuracy. These results underscore the practical viability of our approach.

II. Related Work

The design of smart tires can be divided into contact sensing and noncontact sensing. contact sensing mainly includes accelerometers [12], [13], surface acoustic wave sensors [14], [15], piezoelectric sensors [16], [17], strain sensors [18], [19], and fiber Bragg grating strain sensors [20]. The contact sensors are mounted on the tire's inner wall and can accurately obtain the pressure and tension when the tire is in contact with the ground. However, the spatial resolution of these sensors is low, and since the sensors are pasted on the inner side of the tire, they are easy to detach and malfunction after long-distance movement. Noncontact sensors can acquire tire deformation without touching the tire's surface, which greatly avoids the risk of the sensor falling off and improves the sensor's service life. Typical noncontact sensors include optical sensors [6], [7], [8] and ultrasonic sensors [21]. Most optical sensors use single-point laser sensors or detect the offset of a feature point to obtain contact information. Due to the obstruction of the tire casing, it is difficult for these methods to obtain the texture, deformation, and other information of the contact road surface, which greatly compresses the application range of smart tires, as shown in Table I. To solve this problem, Hu et al. [22] proposed a visual tire that uses transparent acrylic as the casing. While this solution can achieve terrain detection, it uses a visual solution susceptible to the external environment. Furthermore, the acrylic casing not only makes the vehicle prone to skidding but also makes the tire lose its elasticity. When the vehicle passes through some rough, dirty road surface, it is easy to wear on the shell, thus affecting the detection effect.

TABLE I
Comparison of Current Smart Tire Functions

Ref	Pressure	Deformation	Crack	Terrain	Damage
Ref. [5]	✔	✘	✘	✘	✘
Ref. [8]	✔	✔	✘	✘	✘
Ref. [21]	✘	✔	✘	✘	✘
Ref. [23]	✔	✘	✘	✔	✘
Ref. [24]	✘	✘	✘	✔	✘
Ref. [25]	✘	✔	✘	✘	✘
Ours	✔	✔	✔	✔	✔

III. Hardware Design

To improve the sensing ability and mechanical strength, we optimize the tires in terms of structure and material. Besides, to demonstrate the performance of the smart tires, we build an intelligent mobile platform, as shown in Fig. 2.

A. System Structure

The mobile platform is 60-mm long and 29-mm wide and consists of a power system, a sensing system, a control system, and a motion system. To ensure structural stability, the body is made of metal and carbon fiber with a high weighing capacity that can carry adults over 70 kg. The battery is DJI TB47, and the voltage regulator module is ToolkitRC, which can output up to 40-A current. The sensing system consists of Vtire, lidar (OS0-128), and RGB-D camera (Realsense D435i), which enables real-time mapping and terrain classification in different environments. The control system comprises a remote control handle and a Next Unit of Computing NUC. All systems are connected through our designed main control board, as shown in Fig. 3. Vtire not only performs the sensing function but is also used to provide the driving force. Because the front wheels have larger friction, this front-wheel drive improves the platform's ability to cross obstacles. The rear wheel steering configuration makes the platform more agile with a smaller turning radius.

B. Proposed Smart Tire

Like the classic visuotactile sensors, smart tires are composed of a sensing skin, a lighting system, and a vision system. The difference, however, is that the design of smart tires needs to consider their transparency, elasticity, load capacity, and other

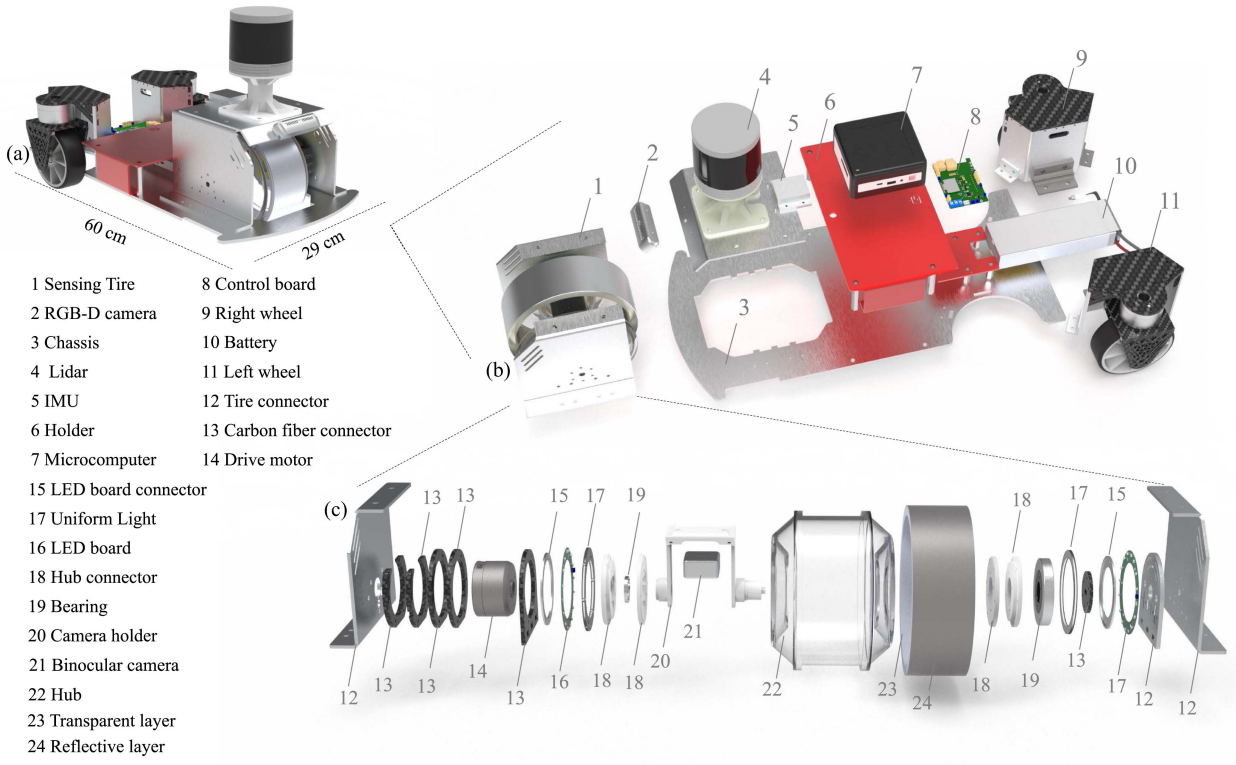

Fig. 2. Hardware structure diagram. (a) Side view of the motion platform. (b) Mobile platform exploded view. (c) VTire exploded view.

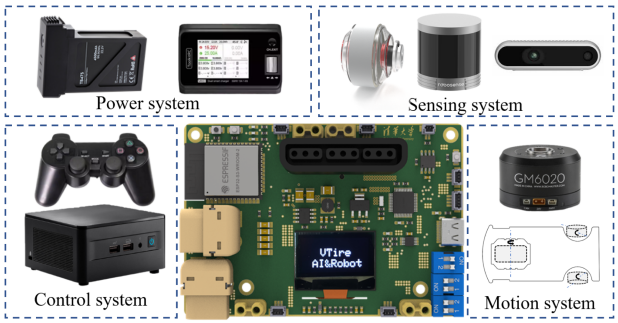

Fig. 3. Mobile platform framework diagram. The mobile platform consists of a power system, a sensing system, a control system, and a motion system.

metrics, making them highly challenging in terms of dimensions, fabrication process, and material selection.

