Visual–Tactile Fusion for Transparent Object Grasping in Complex Backgrounds

Shoujie Li, *Student Member, IEEE*, Haixin Yu, Wenbo Ding, *Member, IEEE*, Houde Liu, *Member, IEEE*, Linqi Ye, *Member, IEEE*, Chongkun Xia, *Member, IEEE*, Xueqian Wang, *Member, IEEE*, and Xiao-Ping Zhang, *Fellow, IEEE*

***Abstract*—The grasping of transparent objects is challenging but of significance to robots. In this article, a visual–tactile fusion framework for transparent object grasping in complex backgrounds is proposed, which synergizes the advantages of vision and touch, and greatly improves the grasping efficiency of transparent objects. First, we propose a multiscene synthetic grasping dataset named SimTrans12 K together with a Gaussian-mask annotation method. Next, based on the TaTa gripper, we propose a grasping network named transparent object-grasping convolutional neural network for grasping position detection, which shows good performance in both synthetic and real scenes. Inspired by human grasping, a tactile calibration method and a visual–tactile fusion classification method are designed, which improve the grasping success rate by 36.7% compared with direct grasping and the classification accuracy by 39.1%. Furthermore, the tactile height sensing module and the tactile position exploration module are added to solve the problem of grasping transparent objects in irregular and visually undetectable scenes. The experimental results demonstrate the validity of the framework.**

***Index Terms*—Complex backgrounds, synthetic transparent object dataset, tactile calibration, transparent object grasping, visual–tactile fusion.**

Manuscript received 4 February 2023; revised 28 April 2023; accepted 30 May 2023. This work was supported in part by the National Key R&D Program of China under Grant 2022YFB4701400/4701402, in part by the 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 Shenzhen Science Fund for Distinguished Young Scholars under Grant RCJC20210706091946001, in part by the Guangdong Special Branch Plan for Young Talent With Scientific and Technological Innovation under Grant 2019TQ05Z111, in part by the National Natural Science Foundation of China under Grant 92248304, Grant U21B6002, Grant 62203260, Grant 62003188, and Grant 62104125, in part by the Natural Science Foundation of Guangdong Province under Grant 2023A1515011773, in part by the China Postdoctoral Science Foundation under Grant 2022M711823, and in part by the Tsinghua Shenzhen International Graduate School-Shenzhen Pengrui Young Faculty Program of Shenzhen Pengrui Foundation under Grant SZPR2023005. This paper was recommended for publication by Associate Editor J. Kelly and Editor M. Yim upon evaluation of the reviewers' comments. *(Shoujie Li and Haixin Yu contributed equally to this work.) (Corresponding authors: Wenbo Ding; Houde Liu.)*

Shoujie Li and Wenbo Ding are with Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China, and also with RISC-V International Open Source Laboratory, Tsinghua-Berkeley Shenzhen Institute, Shenzhen 518055, China (e-mail: lsj20@mails.tsinghua.edu.cn; ding.wenbo@sz.tsinghua.edu.cn).

Haixin Yu, Houde Liu, Chongkun Xia, and Xueqian Wang are with Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China (e-mail: yuhx21@mails.tsinghua.edu.cn; liu.hd@sz.tsinghua.edu.cn; xiachongkun@qq.com; wang.xq@sz.tsinghua.edu.cn).

Linqi Ye is with Institute of Artificial Intelligence, Collaborative Innovation Center for the Marine Artificial Intelligence, Shanghai University, Shanghai 200444, China (e-mail: yelinqi@shu.edu.cn).

Xiao-Ping Zhang is with Tsinghua Shenzhen International Graduate School, Shenzhen 518055, China, and also with the Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: xzhang@ee.ryerson.ca).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TRO.2023.3286071.

Digital Object Identifier 10.1109/TRO.2023.3286071

I. Introduction

TRANSPARENT objects are common in people's daily life, but it is very challenging for robots to accurately detect and grasp them. This is mainly because the appearance of transparent objects changes drastically under different backgrounds, making traditional visual detection prone to failure. Therefore, how to realize the accurate and robust detection of transparent objects toward efficient grasp has attracted tremendous interest in the field of robotics. Representative works include the multimodal transfer learning method [1], the plenoptic sensing approach [2], the transparent depth information complimenting method [3], etc. Nevertheless, these methods usually focus on the detection of transparent objects and assume that the objects are placed in static backgrounds with simple patterns, which is not always the case in practice. Hence, it is of great significance to develop a grasping method for transparent objects that can adapt to various backgrounds, e.g., the objects placed on soft or fluid surfaces, with complex patterns or unpredictable conditions, such as undulating scenes, underwater, and so on.

Thanks to the development of computer vision and deep learning, vision-assisted perception has now become a popular and effective choice for robot interactions and environment explorations. However, the vision-based method cannot work well in dim, reflective, and cloudy conditions. Inspired by the grasping behavior of humans, as shown in Fig. 1, where visual and tactile sensations are collaboratively working toward complicated tasks, a visual–tactile fusion-based framework using the TaTa gripper [4] is proposed in this article for transparent object grasping in complex backgrounds. Here, the tactile sensation is utilized to compensate for the limitation of vision, which not only largely raises the success rate of grasping by 36.7% but also greatly improves the classification accuracy of transparent objects by 39.1%. In addition, the framework can be extended to cover more challenging scenes, such as irregular backgrounds or even visually undetectable scenes. Specifically, the contributions of this work are fourfold.

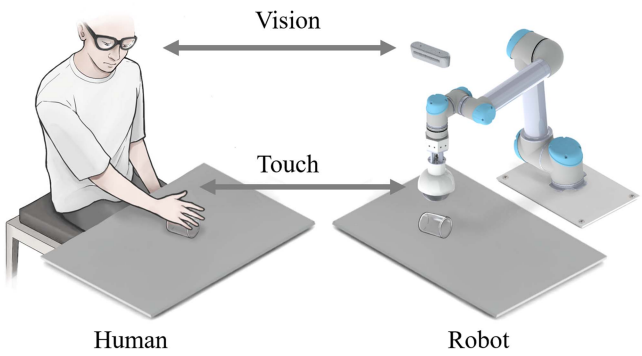

Fig. 1. Visual–tactile fusion framework inspired by human grasping.

Fig. 2. Examples of transparent object dataset. (a) ClearGrasp [3]. (b) Dex-NeRF [5]. (c) LIT [13]. (d) Light-field camera used in LIT dataset.

1) A synthetic transparent object dataset named SimTrans12 K is proposed containing different styles of backgrounds, lighting, and camera positions, which has more complex and abundant background information than the previous transparent object datasets, such as ClearGrasp [3] and Dex-Nerf [5]. Besides, to improve the performance of Sim2Real, we propose a Gaussian-mask method for transparent object-grasping position annotation, which can better represent the position information of transparent objects than the binary ground truth grasping position [1].
2) For the TaTa gripper [4], a generative grasping network named transparent object-grasping convolutional neural network (TGCNN) is proposed, which can achieve transparent object-grasping position detection in complex backgrounds and lighting with training from the synthetic dataset only. Meanwhile, a tactile information extraction algorithm and a visual–tactile fusion-based transparent object classification algorithm are developed to compensate for the visual deviation [6].
3) To realize transparent object grasping in complex backgrounds, we propose a visual–tactile fusion-based transparent object-grasping framework with tactile calibration. Besides, we add the tactile height sensing (THS) module and the tactile position exploration (TPE) module to this framework, which can achieve transparent object grasping in stacking, overlapping, or even visually undetectable scenes. Those scenes are extremely difficult and there are only a few studies before [3], [5], [7], [8], [9].
4) To test the effectiveness of the proposed framework, we carefully design several experiments to extensively compare the performance with several state-of-the-art baseline methods, which indicates that the proposed method has a considerable performance improvement for transparent object grasping and classification. Moreover, we also test the proposed method in some highly difficult scenes, such as stacking, overlapping, undulating, and dynamic underwater environments, which greatly extends the application areas of transparent object grasping.

The rest of this article is organized as follows. The related work is reviewed in Section II. The hardware setup is detailed in Section III. Section IV presents the synthetic data generation, the grasping position detection algorithm, the tactile information extraction algorithm, and the visual–tactile fusion-based classification algorithm. The proposed visual–tactile fusion grasping strategy is presented in Section V. Furthermore, experimental validations are provided in Section VI. Finally, Section VII concludes this article.

II. Related Work

A. Transparent Object Dataset

Xie et al. [10] proposed a transparent dataset Trans 10 K with 10 428 real data, but it only has two limited categories, which was further refined to 11 fine-grained categories of transparent objects in the dataset Trans10K-v2 [11]. Jiang et al. [12] constructed a real-world dataset TRANS-AFF with affordances and depth maps of transparent objects.

With the development of powerful computer graphics simulation tools, researchers have tried to generate the synthetic dataset of the transparent object from simulation considering its low cost, simplicity, and efficiency. Representative works include ClearGrasp [3], a synthetic dataset for depth-completion tasks, Dex-NeRF [5], a synthetic dataset for transparent object detection and localization, and LIT [13], a synthetic dataset for light-field cameras, as shown in Fig. 2.

The implementation of transparent object grasping via Sim2Real puts higher demands on the diversity and validity of the dataset, so we hope that the dataset contains more transparent object data under complex backgrounds and lightness, while ClearGrasp and Dex-NeRF can hardly meet such requirements. Although LIT contains more complex scenes, it contains scenes with low brightness and is designed for light-field cameras, as shown in Fig. 2(d). A new synthetic dataset for the transparent object is proposed in this article. To reduce the discrepancy between the synthetic data and the real scene, we carefully calibrate the parameters of the cameras, lights, and backgrounds in the simulation software to make them as consistent as possible with the real camera, which is neglected in the existing

transparent object synthetic dataset. In addition, we also propose a Gaussian-mask annotation method for transparent objects.

B. Transparent Object Detection

The visual detection methods of transparent objects can be divided into two types: physical feature-based detection methods and deep-learning-based methods.

Traditional methods detect transparent objects based on physical features, such as deformation, reflection, and image gradient changes. Fritz et al. [14] reported an additive latent feature model through the assumption that the texture of transparent objects originates from the background. McHenry et al. [15] proposed a hierarchical support vector machine based glass edge recognition model via the background texture distortion and reflection phenomenon at the glass edge. Maeno et al. [16] used a light-field camera to acquire images and utilized a light-field distortion feature to describe the distortion caused by the refraction of transparent objects.

The development of deep learning paves a new way for transparent object detection. Liu et al. [17] used a convolutional neural network called single shot multibox detector for transparent object detection. Xie et al. [10] proposed a transformer-based segmentation pipeline termed Trans2Seg. Fan et al. [18] applied the transparent object detection to highly dynamic scenes and proposed a recognition tracking network named TransATOM, which can stably track the transparent objects in video. Xu et al. [19] proposed a joint point cloud and depth-completion method, which can complete the depth of transparent objects in cluttered scenes. Zhu et al. [20] presented a novel framework that can complete missing depth given noisy red, green, and blue (RGB)-D inputs.

Deep-learning methods have demonstrated superior robustness to traditional ones, especially for transparent object detection in complex scenarios, showing great application potential.

C. Transparent Object Grasping

Transparent object grasping is another challenging task. Apart from the object position, the optimal grasping position and angle should be considered as well during grasping. We classify transparent object-grasping tasks into different levels of difficulty, as shown in Table I, ranging from the simple case of grasping on a plane to the extremely difficult case of grasping in dynamic underwater environments.

For transparent object grasping, most of the works are performed on planes with a simple background. For example, Weng et al. [1] proposed a multimodal transfer learning method for transparent and reflective object grasping. Sajjan et al. [3] reported a transparent depth-completion method to grasp transparent objects. Liu et al. [21] proposed a keypoint-based method for six-dimensional (6-D) pose estimation of objects using stereo image input, which can easily be applied to transparent object grasping. Ichnowski et al. [5] rendered depth maps of transparent objects using neural radiation fields (NeRF) to infer the geometry of transparent objects and perform plane grasping. Kerr et al. [22] proposed evolving NeRF, leveraging recent speedups in NeRF training and further extending it to rapidly train the NeRF representation concurrently to image capturing. Cao et al. [23] proposed a fuzzy-depth soft grasping algorithm for Tstone-Soft gripper.

TABLE I
RELATED WORK ON TRANSPARENT OBJECT GRASPING

Scenes	Difficulty	Related work
Plane	Normal	Weng et al. [1], Sajjan et al. [3], Liu et al. [21], Ichnowski et al. [5], Kerr et al. [22], Cao et al. [23], Jiang et al. [12], Zhou et al. [13]
Fragments grasping	Medium	N/A
Stacking and overlapping		Zhou et al. [2], Lysenkov et al. [24]
Undulating	High	N/A
Sand		N/A
Underwater		Oberlin et al. [25], Zhou et al. [26]
Highly dynamic underwater	Very high	N/A

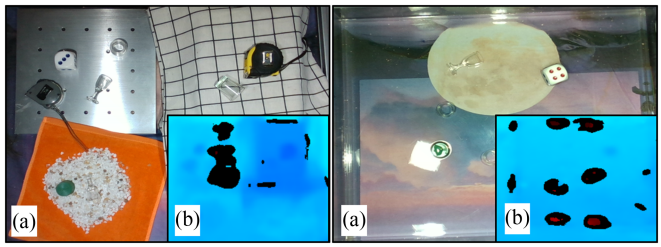

Fig. 3. Detection with RGB and depth cameras. Left: undulating scenes: (a) RGB; and (b) depth images, Right: underwater scenes: (a) RGB and (b) depth images.

