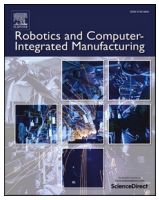
Robotics and Computer–Integrated Manufacturing 84 (2023) 102601

Contents lists available at ScienceDirect

Robotics and Computer-Integrated Manufacturing

journal homepage: www.elsevier.com/locate/rcim



Digital twin-enabled grasp outcomes assessment for unknown objects using visual-tactile fusion perception

Zhuangzhuang Zhang [a], Zhinan Zhang [a,*], Lihui Wang [b], Xiaoxiao Zhu [a], Huang Huang [c], Qixin Cao [a]

[a] School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China
[b] Department of Production Engineering, KTH Royal Institute of Technology, Stockholm 10044, Sweden
[c] Beijing Institute of Control Engineering, Beijing 100191, China

ARTICLE INFO

Keywords:
Grasp outcomes assessment
Visual-tactile perception
Deep learning
Multimodal fusion
Digital twin

ABSTRACT

Humans can instinctively predict whether a given grasp will be successful through visual and rich haptic feedback. Towards the next generation of smart robotic manufacturing, robots must be equipped with similar capabilities to cope with grasping unknown objects in unstructured environments. However, most existing data-driven methods take global visual images and tactile readings from the real-world system as input, making them incapable of predicting the grasp outcomes for cluttered objects or generating large-scale datasets. First, this paper proposes a visual-tactile fusion method to predict the results of grasping cluttered objects, which is the most common scenario for grasping applications. Concretely, the multimodal fusion network (MMFN) uses the local point cloud within the gripper as the visual signal input, while the tactile signal input is the images provided by two high-resolution tactile sensors. Second, collecting data in the real world is high-cost and time-consuming. Therefore, this paper proposes a digital twin-enabled robotic grasping system to collect large-scale multimodal datasets and investigates how to apply domain randomization and domain adaptation to bridge the sim-to-real transfer gap. Finally, extensive validation experiments are conducted in physical and virtual environments. The experimental results demonstrate the effectiveness of the proposed method in assessing grasp stability for cluttered objects and performing zero-shot sim-to-real policy transfer on the real robot with the aid of the proposed migration strategy.

1. Introduction

In the era of industrial automation and smart manufacturing, robots play a central role in various industrial and domestic applications. Meanwhile, robotic grasping as a typical application has been extensively researched [1,2]. However, the current robotic grasping is usually carried out with a fixed gripping force [3,4], which means that the robot performs open-loop grasping and cannot actively adjust its pose and gripping force to enhance the stability of the grasp. Furthermore, it is crucial to endow robots with the ability to grasp objects as gently as humans with minimum force because this can dramatically improve the robot's intelligence to copy the challenges in stochastic and disordered manufacturing scenarios [5]. As a prerequisite for intelligent grasping, grasp outcomes assessment remains an open and challenging task [6]. The grasp outcome is defined as either the success or failure of a grasping trial. The criterion for judgment is whether the object is still stably grasped after the robot is lifted to the desired position.

Before grabbing a random object, humans seamlessly combine the senses of vision and touch to determine if the grasp is stable. During this process, visual feedback provides geometric properties of the object surface, while tactile feedback establishes accurate and intuitive contact conditions between the hand and the object. That is, these two modalities are concurrent and complementary. However, owing to the hardware limitations of tactile sensors and the challenges introduced by the fusion methods [7], the integration of haptic perception into the visual grasping framework has remained a challenge. In recent years, the continued development and improvement of the vision-based tactile sensor (VBTS) has significantly facilitated multimodal grasping research. Because the modal information from VBTS can be easily extracted using the strong feature learning ability of convolutional neural networks (CNNs). Moreover, the integration with visual features is natural using MMFN. Among VBTSs, the GelSight-style (i.e., image

* Corresponding author.
E-mail address: zhinanz@sjtu.edu.cn (Z. Zhang).

https://doi.org/10.1016/j.rcim.2023.102601
Received 19 February 2023; Received in revised form 26 May 2023; Accepted 28 May 2023
Available online 9 June 2023

shading- based) is the most popular type. The representative sensors currently used in robotic manipulation tasks include GelSight [8], GelSlim [9], and DIGIT [10]. A VBTS consists of a contact module, a camera module, and an illumination module. The contact module interacts directly with the environment and is typically made of deformable elastomer. The illumination module that consists of several RGB LEDs is deployed to illuminate the elastomer from different directions. The built-in RGB camera is utilized to capture the deformation of the elastomer and generate high-resolution colored shading images to characterize contact forces and object surface properties. So far, several typical studies regarding grasp outcomes assessment have investigated MMFNs using visual and tactile modalities.

Calandra et al. [6] proposed a multimodal sensing framework for predicting grasp outcomes using tactile and visual input, and the experimental results demonstrated that the visual-tactile model substantially improved the grasping performance. They subsequently proposed a regrasping policy based on real-time grasp-stability evaluation using raw visual-tactile data, and the learned model allows the robot to grasp objects with significantly reduced gripping force [5]. To evaluate the grasp stability of deformable objects, Cui et al. [11] proposed a 3D CNN-based fusion perception network that takes the visual and tactile sequences as input. Additionally, they introduced an MMFN based on the self-attention mechanism to deeply fuse visual and tactile signals [12]. More recently, Kanitkar et al. [13] presented a multimodal dataset that contains tactile and visual data to investigate the grasp outcomes at specific holding poses. However, the above multimodal grasp stability evaluation methods based on visual-haptic fusion still have the following shortcomings. 1) The visual signals for network training are global images containing single object, robot, gripper, and environment, which makes it highly challenging to generalize the trained multimodal model to novel single-object or even cluttered object scenarios. 2) All the data information is collected from the real robot system. But this procedure is time-consuming and high-cost, and the size of the collected dataset is comparatively small. In addition, it is notable that a large and reasonable dataset is a primary prerequisite for deep learning methods.

Inspired by [14], this paper introduces the local point cloud within the gripper as the visual signal and proposes an end-to-end grasp stability evaluation network based on visual-tactile fusion. In the experiments, the raw tactile images are obtained from two GelSight-style high-resolution tactile sensors mounted on a two-finger gripper, while a depth camera captures the 3D point cloud data. As the two modalities are both local information related only to objects, the model trained with a large-size and reasonable dataset is much easier to generalize to novel single-object as well as cluttered object scenarios. To overcome the limitations of collecting large datasets in real-world robotics systems, the digital twin (DT) has proven to be an effective and efficient technique [15]. The DT is a dynamic virtual model of the physical entity digitally, making it feasible to simulate and control reality with the virtual. To accelerate the dataset generation process, in this paper, we first set up a dynamic virtual grasping environment and implement an autonomous visual-tactile dataset collection strategy. Then, aiming to perform zero-shot prediction tasks in the real world, we propose a migration strategy for visual and tactile modalities using domain randomization and domain adaptation to eliminate the sim-to-real transfer gap.

The primary contributions of this paper are in the following aspects: (1) An end-to-end grasp stability evaluation network that uses local visual and tactile inputs is proposed for cluttered and disordered scenarios. (2) A digital twin-enabled robotic grasping system is developed, and a policy for the autonomous collection of large-scale and reasonable visual-tactile datasets is implemented. (3) A migration strategy for visual and tactile modalities in robotic grasping tasks is proposed to eliminate the sim-to-real transfer gap. (4) Extensive validation experiments are conducted in both physical and virtual environments, and the results verify the high efficiency of the proposed method in cluttered object scenarios and zero-shot prediction tasks.

The remainder of this paper is structured as follows. Section 2 introduces the work related to grasping unknown objects, grasp-stability evaluation, and the digital twin of robotic grasping with visual and tactile perception. Section 3 formulates the multimodal grasp-stability evaluation problem. Section 4 describes the proposed end-to-end grasp stability evaluation network, multimodal dataset collection policy, and migration strategy. Sections 5 and 6 present the extensive validation experiments conducted in virtual and physical environments, respectively. An extended discussion of the methodology and experimental results is provided in Section 7. Finally, Section 8 is the conclusion and future work.

2. Related work

This paper addresses the problem of grasp outcomes assessment in cluttered and disordered manufacturing scenarios. Meanwhile, to accelerate the data collection process, we also propose dataset generation and migration strategies based on the digital twin approach. Therefore, the following sections summarize the concepts and work related to grasping unknown objects, grasp-stability evaluation, and digital twins.

2.1. Learning to grasp

The object pose estimation methods represented by template matching can obtain the object's 6D pose with known geometry [16], but it fails to cope with grasping unknown objects in complex and unstructured scenarios. With the rapid progress of deep learning, data-driven approaches greatly facilitate the grasping detection configuration in cluttered and disordered scenarios [17]. Mahler et al. [18] proposed a 2D grasp planning method that generates many candidate grasps from depth images and ranks them with GQ-CNN. Although the grasping success rate is high, the two-stage process is time-consuming. By introducing the fully convolutional networks (FCNs), Morrison et al. [19] presented GG-CNN which takes a deep image as input to predict the quality and grasp pose at every pixel simultaneously. To combine the apparent advantages of both, Hu et al. [20] proposed a new GGS-CNN using FCNs to generate grasp candidates and select the best grasp. Since 2D planar grasp cannot obtain many suitable grasp configurations, 6-DoF grasping methods have attracted much attention. Liang et al. [14] sampled grasps from the 3D point cloud and evaluated their robustness with PointNet [21]. Gou et al. [3] presented RGBD-Grasp to detect 7-DoF grasp including gripper width. By incorporating depth and RGB information, the demand for high-quality depth data could be effectively alleviated. From the above discussion, it can be seen that the current representative grasp detection methods in disordered scenarios only perceive visual information. While the rich haptic feedback is critical for allowing the robot to perform stable and gentle grasp operations. This paper focuses on the seamless integration of haptic and visual sensory data under the 6-DoF grasping framework to achieve the grasp outcomes assessment in cluttered scenarios.

2.2. Grasp-stability evaluation

Tactile sensors have been widely applied to robotic manipulation tasks because they can directly obtain rich contact information regarding robot-environment interactions in the absence of vision. Bekiroglu et al. [22] proposed a probabilistic learning framework to assess grasp stability based on haptic data from pressure-sensitive tactile sensors and machine learning methods. Romano et al. [23] used tactile event cues to drive the transitions between six discrete grasp states. Kwiatkowski et al. [24] combined tactile signals and proprioception to assess grasp stability using CNNs. Veiga et al. [25] predicted the future occurrence of slip from tactile data and modulated the contact forces accordingly. However, as mentioned above, these methods usually employ electronic tactile sensors (ETSs) that are limited in

high-resolution tactile information and tactile feedback, which also hinders the further advancement of robotic tactile sensing performance. Compared with ETSs, VBTSs (e.g., GelSight-style) have significant advantages in terms of high-resolution, high robustness, and visual-tactile fusion. Kolamuri et al. [26] used GelSight sensors to detect the rotational failure of grasp in an early stage and proposed a regrasping policy to improve grasp stability. Si et al. [27] trained a CNN-LSTM model to predict the grasp results by feeding a sequence of tactile images. But these research frameworks do not incorporate visual modality, and visual-tactile data should be concurrent and complementary in the early grasping stage to achieve optimal grasping results. For example, stable grasping of an object at a position far from its center of gravity often requires much greater force than near the center. On the other hand, visual feedback provides an intuitive representation of the geometric properties of an object's surface. Calandra et al. [6] demonstrated that incorporating tactile signals within a multimodal perception framework substantially improves grasping performance. Cui et al. [11] evaluated the grasp state of deformable objects using a 3D CNN-based visual--tactile fusion network. Kanitkar et al. [13] collected a multimodal dataset that contains visual-tactile data to study the effect of different holding poses on grasp stability. However, these studies share similar problems, such as inadequate training datasets and the inability to apply them to cluttered scenarios. This paper proposes an MMFN based on local visual-tactile features to evaluate grasp outcomes for single and cluttered objects. Meanwhile, a digital twin system is proposed to collect a reasonable and large-scale visual-tactile dataset.

2.3. Digital twin

To accelerate the verification of the algorithm, the application of DT provides an appealing avenue [28,40]. A DT model usually consists of virtual entities, physical entities, and the connection between the two, with virtual entities regarded as digital clones of physical entities [29]. Although generating datasets in the virtual world is highly efficient, the distribution shifting between the physical and virtual entities may lead to migration failure, which is also called the sim-to-real gap. This paper focuses on the simulation and migration of visual and haptic modalities.