1) Sensing Skin: The sensing skin is the heart of the tire, which consists of a transparent elastic hub, a transparent layer, and a reflective layer. Transparent elastic wheels not only support the robot body but also have the function of vibration damping. So, transparent wheels need to be elastic, tough, and transparent. After considering silicone, rubber, acrylic, glass, and other materials, we use a polyurethane (PU) material, which has a lower price, higher hardness, and transparency and is widely used in skateboard wheels, omnidirectional wheels, and other fields. The tire adopts nonsealed structures and the manufacturing process of the hub is illustrated in Fig. 4(a). The molds are all made by 3-D printing. PU materials have a high viscosity during the demolding process. To ensure that the surface of the mold is smooth, a special release agent is used, and clear tape is applied to key areas. In addition, we used a water-soluble material [Polyvinyl alcohol] as an internal support [red part in Fig. 4(a)], which softens when dissolved in water so that the hub can be unmolded smoothly.

To ensure the perception ability of the tire, we also design the elastic transparent layer [Fig. 4(b)] and a reflective layer [Fig. 4(c)]. For the elastic transparent layer, we use low-hardness elastic silicone and design a removable mold to realize the transparent layer. We use Eco-Flex 30 as the material for the reflective layer and mix it with silver powder to enhance the texture perception. In addition, to ensure the homogeneity of the outermost layer structure, we design a squeegee to control the thickness of the outermost film of the tire through the squeegee and to make the outermost layer of the tire more homogeneous through rotation. We open-source the production process and hardware, detailed on the website https://sites.google.com/view/vtire.

2) Imaging System: For the vision system design, we consider two options: 1) a multicamera solution [Fig. 5(a)]; and 2) a single-camera structure [Fig. 5(b)]. The multicamera one can monitor the status of the whole tire in real time, but the design structural complexity is high; the main issues are as follows.

a) *Installation difficulties:* Due to the limited space inside the tire, factors, such as the camera's focal length and imaging range must be considered during installation. Since we are using binocular imaging technology, the camera has a minimum detection distance of 12 cm, which

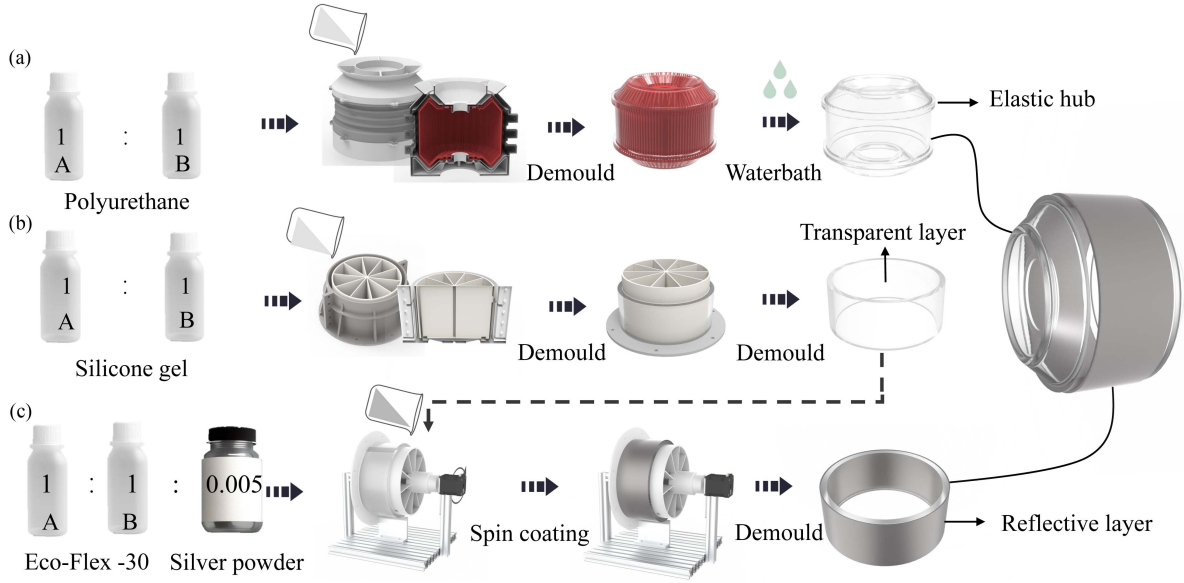

Fig. 4. Fabrication processes of (a) the wheel hub; (b) the transparent layer; and (c) the reflective layer.

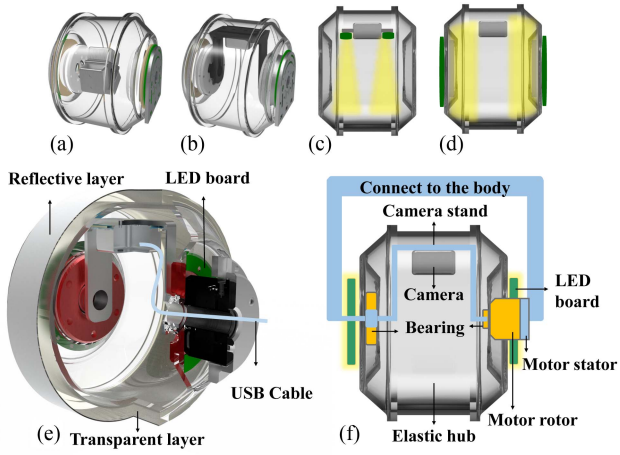

Fig. 5. Comparison of different solutions. (a) Multicamera solution. (b) Single camera solution. (c) Top lighting solution. (d) Side lighting solution. (e) Section view. (f) Connection of components.

is almost the limit for the tire. Moreover, the presence of the tire's central axis will further compress the available space for camera installation when multiple cameras are deployed.

b) *Data transmission difficulties:* In the single-camera solution, the camera does not rotate with the tire, so the camera's data can be stably transmitted to the computer. However, the cameras rotate with the tire when using the multicamera solution, making transmitting data to the computer difficult. On the one hand, the data cables may become entangled due to the tire's rotation, and on the other hand, rotation can affect the stability of the camera's data transmission.

c) *Poor reliability:* With the multicamera solution, the cameras rotate with the tire, and the solid centrifugal force can easily damage the cameras.

For the above reasons, we abandoned the multicamera solution. The single-camera solution uses a bearing + bracket structure to fix the camera on the body and prevent the camera from rotating with the tire. To obtain the depth information inside the tire, we adopt the binocular depth imaging technology, using Intel's Realsense D405 as the image acquisition device and obtaining the texture and deformation information of the tire in real time through the binocular depth reconstruction algorithm. In addition, the camera has a high image rate, which utilizes a global shutter and can reach a frame rate of 90 fps.