Besides, grasping transparent objects in complex scenes, e.g., glass fragments, stacking, overlapping, undulating, sand, and underwater scenes, are more challenging but of practical meaning. First, glass fragments are a type of object with no fixed model, and their detection and grasping pose a significant challenge. Because of its random shape and the presence of more angles, the accuracy of grasping and the universality of grasping tools are highly required, so there is almost no research on the grasping of transparent glass fragments. For overlapping and stacking transparent objects, their texture will be merged with the background, so it is difficult to distinguish them. Zhou et al. [2] proposed a GlassLoc algorithm for grasping pose detection of transparent objects in clutter using plenoptic sensing, which achieves transparent object grasping in stacking scenes, although the experimental setting is simple. Lysenkov and Rabaud [24] proposed a method that can achieve pose estimation for cluttered transparent objects in complex backgrounds, but it applies the pose matching method, which is only applicable to objects within the dataset and cannot solve the grasping of transparent fragments without regular shapes.

Second, object grasping on undulating planes is difficult with RGB cameras because it is hard to estimate the height where the object is placed. As shown in Fig. 3, even by incorporating depth cameras, such a problem cannot be well solved for transparent object grasping. The reason is mainly bifold: on the one hand, the

depth information for transparent objects is inaccurate, and on the other hand, undulating scenes have some shadows, overlaps, and reflective areas, which raises more challenges for transparent object detection. Therefore, the lack of accurate information about transparent objects and the interference of the environmental background is a difficult problem to solve, which is one of the reasons why we use RGB images rather than depth images to achieve transparent grasping position detection in complex backgrounds. Sand is a special undulating scene. In addition to the above problems, its surface is more uneven, where sand particles of different colors will also influence the detection, and sand is also easy to slide during the grasping process. To our knowledge, there is still a lack of studies or experiments for transparent object grasping on undulating scenes or sand environments.

Third, transparent object grasping in underwater scenes is also challenging due to the similar optical properties of water and transparent objects. As shown in Fig. 3, even with a depth camera, transparent objects are still undetectable in water, and there are many reflections on the surface of the water under the illumination of light, making things worse. To solve the problem of underwater transparent object detection, Zhou et al. [26] proposed the plenoptic Monte Carlo localization method for localizing the pose of a translucent object underwater using a Lytro first-generation light-field camera. Similarly, Oberlin and Tellex [25] proposed a formal model for robotic light-field photography, which can turn a calibrated eye-in-hand camera into a time-lapse light-field camera. Although this method can be deployed on conventional cameras, thousands of RGB images from different angles need to be captured for one detection. Furthermore, these methods have not been studied for some dynamic underwater scenes with bubbles, waves, reflections, and complex backgrounds, which are extremely difficult, and even using a light-field camera may probably fail.

In summary, most existing studies focus on grasping transparent objects with known shape in simple scenes, such as on a plane, while several difficult scenes, as listed in Table I, are rarely studied and still remain an open problem.

III. Hardware Setup

The human hand has sensitive tactile perception because its surface is covered with dense tactile nerves [27]. Similarly, various tactile sensors have been designed for robots, such as piezoelectric [28], capacitive [29], triboelectric [30], and piezoresistive sensors [31], but the resolution is still not comparable to human hands. Thanks to the commercialization and miniaturization of the CMOS image sensors, a series of tactile detection devices based on the optical imaging are invented and realize high-resolution sensing with low costs, e.g., GelSight [32], GelSlim [33], and FingerVision [34]. However, such devices are mainly designed for fingertips and cannot acquire the overall contour of the contacted object. In addition, they usually adopted a silicone plus transparent acrylic sheet solution, which has limited deformation capability.

To realize transparent object grasping, a universal soft gripper named TaTa is adopted here, as shown in Fig. 4(a)–(d), which has tactile perception on a large hemispherical surface. Details of the TaTa gripper can be found in our previous article [4]. Meanwhile, we upgrade the previous version of the TaTa gripper by using the camera with a larger imaging range and improving the waterproof ability to achieve better detection performance and durability. TaTa adopts the grasping principle of particle jamming and vision-based tactile detection technology, using the principle of refractive index matching to design a special solid–liquid mixture that looks totally transparent [35], overcoming the interference of internal particles on the internal camera. Hence, it has large-area, high-quality tactile detection ability as well as adaptive grasping ability.

The hardware setup is depicted in Fig. 4(b) and (c). A RealSense D435i camera is fixed on the top frame as the "eye," which can acquire 480×640 image information and the TaTa gripper is attached to the UR5 robotic arm. Two LEDs are used to provide lighting to the platform. We divide the system into five coordinate systems and use the center of the gripping plane as the origin of the world coordinate system O_1. First, we calibrate the intrinsics and extrinsics of the eye camera with a checkerboard [36] to establish the relationship between the camera coordinate system O_3 and the world coordinate system O_1. Second, since the robot arm and the gripping plane are at the same height, the relationship between the arm base coordinate system O_2 and the world coordinate system O_1 is derived by coordinate transformation. Third, the position of the end of the gripper in O_2 can be obtained through the official program interface of the UR5 robot arm so that the arm can be controlled to reach the location of the transparent object captured by the eye camera. Finally, taking the gripper center as the origin and establishing the relationship between the tactile camera coordinate system O_5 of the tactile sensor and the coordinate system O_4 of the gripper end, we can get the offset of the position where the contact between the gripper and the object occurs relative to the origin of the gripper, after which the offset is mapped to the displacement of the end of the arm to achieve tactile calibration.

To verify the capability of handling fragile objects, tests on grasping an egg and a tomato with TaTa are conducted, as shown in Fig. 4(e). Meanwhile, we upgrade the problem of the small imaging range of the previous version of the TaTa gripper by using the camera with a larger imaging range and improving the waterproof ability to achieve better detection performance and durability.

IV. Methodology

This section introduces the algorithms used in our proposed visual–tactile fusion grasping framework. As shown in Fig. 5, to achieve transparent object grasping, we propose a transparent object-grasping position detection algorithm, a tactile information extraction algorithm, and a visual–tactile fusion classification algorithm, respectively. Besides, a Gaussian-mask annotation method is also developed for our synthesized transparent object dataset.

A. Dataset Generation and Annotation

The neural-network-based grasping position detection method requires a large dataset, and hence, it is challenging

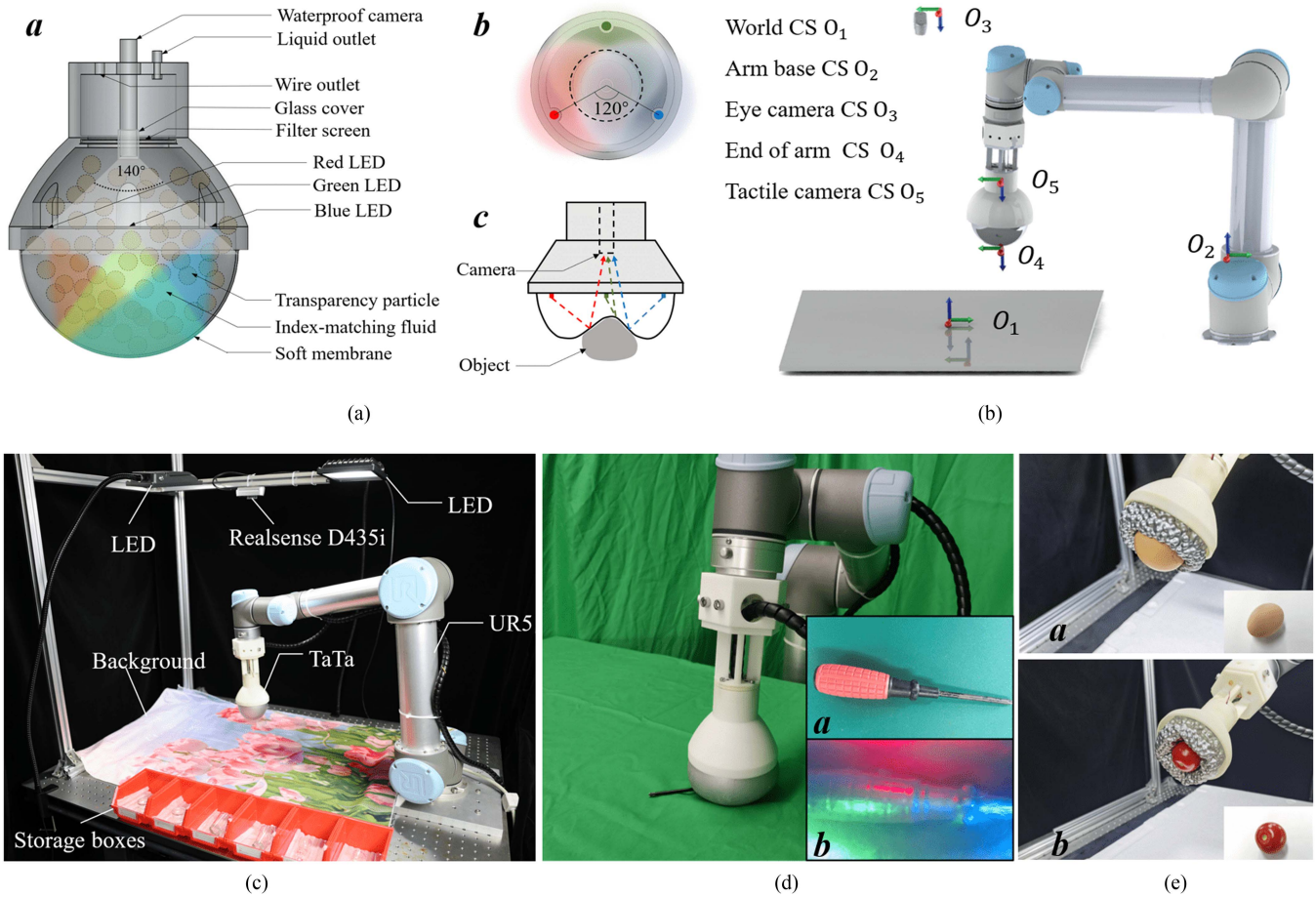

Fig. 4. Hardware system. (A) The structure of TaTa: (a) The schematic diagram of TaTa, (b) The layout of the inside LEDs, (c) The illustration of the inside light path. (B) Coordinate system (CS). (C) Visual–tactile fusion grasping experimental platform. (D) Tactile perception effect test: (a) Screwdriver picture, (b) Perception result. (E) Grasping performance testing: (a) Grasp an egg, (b) Grasp a tomato.

to collect and annotate datasets manually. To tackle this problem, we adopt Blender to make a multibackground transparent object-grasping dataset, SimTrans12 K, which contains 12 000 synthetic images and 160 real images, as illustrated in Fig. 6. The reason to choose Blender is due to its high flexibility and capability to simulate the key features of transparent objects, such as surface reflections, refraction, and soft shadows.

SimTrans12 K contains six types of objects and 2000 different scenes. To obtain sufficiently complex and adequate backgrounds, we cropped some images from videos containing rich home decoration layouts and landscapes as backgrounds. The scene setup for generating the transparent object synthetic dataset is shown in Fig. 6(a). We use Blender 2.90 s physically based cycles renderer with path tracing set to 256 samples pixel, and max light path bounces set to 1024. For glass materials, we set the index of refraction to 1.45 to match the physical glass. In each scene, two light sources are used to illuminate the location of the object and generate reflection spots on the surface of the object. The maximum power of the light is 1000 W and the minimum power is 100 W. A camera is placed above the transparent object, and the acute angle between the camera's optical axis and the z-axis of the world coordinate is varied in the range $[0, \pi/24]$. Camera intrinsics are set the same as the RealSense D435i camera.