Tobin et al. [30] explored the domain randomization technique, which successfully transfers the model trained on the simulated RGB images to real images by randomizing rendering in the simulator. But the RGB modality is sensitive to textures and lighting conditions. On the other hand, models trained with simulated depth modality have been demonstrated to generalize reasonably well for physical scenarios such as intelligent robotic grasping [18,31]. Nevertheless, compared to the visual modality, the VBTS is challenging to simulate. This is because an ideal high-resolution haptic simulator needs to model not only realistic optical properties, but also accurate contact dynamics. To simulate the GelSight-style sensor, researchers from [32] leveraged the Gazebo built-in camera [37] to capture the depth map of the contact area and rendered the RGB image based on Phong's model. The authors in [33] utilized the bidirectional path-tracing algorithm to generate more realistic synthetic GelSight images. However, the realistic result requires large computation consumption. Si et al. [35] proposed Taxim, an example-based method containing optical and marker motion field simulation. This method requires less than 100 contact examples from the real Gelsight sensor to calibrate the simulation model. Wang et al. [34] presented TACTO to simulate the VBTSs, including DIGIT [10] and OmniTact [41]. The proposed simulator used OpenGL to render depth and RGB images of a synchronized scene from PyBullet [36] and achieved a rendering speed of 200 frames per second. Despite the impressive results of the above studies, there is still a gap between synthetic tactile images and real images due to the complexity of modeling optical properties and contact dynamics. For this purpose, domain adaptation techniques are introduced to eliminate the sim-to-real gap. Chen et al. [38] used CycleGAN [42] to train the unpaired data collected from the virtual and real world. However, the physical property of the tactile sensor is neglected. Lin et al. [39] leveraged an image-to-image translation GAN [43] to perform the sim-to-real transfer via real-to-sim image translation. But they only evaluated zero-shot performance in simple scenarios, such as edge-following and surface-following.

This paper attempts to perform zero-shot prediction tasks in disordered grasping scenarios. To this end, we first propose a collection policy for paired tactile data from the virtual and physical world. Then we train a conditional adversarial network built on the pix2pix method [43] to map real tactile images to corresponding virtual ones, aiming to eliminate the gap in both optical properties and contact dynamics. Pix2pix is an image-to-image translation method that uses a U-Net generator and patch-based fully convolutional network discriminator. In addition, several domain randomization techniques are used to make the virtual 3D point clouds match the real ones.

3. Problem statement

This paper aims to learn a function f_θ to predict the grasp outcomes in disordered scenarios based on the visual-tactile information. As shown in Fig. 1, we consider a parallel-jaw gripper in this paper. Grasp pose $\mathcal{G}$ is defined by a tuple:

$$\mathcal{G} = \left(x, y, z, r_x, r_y, r_z\right) \in \mathbb{R}^6 \tag{1}$$

where (x,y,z) denotes the translation vector while (r_x,r_y,r_z) denotes the rotation vector. All sensor observations are represented by a tuple:

$$\mathcal{O} = (\mathcal{P}, \mathcal{I}_L, \mathcal{I}_R) \tag{2}$$

where $\mathcal{P} \in \mathbb{R}^{3\times N}$ is the local point cloud captured by a depth camera while $\mathcal{I}_L \in \mathbb{R}^{3\times H\times W}$ and $\mathcal{I}_R \in \mathbb{R}^{3\times H\times W}$ denote high-resolution tactile images acquired from two VBTSs mounted on the gripper. Let $\mathcal{S}$ denote the current state, including the physical properties of all entities and their mutual relationships.

Given the observations $\mathcal{O}$, we first use the visual encoding network (VEN) and tactile encoding network (TEN) to extract visual and tactile features,

$$\begin{aligned} F_V &= VEN(\mathcal{P}) \\ F_{TL} &= TEN(\mathcal{I}_L) \\ F_{TR} &= TEN(\mathcal{I}_R) \end{aligned} \tag{3}$$

Then a co-attention-based network is employed to fuse the extracted visual and tactile features,

$$F_{V,T} = FN(F_V, F_{TL}, F_{TR}) \tag{4}$$

Finally, a classification module containing fully connected (FC) layers and a loss function is applied to predict the grasp outcomes. More concretely, we use binary labels $\ell \in \{0,1\}$ to characterize the grasp results. A label of 1 indicates that the grasp will be successful. Let $\mathcal{D} = \{(\mathcal{O}_i, \ell_i)\}_{i=1}^N$ denote the training dataset, and θ be the weights of the whole prediction network. Our final objective is to minimize the cross-entropy (CE) loss $\mathcal{L}$ on the training dataset $\mathcal{D}$, resulting in optimal weights,

$$\theta^* = \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{(\mathcal{O}_i,\ell_i)\in\mathcal{D}} \mathcal{L}(\ell_i, f_\theta(\mathcal{O}_i)) \tag{5}$$

4. Methodologies

Grasping in a cluttered and disordered environment is a typical case of smart manufacturing. Therefore, this section mainly studies the multimodal network for grasp outcomes assessment in disordered grasping scenarios, digital twins, and migration strategy. To clearly show the relevance between the various parts, the framework diagram of the proposed method is first introduced in Section 4.1.

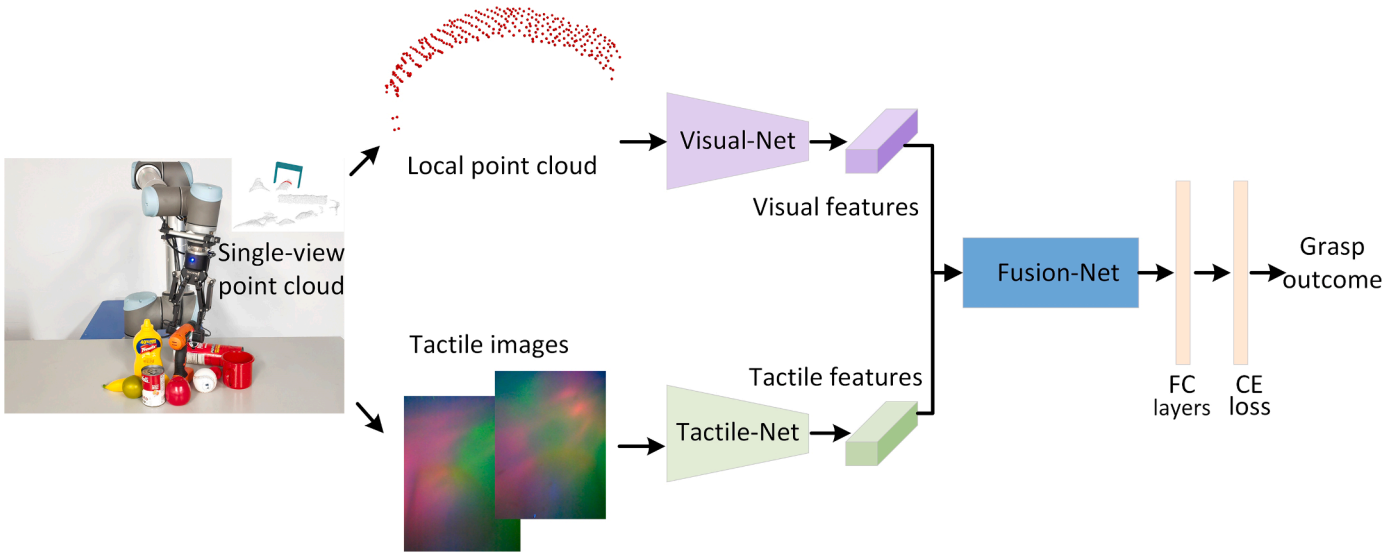


Fig. 1. A simplified framework of the proposed MMFN for evaluating grasping outcomes in cluttered scenarios.

4.1. Overall method

A large and reasonable dataset is a primary prerequisite for data-driven methods. But collecting data from the physical system is high-cost and time-consuming, leading to insufficient data to train deep neural networks. Therefore, this manuscript proposes a digital twin-enabled approach, as shown in Fig. 2. The framework diagram is divided into two parts: offline and online. The offline part shows the dataset acquisition and network training process; the online part is the real-time prediction of the grasping results.

When offline, an object set consisting of scanned models is first gathered. A large-scale visual-tactile dataset $\{(\mathscr{O}_i, \mathscr{A}_i)\}_{i=1}^{N}$ is then generated to train the multimodal fusion network. The resulting evaluation model is applied to real-time prediction tasks. To accelerate the data generation process, a self-supervised autonomous robotic data collection policy is designed. The synthetic dataset collection is carried out in a physics-based simulator PyBullet. At the same time, in order to successfully transfer the trained multimodal policy from the virtual to the real world, several domain randomization techniques are used in the process of generating visual data $\mathscr{P} \in \mathbb{R}^{3\times N}$. In addition, a virtual-real migration policy based on the conditional adversarial network [43] is proposed for tactile data $\mathscr{I} \in \mathbb{R}^{3\times H\times W}$. This policy mainly consists of

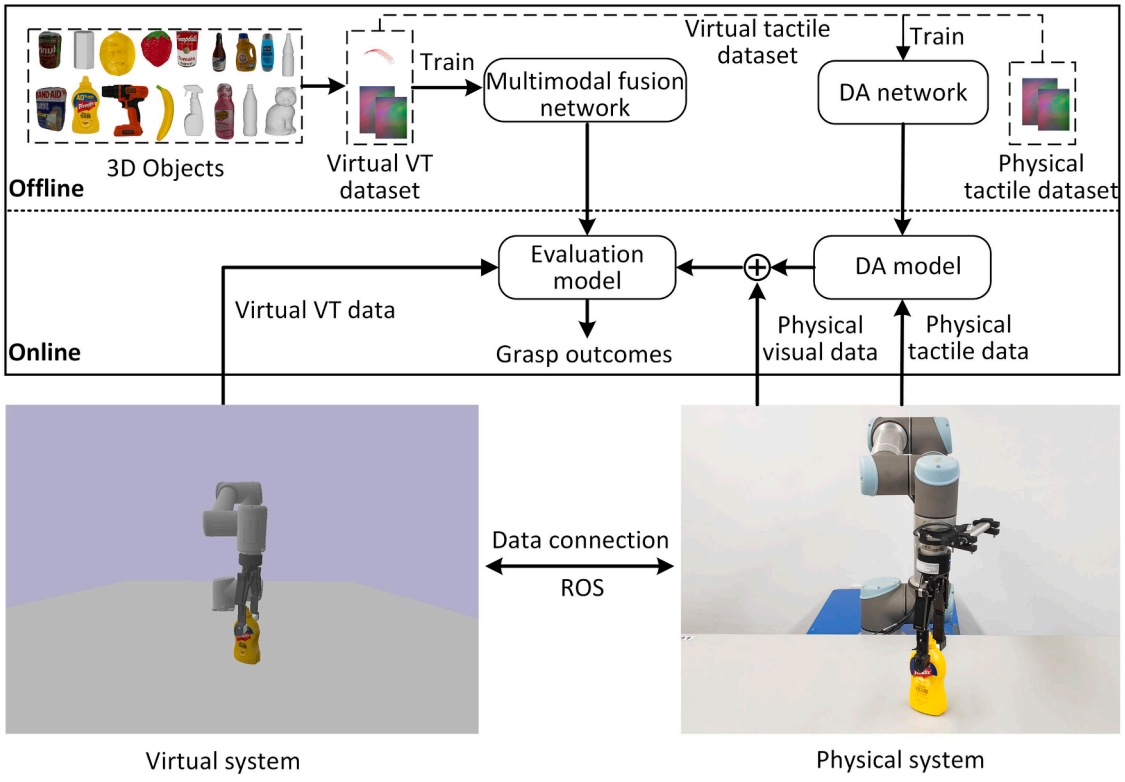


Fig. 2. The framework diagram of the proposed digital twin-based grasp outcomes assessment method. VT and DA are abbreviations of visual-tactile and domain adaptation, respectively.

collecting the paired dataset from the virtual and real world and training the domain adaptation network (i.e., conditional adversarial network). It is worth mentioning that the domain adaptation network maps real tactile images to virtual ones.

When online, the trained multimodal model predicts grasping results in both virtual and real environments. Specifically, this model performs evaluation in the virtual world using raw visual-tactile input, while in the real world, real tactile images need to be translated into virtual images by the trained DA model and then combined with the raw visual data to form the input.