To solve the power supply and data transmission problem, we used hollow motors as VTire drivers so the USB cable could pass through the hollow shaft without getting tangled, as shown in Fig. 5(e). The blue part of Fig. 5(f) represents the part where the VTire is connected to the body. It is fixed during the rotation of the V-tire, while the gray part will rotate driven by the motor rotor. Since the camera stand is fixed to the body, it always keeps the field of view facing the ground. In addition to the difficulty of deployment, there are a few reasons why we went with a single-camera solution as follows.

a) *Cost issue:* The multicamera solution would significantly increase our costs.

b) *Structural issue:* When the tire moves, only a part of the area will be in contact with the ground so that the single camera solution can meet the tactile information collection needs when in contact with the ground and the multicamera solution does not get more contact information.

c) *Application scenario issue:* Our application scenarios are mainly aimed at low-speed scenarios, such as ground crack detection, ground small object search, and other tasks. We have chosen a camera with a frame rate of 90 fps (traditional cameras only have 30 fps), which can meet the needs of these tasks. If it is for high-speed scenarios, then

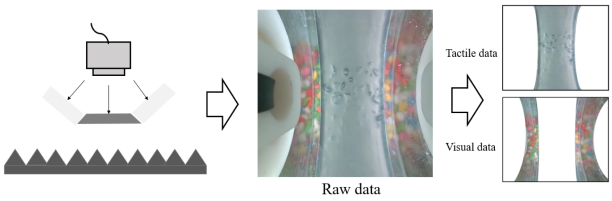

Fig. 6. Schematic of bimodal perception.

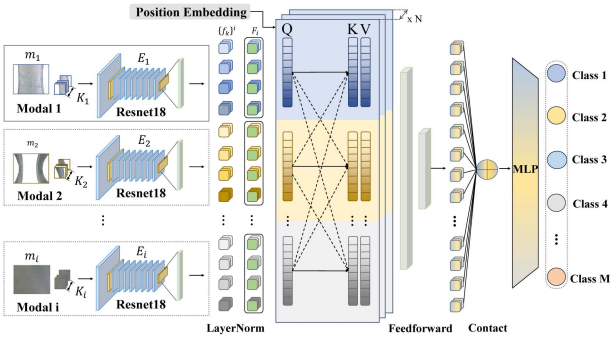

Fig. 7. Transformer-based multimodal terrain classification algorithm.

we can use cameras with higher frame rates, such as event cameras.

3) Lighting System: For the lighting system, we have also attempted several different schemes. To reduce the effect of reflections inside the camera, we choose the solution shown in Fig. 5(d). We can achieve a uniform lighting effect by installing LED light rings on both sides of the tire.

IV. Algorithm Design

A. Multimodal Terrain Classification Algorithm

Vision is the most direct means for robots to perceive the external environment. However, factors, such as smoke, brightness, and dust can significantly impact the visual approach to terrain classification [26]. In contrast, tactile perception provides greater stability in such scenarios. Compared with classic visuotactile sensors, VTire can obtain tactile and visual data at the same time. As shown in Fig. 6, the center area of the tires is covered with sensing skin, and the side area is transparent. The transparent areas allow us to perceive external visual information while acquiring tactile information. In addition, cameras on the outside of the vehicle can also provide clearer visual information compared to the inside of the tire.

To better integrate these data, we propose a multimodal classification algorithm, which adopts a transformer [27] network as the body of the algorithm, as shown in Fig. 7. By adjusting the inputs to the network, it is possible to classify the data directly from a single modality to multiple modalities. Given a series of different modalities m_i, $i = 0, 1, \ldots, M-1$, which may include optical images of the terrain, tactile feedback from the surface, or partial observations, such as dark or smoky images, we employ a set of modality-specific encoders $\{E_i\}$ to extract features from each corresponding modality. Instead of extracting a global feature for each modality, we partition each modality into K_i fragments and use encoder E_i to extract K_i fragmented features. Since data distribution of these modalities may differ significantly, complicating modality fusion and training process, we apply LayerNorm [28] to each modality to alleviate this issue. The process of feature extraction can be formulated as follows:

$$\{\boldsymbol{f}_k\}^i = \text{LayerNorm}(E_i(\text{Seg}(m_i))). \tag{1}$$

To effectively utilize the information within each modality and across different modalities, we employ self-attention and cross-attention mechanisms [29] to fuse the extracted features. Given the obtained features $\{\boldsymbol{f}_k\}^i$, we initially concatenate them along the fragment dimension, followed by concatenation along the modality dimension. Positional embedding is then added to the feature map. From the tokenized features $\boldsymbol{F}$ we can derive the following formulation:

$$\boldsymbol{F} = \begin{bmatrix} \boldsymbol{F}_0 & \boldsymbol{F}_1 & \cdots & \boldsymbol{F}_{M-1} \end{bmatrix} \tag{2}$$

where $\boldsymbol{F}_i$ denotes the concatenated feature of the ith modality. To calculate self-attention and cross-attention, we pass $\boldsymbol{F}$ through a linear layer to derive the queries $\{\boldsymbol{Q}_j\}$, keys $\{\boldsymbol{K}_j\}$, and values $\{\boldsymbol{V}_j\}$ for each attention head $j = 1, 2, \ldots, N$

$$\boldsymbol{Q}_j = \boldsymbol{W}_j^Q \boldsymbol{F} = \begin{bmatrix} \boldsymbol{Q}_{j,0} & \boldsymbol{Q}_{j,1} & \cdots & \boldsymbol{Q}_{j,M-1} \end{bmatrix} \tag{3}$$

$$\boldsymbol{K}_j = \boldsymbol{W}_j^K \boldsymbol{F} = \begin{bmatrix} \boldsymbol{K}_{j,0} & \boldsymbol{K}_{j,1} & \cdots & \boldsymbol{K}_{j,M-1} \end{bmatrix} \tag{4}$$

$$\boldsymbol{V}_j = \boldsymbol{W}_j^V \boldsymbol{F} = \begin{bmatrix} \boldsymbol{V}_{j,0} & \boldsymbol{V}_{j,1} & \cdots & \boldsymbol{V}_{j,M-1} \end{bmatrix} \tag{5}$$

where $\boldsymbol{W}_j^Q$, $\boldsymbol{W}_j^K$, and $\boldsymbol{W}_j^V$ denote the weight matrices for the query, key, and value, respectively. The terms $\boldsymbol{Q}_{j,i}$, $\boldsymbol{K}_{j,i}$, and $\boldsymbol{V}_{j,i}$ denote the query, key, and value of head j corresponding to the ith modality. The attention for head j is then calculated using the following equation:

$$\begin{aligned} \boldsymbol{A}_j &= \boldsymbol{V}_j \times \text{softmax}\left((\boldsymbol{K}_j)^T \boldsymbol{Q}_j / \sqrt{d}\right) \\ &= \boldsymbol{V}_j \times \text{softmax}\left(\begin{bmatrix} (\boldsymbol{K}_{j,0})^T \boldsymbol{Q}_{j,0} & (\boldsymbol{K}_{j,0})^T \boldsymbol{Q}_{j,1} & \cdots \\ (\boldsymbol{K}_{j,1})^T \boldsymbol{Q}_{j,0} & (\boldsymbol{K}_{j,1})^T \boldsymbol{Q}_{j,1} & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} / \sqrt{d} \right) \end{aligned} \tag{6}$$

where d is the dimension of the multimodal features. The self-attention and cross-attention mechanisms are as follows: The diagonal elements represent self-attention, where self-generated queries are used to draw self-generated keys. Conversely, the nondiagonal elements represent cross-attention, wherein queries and keys are generated from different modalities to extract features. Subsequently, we pass the attention outputs through a feedforward network and concatenate all tokens into a fused feature. In the end, a classification head is employed to produce the output.