Based on the information obtained during rendering, ground truth labels are generated for training. Generative models, such as the generative grasping CNN (GGCNN) [37], rely on the same binary ground truth generation during training. However, binary ground truth labels treat the edge and the center of annotation with the same weight, which is easy to make the grasping position deviate from the optimal center and come to the edge of the object. It can lead to a declined grasping success rate and even damage the object. To improve the reliability of dataset annotation, we propose a transparent object-grasping position annotation method based on Gaussian distribution and the transparent object mask (Gaussian-mask). Previously, Cao et al.[38] proposed the idea of using Gaussian distribution for object-grasping position annotation. A Gaussian distribution rectangular box was adopted for the annotation of the gripping position, which works well for ordinary objects. However, the grasping position detection of transparent objects is much more challenging since the texture of transparent objects changes with the backgrounds. Fortunately, although the texture properties of transparent objects may change dramatically, their boundary

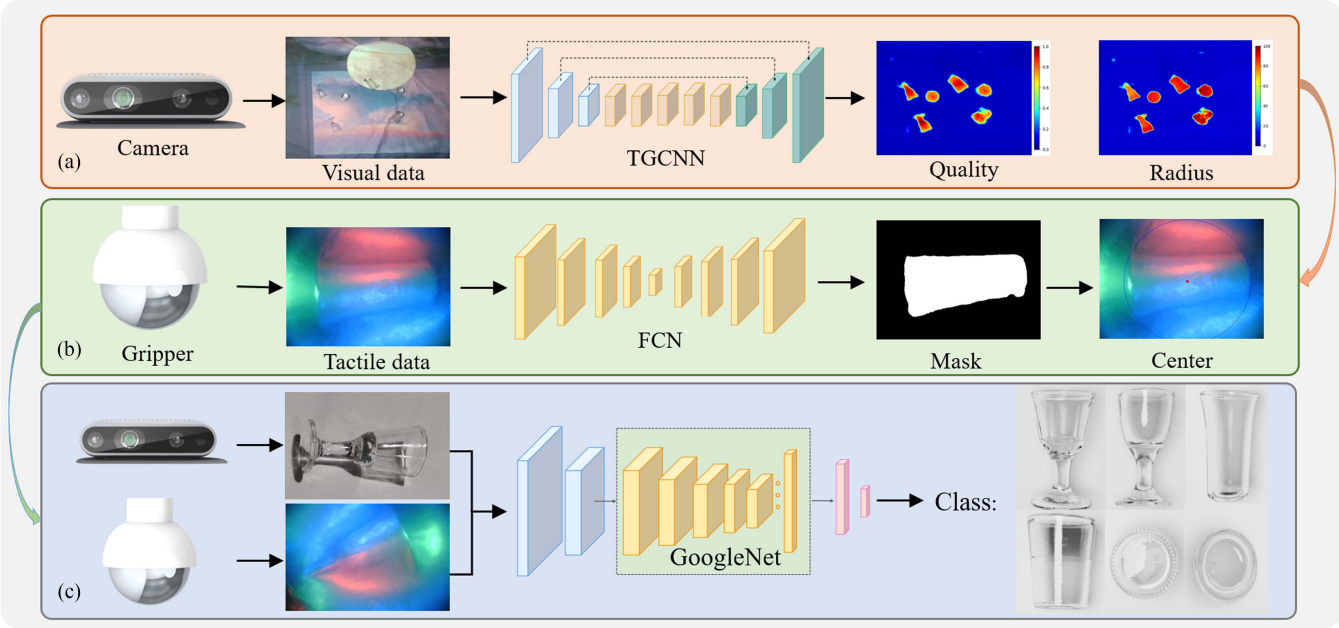

Fig. 5. Visual–tactile fusion framework for transparent object grasping. (a) Grasping position detection. (b) Tactile information extraction. (c) Visual–tactile fusion classification.

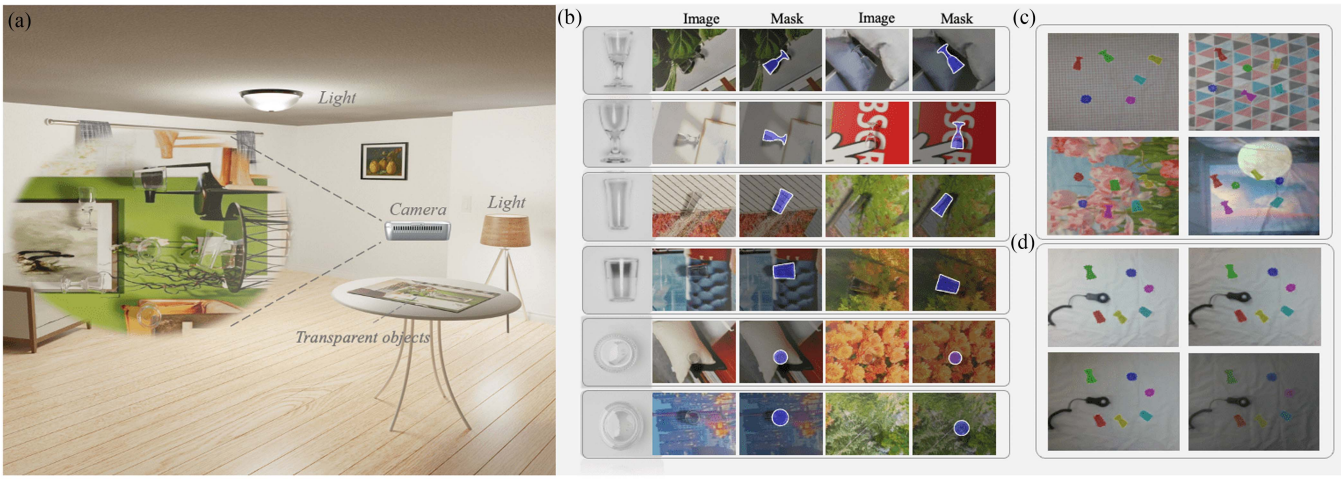

Fig. 6. SimTrans12 K dataset. (a) Scene setup for generating the transparent object synthetic dataset using Blender. (b) Synthetic dataset of transparent objects in different backgrounds. (c) Real dataset in different backgrounds. (d) Real dataset in different brightnesses.

information is relatively stable [10]. Therefore, instead of using rectangular box annotation, we directly use the mask of the transparent object itself as the grasping contour and use Gaussian distribution to represent the optimal grasping position, which makes full use of the boundary information of the transparent object.

For the TaTa gripper, which achieves gripping by wrapping the whole object, we can use the center of the transparent object as the optimal grasping position. To make the grasping position as close as possible to the object center, a grasping quality distribution map that satisfies a Gaussian distribution from the center of the annotated object is generated. In this way, the point near the object center has a higher grasping quality than the point away from the center, so it is easier for the gripper to select the object center for grasping. Positioning in the object center helps TaTa for better tactile detection and also reduces the probability of potential damage to the object during grasping. To make the Gaussian-mask annotation adapt to different objects and camera positions in the scene, the two farthest points on the object on the same x–y-plane are selected. Then, the two points are projected onto the rendered image. The center of the line connecting the two points is chosen as the center of the Gaussian distribution, and half the distance between the two points is determined as the Gaussian distribution radius.

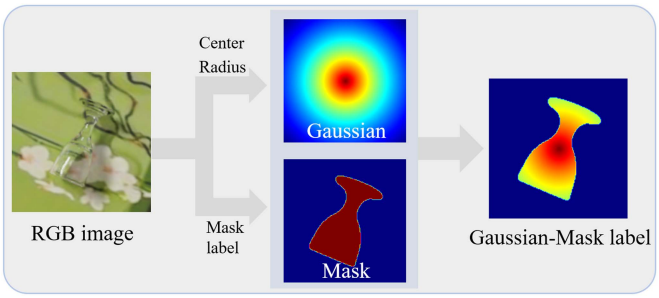

Fig. 7. Gaussian-mask annotation process of transparent objects.

However, the Gaussian-mask annotation method also has some limitations. This method uses the object's mask as the annotation frame and is more suitable for objects whose center is located on the object. The Gaussian-mask annotation process is shown in Fig. 7. The object center, Gaussian distribution radius, and Gaussian-mask labels are obtained directly from the RGB image, which is further processed to obtain the Gaussian representation ground truth labels. Thanks to the Gaussian-mask annotation, the grasping network could regress to a more accurate grasping center.

B. Transparent Object-Grasping Position Detection

Due to the unique optical properties, the appearance of transparent objects is easily disturbed by the backgrounds. In many cases, it is difficult for even humans to directly distinguish the types of transparent objects under different backgrounds, which makes the classification and grasping of transparent objects by vision difficult. Therefore, we adopt the generative grasping model, which can directly generate the grasping positions from images without recognition and classification. The model is expected to learn the general properties of transparent objects and then apply them to detect and grasp unseen transparent objects in complex backgrounds.

1) Network Architecture: Fig. 8 shows the proposed TGCNN model, which is a generative architecture taking in a three-channel RGB image and generating pixelwise grasps in the form of two images. The three-channel RGB image is passed through convolutional layers, residual layers, and convolution transpose layers to generate two images. Each residual layer contains two convolutional layers, two batch normalization layers, and a shortcut connection. At the same time, the skip connections in the network enable the network to obtain more hierarchical information fusion, which makes the network more effectively combine scene information to detect transparent objects. TGCNN has 2 145 154 parameters with a trained model size of 8.23 MB. The output image of the network includes grasping quality and grasping radius, and these two parameters can guide the grasping for the TaTa gripper.

2) Grasp Definition: A grasp perpendicular to the x–y-plane is defined as $\mathbf{g}_r = (\mathbf{p}, r)$. The grasp is described by the projection of the center position of the gripper $\mathbf{p} : (x, y)$ onto the x–y-plane and the height h between the gripper center and the x–y-plane in Cartesian coordinates. Since the shape of the gripper is hemispherical, the flexibility of the gripper enables it to be deformed. For the same gripper and object, the lower the gripper center from the x–y-plane after contact with the object, the more the gripper is compressed and the larger the contact area between the gripper and the object will be.

Since the contact surface is similar to a circle, the index grasping radius r was used to describe the contact area sizes caused by different heights h of the gripper. The influence of different grasping radius r on detection is shown in Fig. 9. A scalar quality measure q, representing the chances of grasp success, is added to the pose. To further improve the grasping efficiency of the gripper, an adaptive height-dropping method (AHD) is proposed. AHD can determine the distance between the gripper and the detection surface according to the size of the object. We use a small drop height for small objects and use a larger drop height for large objects. Because for small objects, a smaller drop height can make the gripper obtain complete tactile detection information. While for a large object, a larger drop height can obtain a more complete tactile image of the object. The generated grasping radius is smaller when the object detection uncertainty is large, which ensures a balance between safety and detection efficiency when detecting in highly uncertain environments.

Assume that we want to detect grasps through an RGB image $\mathbf{P} \in \mathbb{R}^{m\times n\times 3}$, with known camera intrinsic parameters. In the image coordinate system, a grasp is described by

$$\mathbf{g}_i = (\mathbf{s}, r_i, q) \tag{1}$$

where $\mathbf{s} = (u, v)$ is the center and r_i is the grasping radius in image coordinates. In order to perform the grasp in image space on the robot, we can convert the image coordinates to the robot's frame of reference by the following transformation:

$$\mathbf{g}_r = \xi_{RC}(\xi_{CI}(\mathbf{g}_i)) \tag{2}$$

where ξ_{RC} is a transformation from the camera frame to the world frame and ξ_{CI} is a transformation from 2-D image coordinates to the 3-D camera frame.

The above notation can represent multiple grasps in an image. The collective group of all grasps can be denoted as follows:

$$\mathbf{G} = (\mathbf{R}, \mathbf{Q}) \in \mathbb{R}^{m\times n\times 2} \tag{3}$$

where $\mathbf{R}$ and $\mathbf{Q} \in \mathbb{R}^{m\times n}$ contain the values of grasping radius r_i and quality measure q, respectively, at each pixel $\mathbf{s}$.

Grasp candidates $\mathbf{g}_i$ are wanted to create directly by calculating the RGB images, so a mapping ϕ from RGB images to grasp map in the image coordinates was defined: $\phi(\mathbf{P}) = \mathbf{G}$. From $\mathbf{G}$, the best visible grasp in the image space $\mathbf{g}_i^* = \max_{\mathbf{Q}} \mathbf{G}$ can be calculated, and the equivalent best grasp in world coordinates $\mathbf{g}_r^*$ can be obtained as well. For the case of multiple objects in the same scene, we will sort the visible grasps according to the quality measure q and select the first k visible grasps (k is manually specified).

Huber loss is used for network training, written as follows:

$$\mathcal{L}\left(\mathbf{G}_i, \hat{\mathbf{G}}_i\right) = \begin{cases} 0.5(\|\mathbf{G}_i - \hat{\mathbf{G}}_i\|_{\mathrm{F}})^2, & \text{if } \|\mathbf{G}_i - \hat{\mathbf{G}}_i\|_1 < 1 \\ \|\mathbf{G}_i - \hat{\mathbf{G}}_i\|_{1,1} - 0.5, & \text{otherwise.} \end{cases} \tag{4}$$

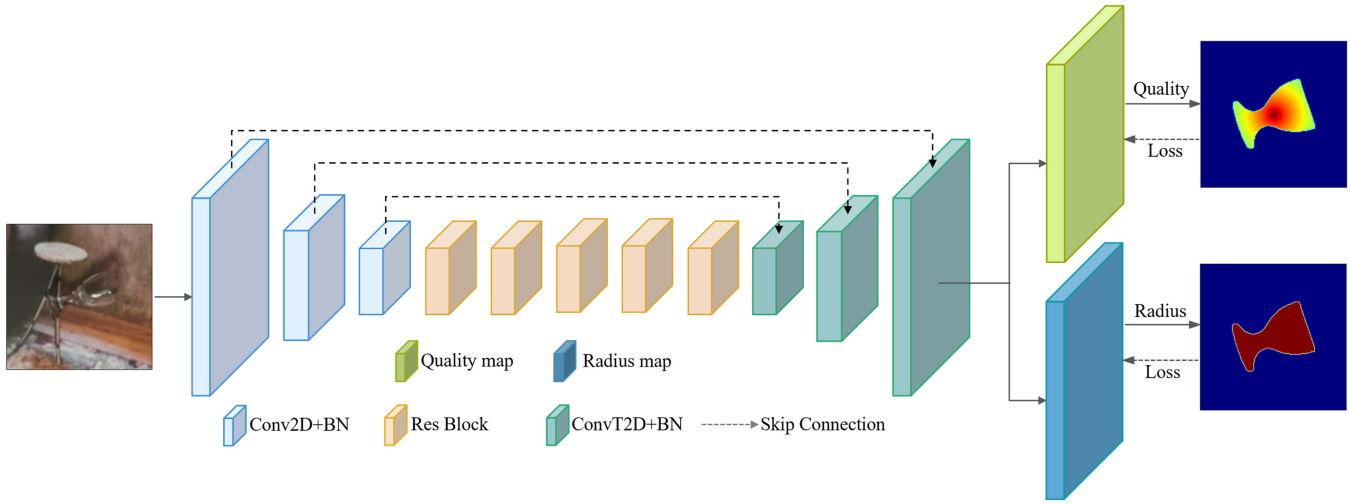

Fig. 8. Network architecture of the proposed TGCNN.