4.2. Visual-tactile fusion network

When attempting to pick up objects, humans instinctively and seamlessly combine visual and haptic information to determine whether the current grasp will be successful. However, the limitations of sensor hardware and fusion algorithms make this highly challenging for current robots. Different works [5,6,11,12,13] have investigated the grasp outcomes assessment problem using visual-tactile fusion perception. However, they all employ the global images as the visual signals, which hampers the generalization of the trained multimodal model to novel single-object or even cluttered object scenarios. Therefore, this manuscript proposes an end-to-end multimodal fusion framework that integrates local visual and haptic inputs to facilitate generalization to cluttered scenes, as shown in Fig. 3. The visual signal is the object's local point cloud within the gripper and a PointNet [21] like network is introduced to extract the features of 3D spatial points. The raw tactile images are captured from two GelSight-style tactile sensors mounted on a two-finger gripper, and we use ResNet-18 [44] to produce the feature vectors. We further investigate the contributions and mutual relations of individual sensory modalities by a co-attention mechanism, as shown in Fig. 4. That is to say, these three global feature vectors are weighted and fused rather than directly concatenated. The architectures of these three partial networks are described as follows.

Visual-Net: In contrast to the global image, the raw 3D point cloud determined by the grasp configuration is only relevant to the object. In addition, the point cloud also represents geometric features more intuitively. Therefore, the point cloud of the object within the gripper closing area is selected as the visual signal. We adopt PointNet like architecture to process the raw 3D point cloud. The network takes n 3D points as input, and then two joint alignment networks are used to align input points and point features. The network subsequently uses multiple 1D convolutional layers to learn the spatial encoding of each point. Finally, global features are aggregated by max pooling and fed into an FC layer to produce a feature vector of 512.

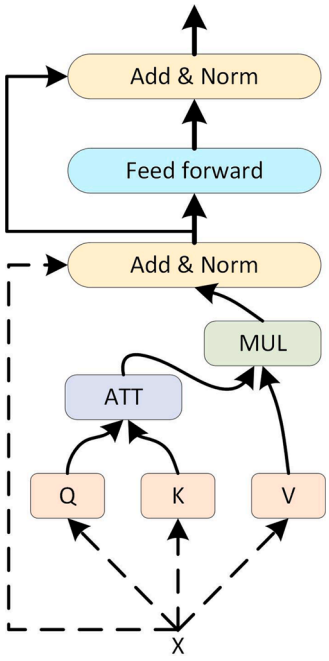


Fig. 4. The network architecture of the co-attention mechanism. Given an input sequence X, Q (Query), K (Key), and V (Value) are three embeddings produced by $\{Q, K, V\} = \{XW^Q, XW^K, XW^V\}$, where W^Q, W^K, and W^V are projection matrices. ATT denotes operation $\text{softmax}(QK^T/\sqrt{d_k})$, while MUL represents matrix multiplication.

Tactile-Net: We use ResNet-18 architecture to extract features from tactile signals, and both CNN networks share the same parameters. Furthermore, both tactile channels generate a 512-length feature vector. To initially match the distribution of virtual images to real readings, a 2D Gaussian filter $G(x, y)$ is first applied over real-time virtual image $\mathscr{I}$,

$$G(x,y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{6}$$

$$\mathscr{I} = \mathscr{I} * G \tag{7}$$

where $*$ denotes convolution operation, σ is the standard deviation. Then the virtual difference image is added to the real background image to form the training input. Specifically, the background is the image when the sensor is not in contact with the environment.

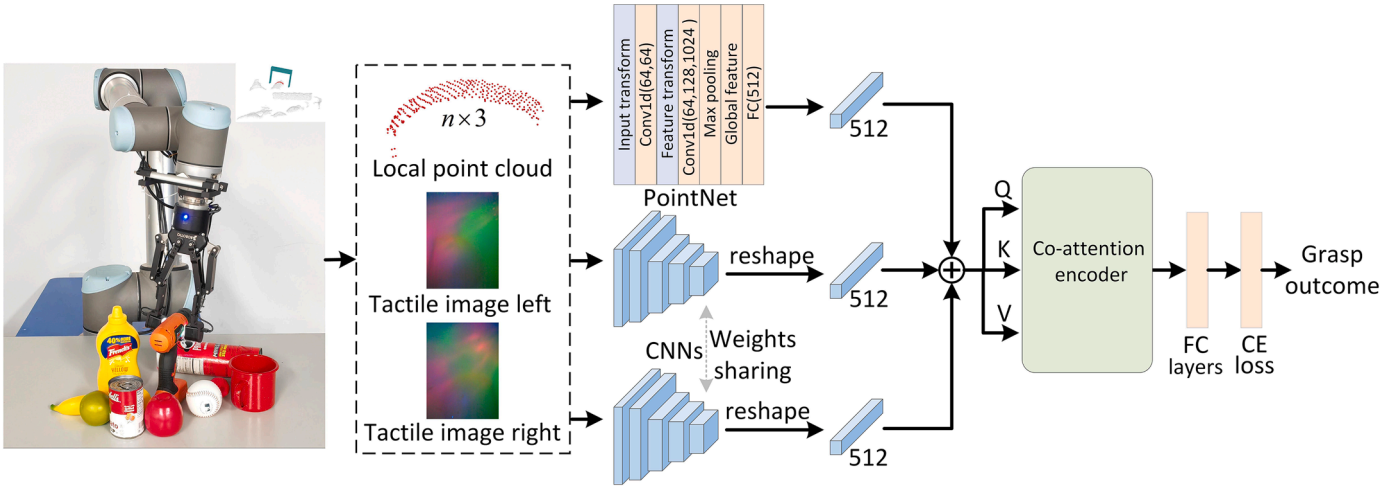


Fig. 3. The proposed multimodal prediction network that is trained on virtual data and ultimately applied to real-world scenarios.

$$\mathcal{I} = \mathcal{I} - \mathcal{I}_{sim,background} + \mathcal{I}_{real,background} \tag{8}$$

Fusion-Net: We denote the global features of three channels as $X = [F_V, F_{TL}, F_{TR}]$. The co-attention mechanism is computed as follows,

$$Z = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{9}$$

where in our case, the query $Q = XW_Q$, key $K = XW_K$, and value $V = XW_V$ have the same size, $W_Q \in \mathbb{R}^{d\times d_k}$, $W_K \in \mathbb{R}^{d\times d_k}$, $W_V \in \mathbb{R}^{d\times d_v}$ are weights. In the experiment, we set $d = d_k = d_v = 512$. The fully connected feed-forward network consists of two layers, which can be represented as, $FFN(Z) = \sigma(ZW_1 + b_1)W_2 + b_2$, where σ is the ReLU activation function. The residual connection with layer normalization $Z \leftarrow \text{LayerNorm}(Z + \text{sublayer}(Z))$ is applied to the output of each sublayer. The weighted vectors are subsequently fed to two FC layers and CE loss to complete the prediction of grasping results.

Training: The local 3D point cloud within the gripper closing area is sampled to a fixed number of 750 points before being fed to the network, and samples with less than 200 points will be discarded. For the tactile image with size $320 \times 240 \times 3$, we resize it to $256 \times 256 \times 3$ and randomly sample $224 \times 224 \times 3$ crops for data augmentation. To speed up the training, the ResNet-18 model pre-trained on ImageNet is employed. We train 10 epochs on the full network using Adam optimizer with a learning rate of 1×10^{-4} and batch size of 24. The experiments are implemented on Ubuntu 18.04 with one NVIDIA GTX1080Ti GPU and a 2.10 GHz Intel Xeon E5–2620 CPU.

4.3. Digital twin system

To collect a large and reasonable multimodal dataset, a digital twin system for robotic grasping is developed. Fig. 5(a) shows the entities in the real world, including a UR10 robot, a Robotiq 2-Finger adaptive gripper, a RealSense SR305 depth camera, two high-resolution tactile sensors, and an object to be grasped. We choose DIGIT [10] as tactile sensing hardware in the real world because of its easy integration with the gripper and ease of operation. At the same time, we use TACTO [34] to simulate DIGIT in the virtual environment. In addition, OpenGL integrated in PyBullet is utilized to render 3D point clouds. All simulated hardware has the same CAD dimensions as in the real environment and is imported via URDF. The robot operating system (ROS) [46] is utilized as a communication middleware between virtual and real environments.

The coordinate systems used in the grasping are marked in Fig. 5(b), and they mainly consist of the base coordinate system (B), the gripper or control coordinate system (G), the canonical grasp reference frame (C), and the object reference frame (O). When performing grasping, C is required to coincide with the grasp configuration represented in B. G is the control coordinate system which means the grasping homogeneous matrix is the transformation from B to G.

4.4. Dataset collection

Collecting data in the real world is time-consuming and laborious, making it extremely challenging to generate a large-scale dataset, which also has become a bottleneck restricting the further development of data-driven approaches. In this section, we propose a multimodal data collection policy and use the digital twin of robotic grasping to generate a large-scale visual-tactile dataset. This dataset $\mathcal{D} = \{(\mathcal{O}_i, \mathcal{C}_i)\}_{i=1}^{N}$ contains tactile images and rendered point clouds for each grasp configuration.

To generate the visual-tactile dataset, we first select an object set that has 130 3D scanned models from real-life objects. This set includes 40 objects from the YCB dataset [47], 40 objects from the BigBird dataset [48], and 50 objects from the KIT dataset [49]. Some of these objects are listed in Fig. 6. Moreover, 20 primitive objects are specifically designed to learn more basic grasping experiences. Primitive objects refer to simple geometric shapes that can approximate various forms and shapes. We select four primitive shapes: hexahedrons, cylinders, spheres, and cones based on how robots grasp everyday objects. Each shape includes five different scaling sizes. The training object set consists of 120 objects, and the rest are used as test set.

The overall process framework for visual-tactile dataset generation is constructed, as shown in Fig. 7. The first step is to sample grasp configurations offline for a single object. Before sampling, the origin of the object's reference frame is set to the mesh center of mass, and the coordinate axes are calculated with Principal Component Analysis (PCA) on the mesh vertices. In the grasp sampler module, each mesh in OBJ format is first converted to a signed distance function (SDF). Then we sample 400 force closure grasp candidates for each object using the antipodal grasp sampling method [50], and the sampled grasp configurations are represented in the object's reference frame. For simulation, we use the MeshPy library to calculate the stable poses for each object on a table surface and make corrections in the PyBullet gravity environment, discarding entries that do not make sense (i.e., unstable poses). Twenty collision-free grasps are subsequently computed for each stable

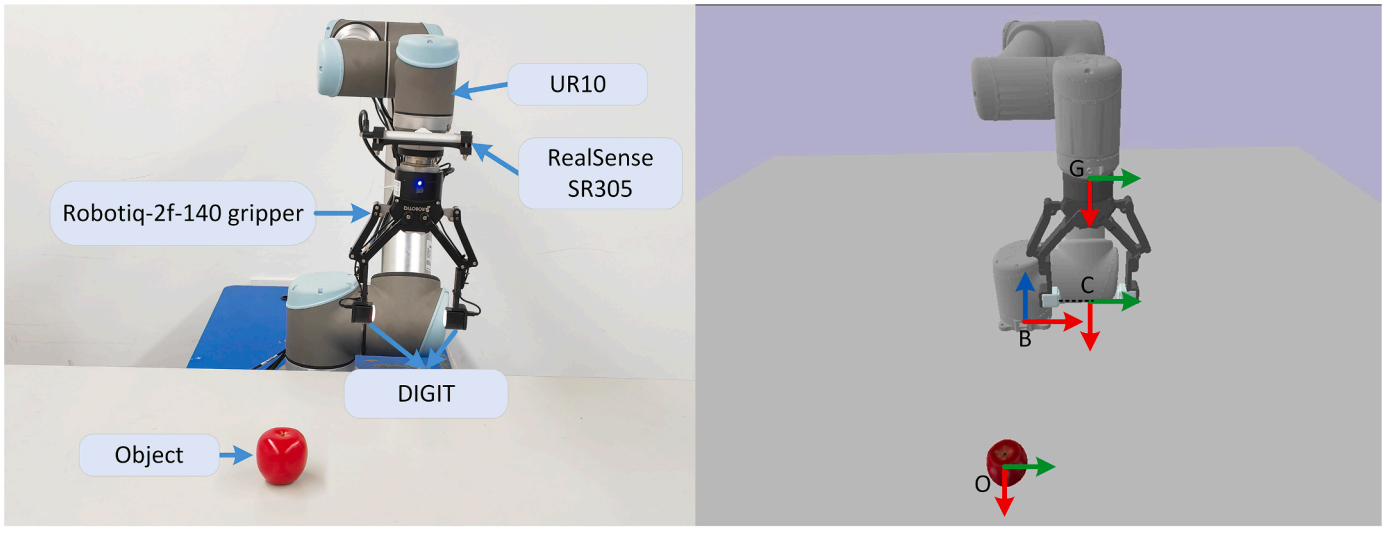


Fig. 5. Digital twin system. (a) Physical grasping platform (b) Virtual grasping platform.