In the practical implementation, we preprocess the raw images into 16×16 patches and select ResNet18 [30] as the backbone for the encoder E_i. This choice is due to ResNet18's efficiency in extracting diverse features and its capacity for facilitating rapid training. We configure the feature dimension to be 384. To promote effective modality fusion, we utilize four attention

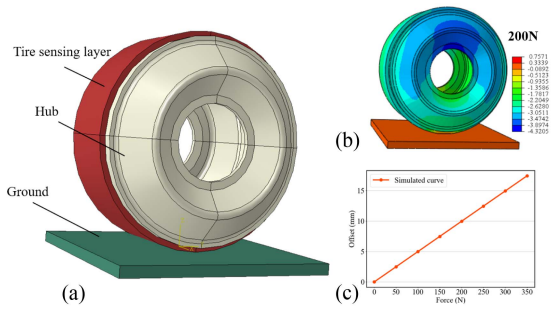

Fig. 8. Simulation results. (a) Initial FEA simulation configuration. (b) Displacement deformation under 200 N. (c) Simulated relationship between deformation and force.

heads, ensuring diverse attention patterns across modalities. We also find that a single attention block suffices for all tasks. For the classification task, we employ a straightforward MLP comprising a hidden layer followed by a softmax layer. During training, we minimize the cross-entropy loss function using the batch size of 32.

B. Load Weight Sensing Algorithm

The load weight of a robot is directly correlated with the deformation experienced by its tires. To explore this relationship, we adopt the Abaqus/standard FEA simulation. This approach allowed us to model the tire accurately under various load conditions. The loading process's configuration and the simulation setup are depicted in Fig. 8(a).

1) *Material Properties:* Based on the experimental measurement results, the mechanical parameters of the hub and tire sensor layer are determined as $E = 0.1973$ MPa, $\nu = 0.48$, and $E = 24.06$ MPa, $\nu = 0.49$, respectively. Notably, the ground is considered a rigid body in the simulation.
2) *Interaction and Loading:* Interactions between the hub and sensor layer and between the hub and ground are defined as constraints within the model. The applied load is centrally located on the hub and directed toward the ground.
3) *Mesh:* For the mesh, we employ a simplified integration hexahedral eight-node element (C3D8R) across the model's three components.
4) *Simulated Results:* We integrate Python scripting with Abaqus to dynamically adjust the applied force, enabling us to derive the deformation-force relationship.

Here, the deformation field at an applied force of 200 N is depicted in Fig. 8(b), while Fig. 8(c) illustrates the corresponding deformation-force curve.

C. Floor Crack and Contact Object Segmentation Algorithm

Compared with vision, touch is more stable and robust. It is difficult to detect some small or transparent objects using only vision. Besides, crack detection could be challenging with some richly textured floors. To solve this problem, we adopt the FCN [31] to segment the areas of VTire in contact with objects, of which the network structure is shown in Fig. 9.

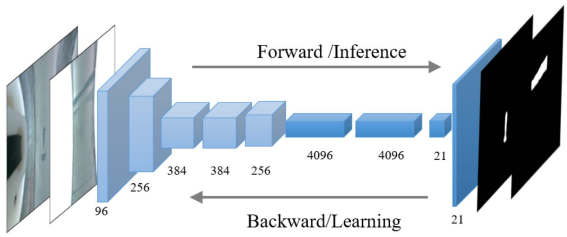

Fig. 9. Floor crack and contact object segmentation algorithm.

TABLE II
TEST RESULTS UNDER DIFFERENT MODAL AND NETWORK CONDITIONS

	TO	VO	RVT	SVT
ResNet [30]	**92.7%**/93.4%	59.0%/60.4%	82.3%/83.6%	92.4%/94.0%
LSTM [32]	86.3%/86.7%	74.4%/77.2%	84.8%/87.5%	82.4%/85.7%
ViT [33]	64.2%/64.9%	71.6%/72.2%	95.0%/95.6%	86.4%/86.9%
ExViT [34]	77.6%/79.0%	66.0%/71.2%	90.2%/91.5%	88.5%/90.1%
MMVTT	92.6%/**93.8%**	**77.9%/77.9%**	**96.7%/97.5%**	**98.1%/98.7%**

Bold font indicates the best value in each column.

V. EXPERIMENTS

A. Bimodal Terrain Classification (VTire Bimodal Data)

We collect tactile and visual data from the tire's contact with 12 different terrains to validate the effectiveness of bimodal tires and multimodal sensing networks. We capture 150 images for each terrain, as shown in Fig. 10. These terrains contain rubber tracks, painted roads, brick roads, lawns, and gravel roads made of different colored and sized stones. To better demonstrate the effectiveness of the system, we design different comparison experiments.

First, to test the effectiveness of the smart tires, we compare the classification accuracy in different modalities. We process and divide the collected dataset into the following sets.

1) *Tactile Data Only (TO):* Raw data are segmented to focus solely on the tactile region.
2) *Visual Data Only (VO):* Raw data are segmented to focus solely on the transparent and visible region.
3) *Raw VisuoTactile Data (RVT):* Raw data encompass both the tactile region and visible region.
4) *Segmented VisuoTactile Data (SVT):* Raw data are segmented into tactile region and visible region.

For all cases, we split the dataset into 70% for training and 30% for evaluation. To simulate noise caused by mud on the transparent region, we add salt-and-pepper noise to the visual modality. We train our network on an Intel(R) Xeon(R) Gold 5218 with a single GeForce RTX A6000 for 80 epochs. The learning rate is 2e-5, and we repeated the experiment on three different seeds. The training results and related indicators are shown in the last row of Table II and Fig. 10(f). It can be seen from the results that the classification method with bimodal fusion has higher accuracy, and SVT also achieves a better classification performance than RVT.

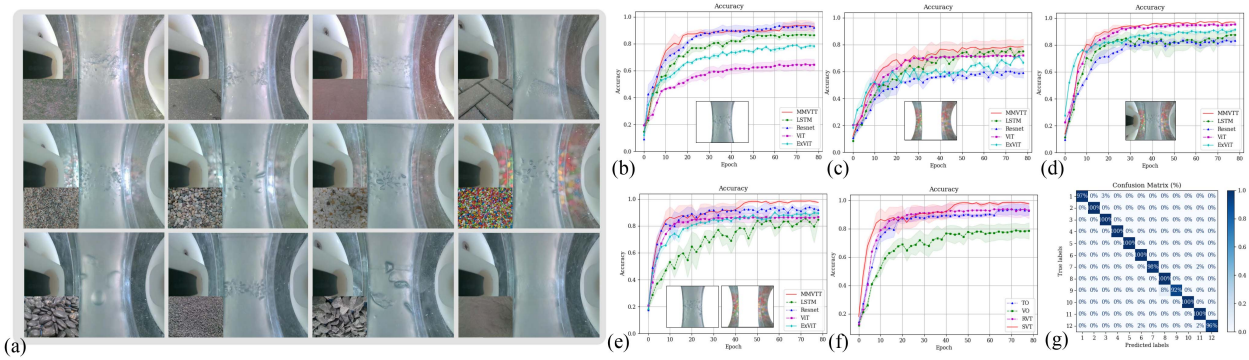

Fig. 10. Bimodal terrain classification. (a) Raw visuotactile data in different terrains. (b) Classification result of tactile data only (TO). (c) Classification result of the sensor's visual data only (VO). (d) Classification result of the raw visuotactile data (RVT). (e) Classification result of segmented visuotactile data (SVT). (f) Classification result of our proposed network in different modalities. (g) Confusion matrix of our proposed network for segmented visual image data.