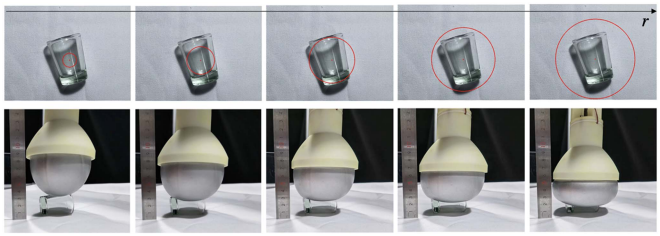

Fig. 9. Influence of different grasping radius r on detection.

Here, $\mathbf{G}_i$ denotes the grasp candidate, which can be generated by the network, and $\hat{\mathbf{G}}_i$ is the ground truth grasp. $\|\cdot\|_{\mathrm{F}}$ and $\|\cdot\|_{1,1}$ represent the Frobenius and "entrywise" $L1$ matrix norms, respectively.

C. Tactile Information Extraction Algorithm

To obtain the contact information between the gripper and the object, we extract the contour of the contact area using fully convolutional networks (FCNs). Compared with the frame difference method [39] and optical flow method [40], the FCN-based tactile feature extraction algorithm has stronger robustness and can still obtain clear contact information even if the internal optics of the sensor are changed. We acquired 160 images of the object in contact with the gripper as a training set and annotated the data at the pixel level, and some results are illustrated in Fig. 10(a). After 60 rounds of training, the segmentation accuracy achieves 98%.

D. Visual–Tactile Fusion Classification

Transparent objects have little visual information and the surface pattern changes with the backgrounds as well as the lighting conditions, making it difficult to classify by vision only. To solve this problem, a vision–tactile fusion method for transparent object classification is proposed, where RGB and tactile images are concatenated together for classification using GoogleNet, as depicted in Fig. 11. We collected 1200 data of six objects with different backgrounds, such as reflections, patterns, colors, overlapping, and stacking scenes as the training set and 600 data as the test set, as shown in Fig. 10(b).

To test the performance of the algorithm, we compare the visual classification and visual–tactile fusion classification algorithms. The visual classification accuracy is 59.3%, while the visual–tactile fusion classification accuracy reaches 98.4%, which increases the classification success rate by 39.1%.

V. Grasping Strategy

Based on the algorithms proposed in Section V, this section explains how to integrate them to accomplish transparent object grasping in different scenes, which forms the high-level grasping strategy for our visual–tactile fusion framework. We decompose a grasping task into three subtasks, i.e., object classification, grasping position determination, and grasping height determination. Each subtask can be conducted by vision, touch, or fusion. Similar to humans, when vision can directly obtain the precise position of the object, we can control the hand to directly reach the object and complete the grasp, as shown in Fig. 12(a). When the vision cannot accurately obtain the object's position information, we will use the tactile perception of the hand to slowly adjust the grasping position after obtaining the object's general position information until it touches the object and reaches the appropriate grasping position, as shown in Fig. 12(b). For object grasping in visually limited situations, as shown in Fig. 12(c), we will use the hand's rich tactile nerve to search for the position of the object in a wide range, which obviously wastes more time but is an effective way to solve the object grasping in these special scenes.

Inspired by human grasping strategies, we divide transparent object-grasping tasks into three types: planes with complex backgrounds, irregular scenes, and visually undetectable scenes, as shown in Fig. 13. In the first type, where vision is very effective

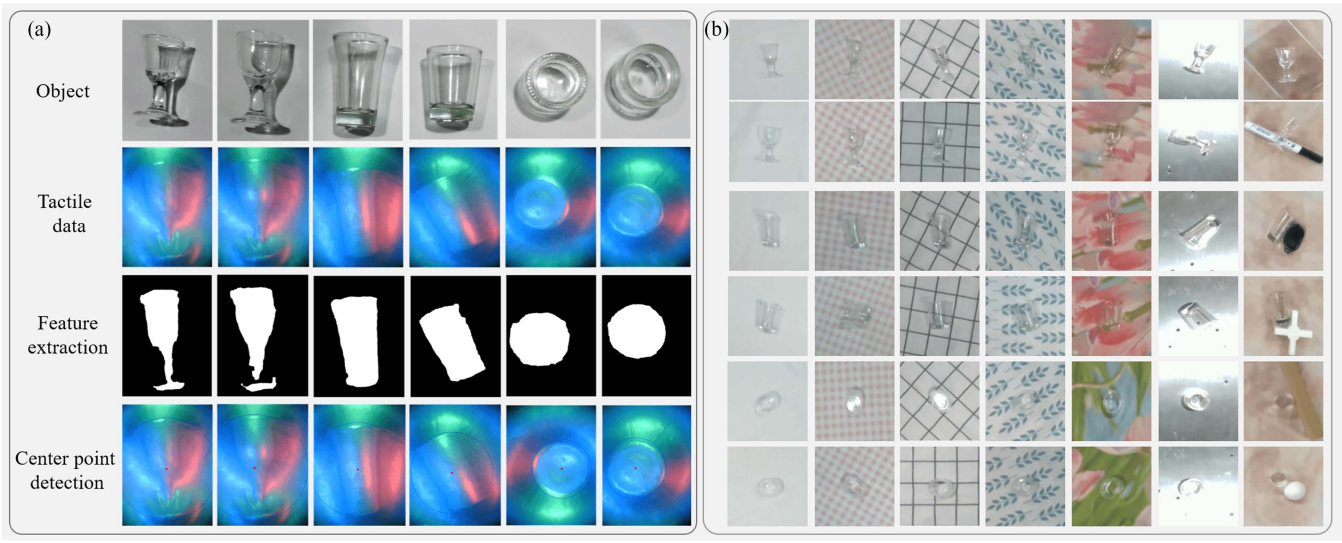

Fig. 10. Visual–tactile fusion classification dataset. (a) Tactile data, FCN feature extraction, and center point detection results. (b) Visual classification dataset.

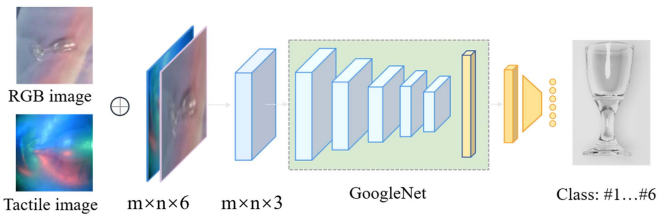

Fig. 11. Visual–tactile fusion classification framework for transparent objects.

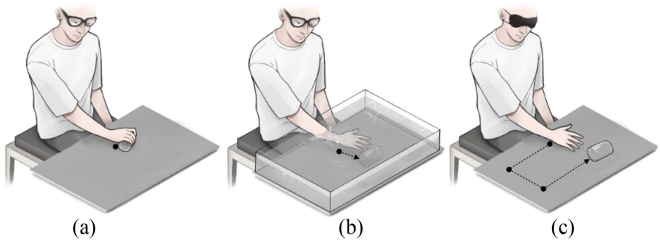

Fig. 12. Human grasping strategies in different scenes (The black curve indicates the movement path of the hand). (a) Grasping objects in clear view. (b) Grasping transparent objects underwater. (c) Grasping objects in visually undetectable scenes.

and plays a key role, we use vision-first grasp. In the second type, where vision and touch can work synergistically, we use vision–tactile grasp. While in the last type, where vision may fail and touch becomes dominant in the task, we use touch-first grasp. Details of the three grasping strategies are introduced as follows.

A. *Planes With Complex Backgrounds—Vision-First Grasp*

Grasping objects on a plane can be achieved by visual detection [41], [42], but the texture information of transparent objects changes with the background, so grasping transparent objects

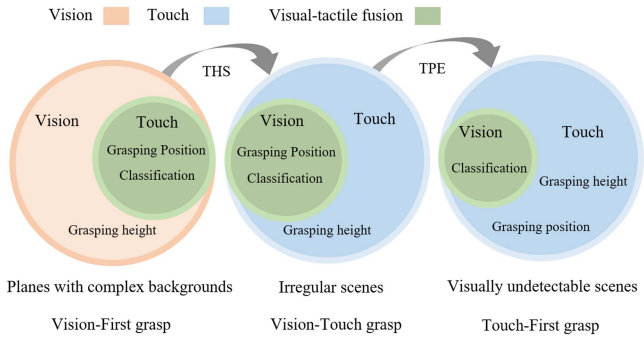

Fig. 13. Grasping strategies in different scenes. The orange, blue, and green colors represent the functions of visual detection, tactile, and visual–tactile fusion, respectively. The framework can be adapted to different scenes by adjusting the grasping strategies.

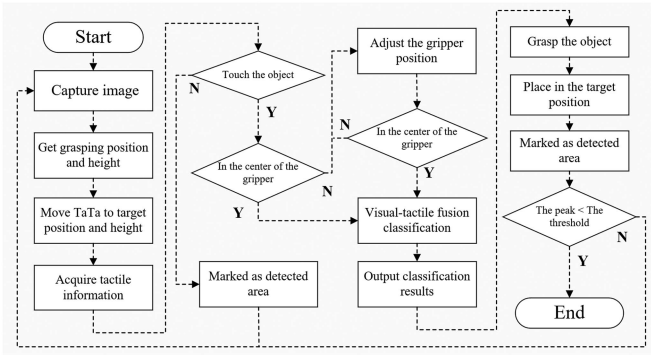

Fig. 14. Flowchart of transparent object grasping on a plane with complex backgrounds.

in complex backgrounds is challenging even on a plane. To tackle this, we propose a strategy for transparent object grasping with visual–tactile fusion, as depicted in Fig. 14, which mainly includes three steps.

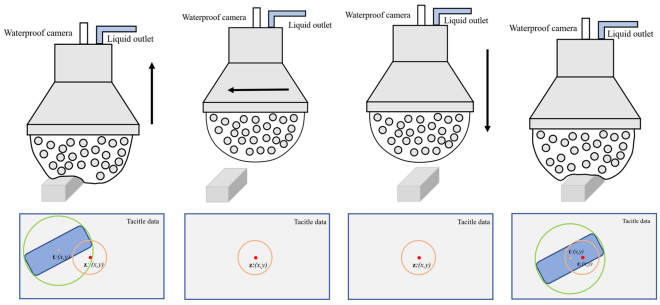

Fig. 15. Tactile calibration grasping process. First, the gripper touches the object, the center of the object outline does not appear in the gripper center, and the displacement between the object and the gripper is calculated as $\mathbf{d} : (x, y)$. Second, lift the gripper. Third, move the gripper with a distance of $\mathbf{d}$. Fourth, the gripper touches the object again. If the object center coincides with the gripper center, then calibration is completed.

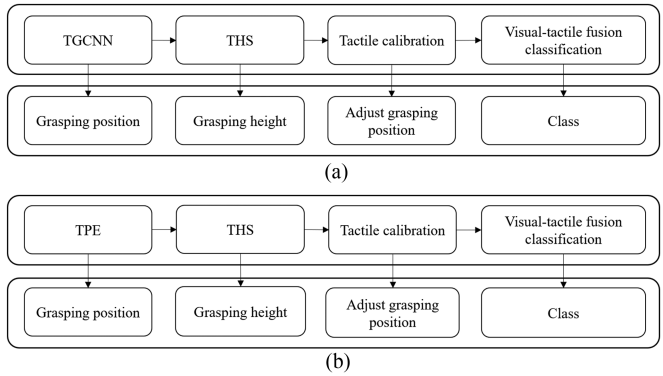

Fig. 16. Transparent object-grasping process. (a) In irregular scenes. (b) In visually undetectable scenes.

a) Use TGCNN to obtain the grasping position and height of the transparent object as the target position and control the gripper to reach the target position.

b) Use tactile information to check if the gripper contacts the object. If not, mark it as a wrong detection point and proceed to the next position. If yes, the tactile information will be applied to further adjust the gripping position. Here, to get a more precise grasping position, we propose a tactile calibration method, as illustrated in Fig. 15. The method first uses a tactile information feature extraction network to segment the contact area [43], which has been introduced in Section IV. Then, use the minimum outer circle detection algorithm to obtain the circle center of the contact area, calculate the position relationship between the gripper center $\mathbf{z} : (x, y)$ and the minimum outer circle center $\mathbf{t} : (x, y)$ obtained by the tactile feature extraction $\mathbf{d} = \mathbf{z} - \mathbf{t}$, and control the gripper to move the distance of $\mathbf{d}$. During this stage, the tactile calibration algorithm will continue running until the object locates in the gripper center.

c) Use the visual–tactile fusion framework to classify and place the object at a given place. Finally, the area is marked as detected and will not be revisited. The use of the grasping position marker prevents invalid revisiting, especially when there are many interference areas in visual detection results.