Fig. 6. Partial representative objects from datasets YCB, BigBird, KIT, and primitive objects, respectively.

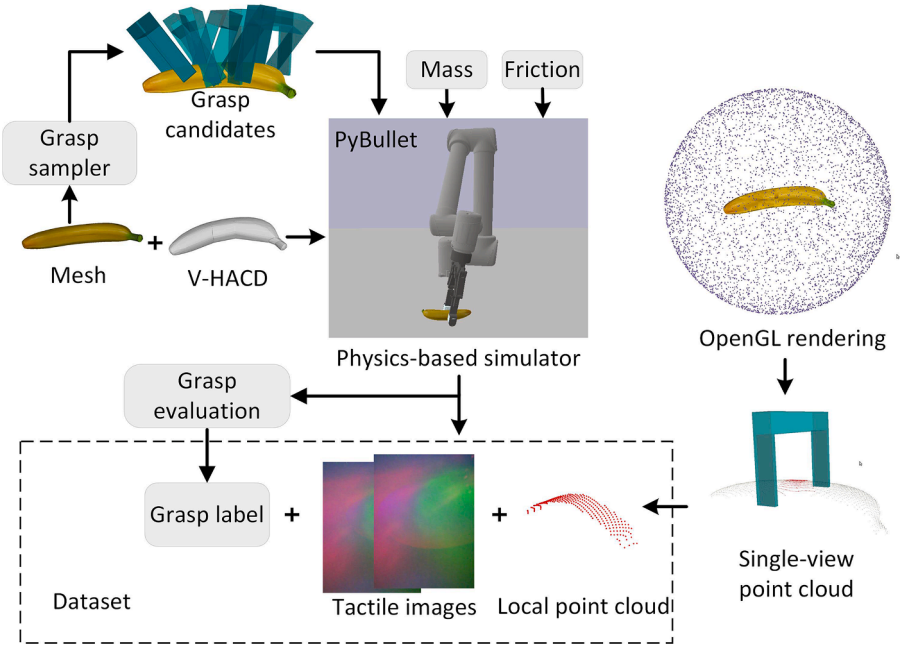


Fig. 7. Overall process framework for visual-tactile dataset generation.

pose on a planar worksurface. The resulting grasps corresponding to each stable pose are stored in the HDF5 file.

The second step is to set up the virtual environment. To improve collision detection efficiency, the volumetric hierarchical approximate convex decomposition (V-HACD) is used as a collision model to approximate the convex hull of the object, while OBJ mesh serves as the visual representation. To simplify the problem, we assume that the grasping is a quasi-static process with a Coulomb friction model, and the friction coefficient is set to $\mu = 0.6$ for objects and gripper in physical trials. In addition, we assign a mass of 0.5 kg at the objects' center.

The third step is to perform grasping trials to collect the dataset. When resetting the scene, the end of the gripper will be at a height of 0.5 m from the ground. Then each object will be loaded with all stored stable poses in order, and each stable pose comes with 20 grasp configurations. For each grasp configuration, the robot will perform 30 grasping trials, starting with a gripping force of 5N and rising at an interval of 1N. After closing the gripper, images from two tactile sensors are recorded. For point cloud rendering, we use a scene without the robot, and each object will be loaded into PyBullet as before. Then 100 sets of point clouds from different locations are rendered for each stable pose using an OpenGL camera model. More concretely, firstly, camera positions are randomly sampled from a spherical surface with a height $z > 0$, radius $r = 0.5m$, and center at the origin of the object's reference frame. Secondly, the camera points toward the object's center and renders all the single-view point clouds for each stable pose. All single-view point clouds are subjected to plane segmentation and outlier removal processing and paired with 20 configurations corresponding to the same stable pose. In other words, each configuration matches 100 rendered point clouds.

Finally, collision detection is performed between the processed point cloud and the gripper model with the grasp configuration registered to obtain the point cloud within the gripper, and the resulting local point cloud will be transformed into the local grasp coordinate that coincides with the canonical grasp reference frame (C) when grasping. This eliminates the effect of different camera positions along the radius producing different point clouds. Thus the sphere radius r can be changed to other values.

The labels for grasping trials are automatically generated, and the label generation process is shown in Fig. 8. The entire process consists of five steps. First, read a grasp configuration corresponding to this stable pose from the HDF5 file in sequence, and transform it into the base coordinate system B. The robot then reaches and closes the gripper with the desired force. At the same time, the tactile images are recorded. To enhance the robustness of the grasp outcomes, after lifting the object, the robot performs a shaking action to check if the object is still stable. Finally, if the object is still in hand (i.e., the object's height along the z-axis $z > h$), the label $\ell = 1$. Otherwise, $\ell = 0$. The dataset is collected with 16 processes to obtain a total of 1.38 $\times 10^7$ groups of data, where the ratio of positive to negative labels is around 6 to 4. We summarize the proposed dataset generation method in Algorithm 1.

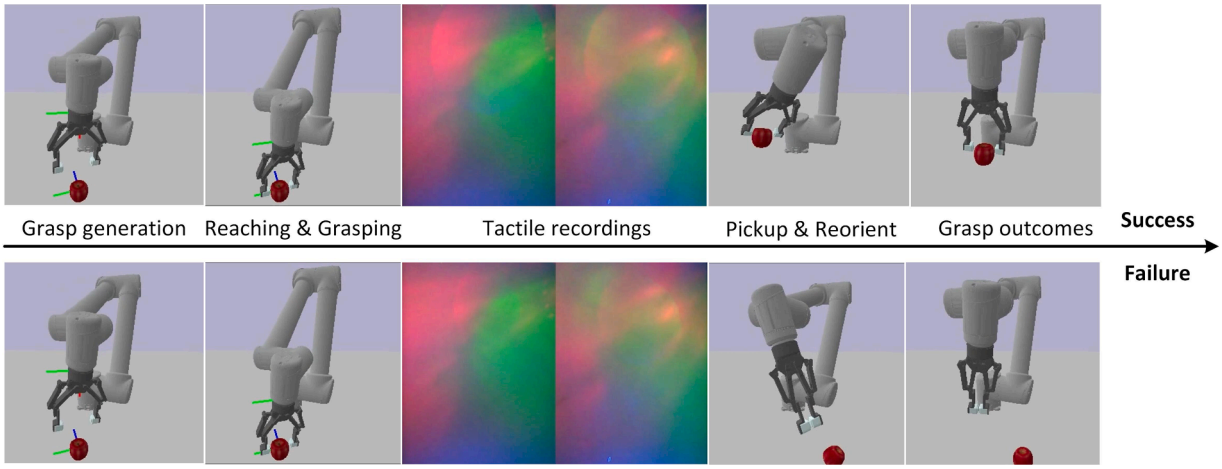


Fig. 8. Label generation process. This figure shows the grasp outcomes with gripping forces of 15N and 5N, respectively, given the same grasp configuration.

Algorithm 1
Visual-tactile dataset generation.

Require: object's grasp configuration set $\mathcal{G}$.
Read the number of stable poses p.
for i in range(p) **do**
  gripping force $f = 5$N.
  Get 100 local point clouds pc.
    for j in range(20) **do**
      for k in range(30) **do**
        Execute the jth grasp with f.
        Record tactile images $\mathcal{I}_{jk}$.
        Lift, shake, and get label ℓ.
        Pair (pc, $\mathcal{I}_{jk}$, ℓ)
        f+= 1.
      end for
    end for
end for

4.5. Migration strategy

Compared to the visual modality, the simulation for image shading-based haptic is significantly challenging due to the complex optical properties and contact dynamics model. Despite the impressive results of recent studies [32,35] on the simulation of vision-based tactile sensors, the gap between virtual and real images still exists. This paper proposes a data-driven migration strategy for robotic grasping scenarios. Overall, this strategy consists of paired dataset collection and conditional adversarial network training.

The paired data collection process is shown in Fig. 9. Firstly, we select 20 objects from the YCB [47] dataset and gather real objects together with the corresponding 3D scanned models. For each real object, we manually place a few stable poses. After that, we use a surface-based matching method to estimate the object's 6D pose. Specifically, this method requires only a single-view point cloud of the object observed by the robot at the initial position and a 3D model of the object, such as OBJ, PLY, etc. Moreover, this method is implemented using the vision software HALCON [51] for real-time pose estimation. Secondly, the estimated pose is expressed in the base coordinate system (B) and synchronized to the virtual world via ROS service. In the virtual world, 20 collision-free grasps corresponding to each object's pose are filtered from the set containing 400 grasps. Finally, these grasps are synchronized sequentially to the real scene. For each synchronized grasp, the robots in both environments will perform 20 grasping trials, starting with a gripping force of 10N (the minimum gripping force of robotiq-2f-140 gripper) and rising at an interval of 1N. When closing the gripper, the paired images ($\mathcal{I}_r$,$\mathcal{I}_v$) from real and virtual environments are recorded. Furthermore, for different grasps, the object's 6D pose will be re-estimated and synchronized to eliminate the position deviation that may be caused by the previous grasp. The final dataset contains a total of 23,600 paired data. We split the dataset into training (80%) and test (20%) sets. The paired data collection policy is summarized in Algorithm 2.

The migration strategy aims to learn a mapping $f = G(x)$ from real tactile images to corresponding virtual ones, with which the MMFN can be trained using large amounts of virtual data and transferred to the real

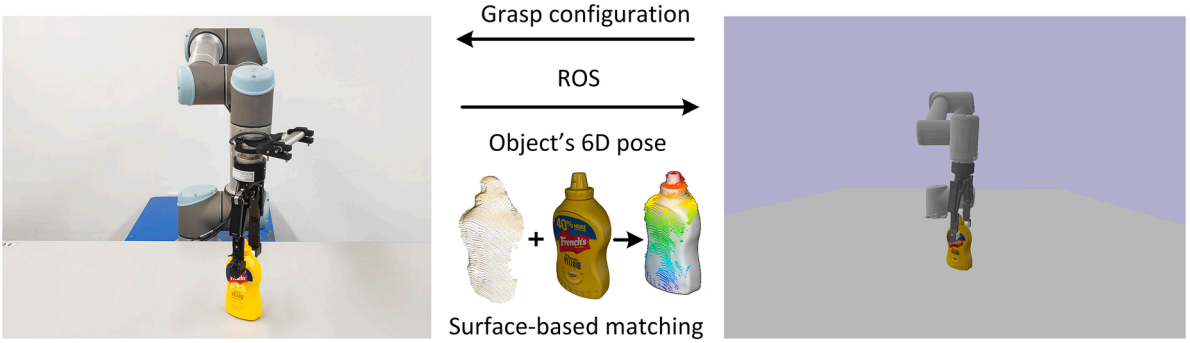


Fig. 9. Paired data collection process. During this process, the object's 6D pose from the real world is synchronized to the virtual world via ROS service. To that end, the pose is estimated by surface-based matching and expressed in the base coordinate system (B). After that, the grasp configuration generated in the virtual world is also synchronized to the real world.

Algorithm 2
Paired tactile data collection policy.