Second, to further validate the effectiveness of our proposed network, we compare it with current classical classification algorithms. Specifically, we benchmark our method against two baselines as follows.

1) *ResNet:* We utilize a pretrained ResNet50 to extract global features for each modality. These global features are then concatenated and passed through a classification head.
2) *LSTM* [32]: We employ a ResNet18 to extract patched features, which are subsequently processed by an LSTM to achieve a fused feature. The fused feature is passed through a classification head to produce the final output.
3) *ViT* [33]: We use a Vision Transformer as another baseline for encoder comparison, leveraging the same transformer backbone as our method.
4) *ExViT* [34]: A multimodal classification method that replaces our ResNet encoder with a CNN.

Our method, multimodal visotactile transformer (MMVTT), is conceptually similar to the LSTM approach but replaces the recurrent structure with an attention block. To ensure a fair comparison, we design the classification heads of these networks to be as similar as possible. The parameters of MMVTT, ExVit, ViT, LSTM, and ResNet are approximately 14 M, 11 M, 14 M, 13 M, and 18 M, respectively. All three networks are trained on the multimodal dataset with a learning rate of 2e-5 for 80 epochs. In addition, we conduct the experiments using three different random seeds to minimize the likelihood of incidental results.

The training results and related indicators are shown in Table II and Fig. 10. From the results, we can see that our proposed network outperforms all baselines (Resnet, LSTM, ViT and ExViT) in both last-10-epoch-average and maximum classification accuracy in most cases, especially in bimodal classification, proving our proposed network's effectiveness in dealing with multimodal data.

B. Multimodal Terrain Classification (VTire Bimodal Data + External Visual Data)

The most common method for terrain recognition is visual processing because the visual information has a greater detection distance and range. Still, for some smoke, darkness, and other scenes, the visual detection effect will receive a great impact, but the tactile information has better stability. Therefore, we think that the visual-tactile fusion approach can solve the problem of terrain perception in complex scenes. To prove this, we collect visual information from 12 different scenes of normal, smoke, and darkness and compare it with the effect of terrain classification of smart tires. Among them, the data for the smoke scene was collected using a lens wrapped in a semitransparent film [0.5-mm thickness expanded polyethylene material]. We capture 150 images for each terrain, as shown in Fig. 11(a).

In this experiment, we consider three modalities of inputs as follows.

1) *External Visual Only (EVO):* Only external vision is used for classification (Although the smart tire itself has visual perception capabilities, its perception is often fuzzy with a limited viewing angle. Therefore, we consider adding external vision to enhance detection accuracy further.).
2) *External Visual Data + Tactile Data (EVT):* Both external vision and segmented tactile region data are employed.
3) *External Visual Data + Segmented VisuoTactile Data (EVVT):* All available modalities are utilized. The multimodal data were randomly split into training and validation datasets with a 7:3 ratio.

To assess the capability of each modality, we employed MMVTT across the three input modes mentioned, as MMVTT has demonstrated superior performance, thereby mitigating potential assessment bias that could arise from the use of less effective models. Each configuration was run with three random initializations for 80 epochs, using a learning rate of 2e-5. As illustrated in the last row of Table III and Fig. 11(b), the results demonstrate that the EVVT configuration achieves the highest last-10-epoch-average and maximum accuracy with efficient modality fusion.

Furthermore, we compare our method with the previously mentioned baselines: ResNet, LSTM, ViT, and ExViT. We utilized all modalities as input, corresponding to the EVVT input mode, and split the data into a training ratio of 0.7. The learning rate is set to 2e-5, and the training process is repeated three times. Table III and Fig. 11 present the training and evaluation

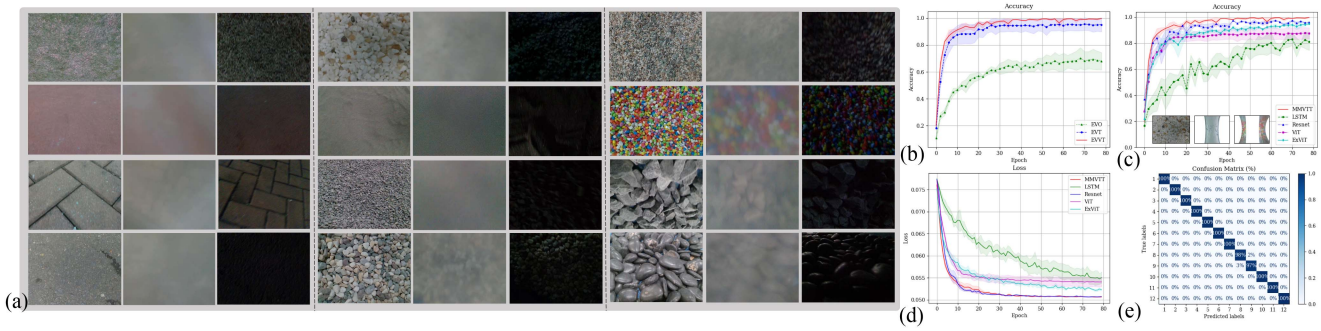

Fig. 11. Multimodal terrain classification. (a) Visual images detected by an external camera under sunny, smoky, and dark conditions. (b) Classification result of our proposed network in different modalities (EVO: external visual only, EVT: external visual data + tactile data, EVVT: external visual data + segmented visual data + tactile data.). (c) Classification result of different networks in segmented visuotactile data with EVVT. (d) Loss of different networks in segmented visuotactile data with EVVT. (e) Confusion matrix of our proposed network for EVVT.

TABLE III
TEST RESULTS UNDER DIFFERENT MODAL AND NETWORK CONDITIONS

	EVO	EVT	EVVT
ResNet	-	-	96.1%/97.5%
LSTM	-	-	79.5%/83.0%
ViT	-	-	87.4%/87.7%
ExViT	-	-	93.6%/94.8%
MMVTT	68.5%/70.2%	95.1%/95.6%	**99.2%/99.7%**

Bold font indicates the best value in each column.