B. Irregular Scenes—Vision-Touch Grasp

Compared with grasping objects on planes, it is more challenging to grasp transparent objects in irregular scenes, such as overlapping, stacking, and undulating scenes. For stacking and overlapping scenes, it is difficult to separate two objects with similar textures by RGB vision detection either for transparent or nontransparent objects. And for undulating surfaces, it is difficult to obtain the precise grasping height of the object simply by using an RGB camera.

So, we add the THS module that uses tactile sensing to adjust the grasp height. The implementation process is shown in Fig. 16(a). First, we still use TGCNN to get the grasping position of the transparent object, but the height of the object cannot be determined. Therefore, when the gripper reaches the specified position, the THS module will be activated. The gripper is controlled to keep exploring downward until it touches the object or reaches the lowest point, and then complete grasping the object through tactile calibration. The THS module not only enables the grasping of objects in undulating scenes but also solves the problem of grasping transparent objects in overlapping and stacking scenes.

C. Visually Undetectable Scenes—Touch-First Grasp

Although vision is a powerful detection method, it may fail in some scenes, such as transparent object detection in highly dynamic underwater scenes. Because water and transparent objects have similar optical properties, the water flow and ripples will result in difficulties to detect the grasping position by vision. For transparent objects, we define scenes, such as highly dynamic underwater, darkness, and smoke, as visually undetectable scenes.

To achieve grasping in visually undetectable scenes, we add a TPE module as introduced. The implementation process is shown in Fig. 16(b). First, it uses touch to search for the transparent object in a specific range. When in contact with the object, it uses THS to determine the grasping height and tactile calibration to determine the grasping position. Finally, it applies visual–tactile fusion for classification. Therefore, when vision is not effective, we can use touch to obtain both the grasping height and position, such as human grasping in the dark. The advantage of this method is that object grasping can still be achieved even without vision, while the disadvantage is that it is inefficient and may fail to find the object when the exploration area is too large.

VI. Experiments

This section presents the experimental results of the proposed algorithms and visual–tactile fusion grasping framework. First, to test the effectiveness of our proposed transparent object dataset, the annotation method, and the grasping position detection network, we conduct synthetic data detection experiments (*Exp. 1*) and transparent object-grasping position detection experiments under different backgrounds (*Exp. 2*) and brightness (*Exp. 3*). Second, to verify the effectiveness of the

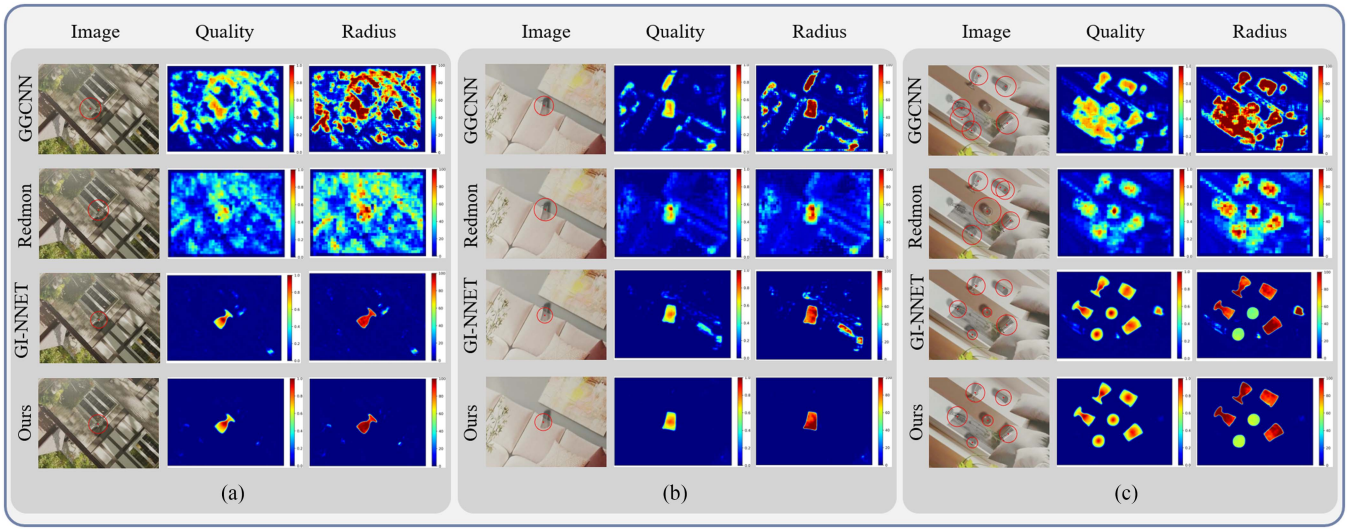

Fig. 17. Synthetic dataset detection results. (a) Imagewise evaluation in unseen backgrounds. (b) Objectwise evaluation with unseen objects. (c) Multiobject evaluation in cluttered scenes.

visual–tactile fusion grasping framework, transparent object classification grasping experiments (*Exp. 4*) and transparent fragment grasping experiments (*Exp. 5*) are designed. Third, we design transparent object-grasping experiments in irregular (*Exp. 6*) and visually undetectable scenes (*Exp. 7*) to test the effectiveness of the framework after adding the THS module and the TPE module. The failed trials and limitations are provided and discussed as well.

TABLE II
DETECTION RESULTS IN UNSEEN BACKGROUNDS

Algorithm	Accuracy (Gaussian-mask)	Accuracy (Binary)
GGCNN [37]	67.2%	27.2%
Redmon [46]	59.3%	42.5%
GI-NNET [47]	89.3%	25.1%
Ours	94.2%	25.3%

A. *Exp. 1: Object Detection With Synthetic Data*

To evaluate the performance of the TGCNN algorithm, we apply the index of grasping overlap degree (GOD) to measure if a detection is successful, similar to Saxena et al. [44] and Jiang et al. [45]. If a calculated grasp circle and the label mask share an intersection (i.e., the GOD) greater than 45%, a detection is considered to be correct.

Most of the current research articles on transparent object grasping, such as ClearGrasp, Dex-NerF, and LIT, are based on depth complementation and pose estimation of RGB-D information, which cannot be directly compared with our algorithms that are based on RGB data. To make a comparison, we consider the currently more mainstream generative grasp networks, such as GGCNN [37], Redmon [46], and GI-NNET [47]. Since most of these networks are designed for parallel two-finger grippers based on RGB-D input, modifications are needed to enable them to operate on our proposed dataset—we change the input data to RGB and change their output from the width and angle to the radius.

In the experiments, we evaluate TGCNN from multiple aspects:

1) imagewise evaluation with unseen backgrounds;
2) objectwise evaluation with unseen objects;
3) evaluation of Gaussian representation; and
4) multiobject evaluation in cluttered scenes.

1) Imagewise Evaluation in Unseen Backgrounds: For transparent objects, changes in the background can greatly affect their visual features and may result in recognition errors and grasping failures. To evaluate the performance of TGCNN, we select six objects for training and then test the detection accuracy on unseen backgrounds. The training set contains 4000 images and the testing set contains 1000 images with different backgrounds from the training dataset. Although the proposed synthetic data rendering scheme can generate a large number of transparent object data easily, we hope that the network can learn with not too much data.

As a result, TGCNN successfully detects a total of 942 objects within the test set, with an accuracy of 94.2%. The results are compared with some currently well-known and open-source algorithms, as shown in Table II. Accuracy (Gaussian-mask, %) in Table II means that we use the Gaussian representation of the label, and accuracy (Binary, %) means that we use the binary representation of the label. Fig. 17(A) shows the detection results of each algorithm, indicating that our algorithm has better performance in unseen backgrounds.

2) Objectwise Evaluation With Unseen Object: Besides the good performance under new backgrounds, TGCNN can also achieve grasping position detection for unseen objects. To test this, we use two objects from the dataset as the training set and the remaining four objects as the testing set. Furthermore, we also included four objects from the LIT dataset [13] in the test

TABLE III
DETECTION RESULTS WITH UNSEEN OBJECTS

Algorithm	Accuracy (Gaussian-mask)	Accuracy (Binary)
GGCNN [37]	83.8%	44.2%
Redmon [46]	48.6%	30.1%
GI-NNET [47]	96.1%	61.2%
Ours	99.2%	70.9%

TABLE IV
DETECTION RESULTS WITH REAL DATA

Algorithm	Accuracy (Different backgrounds)	Accuracy (Different brightness)
GGCNN [37]	51.8%	70.9%
Redmon [46]	73.2%	73.2%
GI-NNET [47]	50.7%	48.0%
Ours	91.6%	84.0%

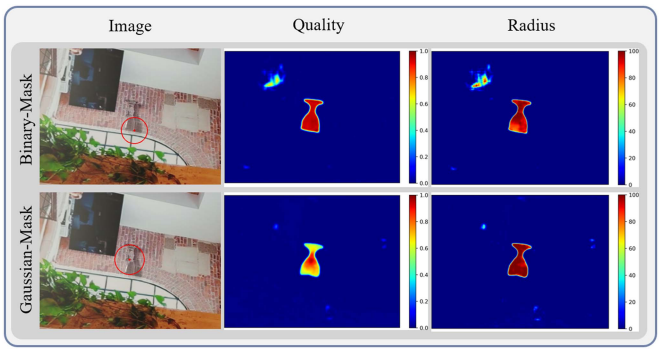

Fig. 18. Detection comparison between Gaussian and binary representations.

set to further guarantee the generalization of the dataset. The training set contains 4000 images and the testing set contains 1000 images with unseen objects but the same backgrounds as the training set. As can be seen from Table III, TGCNN outperforms other algorithms in the detection accuracy of unseen objects. Fig. 17(b) shows the detection results of TGCNN and other algorithms.

3) Evaluation of Gaussian Representation: During the experiments, we found that the utilization of Gaussian representation in the annotation plays a key role in improving the detection accuracy, either for TGCNN or other algorithms. The grasping position detection results of each algorithm using Gaussian-mask and binary representations are listed in two columns, as shown in Tables II and III. As can be observed, by introducing Gaussian representation, the accuracy of all algorithms is greatly improved compared with binary representation. With binary representation, our algorithm does not always perform the best because TGCNN is sensitive to the boundary, while other algorithms are not. It can be seen from Fig. 18 that TGCNN with binary representation locates the grasping center at the edge of the object, resulting in an unsatisfactory initial tactile detection position. However, with Gaussian representation, the grasping position is guided to the center, allowing tactile calibration to get a better initial position.

4) Multiobject Evaluation in Cluttered Scenes: Besides predicting the optimal grasping of unseen objects, the robustness of TGCNN is also reflected in the ability to predict the grasping of multiple objects in cluttered scenes. In the experiment, the training set contains 4000 images with a single object, and another 1000 images with multiple objects in the clutter are used for testing. In each test, we randomly change the object type, object position, camera position, and scene background in the scene (the scene backgrounds appeared in the training set). The comparison of different algorithms is shown in Fig. 17(c). It can be seen that, although TGCNN is only trained on a dataset with a single object, it can effectively predict the grasping position of multiple objects with better performance than other algorithms.

B. Exp. 2: Grasping Position Detection in Different Backgrounds

To verify the grasping position detection performance of TGCNN in real scenes, we select 12 backgrounds with different features, including 6 colored backgrounds, 4 patterned backgrounds, and 2 scenic backgrounds, as shown in the first row in Fig. 19. The six objects in SimTrans12 K are used for experiments. A total of 4000 synthetic data of two transparent objects are selected as the training set, 110 real data of six transparent objects as the test set, which contains about 600 labels, and GOD is used to quantify the detection performance. The performance comparison of GGCNN, Redmon, GI-NNET, and TGCNN trained under the same dataset is shown in Fig. 19 and Table IV. The results reveal that all networks have good detection performance under a solid colored background [see Fig. 19(a)]. While in the patterned and scenic backgrounds [see Fig. 19(b)], the GGCNN [37], Redmon [46], and GI-NNet [47] algorithms produce more noise in the grasping position, whereas TGCNN still maintains good performance.

Compared with GI-NNET, TGCNN has a larger number of parameters. Thanks to the application of residual layers [48] and skip layer connections [49], we can increase the network depth while preventing the network from overfitting. In addition, TGCNN is a grasping network specially designed for jamming grippers, and we make some adjustments and optimizations in the number of layers and blocks of the network, so the TGCNN network has better detection results compared with GI-NNET.