Require: object's grasp configuration set $\mathcal{G}$. Object set $\mathcal{O}$.
for i in range(20) **do**
  Manually select n stable poses for the ith object.
  for j in range(n) **do**
    Estimate the object's 6D pose p.
    Sync to the virtual world.
    Filter 20 collision-free grasps.
    for k in range(20) **do**
      Sync kth grasp to real world.
      20 grasping trials in virtual world.
      20 grasping trials in real world.
      Pair $(\mathcal{I}_r, \mathcal{I}_v)$.
    end for
  end for
end for

world. The migration training network is built on the pix2pix method [43], as shown in Fig. 10. The pix2pix method is proposed for image-to-image translation and consists of a U-Net architecture as the generator G and a patch-based fully convolutional network as the discriminator D. For our task, the objective of the discriminator D is to reveal the differences between virtual and generated images, while the generator G is to translate real images to virtual-like images to fool the discriminator D. The objective of conditional GANs can be expressed as,

$$G^* = \arg\ \min_G \max_D \mathscr{L}_{cGAN}(G, D) \tag{10}$$

where the regular adversarial loss is derived as,

$$\mathscr{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_x[\log(1 - D(x, G(x)))] \tag{11}$$

However, it often leads to model collapse, making training hard. Therefore, we adopt LSGAN [52] loss for training, which can generate higher-quality images and is more stable. The discriminative loss and the generative loss are defined as follows,

$$L_{LSGAN}(D) = 1/2\mathbb{E}_{x,y}\left[(D(x, y) - 1)^2\right] + 1/2\mathbb{E}_x\left[D(x, G(x))^2\right] \tag{12}$$

$$L_{LSGAN}(G) = 1/2\mathbb{E}_x\left[(D(x, G(x)) - 1)^2\right] \tag{13}$$

The L1 loss is also added to reduce the difference between ground truth y and the predicted result $G(x)$ and can be formulated as,

$$L_1(G) = \mathbb{E}_{(x,y)} \| y - G(x) \|_1 \tag{14}$$

Thus we get the final objective functions,

$$\min_D L(D) = L_{LSGAN}(D) \tag{15}$$

$$\min_G L(G) = L_{LSGAN}(G) + \lambda L_1(G) \tag{16}$$

To optimize networks, we set $\lambda = 10$ for L1 loss and train the model with the Adam solver with a learning rate of $2e^{-4}$. We train the model with a batch size of 32 and 10 epochs.

For visual modality, depth has proven to generalize fairly well for sim-to-real problems [18,31]. Therefore, when rendering point clouds, we augment depth maps with zero-mean Gaussian Process noise ϵ and multiplicative gamma noise α similar to [18], which can be expressed as $d = \alpha d + \epsilon$. Then the point cloud $\mathscr{P} \in \mathbb{R}^{3 \times N}$ can be computed by back-projecting a depth map d using real camera intrinsics.

5. Simulation experiments

To validate the effectiveness of the proposed multimodal sensing framework in the single object and cluttered grasping scenarios, extensive experiments are conducted in virtual and real environments. In this section, we perform several comparative and ablation studies in the virtual world, aiming to answer two questions. 1) Can the proposed multimodal approach effectively evaluate grasping results in cluttered scenarios? 2) Is the trained multimodal model suitable for delicate grasping experiments, e.g., gripping force optimization?

5.1. Experimental design

We first compare the predictive performance of the proposed method with five baselines: RGB only and Depth only mask out tactile input and only take global RGB image and depth map as input of ResNet-18, respectively. The depth map is converted to a 3-channel color image before being fed into the network. As shown in Fig. 11, the global image includes mainly a single object and gripper to avoid the influence of irrelevant factors and is collected synchronously with the tactile data. When capturing data, the camera is fixed in the world coordinate system, pointing to the object's center, and the mounting position is randomized within a specific spatial range. Moreover, the resulting dataset size is the same as that generated in Section 4.4. Point cloud only masks

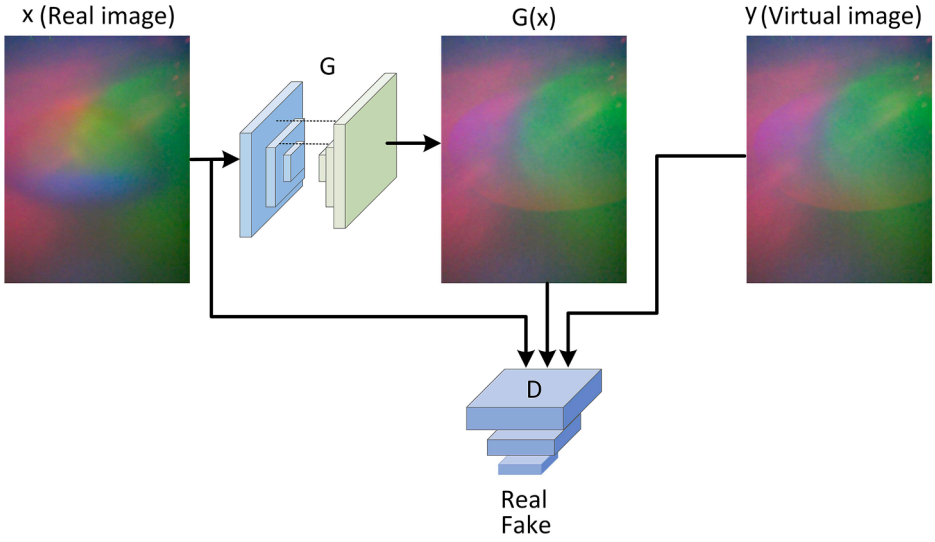


Fig. 10. Sim-to-real via real-to-sim: Transfer trained multimodal model to the real world by translating the real tactile image into the corresponding virtual one with the conditional adversarial network.

Fig. 11. The rendered global RGB and depth map. The height and width sizes of the image are 480 and 640, respectively.

out the tactile input, and the PointNet-like network is used to process the local point cloud. Tactile only masks out the visual information and the tactile images are fed to ResNet-18. Ours† removes the co-attention layer. In the network, the features of the three channels are concatenated as input to the FC layer.

Besides these baselines, we also choose a representative research, Calandra et al. [6], in which ResNet-50 architecture extracts features from global RGB images and tactile signals to predict grasp outcomes. To be consistent with other comparison methods, we replace ResNet-50 with ResNet-18 and keep other things unchanged. Ours denotes the visual-tactile fusion framework proposed in Section 4.2. The predictive power of the different models is verified using the 3-fold cross-validation method. Firstly, all data are randomly shuffled and divided into three parts on average. Then, one is used as the validation set and the other two as the training set in turn. Finally, we choose Accuracy (A), Precision (P), Recall (R), and F1 score (F1) as evaluation metrics, which are defined as follows:

$$A = \frac{TP + TN}{TP + FN + FP + TN} \tag{17}$$

$$P = \frac{TP}{TP + FP} \tag{18}$$

$$R = \frac{TP}{TP + FN} \tag{19}$$

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{20}$$

where *TP*, *FP*, *TN*, and *FN* represent true positive, false positive, true negative, and false negative, respectively.

Testing on the dataset gives an intuitive indication of the predictive power of different models, but what we ultimately want to know is how well the trained model performs in the grasping trials. To this end, we conduct three experiments with the seven methods mentioned above, and the optimal weights generated in the cross-validation experiments are retrained on the remaining data. Firstly, we select ten objects from the training object set, as shown in Fig. 13. We perform 50 grasping trials for each object and use the prediction accuracy to measure the performance. Moreover, the gripping force is randomly sampled from 5N to 30N during grasping. Most of the existing grasp outcome assessment methods based on visual-haptic fusion are oriented to 2D planar grasp, while this paper focuses on the most widely used 6-DoF grasping framework. Thus we adopt GPG [45] to sample 6-DoF grasps from the single-view point cloud. GPG is a solution that allows fast sampling of parallel grasps from the 3D unknown point cloud. In each trial, we drop the object at a certain height in a random pose and sample a collision-free grasp configuration. Then the grasp is executed with a desired force. Before lifting, the grasp result is evaluated by the corresponding model. The criterion for successful grasping is that the object remains stable within the gripper after being lifted to a certain height. Secondly, we select ten objects from the test object set to verify the generalization capabilities, as shown in Fig. 14. The experimental procedure is the same as the above steps. Finally, To verify the effectiveness of the proposed method in cluttered scenarios, three scenes are designed, as shown in Fig. 15. Scene 1 consists of 10 objects from the training object set, scene 2 is composed of 10 objects from the test object set, and scene 3 is created by taking five objects from each of the previous two scenes. Each scene is only initialized once, and the objects in the scene are arranged randomly. We also conduct 50 grasping tests with the gripping force ranging from 5N to 30N. The object will be remixed into the remaining part after each grasp.

The grasping strategy can be adjusted based on the real-time grasp-stability evaluation to optimize gripping force while maintaining grasp success. In this paper, we design a simplified rule to demonstrate the feasibility of the proposed visual-tactile fusion approach for delicate grasping experiments. Specifically, for a given grasp, let the robot start with a minimum force of 5N, increase by 1N each time, and the maximum force is 30N. Throughout the process, the robot's grasp pose is fixed. When the model predicts that the grasping result in the current state is successful or the gripping force reaches the maximum value, the robot will execute a lifting action. To test this hypothesis, we first select object 1 in Fig. 14 and scene 2 in Fig. 15 as the unknown single-object and cluttered object scenarios, respectively. Then, each scene is initialized once and the object's location information is recorded simultaneously. We sample 10 grasp configurations for each scene, and 50 grasping trials are carried out for each configuration. The average value of the final gripping force and the grasp success rate are recorded. Meanwhile, grasping with a fixed gripping force of 5N and 30N is the control group. After each trial, the scene will be restored to its initial state to ensure consistency of environmental information.

5.2. Experimental results

The 3-fold cross-validation results are summarised in Table 1, and the prediction accuracy is presented in Fig. 12(a). Fig. 12(b) shows the inference time of each model for a single sample. It can be seen from the figure that the network with point cloud as input is more efficient than the network with image as input. From Fig. 12(a), we see that RGB only and Depth only perform worst and have comparable predictive power.

Table 1
Cross-validation performance of the different models. The mean and standard deviation are reported.

	Precision (%)	Recall (%)	F1 score (%)
RGB only	82.1 ± 1.3	76.7 ± 0.8	79.3 ± 1.0
Depth only	82.8 ± 0.7	80.0 ± 0.5	81.4 ± 1.2
Point cloud only	84.5 ± 0.3	81.7 ± 0.6	83.1 ± 0.9
Tactile only	86.2 ± 1.5	83.3 ± 1.8	84.7 ± 0.6
Calandra et al. [6]	93.1 ± 1.1	90.0 ± 0.9	91.5 ± 1.4
Ours †	94.8 ± 0.8	91.7 ± 1.3	93.2 ± 0.7
Ours	94.9 ± 0.3	93.3 ± 0.9	94.1 ± 0.4

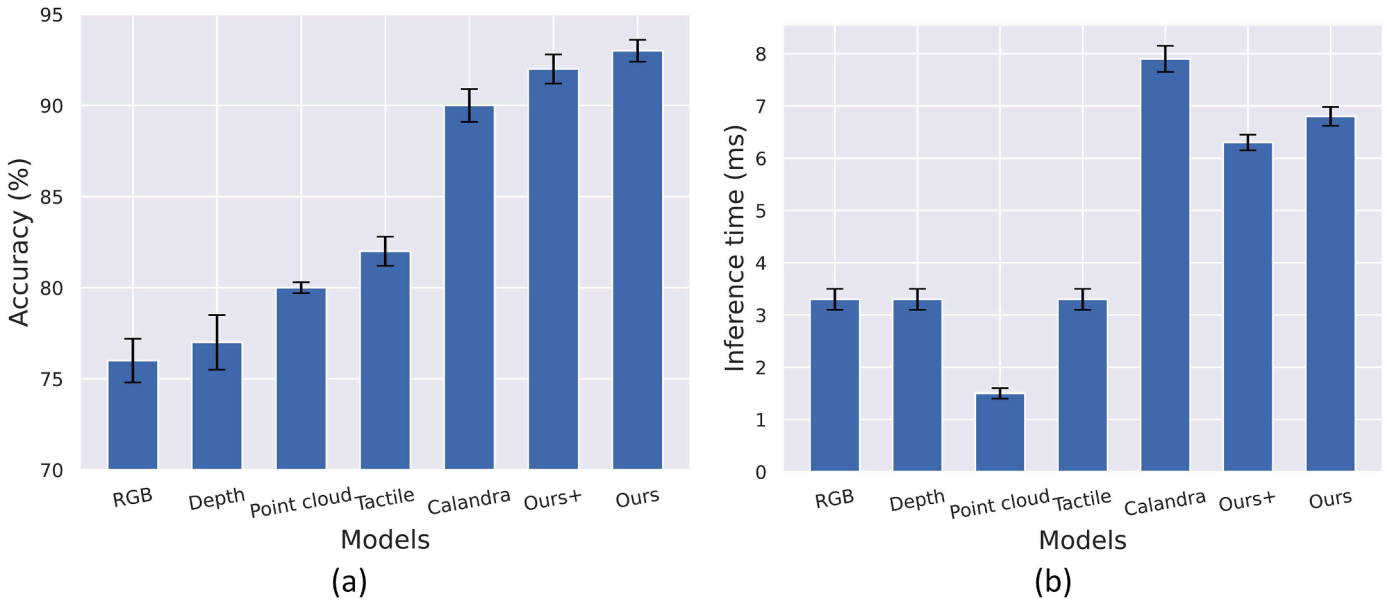


Fig. 12. Test results of different models. a): Cross-validation accuracy of the different models, where x-axis is the abbreviation of the methods listed in Table 1 in that order. (b): The inference time of each model for a single sample, and the inference process is executed on the GPU. The mean and standard deviation are reported in both figures.