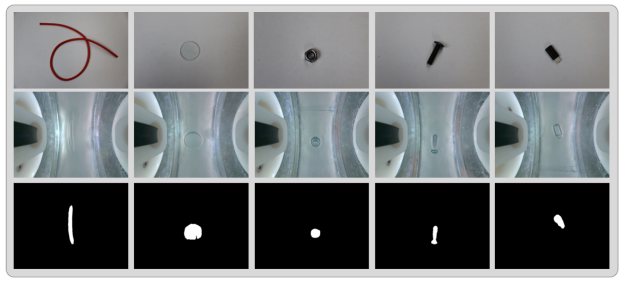

Fig. 12. Segmentation results for contacting objects. From left to right: wire, pane, nut, screwdriver, USB converter.

results. Notably, our network outperforms the other baselines and achieves an accuracy exceeding 99% . To test the classification of tires at different weights, we test five different loads (0, 10, 20, 30, 35 kg) on three different terrains. We tested the data using a classification network, and the classification effect is the same compared to the previous results.

C. Object Search Experiment on the Ground

Finding objects dropping on the ground is a big pain point for humans, especially for transparent objects. Because transparent objects have special optical properties, not only do they lack texture, but their color changes with the background. Compared with vision, although touch has a smaller detection area, it can get more stable contact information by touching objects (independent of background, and optical properties). However, VTire can solve this problem well with its powerful tactile perception ability. By combining VTire with sweeping robots and wheeled robots, they can detect the cleanliness of the floor and the presence of foreign objects as they move.

We design an object search experiment to verify the function of VTire on search. First, we collect data from five different objects, namely, rope, lens, nut, screw, and USB converter. A total of 150 images of tires in contact with objects are collected and then manually annotated. After completing the design of the dataset, we train the FCN. After 60 rounds of training, the segmentation accuracy can reach 99% , and the object search results are shown in Fig. 12.

In addition, we also conduct object search experiments in real scenarios, where objects are randomly thrown on the ground and searched using VTires. After 50 experimental tests, the success rate of finding the object successfully is 98% , indicating the potential of VTires in the field of searching.

D. Cracks Detection Experiment

Besides object searching, crack searching can be a very rewarding endeavor. Floor tiles with a variety of textures are often integrated with cracks, which can cause significant interference with visual detection [35], [36], as shown in Fig. 13(a) , while tactile perception can avoid the impact of the pattern on detection. Combining home robots with VTire can be used to inspect flooring equipment in real time, in the renovation industry to check the quality of floor coverings, and in transportation to assess road strength, integrity, and other indicators.

To test the effectiveness of VTire on crack detection, we design a crack detection experiment. First, we collect fragments of different sizes of cracks. Then, we capture 120 images of tires in contact with cracks. After that, we annotate the image at the pixel level, and unlike object searching, we mask the visuotactile part of the tires to prevent other regions from influencing the detection of the tactile region, as crack detection is much more difficult. Finally, we train the FCN using the dataset, and the results obtained are shown in Fig. 13; after 60 rounds of training, the training accuracy can reach 98% , and

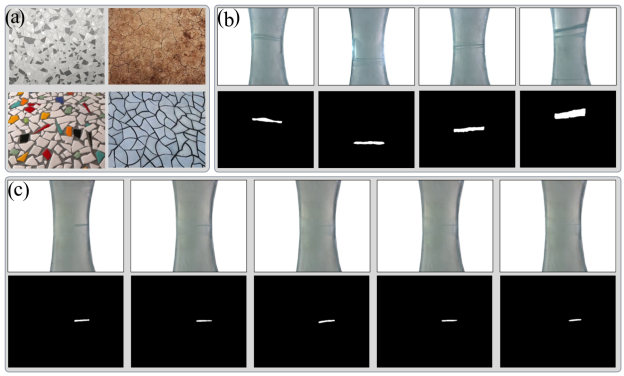

Fig. 13. Ground cracked detection results. (a) Floor tiles with a cracked decorative pattern. (b) Segmentation results in broken ground. (c) Perception effect of needles of different thicknesses. From left to right: 0.5 mm; 0.4 mm; 0.3 mm; 0.25 mm; 0.2 mm.

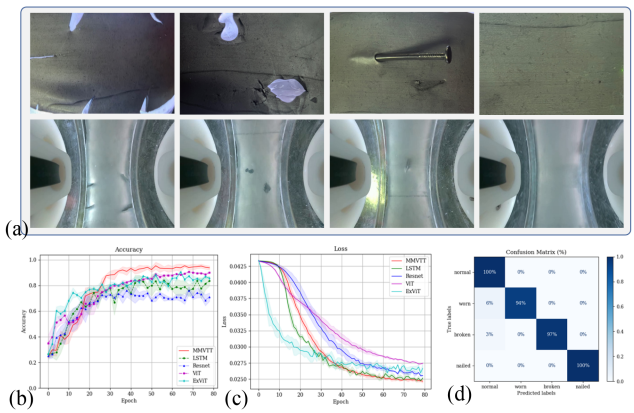

Fig. 14. Damage detection. (a) Common tire damage: cracks, irregular wear, punctures, and normal tires (from left to right). (b) Classification accuracy of different networks. (c) Loss curve of different networks. (d) Confusion matrix for the classification result of our proposed network.

the cracks search results are shown in Fig. 13(b). In addition, we use needles of different thicknesses to quantify the tactile resolution of the tires, which is tested to 0.2 mm, as shown in Fig. 13(c).

E. Tire Damage Detection Experiment

Tires are susceptible to damage due to contact with sharp objects or long-distance driving while the vehicle is driving. Compared with traditional smart tires, VTire can detect damage in real time, such as crack, abrasion, and nail penetration, thanks to high-resolution tactile data.

We experiment with different kinds of discrimination to evaluate the capability of detecting damage. We consider three types of damage: 1) cracks; 2) irregular wear; 3) punctures, and a normal state. We collect 120 images for each state of tile and split 70% for training. We also add pepper and salt noise to mimic the possible effects of the real environment. MMVTT, ExViT, ViT, LSTM, and ResNet are employed to learn from the training data. We choose a learning rate of 2e-5 and run the experiment on three random seeds for 80 epochs. Fig. 14 illustrates the training

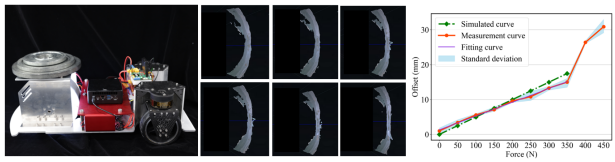

Fig. 15. Load weight perception experiment. (Left) Weight test scenario; (Middle) Depth information of tires under different loads; (Right) Corresponding curve between offset and load.

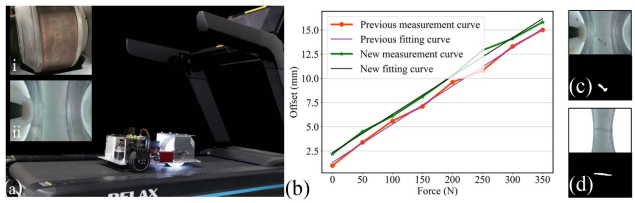

Fig. 16. Performance test experiment. (a) Durability test platform. (i) Tire appearance after testing; (ii) Internal image after testing. (b) Load test data after durability test. (c) Object search data after durability test. (d) Crack detection data after durability test.

results and related metrics. We find that MMVTT can detect the damage accurately and correctly classify the type of damage in more than 97% of cases.