C. Exp. 3: Grasping Position Detection Under Different Brightness

Besides the background, the light condition is also an important factor affecting the detection accuracy. In this experiment, we test the impact of lightness on transparent object detection by changing the brightness (from 151 lx to 2500 lx measured by a Lux meter). A total of 4000 synthetic data of two transparent objects are selected as the training set and 50 real data of six transparent objects at different brightness as the test set. The detection results of the four networks are shown in Fig. 20 and Table IV. From the experimental results, we can see that the detection results of Redmon and TGCNN are relatively stable, but Redmon has more noise. The GGCNN and GI-NNET

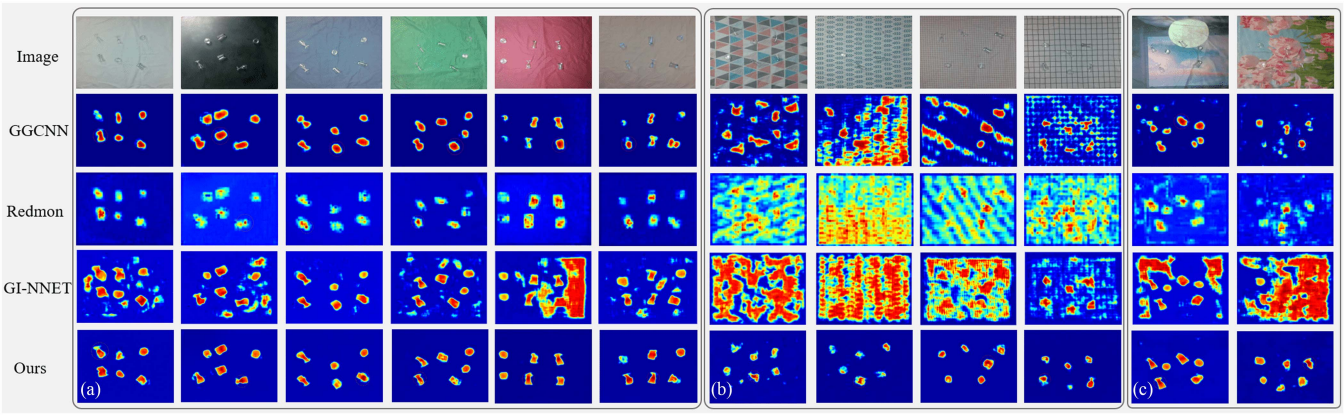

Fig. 19. Comparison of the four algorithms for grasping position detection in different backgrounds. (a) Colored. (b) Patterned. (c) Scenic backgrounds.

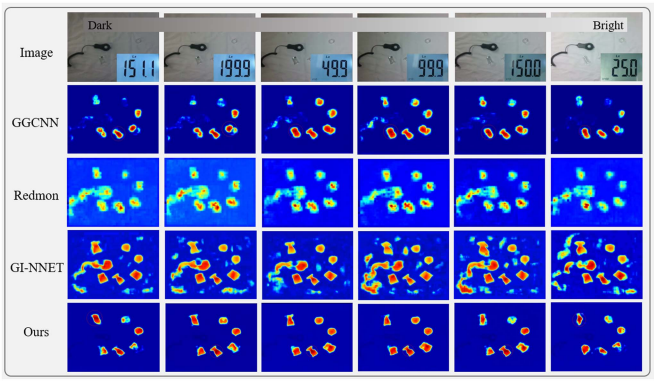

Fig. 20. Comparison of the four algorithms for grasping position detection in different brightness.

networks have a stable detection effect when the brightness of the light is in the 199.9–999 lx interval. When the interval is exceeded, the detection results will be affected, for example, some objects in the detection results of GGCNN will not be detected, and the detection results of GI-NNET will have more noise information.

The experimental results show that the proposed method has good detection performance in real environments of different backgrounds and lighting conditions even though the network is trained using only synthetic data. In addition, we also test the influence of camera height and light position on the detection accuracy. It shows that when the camera height ranges between 35 and 120 cm, TGCNN maintains high detection accuracy. And under relatively uniform light conditions, the light position does not have an obvious impact on the detection performance.

D. Exp. 4: Grasping and Classification on Planes With Complex Backgrounds

To verify the effectiveness of the proposed transparent object grasping and classification framework, grasping and classification experiments are carried out. The selected objects are the same as in Fig. 10, including an angled wine glass and a smooth wine glass, a girdled water glass and a normal water glass, and a medicine bottle with a textured bottom and a smooth medicine bottle, which have slippery surfaces and similar shapes, and are difficult to both grasp and classify.

TABLE V
EXPERIMENTAL RESULTS OF TRANSPARENT OBJECT GRASPING AND CLASSIFICATION

Algorithm	Grasping success rate		Classification success rate		No. of tactile calibration	
Background	Pink	Moon	Pink	Moon	Pink	Moon
GGCNN [37]	93%	90%	95%	93%	2	4
Redmon [46]	93%	88%	97%	95%	1	4
GI-NNET [47]	87%	75%	95%	93%	3	5
Ours	98%	93%	98%	93%	1	2

The experimental procedure is shown in Fig. 21. Two backgrounds are used in the experiment—the pink background, which is relatively simple, and the moon background, which has various colors and complex textures. We compare the performance of GGCNN [37], Redmon [46], GI-NNET [47], and TGCNN. For each algorithm, we choose 3 objects randomly placed on the table each time and conduct 20 experiments on each background with a total of 60 grasping experiments on each background.

The experimental process is divided into three parts: grasping position detection, tactile calibration, and visual-touch fusion classification. In the grasping position detection stage, the image is acquired using RealSense D435i, and the transparent object-grasping position and height are output using TGCNN. After getting to the grasping position, the gripping position will be adjusted using a tactile calibration algorithm. After reaching the optimal grasping position, the object will be classified using the visual-touch fusion classification algorithm and placed in the target location. Finally, we compare the peak value of the output of the grasping position detection network with a preset threshold, repeat the above operation if it is greater than the threshold, and end the grasping if less than the threshold.

The experimental results are shown in Table V. The classification success rate in the table indicates the classification success rate in the case of successful grasping. In addition, the

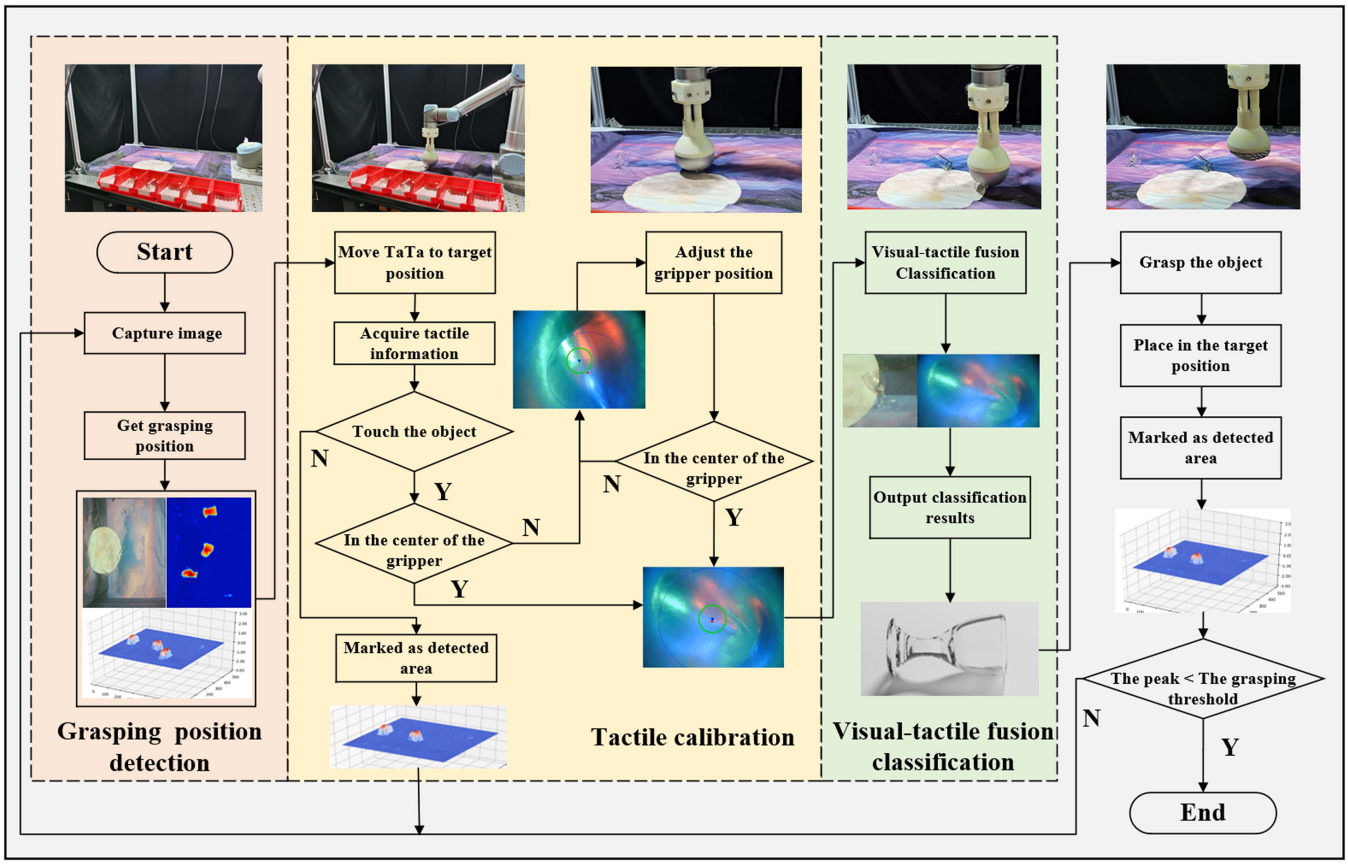

Fig. 21. Flowchart of visual–tactile fusion-based transparent object grasping and classification.

number of tactile calibrations indicates the number of calibrations performed in each experiment in the case of successful grasping, which reflects the grasping position detection accuracy (the number of calibrations is less when the accuracy is higher). The experimental results show that all four networks have good detection performance in the pure color background, while in complex backgrounds, TGCNN has a better performance in terms of grasping success rate and the tactile calibration number. Even in the case of poor grasping position detection, the tactile calibration algorithm in the framework still has a certain probability to achieve the grasping of transparent objects. In addition, we have also compared the detection effects of visual–tactile fusion classification with visual-only classification in the real experiments, and obtained results similar to the algorithm introduction section, with a detection accuracy improvement of 39%.

E. Exp. 5: Transparent Fragment Grasping

Once a transparent object is broken, a large number of fragments will be produced, which have irregular shapes and various sizes and are difficult to grasp. To test the effectiveness of the visual–tactile fusion grasping framework, transparent fragment grasping experiments are performed, suggesting that tactile sensing has an important enhancement to the grasping success rate.

TABLE VI
GRASPING SUCCESS RATES WITH AND WITHOUT TACTILE CALIBRATION

Algorithm	Tactile calibration grasping			Direct grasping		
Background	Yellow	Grid	Flower	Yellow	Grid	Flower
GGCNN [37]	95%	85%	80%	80%	30%	25%
Redmon [46]	95%	75%	80%	65%	15%	20%
GI-NNET [47]	95%	85%	65%	60%	30%	15%
Ours	100%	90%	95%	85%	40%	50%

The transparent fragments used in the experiment are shown in Fig. 22(a), which are some glass fragments with jagged surfaces, further increasing the difficulty of grasping position detection. The fragment grasping process omits the classification process compared with *Exp. 4* but places higher demands on the grasping process, and the experimental process is shown in Fig. 22(b)–(h). To test the tactile calibration algorithm, grasping experiments with and without tactile calibration are performed. When tactile calibration is disabled, the gripper will grasp directly, without further adjustment of the grasping position. We compare the performance of GGCNN [37], Redmon [46], GI-NNET [47], and TGCNN. Each algorithm is tested in the yellow, grid, and flower backgrounds, and 20 visual–tactile fusion grasps and 20 direct grasps are performed in each background.

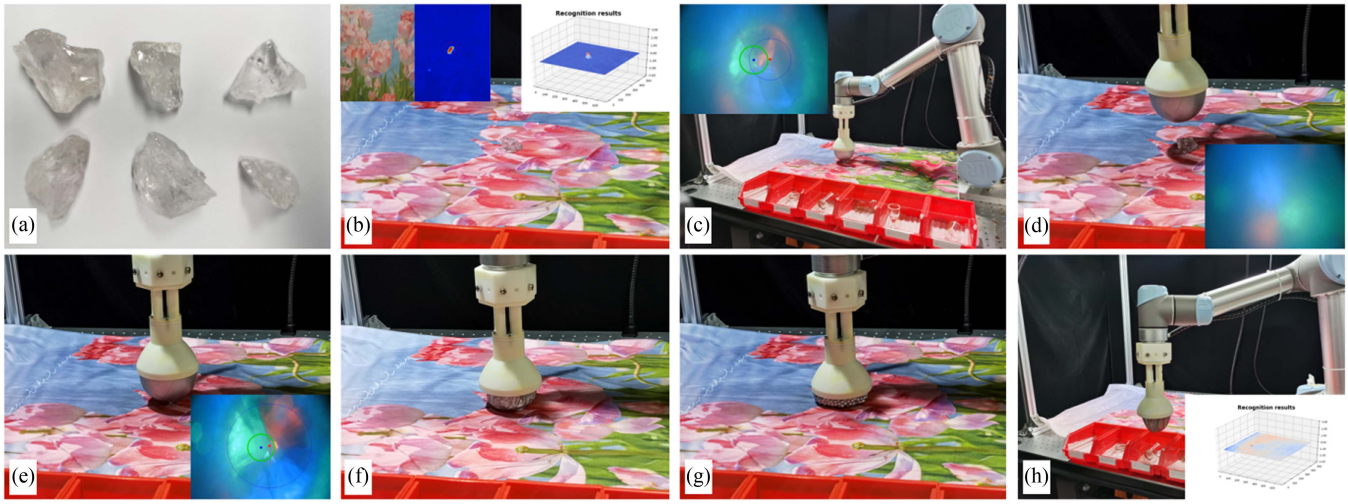

Fig. 22. Transparent fragment grasping experiment based on visual–tactile fusion. (a) Transparent fragments. (b) Get the grasping position and height. (c) Contact with the object and detect the center of its contour is not in the gripper center. (d) Adjust the position of the gripper. (e) Touch the object again, whose contour center coincides with the gripper center. (f) and (g) Grasp object. (h) Place the object in the specified position.