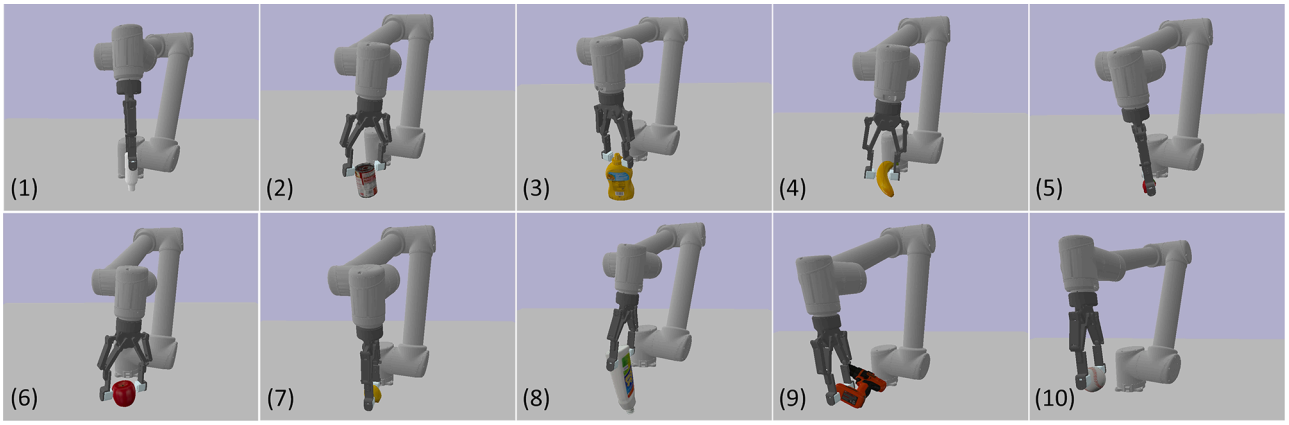


Fig. 13. Grasp outcomes evaluation tasks on ten objects from the training object set.

Point cloud only outperforms the previous two methods by two percentage points, which indicates that local vision generalizes better than global vision. The performance of Tactile only is second only to multimodal fusion methods. This suggests that haptic feedback is more crucial than vision in predicting the grasp outcomes. Furthermore, the predictive power of the multimodal approaches is substantially improved compared to the unimodal methods. That is to say, integrating vision and haptics is essential for stable and gentle grasp operations. Additionally, the proposed full model Ours performs best and achieves a degree of improvement compared to Calandra et al. [6] and Ours†.

To test the performance of multiple models in grasping trials, we first experiment with scenes consisting of single objects from the training and test object sets. The prediction accuracy results are shown in Tables 2 and 3. These approaches that take the unimodal vision as input (i.e., RGB only, Depth only, and Point cloud only) perform reasonably poorly. The main reason is that the gripping force changes between 5N and 30N during grasping trials, but the visual modality cannot perceive such slight variations. Comparatively, Tactile only shows relatively good prediction accuracy and generalization ability. This also again demonstrates the importance of haptics in delicate grasping tasks. The model

Table 2
The prediction accuracy (%) on ten objects from the training object set.

	obj. 1	obj. 2	obj. 3	obj. 4	obj. 5	obj. 6	obj. 7	obj. 8	obj. 9	obj. 10
RGB only	52	42	30	88	68	84	70	32	42	94
Depth only	42	58	32	62	22	82	26	44	50	92
Point cloud only	52	62	68	72	78	82	54	42	52	90
Tactile only	62	84	58	86	68	80	72	46	48	96
Calandra et al. [6]	100	100	100	100	70	100	78	82	68	100
Ours †	96	100	72	100	100	100	100	94	86	100
Ours	100	100	76	100	100	100	100	90	84	100

Table 3
The prediction accuracy (%) on ten objects from the test object set.

	obj. 1	obj. 2	obj. 3	obj. 4	obj. 5	obj. 6	obj. 7	obj. 8	obj. 9	obj. 10
RGB only	42	38	44	78	52	48	24	36	28	30
Depth only	32	46	40	70	64	42	18	44	34	46
Point cloud only	44	30	38	78	60	52	16	38	36	42
Tactile only	68	62	60	80	76	72	34	66	74	68
Calandra et al. [6]	84	72	74	90	96	82	52	90	86	76
Ours †	92	86	90	100	100	94	48	88	90	84
Ours	100	84	92	100	100	100	46	92	100	88

Ours achieves the highest accuracy and has strong generalization capability to unknown objects. It is worth mentioning that the method Calandra et al. [6] shows similar performance compared to Ours in single-object scenarios. In addition, all methods perform poorly for objects with complex shapes and difficult to grasp, such as object 7 (i.e., glass cup) in Fig. 14. Then we test these methods in cluttered scenarios, as shown in Fig. 15. The results are presented in Table 4. The proposed method Ours that makes full use of local point cloud and tactile information achieves high prediction accuracy, while methods with global visual input hardly work. Because the dataset is collected in single-object scenes, the global visual information differs too much when facing cluttered scenes.

We use the average gripping force and grasp success rate to evaluate the performance of delicate grasping experiments. The results are shown in Figs. 16 and 17. For most grasping trials, the fixed policy with 30N gripping force consistently achieves the highest grasp success rate. In other words, a higher force will result in a more stable grasp in most cases. On the other hand, our proposed multimodal policy allows the robot to grasp objects with significantly reduced force while maintaining a comparable success rate. Furthermore, for somewhat unstable grasp configurations, such as grasp 1 in Fig. 17, the fixed policy with 30N leads instead to a very low success rate, while the regrasping policy based on the grasp-stability evaluation achieves a relatively high success rate.

The above experimental results demonstrate the feasibility of our proposed multimodal prediction framework in cluttered and disordered scenarios. Meanwhile, the delicate grasping experimental results show that our approach can be efficiently applied to regrasping policy for single-object and cluttered scenarios.

6. Real robot experiments

Guided by the simulation results, several grasping experiments are designed in the real world to evaluate the generalization capabilities of the proposed method.

6.1. Experimental design

To assess the performance of the final trained model in the actual environment, we first select five objects from the training set and five unseen objects as single-object scenes, as shown in Fig. 18. For each object, 50 grasping trials are conducted, with gripping forces sampled from 10N to 30N. In addition, three cluttered scenarios are designed, and the same tests as before are performed, as shown in Fig. 19. Scenes 1

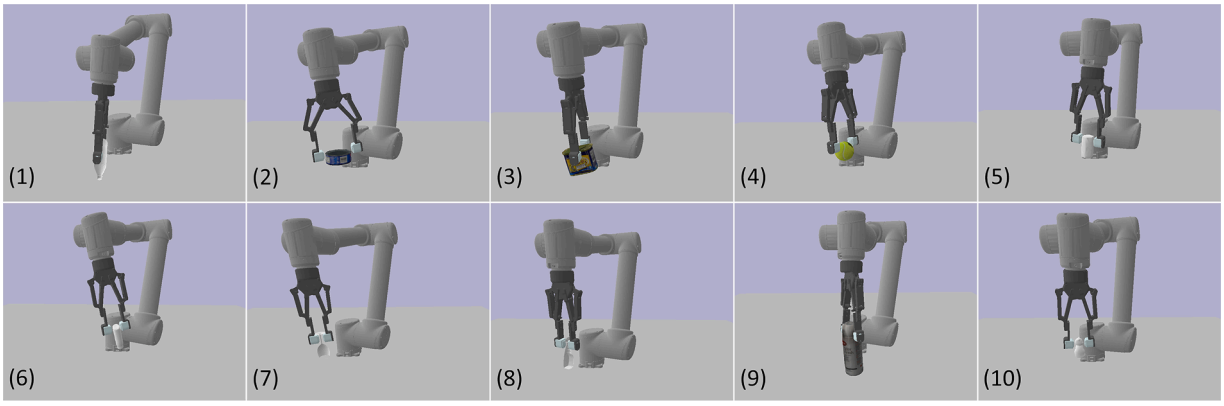


Fig. 14. Grasp outcomes evaluation tasks on ten objects from the test object set.

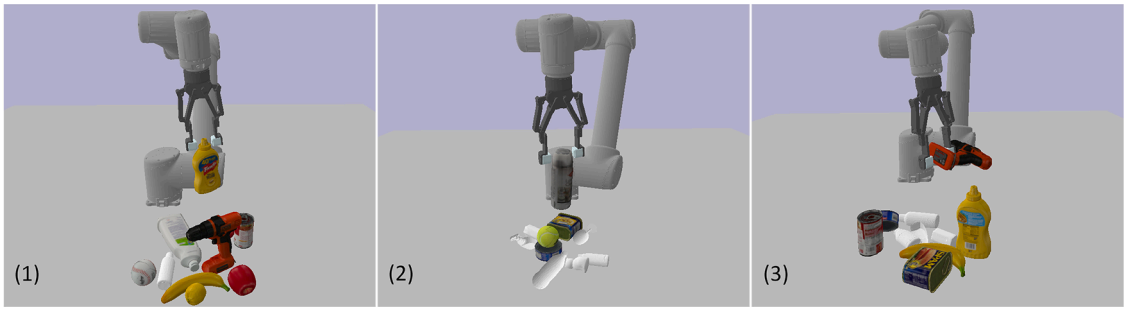


Fig. 15. Grasping results prediction in cluttered scenarios. Scene 1 has ten objects from the training object set, scene 2 includes ten objects from the test object set, and scene 3 is created by taking five objects from each of the previous two scenes.

Table 4
Prediction accuracy results (%) in three cluttered grasping scenarios.

	Scene. 1	Scene. 2	Scene. 3
RGB only	12	10	16
Depth only	20	14	18
Point cloud only	56	48	46
Tactile only	62	58	52
Calandra et al. [6]	34	26	32
Ours †	86	82	78
Ours	90	84	82

and 2 include 5 and 10 objects from the training object set, respectively, to verify the effect of different clutter levels on the algorithm, and scene 3 consists of 5 unseen objects. Each scene is only initialized once, and the objects in the scene are arranged randomly. For each trial, the GPG algorithm generates a collision-free grasp configuration from the single-view point cloud. All objects in the scene may be grasped every time, regardless of their height. The object will be remixed into the remaining part after each grasp. The prediction accuracy defined in Eq. (17) is utilized to assess the predictive performance of the model.

Furthermore, we also evaluate the effect of the proposed migration strategy. W-GAN and WO-GAN represent with and without migration strategies in grasping tasks, respectively. It is worth noting that the transfer strategy here is to address the problem of domain adaptation for tactile images. In actual experiments, we deploy and test two models (i.e., Ours† and Ours). Finally, to visually evaluate the performance of the migration strategy based on the conditional gan network, we use the trained conditional generation model to generate virtual-like images on the evaluation set (the paired dataset collected in Section 4.5) and the SSIM score metric is employed to measure the similarity with the corresponding rendered virtual images.