F. Load Weight Perception Experiment

To evaluate the load weight perception performance of the tires, we build a test rig, as shown in Fig. 15(left), where we use dumbbell pieces as loads to test the tires' load capacity. After obtaining the load limits of the tires, we conduct load capacity perception experiments on the tires. Based on the force sensing algorithm proposed above, we fit the curve between force and offset using the equation. From Fig. 15(right), we can see that the load is linearly related to the offset when the tire is 0–35 kg, and the results are very close to the results of FEA. We test ten weights, each with five measurements, with a weight perception accuracy of 0.75 kg. After exceeding 35 kg, the tire will undergo a sudden change due to breaking the load limit of the tire.

G. Durability Test

Durability is an important indicator of a tire. However, the special properties of a specialty tire often limit its durability; for example, street car tires can last about 15 000 km, but the life of an F1 tire is between 60 and 120 km [37], [38]. To test the VTire's durability, we set up a simple test system on a treadmill. The system is driven continuously for 200 km at a speed of 10 km/h, as shown in Fig. 16. After the durability test, we reevaluated the tire's performance in terrain classification, object search, crack detection, and load sensing experiments to verify its stability and durability.

Specifically, after the 200-km durability test, the VTire maintained consistent performance in all key metrics. The terrain classification accuracy can up to 98% , the object search experiment still provided clear contour information, and the crack detection resolution was stable at 0.2 mm. For load sensing,

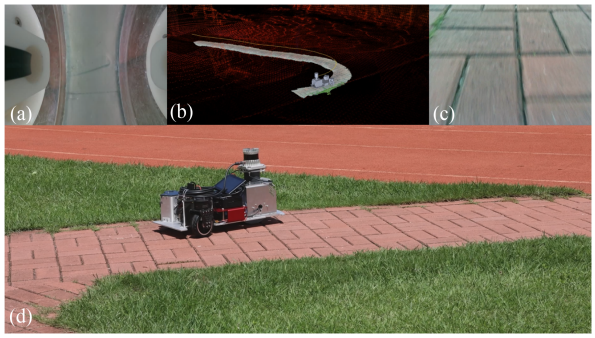

Fig. 17. Outdoor scene test experiment. (a) Visuotactile image. (b) Camera-LiDAR mapping result. (c) External visual image. (d) Test scene.

although prolonged use caused a slight shift in the fitting curve, recalibration restored the detection accuracy to its original level. These results confirm that the VTire retains high performance even after extended use. In addition, the tire adopts a modular design so that when the sensing skin is damaged, only the outermost sensing layer needs to be replaced, not the entire tire, which reduces the cost of tire maintenance.

H. Outdoor Scene Test

Besides the indoor scene, we test it in an outdoor scene, as shown in Fig. 17. We implement mapping using the RealSense D435i fused with LiDAR and employ the multimodal classification algorithm to detect the terrains. We test terrains, such as grass, road, tartan track, and brick, achieving good detection, and mapping results. (Details can be referred to the video file.)

VI. Conclusion

This article proposes a bimodal smart visuotactile trie named VTire, which can obtain large-area and high-resolution tactile and visual data. We optimize the material and structure to obtain highly elastic, durable, and transparent tires and propose a complete set of preparation processes. The tires can withstand loads exceeding 35 kg and have a durability of over 200 km. Algorithmically, we propose a transformer-based multimodal classification algorithm, a load detection based on FEA, and a contact segmentation algorithm. In addition, to validate the algorithm's performance, we build a mobile platform and propose a set of visual and tactile datasets in different terrain and visibility situations. After experimental validation, the accuracy of our proposed multimodal classification method can reach 99.2% in terrain classification, 98% success rate in object search, and 0.75-kg accuracy in weight sensing, which proves that VTire has a high application value in terrain sensing and object searching scenarios.

The tire also has some limitations. Due to our single-camera solution, the tire is unsuitable for high-speed scenes. Future work will focus on integrating advanced vision systems, applying the tire to wheeled robots, and exploring new materials, real-time algorithms, and autonomous robot integration for complex terrains.

References

[1] B. Yang, Q. Sun, R. Fu, C. Wang, Y. Guo, and L. Zhou, "A model predictive control-based electronic differential control strategy for distributed-drive buses considering the reduction of tire wear," *IEEE/ASME Trans. Mechatron.*, early access, Aug. 26, 2024, doi: 10.1109/TMECH.2024.3440312.
[2] Y. Wang, T. Chen, X. Rong, G. Zhang, Y. Li, and Y. Xin, "Design and control of SKATER: A wheeled-bipedal robot with high-speed turning robustness and terrain adaptability," *IEEE/ASME Trans. Mechatron.*, vol. 30, no. 2, pp. 1310–1321, Apr. 2025.
[3] H. Lee and S. Taheri, "Intelligent tires? A review of tire characterization literature," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 2, pp. 114–135, Summer 2017.
[4] N. Xu, H. Askari, Y. Huang, J. Zhou, and A. Khajepour, "Tire force estimation in intelligent tires using machine learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 4, pp. 3565–3574, Apr. 2022.
[5] D. Maurya et al., "3D printed graphene-based self-powered strain sensors for smart tires in autonomous vehicles," *Nature Commun.*, vol. 11, no. 1, 2020, Art. no. 5392.
[6] A. J. Tuononen, "Laser triangulation to measure the carcass deflections of a rolling tire," *Meas. Sci. Technol.*, vol. 22, no. 12, 2011, Art. no. 125304.
[7] A. Tuononen, "On-board estimation of dynamic tyre forces from optically measured tyre carcass deflections," *Int. J. Heavy Veh. Syst.*, vol. 16, no. 3, pp. 362–378, 2009.
[8] R. Matsuzaki, N. Hiraoka, A. Todoroki, and Y. Mizutani, "Optical 3D deformation measurement utilizing non-planar surface for the development of an "intelligent tire"," *J. Solid Mechan. Materials Eng.*, vol. 4, no. 4, pp. 520–532, 2010.
[9] A. C. Abad and A. Ranasinghe, "Visuotactile sensors with emphasis on gelsight sensor: A review," *IEEE Sensors J.*, vol. 20, no. 14, pp. 7628–7638, Jul. 2020.
[10] C. Lu, K. Tang, M. Yang, T. Yue, H. Li, and N. F. Lepora, "DexiTac: Soft dexterous tactile gripping," *IEEE/ASME Trans. Mechatron.*, vol. 30, no. 1, pp. 333–344, Feb. 2025.
[11] S. Cui et al., "Gelstereo BioTip: Self-calibrating bionic fingertip visuotactile sensor for robotic manipulation," *IEEE/ASME Trans. Mechatron.*, vol. 29, no. 4, pp. 2451–2462, Aug. 2024.
[12] K. B. Singh and S. Taheri, "Estimation of tire–road friction coefficient and its application in chassis control systems," *Syst. Sci. Control Eng.*, vol. 3, no. 1, pp. 39–61, 2015.
[13] B. H. G. Barbosa, N. Xu, H. Askari, and A. Khajepour, "Lateral force prediction using Gaussian process regression for intelligent tire systems," *IEEE Trans. Syst., Man, Cybern. Syst.*, vol. 52, no. 8, pp. 5332–5343, Aug. 2022.
[14] H. Oh, K. Lee, K. Eun, S.-H. Choa, and S. S. Yang, "Development of a high-sensitivity strain measurement system based on a SH SAW sensor," *J. Micromechanics Microengineering*, vol. 22, no. 2, 2012, Art. no. 025002.
[15] X. Zhang, Z. Wang, L. Gai, Y. Ai, and F. Wang, "Design considerations on intelligent tires utilizing wireless passive surface acoustic wave sensors," in *Proc. 5th World Congr. Intell. Control Autom.*, 2004, pp. 3696–3700.
[16] J. Yi, "A piezo-sensor-based "smart tire" system for mobile robots and vehicles," *IEEE/ASME Trans. Mechatron.*, vol. 13, no. 1, pp. 95–103, Feb. 2008.
[17] X. Sun, Z. Quan, Y. Cai, L. Chen, and B. Li, "Direct tire slip angle estimation using intelligent tire equipped with PVDF sensors," *IEEE/ASME Trans. Mechatron.*, vol. 30, no. 2, pp. 1190–1200, Apr. 2025.
[18] M. F. Mendoza-Petit, D. García-Pozuelo, V. Díaz, and O. Olatunbosun, "A strain-based intelligent tire to detect contact patch features for complex maneuvers," *Sensors*, vol. 20, no. 6, 2020, Art. no. 1750.
[19] J. Yunta, D. Garcia-Pozuelo, V. Diaz, and O. Olatunbosun, "Influence of camber angle on tire tread behavior by an on-board strain-based system for intelligent tires," *Measurement*, vol. 145, pp. 631–639, 2019.
[20] R. Gubaidullin, T. Agliullin, O. Morozov, A. Z. Sahabutdinov, and V. Ivanov, "Microwave-photonic sensory tire control system based on FBG," in *Proc. Syst. Signals Generating Process. Field Board Commun.*, 2019, pp. 1–6.
[21] R. G. Longoria, R. Brushaber, and A. Simms, "An in-wheel sensor for monitoring tire-terrain interaction: Development and laboratory testing," *J. Terramechanics*, vol. 82, pp. 43–52, 2019.
[22] L. Hu et al., "Terrain classification using inside-wheel cameras based on wheel-terrain interaction characteristics," in *Proc. IEEE Int. Conf. Robot. Biomimetics*, 2023, pp. pp. 1–6.
[23] H.-J. Kim et al., "A road condition classification algorithm for a tire acceleration sensor using an artificial neural network," *Electronics*, vol. 9, no. 3, 2020, Art. no. 404.