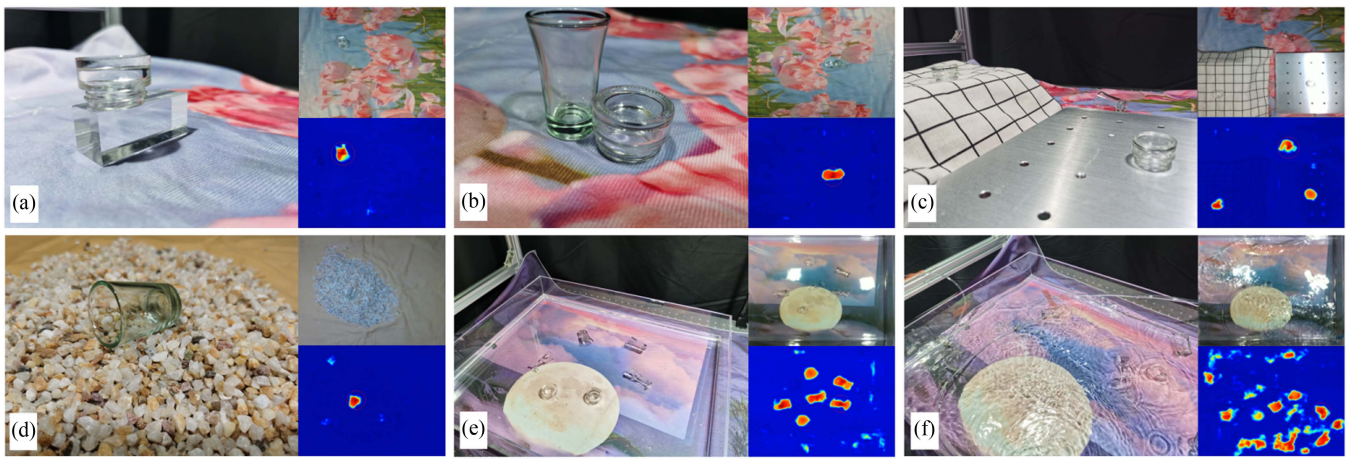

Fig. 23. Transparent object in complicated scenes: (a) overlapping; (b) stacking; (c) undulating; (d) sand; (e) underwater; and (f) highly dynamic underwater scenes.

It can be seen from Table VI that, in the yellow background, the detection accuracy remains high in direct grasping because the grasping position can be determined more accurately by a vision in the yellow background compared with grid and flower backgrounds. When TGCNN is adopted as the transparent object-grasping position detection network, the tactile calibration method can improve the grasping success rate by 15%, 50%, and 45% under the yellow, grid, and flower backgrounds, separately, and the overall grasping success rate by 36.7%, showing the feasibility of the framework for transparent fragment grasping.

F. Exp. 6: Grasping in Irregular Scenes

Compared with grasping on a plane, it is more challenging to grasp transparent objects in irregular scenes, such as overlapping, stacking, undulating, and sand [see Fig. 23(a)–(d)], where the grasping position and height are difficult to obtain. To solve this problem, we add the THS module based on the previous framework to obtain the height where the object is located by tactile. As shown in Fig. 23(a)–(e), to verify the grasping effect of the THS module in irregular scenes, we conduct experiments on the grasping of transparent objects in the case of stacking and overlapping, as well as the grasping of transparent objects in special scenes, such as undulating surfaces, sand, and underwater.

To demonstrate the experimental process, we designed two representative scenes, the first with stacking and overlapping problems, and the second with undulating areas, reflective areas, and sand, as shown in Figs. 24 and 25. We conduct 20 grasping experiments in each scene, and the overall success rate can reach more than 90%, which shows the feasibility of the method for grasping transparent objects on irregular planes. More experimental procedures can be

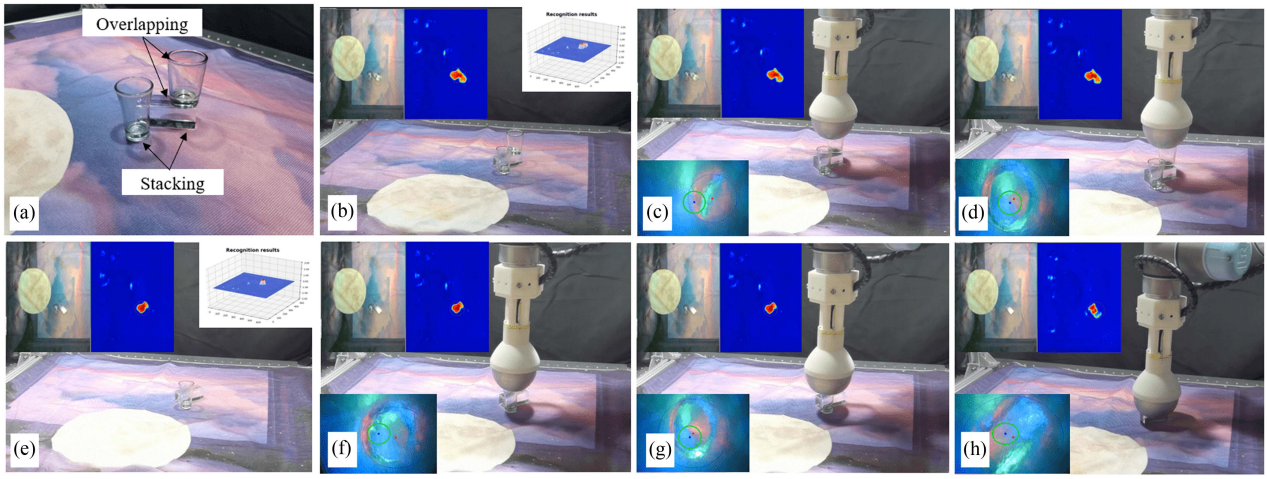

Fig. 24. Transparent object-grasping process in stacking and overlapping scenes. (a) Experimental setup. (b) Get the grasping position. (c) Use the THS module to search objects and contact the first object. (d) Adjust the grasping position with the tactile calibration module and grasp the object. (e) Get the new grasping position after completing the first object grasping. (f) Use the THS module to search objects and contact the second object. (g) Adjust the grasping position with the tactile calibration module and grasp the object. (h) Get the new grasping position and grasp the third object.

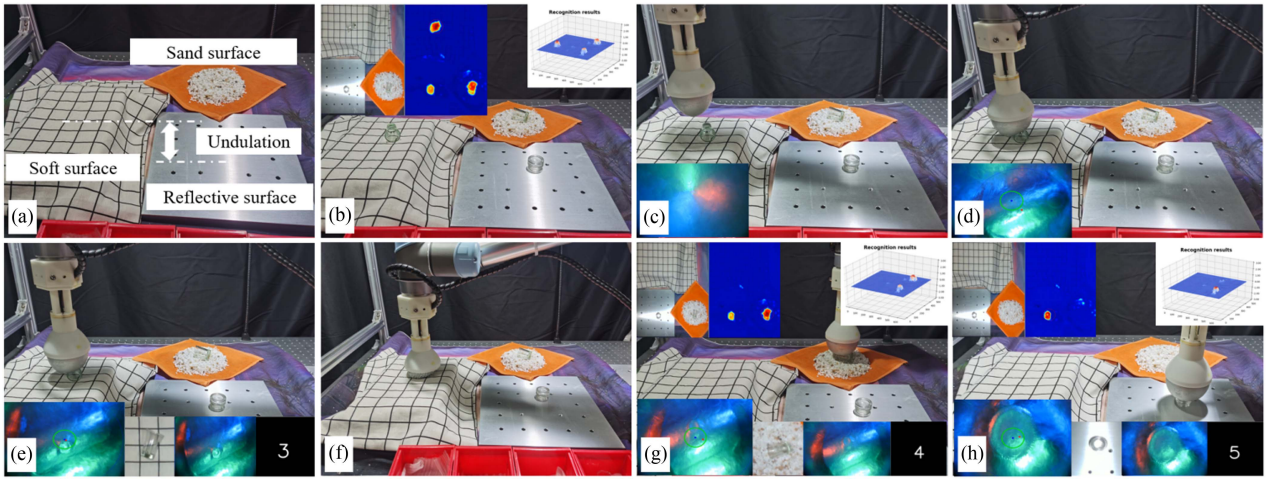

Fig. 25. Transparent object-grasping process in undulating and sand scenes. (a) Experimental setup. (b) Get the grasping position. (c) Arrive at the first object position. (d) Use the THS module to obtain the object height. (e) Adjust the grasping position with the tactile calibration module and use the visual–tactile fusion algorithm to classify the object. (f) Finish the first object grasp. (g) Grasp and classify the second object. (h) Grasp and classify the third object.

found on the website https://sites.google.com/view/visual-tactilefusion.

G. Exp. 7: Grasping in Visually Undetectable Scenes

Finally, we test transparent object grasping in highly dynamic underwater scenes, where the object becomes visually undetectable, as shown in Fig. 23(f). In this case, the touch-first grasp strategy is applied, which incorporates the TPE module. We assume that the object will not be moved by the water wave. The experimental procedure is shown in Fig. 26. Through the experiment, we find that the exploration step length (distance between two exploration positions) of the gripper has a significant impact on the success rate of grasping, so we conduct a comparison experiment with exploration step lengths of 5 cm, 10 cm, and 15 cm, respectively, and 20 experiments are conducted for each step length. The results are shown in Table VII, where the average time consumed refers to the time consumed to successfully find the object, and the failure cases are not counted. From the results, we can see that the smaller the step size, the higher the success rate of grasping, but also the more time consumed. Besides, we compare the grasping experiments in three environments, i.e., plane, irregular, and

TABLE VII
EXPERIMENTAL RESULTS FOR GRASPING IN DYNAMIC UNDERWATER

Exploration step length	5 cm	10 cm	15 cm
Grasping success rate	90%	65%	25%
Average time consumption	121 s	78 s	52 s

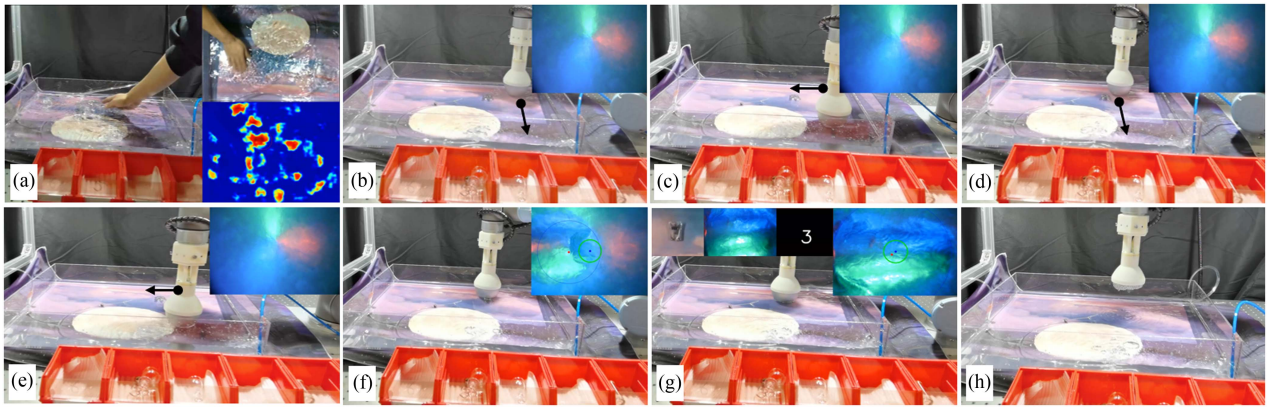

Fig. 26. Transparent object grasping in highly dynamic underwater scenes. (a) Transparent object-grasping position detection results. (b)–(e) Explore the transparent object within a specific area using tactile perception. (f) Adjust the grasping position after contacting the object. (g) Visual–tactile fusion classification. (h) Finish grasping.

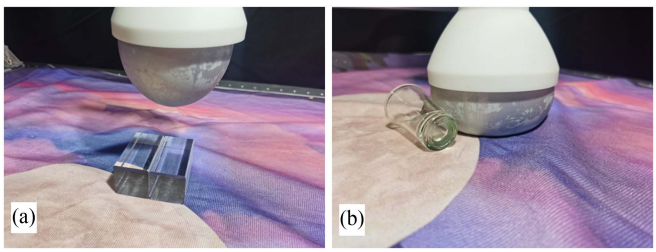

Fig. 27. Some failed grasping trials. (a) Grasping trial for two flat transparent objects with the same height. (b) Transparent object collides with the gripper, thus causing the object to slide.

visually undetectable scenes. In each scene, we conduct 20 grasping experiments, and the average time consumption from the beginning to reach the appropriate grasping position is 22 s, 32 s, and 121 s (the exploration step is 5 cm in a high-dynamic underwater scene) in three environments, respectively.

H. Failed Trials and Limitations

There are some failed trials in the experiments, as shown in Fig. 27. The first case is to grasp two close transparent objects with the same height. It sometimes fails mainly because such transparent objects have not only the same texture but also similar height information. The second case is that if the grasping position detection error is close to the radius of sensing of the gripper, the edge of the gripper will easily collide with the transparent object and cause the object to slide. In this case, the grasping may also fail. However, as long as the deviation of the detected gripping position from the actual gripping position do not exceed the radius of the gripper, almost no slipping and failure will occur.