6.2. Experimental results

The results of the grasp-stability evaluation for single objects are shown in Table 5. It can be seen that the trained full model Ours achieves an average prediction accuracy of 75.4% compared to 71.6% for Ours†. The results indicate that the full model with the migration strategy can be effectively generalized to the real world, and the prediction value for most objects exceeds 80%. However, our current model is also challenging to handle objects with complex surfaces and shapes, such as objects 6 and 7 (i.e., cup and rock). Due to the complex shape of the cup,

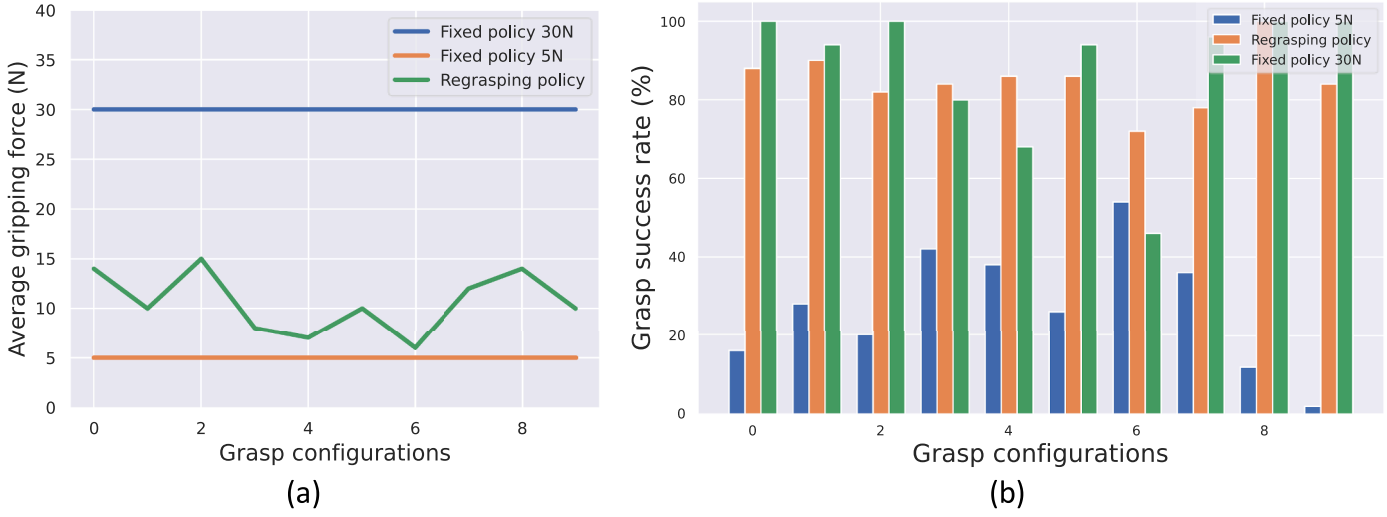


Fig. 16. Delicate grasping experiments in the single-object scene. (a): Average gripping force (b): Grasp success rate, which is defined as the percentage of successful grasps to the total number of grasps.

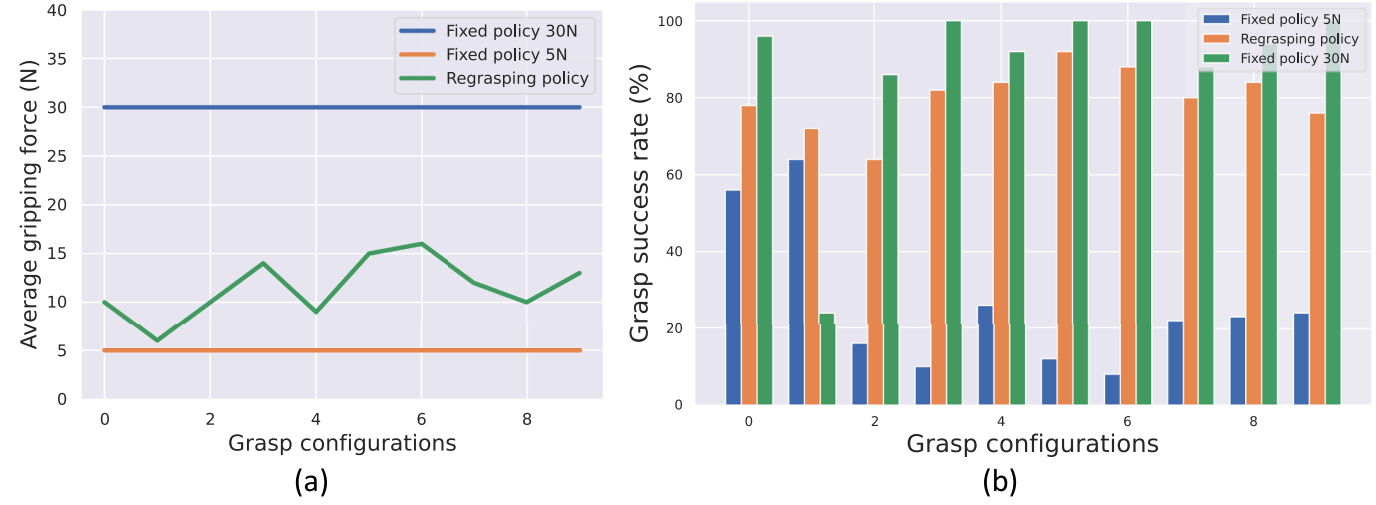


Fig. 17. Delicate grasping experiments in the cluttered scene. (a) Average gripping force (b) Grasp success rate, which is defined as the percentage of successful grasps to the total number of grasps.

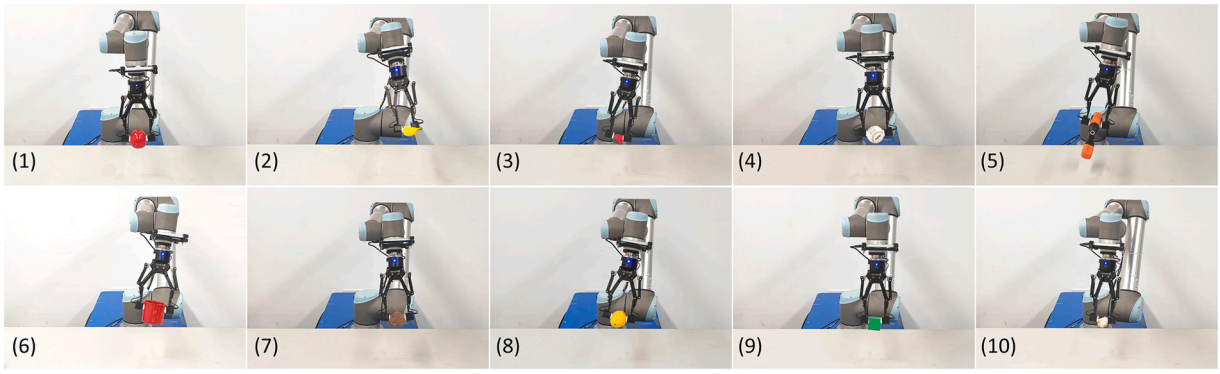


Fig. 18. Single-object scenes tested on the real robot, where the first five are from the training set and the last five are unseen objects.

Fig. 19. Grasp-stability evaluation in cluttered scenes. Scenes 1 and 2 consist of five and ten objects from the training set, respectively, and scene 3 includes five unseen objects.

Table 5
Prediction accuracy results (%) in single-object scenes. The bolded ones are unseen objects.

	obj. 1	obj. 2	obj. 3	obj. 4	obj. 5	**obj. 6**	**obj. 7**	**obj. 8**	**obj. 9**	**obj. 10**
Ours † (WO-GAN)	62	48	42	60	44	22	16	42	36	32
Ours (WO-GAN)	58	52	44	56	42	16	18	48	40	38
Ours † (W-GAN)	80	94	78	84	72	42	46	78	70	72
Ours (W-GAN)	86	92	82	90	74	40	52	84	78	76

it is relatively difficult to grasp. Moreover, there are very few point clouds when the cup is viewed from top to bottom, which leads to a decrease in the network's judgment ability. The rock, on the other hand, has complex surface, resulting in significant differences between the tactile images and the training set. This decreases the network's ability to generalize. In most cases, Ours outperforms Ours†, which proves that the attention mechanism can improve the predictive performance of multimodal networks. In all grasp tests, models without transfer strategy perform poorly because of the large gap between the real tactile image and the virtual one. The domain adaptation strategy eliminates the gap between the actual tactile and virtual images.

It can be seen from Table 6 that our proposed model can be effectively generalized to cluttered scenes and is largely independent of the clutter level, which is not achievable with current global vision-based approaches. For the validation of the migration strategy, we calculate the SSIM scores between generated and simulated tactile images on the validation set, and the average reaches 0.992. This proves that the similarity between the two is very high. We also present five sets of paired tactile images in Fig. 20, and the corresponding SSIM scores are shown in Fig. 21.

Table 6
Prediction accuracy results (%) in three cluttered grasping scenarios.

	Scene. 1	Scene. 2	Scene. 3
Ours † (WO-GAN)	48	46	32
Ours (WO-GAN)	52	54	30
Ours † (W-GAN)	82	76	68
Ours (W-GAN)	84	82	64

In summary, the robot effectively achieves grasp outcomes assessment in cluttered scenes with the help of the proposed multimodal network and dataset collection policy. Furthermore, this process takes full advantage of the concurrency and complementarity of visual and haptic senses. The developed digital twin-enabled robotic grasping system allows us to quickly acquire large and reasonable multimodal datasets, while the proposed migration strategy can realize zero-shot sim-to-real policy transfer on the real robot.

7. Discussion

With the proposed visual-tactile fusion network and data collection policy, this paper addresses the problem of grasp-stability evaluation for single and cluttered objects. There are three topics worthy of further discussion regarding the content of this manuscript.

The first is about the properties of the object. Since current simulators, e.g., PyBullet, are challenging to simulate fragile and deformable objects, and the current tactile simulation only supports rigid objects, this paper validates the proposed method using high-stiffness objects. In essence, the proposed multimodal sensing framework learns a general object-oriented grasping experience. Therefore, the authors believe extending the model to non-rigid objects would be feasible by fine-tuning or adding restrictions. In future work, we will continue to explore the potential application scope.

The second is related to the experimental conditions and results. In this paper, to simplify the problem, we set the friction coefficient and object mass to a fixed value. Randomizing these dynamic parameters in

a reasonable range could further bridge the gap between simulation and the real world. In addition, the training object set used in the experiments is composed of 3D scanned models from real-life objects, and their surfaces are relatively simple and smooth. As a result, the trained model is not competent for objects such as rocks with complex surfaces and cups that are difficult to grasp.

The third is the application of the grasp-stability evaluation methodology based on visual-tactile fusion perception. For example, when grasping an unknown object in an unstructured environment, we would like to find an optimal strategy that allows the robot to grasp it successfully with minimal force through a small amount of exploration. In this paper, a delicate grasping experiment has been conducted, and the experimental results demonstrate the effectiveness of the proposed model for such grasping strategy. In other words, the robot will determine whether or not to lift the object based on the current assessment of grasp stability. However, the human-designed strategy is limited [5]. Robots lack active exploration and often require many steps to find the optimal grasp. With the development of deep neural networks, deep reinforcement learning is considered the most viable method to endow robots with the capability to explore unknown environments. However, learning strategies from scratch is highly time-consuming. Thus, multimodal representation-based reinforcement learning is the most likely way to quickly obtain the optimal grasping strategy. This representation is generated by supervised learning (i.e., like the proposed multimodal network in this paper). The agent guided by reinforcement learning actively adjusts the grasp pose and gripping force according to the current multimodal input to obtain the optimal policy quickly. For future work, we plan to introduce reinforcement learning to achieve minimum force grasping of unknown objects based on current work.

8. Conclusion and future work

This paper proposes a digital twin-based method that takes full advantage of visual and tactile perception to solve the problem of grasp outcomes assessment in cluttered scenarios. Firstly, we propose a multimodal prediction network to fuse visual and haptic data. Inspired by [14], we introduce the object's local point cloud within the gripper as the visual signal to avoid the influence of other information in the environment on generalization performance. Two GelSight-style tactile sensors (DIGIT) provide high-resolution tactile images. Secondly, we develop a digital twin system for robotic grasping to collect a large-scale visual-haptic dataset, and we also propose a self-supervised autonomous collection policy to accelerate the generation process. Thirdly, to deploy the trained model directly on the real robot, we propose a migration strategy based on conditional adversarial networks. Finally, extensive validation experiments are conducted in virtual and real environments.

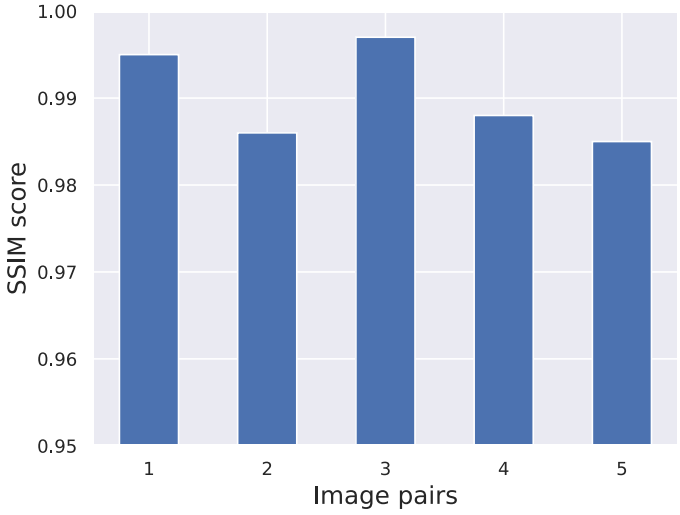


Fig. 21. SSIM scores between the generated and virtual images for each group in Fig. 20.