[24] S. Khaleghian and S. Taheri, "Terrain classification using intelligent tire," *J. Terramechanics*, vol. 71, pp. 15–24, 2017.
[25] K. Eun, K. J. Lee, K. K. Lee, S. S. Yang, and S.-H. Choa, "Highly sensitive surface acoustic wave strain sensor for the measurement of tire deformation," *Int. J. Precis. Eng. Manuf.*, vol. 17, pp. 699–707, 2016.
[26] J. Jiang, G. Cao, A. Butterworth, T.-T. Do, and S. Luo, "Where shall i touch? Vision-guided tactile poking for transparent object grasping," *IEEE/ASME Trans. Mechatron.*, vol. 28, no. 1, pp. 233–244, Feb. 2023.
[27] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.
[28] J. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.
[29] Y. Chen, A. Sipos, M. V. der Merwe, and N. Fazeli, "Visuo-tactile transformers for manipulation," in *Proc. Conf. Robot Learn.*, 2022, pp. 2026–2040.
[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 770–778.
[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.
[32] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W.-K. Wong, and W. Chun Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Neural Inf. Process. Syst.*, 2015, pp. 802–810.
[33] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[34] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.
[35] K. Hu, Z. Chen, H. Kang, and Y. Tang, "3D vision technologies for a self-developed structural external crack damage recognition robot," *Automat. Construction*, vol. 159, 2024, Art. no. 105262.
[36] Y. Tang, S. Qi, L. Zhu, X. Zhuo, Y. Zhang, and F. Meng, "Obstacle avoidance motion in mobile robotics," *J. Syst. Simul.*, vol. 36, no. 1, pp. 1–26, 2024.
[37] Durability of tires, Feb. 5, 2019, [Online]. Available: https://www.deccanherald.com/sports/f1-racing/lasting-little-60km-tyres-are-716735.html
[38] S. L. Weissman, J. L. Sackman, D. Gillen, and C. Monismith, "Extending the lifespan of tires," Sympletic Eng. Corporation, Berkeley, CA, USA, 2003.

Shoujie Li (Graduate Student Member, IEEE) 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 working toward the Ph.D. degree with Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

His research interests include tactile perception, grasping, and machine learning.

Jianle Xu received the B.S. degree in agricultural mechanization and automation from Hainan University, Hainan, China, in 2023. He is currently working toward the M.S. degree with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

His research interests include robot dexterous hands and electronic devices.

Tong Wu received the B.S. degree in electronic engineering, in 2023, from Tsinghua University, Beijing, China, where he is currently working toward the Ph.D. degree with Institute of Data and Information, Shenzhen International Graduate School, Tsinghua, Shenzhen, China.

His research interests include robot manipulation, multimodal sensing, and embodied intelligence.

Yang Yang is currently working toward the B.S. degree in engineering mechanics with Sichuan University, Chengdu, China.

His research interests include tactile sensing and robotic dexterous manipulation.

Yanbo Chen received the B.S. degree in automation from the Harbin Institute of Technology, Shenzhen, China, in 2023. He is currently working toward the M.S. degree with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen.

His research interests include SLAM and autonomous navigation for robots.

Xueqian Wang (Member, IEEE) received the B.E. degree in mechanical design, manufacturing, and automation from the Harbin University of Science and Technology, Harbin, China, in 2003, the M.Sc. degree in mechatronic engineering and the Ph.D. degree in control science and engineering from the Harbin Institute of Technology (HIT), Harbin, China, in 2005 and 2010, respectively.

From 2010 to 2014, he was the Postdoc Research Fellow with the HIT. He is currently a Professor and the Leader of the Center of Intelligent Control and Telescience, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. His research interests include dynamics modeling, control, and teleoperation of robotic systems.

Wenbo Ding (Member, IEEE) received the B.S. and Ph.D. 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 Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, where he leads the Smart Sensing and Robotics group. His research interests are diverse and interdisciplinary, which include self-powered sensors, energy harvesting, and wearable devices for health and robotics with the help of signal processing, machine learning, and mobile computing.

Dr. Ding was the recipient of many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva and the IEEE Scott Helt Memorial Award.

Xiao-Ping Zhang (Fellow, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, Beijing, China, in 1992 and 1996, respectively, and the MBA (Hons.) degree in finance, economics and entrepreneurship from the University of Chicago Booth School of Business, Chicago, IL, USA, in 2008.

He was the founding Dean of the Institute of Data and Information with Tsinghua Shenzhen International Graduate School (SIGS), where he is currently a Chair Professor. 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.