VII. Conclusion

To solve the challenging problem of detecting, grasping, and classifying transparent objects, in this article, a visual–tactile fusion framework based on the synthetic dataset was proposed. First, we used the Blender simulation engine to render synthetic datasets rather than manually annotated datasets. Besides, we used Gaussian mask instead of the traditional binarized annotation to make the generation of the grasping position more accurate. To achieve grasping position detection for transparent objects, an algorithm named TGCNN was proposed and multiple comparative experiments were conducted, which show that the algorithm can achieve good detection under different backgrounds and lighting conditions even when trained with only synthetic datasets. Considering the grasping difficulty caused by the limitation of visual detection, we proposed a tactile calibration method combined with the soft gripper TaTa to improve the grasping success rate by adjusting the grasping position with tactile information. The method improved the grasping success rate by 36.7% compared with vision-only grasping. Furthermore, to solve the classification problem of transparent objects in complex scenes, a transparent object classification method based on visual–tactile fusion was proposed, which improves the accuracy by 39.1% compared with the vision-only-based classification. In addition, to achieve transparent object grasping in irregular and visually undetectable scenes, we proposed the THS and TPE modules, which can compensate for the problem of transparent object grasping in the absence of visual information. Extensive experiments were designed systematically and the results verified the effectiveness of the proposed framework in various complex scenarios, including stacking, overlapping, undulating, sand, underwater scenes, etc. We believed that the proposed framework can also be applied to object detection in low-visibility environments, such as smoke and murky underwater, where tactile perception can compensate for the shortcomings of visual detection and improve classification accuracy by using visual–tactile fusion.

References

[1] T. Weng, A. Pallankize, Y. Tang, O. Kroemer, and D. Held, "Multi-modal transfer learning for grasping transparent and specular objects," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 3791–3798, Jul. 2020.

[2] Z. Zhou, T. Pan, S. Wu, H. Chang, and O. C. Jenkins, "Glassloc: Plenoptic grasp pose detection in transparent clutter," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 4776–4783.

[3] S. Sajjan et al., "ClearGrasp: 3D shape estimation of transparent objects for manipulation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2020, pp. 3634–3642.

[4] 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 *Proc. Int. Conf. Robot. Autom.*, 2022, pp. 6151–6157.

[5] J. Ichnowski, Y. Avigal, J. Kerr, and K. Goldberg, "Dex-NeRF: Using a neural radiance field to grasp transparent objects," in *Proc. Conf. Robot Learn. (CoRL)*, 2022, pp. 526–536.

[6] V. Seib, A. Barthen, P. Marohn, and D. Paulus, "Friend or foe: Exploiting sensor failures for transparent object localization and classification," in *Proc. Int. Conf. Robot. Mach. Vis.*, 2017, vol. 10253, pp. 94–98.

[7] Q. Dai et al., "Domain randomization-enhanced depth simulation and restoration for perceiving and grasping specular and transparent objects," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 374–391.

[8] K. Huang, L.-T. Jiang, J. R. Smith, and H. J. Chizeck, "Sensor-aided teleoperated grasping of transparent objects," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 4953–4959.

[9] J. Chang et al., "GhostPose: Multi-view pose estimation of transparent objects for robot hand grasping," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 5749–5755.

[10] E. Xie, W. Wang, Wenhai Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 696–711.

[11] E. Xie et al., "Segmenting transparent objects in the wild with transformer," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, 2021, pp. 1194–1200.

[12] J. Jiang, G. Cao, T.-T. Do, and S. Luo, "A4T: Hierarchical affordance detection for transparent objects depth reconstruction and manipulation," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 9826–9833, Oct. 2022.

[13] Z. Zhou, X. Chen, and O. C. Jenkins, "LIT: Light-field inference of transparency for refractive object localization," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4548–4555, Jul. 2020.

[14] M. Fritz, G. Bradski, S. Karayev, T. Darrell, and M. Black, "An additive latent feature model for transparent object recognition," in *Advances in Neural Information Processing Systems*, vol. 22. Red Hook, NY, USA: Curran Associates, Inc., 2009.

[15] K. McHenry, J. Ponce, and D. Forsyth, "Finding glass," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2005, vol. 2, pp. 973–979.

[16] K. Maeno, H. Nagahara, A. Shimada, and R.-I. Taniguchi, "Light field distortion feature for transparent object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2786–2793.

[17] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 21–37.

[18] H. Fan et al., "Transparent object tracking benchmark," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10714–10723.

[19] H. Xu, Y. R. Wang, S. Eppel, A. Aspuru-Guzik, F. Shkurti, and A. Garg, "Seeing glass: Joint point-cloud and depth completion for transparent objects," in *Proc. 5th Annu. Conf. Robot Learn.*, 2022, pp. 827–838.

[20] L. Zhu et al., "RGB-D local implicit function for depth completion of transparent objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4647–4656.

[21] X. Liu, R. Jonschkowski, A. Angelova, and K. Konolige, "Keypose: Multi-view 3D labeling and keypoint estimation for transparent objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11599–11607.

[22] J. Kerr et al., "Evo-NeRF: Evolving NeRF for se-quential robot grasping of transparent objects," in *Proc. Conf. Robot Learn. (CoRL)*, 2023, pp. 353–367.

[23] H. Cao, J. Huang, Y. Li, J. Zhou, and Y. Liu, "Fuzzy-depth objects grasping based on FSG algorithm and a soft robotic hand," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 3948–3954.

[24] I. Lysenkov and V. Rabaud, "Pose estimation of rigid transparent objects in transparent clutter," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2013, pp. 162–169.

[25] J. Oberlin and S. Tellex, "Time-lapse light field photography for perceiving transparent and reflective objects," in *Proc. Robot., Sci. Syst.*, 2017, pp. 1–10.

[26] Z. Zhou, Z. Sui, and O. C. Jenkins, "Plenoptic Monte Carlo object localization for robot grasping under layered translucency," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1–8.

[27] J. Dargahi and S. Najarian, "Human tactile perception as a standard for artificial tactile sensing—A review," *Int. J. Med. Robot. Comput. Assist. Surg., MRCAS*, vol. 1, no. 1, pp. 23–35, 2004.

[28] J. Zhang et al., "Finger-inspired rigid-soft hybrid tactile sensor with superior sensitivity at high frequency," *Nature Commun.*, vol. 13, no. 1, 2022, Art. no. 5076.

[29] B. W. An, S. Heo, S. Ji, F. Bien, and J.-U. Park, "Transparent and flexible fingerprint sensor array with multiplexed detection of tactile pressure and skin temperature," *Nature Commun.*, vol. 9, no. 1,2018, Art. no. 2458.

[30] Z. Song et al., "A flexible triboelectric tactile sensor for simultaneous material and texture recognition," *Nano Energy*, vol. 93, 2022, Art. no. 106798.

[31] S. Stassi, V. Cauda, G. Canavese, and C. F. Pirri, "Flexible tactile sensing based on piezoresistive composites: A review," *Sensors*, vol. 14, no. 3, pp. 5296–5332, 2014.

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

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

[34] C. Trueeb, C. Sferrazza, and R. D'Andrea, "Towards vision-based robotic skins: A data-driven, multi-camera tactile sensor," in *Proc. IEEE 3rd Int. Conf. Soft Robot.*, 2020, pp. 333–338.

[35] J. W. Goodman, *Statistical Optics*. Hoboken, NJ, USA: Wiley, 2015.

[36] A. Fetić, D. Jurić, and D. Osmanković, "The procedure of a camera calibration using camera calibration toolbox for MATLAB," in *Proc. IEEE 35th Int. Conv. MIPRO*, 2012, pp. 1752–1757.

[37] D. Morrison, P. Corke, and J. Leitner, "Learning robust, real-time, reactive robotic grasping," *Int. J. Robot. Res.*, vol. 39, no. 2-3, pp. 183–201, 2020.

[38] H. Cao, G. Chen, Z. Li, Q. Feng, J. Lin, and A. Knoll, "Efficient grasp detection network with gaussian-based grasp representation for robotic manipulation," *IEEE/ASME Trans. Mechatronics*, 2022.

[39] N. Singla, "Motion detection based on frame difference method," *Int. J. Inf. Comput. Technol.*, vol. 4, no. 15, pp. 1559–1565, 2014.

[40] T. Brox, C. Bregler, and J. Malik, "Large displacement optical flow," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 41–48.

[41] C. Choi, W. Schwarting, J. DelPreto, and D. Rus, "Learning object grasping for soft robot hands," *IEEE Robot. Autom. Lett.*, vol. 3, no. 3, pp. 2370–2377, Jul. 2018.

[42] J. Mahler et al., "Dex-Net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robot., Sci. Syst.*, 2017, pp. 1–10.

[43] 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.

[44] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. J. Robot. Res.*, vol. 27, no. 2, pp. 157–173, 2008.

[45] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from RGBD images: Learning using a new rectangle representation," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2011, pp. 3304–3311.

[46] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2015, pp. 1316–1322.

[47] P. Shukla, N. Pramanik, D. Mehta, and G. C. Nandi, "Generative model based robotic grasp pose prediction with limited dataset," *Appl. Intell.*, vol. 52, pp. 9952–9966, 2022.

[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[49] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal, "The importance of skip connections in biomedical image segmentation," in *Int. Workshop Deep Learn. Med. Image Anal.,*, 2016, pp. 179–187.

Shoujie Li (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 in data science and information technology with Tsinghua-Berkeley Shenzhen Institute, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

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

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 electronic information with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China.

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

Wenbo Ding (Member, IEEE) received the B.S. (Hons.) and Ph.D. (Hons.) degrees in electrical engineering from Tsinghua University, Beijing, China, in 2011 and 2016, respectively.

He worked as a Postdoctoral Research Fellow with Georgia Tech under the supervision of Professor Z. L. Wang from 2016 to 2019. He is currently an Associate Professor and the Ph.D. supervisor with Tsinghua-Berkeley Shenzhen Institute, Institute of Data and Information, Shenzhen International Graduate School, Tsinghua University, Shenzhen, China, where he leads 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 a recipient of many prestigious awards, including the Gold Medal of the 47th International Exhibition of Inventions Geneva and the IEEE Scott Helt Memorial Award.

Houde Liu (Member, IEEE) received the B.S. degree in automation from the Huazhong University of Science and Technology, Wuhan, China, in 2007, and the M.S. and Ph.D. degrees in control science and engineering from the Harbin Institute of Technology, Harbin, China, in 2009 and 2013, respectively.

He is currently a Full Professor with the Centre for Artificial Intelligence and Robotics, Tsinghua Shenzhen International Graduate School, Tsinghua University, Beijing, China. His research interests include space robots, and motion planning and control.

Linqi Ye (Member, IEEE) received the 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 with Cornell University, Ithaca, NY, USA. From 2019 to 2022, he was a Postdoctor with Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. He is currently an Associate Professor with Shanghai University, Shanghai, China. His research interests include humanoid robots and reinforcement learning.

Chongkun Xia (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from Northeastern University, Shenyang, China, in 2021.

He is a Postdoctor with the Centre for Artificial Intelligence and Robotics, Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. His research interests include robotics, motion planning, and machine learning.

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, and 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 June 2010 to February 2014, he was the Postdoctoral Research Fellow with 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.

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

Since Fall 2000, he has been with the Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, ON, Canada, where he is currently a Professor and the Director of Communication and Signal Processing Applications Laboratory. He has served as the Program Director of graduate studies. He is cross-appointed to the Finance Department, Ted Rogers School of Management, Ryerson University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is a frequent Consultant for biotech companies and investment firms. His research interests include sensor networks and Internet of Things, machine learning, statistical signal processing, image and multimedia content analysis, and applications in Big Data, finance, and marketing.

Dr. Zhang is Fellow of the Canadian Academy of Engineering and the Engineering Institute of Canada, and a registered Professional Engineer in Ontario, Canada, and a member of Beta Gamma Sigma Honor Society. He is the general Co-Chair of the IEEE International Conference on Acoustics, Speech, and Signal Processing, in 2021. He is the general Co-Chair of 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the general Co-Chair of 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He was an elected member of the ICME Steering Committee. He is the General Chair of the IEEE International Workshop on Multimedia Signal Processing, in 2015. He is the Publicity Chair of the International Conference on Multimedia and Expo 2006 and the Program Chair of International Conference on Intelligent Computing in 2005 and 2010. He served as a Guest Editor for *Multimedia Tools and Applications* and the *International Journal of Semantic Computing*. He was a tutorial speaker at the 2011 ACM International Conference on Multimedia, the 2013 IEEE International Symposium on Circuits and Systems, the 2013 IEEE International Conference on Image Processing, the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, the 2017 International Joint Conference on Neural Networks, and the 2019 IEEE International Symposium on Circuits and Systems. He is the Editor-in-Chief for the IEEE JOURNAL OF SELECTED TOPICS IN SIGNAL PROCESSING. He is the Senior Area Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING. He served as the Senior Area Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING and an Associate Editor for the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON SIGNAL PROCESSING, and IEEE SIGNAL PROCESSING LETTERS. He was a recipient of 2020 Sarwan Sahota Ryerson Distinguished Scholar Award—the Ryerson University highest honor for scholarly, research, and creative achievements. He is selected as an IEEE Distinguished Lecturer by the IEEE Signal Processing Society for the term 2020–2021, and by the IEEE Circuits and Systems Society for the term 2021–2022.