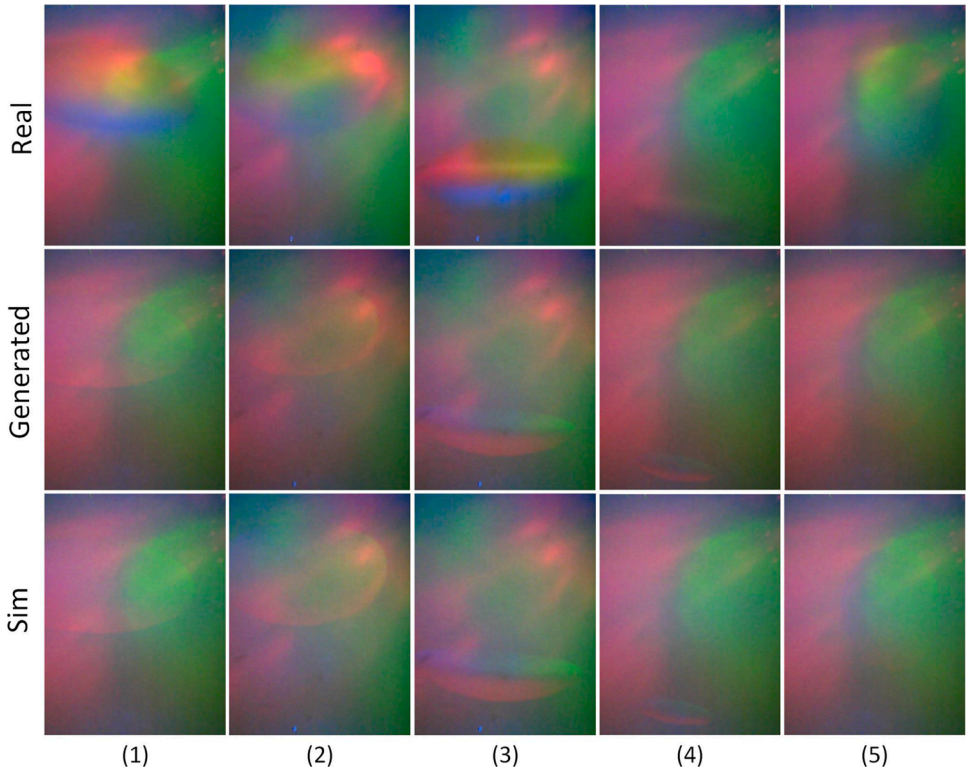


Fig. 20. Five sets of paired tactile images from the validation set. Each group consists of a real image acquired from the DIGIT sensor, a virtual-like image generated by a conditional GAN network, and a virtual image rendered by TACTO.

The experimental results demonstrate that the proposed method can effectively predict the grasp-stability in cluttered scenarios and be applied to regrasping policy.

In future work, we will explore how to extend our approach to manipulate deformable and fragile objects. At the same time, the regrasping policy using multimodal representation-based reinforcement learning is also an important research issue.

CRediT authorship contribution statement

Zhuangzhuang Zhang: Investigation, Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing, Resources, Data curation, Visualization. **Zhinan Zhang:** Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Lihui Wang:** Conceptualization, Writing – review & editing. **Xiaoxiao Zhu:** Writing – original draft, Resources. **Huang Huang:** Funding acquisition. **Qixin Cao:** Supervision, Project administration.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

This research is supported by China's National Key Research and Development Program under Grant 2018AAA0102700.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.rcim.2023.102601.

References

[1] Z. Liu, Q. Liu, W. Xu, L. Wang, Z. Zhou, Robot learning towards smart robotic manufacturing: a review, Robot. Comput. Integr. Manuf. 77 (2022), 102360, https://doi.org/10.1016/j.rcim.2022.102360.

[2] Í. Elguea-Aguinaco, A. Serrano-Muñoz, D. Chrysostomou, I. Inziarte-Hidalgo, S. Bøgh, N. Arana-Arexolaleiba, A review on reinforcement learning for contact-rich robotic manipulation tasks, Robot. Comput. Integr. Manuf. 81 (2023), 102517, https://doi.org/10.1016/j.rcim.2022.102517.

[3] M. Gou, H.S. Fang, Z. Zhu, S. Xu, C. Wang, C. Lu, Rgb matters: learning 7-dof grasp poses on monocular rgbd images, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2021, pp. 13459–13466, https://doi.org/10.1109/ICRA48506.2021.9561409.

[4] T. Zhang, C. Zhang, T. Hu, A robotic grasp detection method based on auto-annotated dataset in disordered manufacturing scenarios, Robot. Comput. Integr. Manuf. 76 (2022), 102329, https://doi.org/10.1016/j.rcim.2022.102329.

[5] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E.H. Adelson, S. Levine, More than a feeling: learning to grasp and REGRASP using vision and touch, IEEE Robot. Autom. Lett. 3 (2018) 3300–3307, https://doi.org/10.1109/LRA.2018.2852779.

[6] R. Calandra, A. Owens, M. Upadhyaya, W.Z. Yuan, J. Lin, E.H. Adelson, S. Levine, The feeling of success: does touch sensing help predict grasp outcomes?, in: Proceedings of the Conference on Robot Learning, 2017, pp. 314–323, https://doi.org/10.48550/arXiv.1710.05512.

[7] S. Zhang, Z. Chen, Y. Gao, W. Wan, J. Shan, H. Xue, F. Sun, Y. Yang, B. Fang, Hardware technology of vision-based tactile sensor: a review, IEEE Sens. J. 22 (2022) 21410–21427, https://doi.org/10.1109/JSEN.2022.3210210.

[8] W. Yuan, S. Dong, E.H. Adelson, Gelsight: high-resolution robot tactile sensors for estimating geometry and force, Sensors 17 (2017) 2762, https://doi.org/10.3390/s17122762.

[9] I.H. Taylor, S. Dong, A. Rodriguez, GelSlim 3.0: high-resolution measurement of shape, force and slip in a compact tactile-sensing finger, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2022, pp. 10781–10787, https://doi.org/10.1109/ICRA46639.2022.9811832.

[10] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V.R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, R. Calandra, Digit: a novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation, IEEE Robot. Autom. Lett. 5 (2020) 3838–3845, https://doi.org/10.1109/LRA.2020.2977257.

[11] S. Cui, R. Wang, J. Wei, F. Li, S. Wang, Grasp state assessment of deformable objects using visual-tactile fusion perception, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2020, pp. 538–544, https://doi.org/10.1109/ICRA40945.2020.9196787.

[12] S. Cui, R. Wang, J. Wei, J. Hu, S. Wang, Self-attention based visual-tactile fusion learning for predicting grasp outcomes, IEEE Robot. Autom. Lett. 5 (2020) 5827–5834, https://doi.org/10.1109/LRA.2020.3010720.

[13] S. Kanitkar, H. Jiang, W. Yuan, PoseIt: a visual-tactile dataset of holding poses for grasp stability analysis, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2022, pp. 71–78, https://doi.org/10.1109/IROS47612.2022.9981562.

[14] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, J. Zhang, Pointnetgpd: detecting grasp configurations from point sets, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2019, pp. 3629–3635, https://doi.org/10.1109/ICRA.2019.8794435.

[15] L. Li, B. Lei, C. Mao, Digital twin in smart manufacturing, J. Ind. Inf. Integr. 26 (2022), 100289, https://doi.org/10.1016/j.jii.2021.100289.

[16] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, N. Navab, Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes, in: Proceedings of the Asian Conference on Computer Vision, 2012, pp. 548–562, https://doi.org/10.1007/978-3-642-37331-2_42.

[17] J.P.C. de Souza, L.F. Rocha, P.M. Oliveira, A.P. Moreira, J. Boaventura-Cunha, Robotic grasping: from wrench space heuristics to deep learning policies, Robot. Comput. Integr. Manuf. 71 (2021), 102176, https://doi.org/10.1016/j.rcim.2021.102176.

[18] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J.A. Ojea, K. Goldberg, Dex-net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics, in: Proceedings of the Robotics: Science and Systems, 2017, https://doi.org/10.15607/RSS.2017.XIII.058.

[19] D. Morrison, P. Corke, J. Leitner, Closing the loop for robotic grasping: a real-time, generative grasp synthesis approach, in: Proceedings of the Robotics: Science and Systems, 2018, https://doi.org/10.15607/RSS.2018.XIV.021.

[20] W. Hu, C. Wang, F. Liu, X. Peng, P. Sun, J. Tan, A grasps-generation-and-selection convolutional neural network for a digital twin of intelligent robotic grasping, Robot. Comput. Integr. Manuf. 77 (2022), 102371, https://doi.org/10.1016/j.rcim.2022.102371.

[21] C.R. Qi, H. Su, K. Mo, L.J. Guibas, Pointnet: deep learning on point sets for 3d classification and segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 652–660, https://doi.org/10.1109/CVPR.2017.16.

[22] Y. Bekiroglu, J. Laaksonen, J.A. Jorgensen, V. Kyrki, D. Kragic, Assessing grasp stability based on learning and haptic data, IEEE Trans. Robot. 27 (2011) 616–629, https://doi.org/10.1109/TRO.2011.2132870.

[23] J.M. Romano, K. Hsiao, G. Niemeyer, S. Chitta, K.J. Kuchenbecker, Human-inspired robotic grasp control with tactile sensing, IEEE Trans. Robot. 27 (2011) 1067–1079, https://doi.org/10.1109/TRO.2011.2162271.

[24] J. Kwiatkowski, D. Cockburn, V. Duchaine, Grasp stability assessment through the fusion of proprioception and tactile signals using convolutional neural networks, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, pp. 286–292, https://doi.org/10.1109/IROS.2017.8202170.

[25] F. Veiga, J. Peters, T. Hermans, Grip stabilization of novel objects using slip prediction, IEEE Trans. Haptics 11 (2018) 531–542, https://doi.org/10.1109/TOH.2018.2837744.

[26] R. Kolamuri, Z. Si, Y. Zhang, A. Agarwal, W. Yuan, Improving grasp stability with rotation measurement from tactile sensing, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2021, pp. 6809–6816, https://doi.org/10.1109/IROS51168.2021.9636488.

[27] Z. Si, Z. Zhu, A. Agarwal, S. Anderson, W. Yuan, Grasp stability prediction with sim-to-real transfer from tactile sensing, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2022, pp. 7809–7816, https://doi.org/10.1109/IROS47612.2022.9981863.

[28] Z. Huang, M. Fey, C. Liu, E. Beysel, X. Xu, C. Brecher, Hybrid learning-based digital twin for manufacturing process: modeling framework and implementation, Robot. Comput. Integr. Manuf. 82 (2023), 102545, https://doi.org/10.1016/j.rcim.2023.102545.

[29] F. Tao, Q. Qi, Make more digital twins, Nature 573 (2019) 490–491, https://doi.org/10.1038/d41586-019-02849-1.

[30] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, P. Abbeel, Domain randomization for transferring deep neural networks from simulation to the real world, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2017, pp. 23–30, https://doi.org/10.1109/IROS.2017.8202133.

[31] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, K. Goldberg, Segmenting unknown 3D objects from real depth images using mask R-CNN trained on synthetic data, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2019, pp. 7283–7290, https://doi.org/10.1109/ICRA.2019.8793744.

[32] D.F. Gomes, P. Paoletti, S. Luo, Generation of gelsight tactile images for sim2real learning, IEEE Robot. Autom. Lett. 6 (2021) 4177–4184, https://doi.org/10.1109/LRA.2021.3063925.

[33] A. Agarwal, T. Man, W. Yuan, Simulation of vision-based tactile sensors using physics based rendering, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2021, pp. 1–7, https://doi.org/10.1109/ICRA48506.2021.9561122.
[34] S. Wang, M. Lambeta, P. Chou, R. Calandra, TACTO: a fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors, IEEE Robot. Autom. Lett. 7 (2022) 3930–3937, https://doi.org/10.1109/LRA.2022.3146945.
[35] Z. Si, W. Yuan, Taxim: an example-based simulation model for gelsight tactile sensors, IEEE Robot. Autom. Lett. 7 (2022) 2361–2368, https://doi.org/10.1109/LRA.2022.3142412.
[36] E. Coumans, Y. Bai, Pybullet, a python module for physics simulation for games, robotics and machine learning, https://pybullet.org/wordpress/, 2022 (accessed 15 November 2022).
[37] N.P. Koenig, A. Howard, Design and use paradigms for gazebo, an open-source multi-robot simulator, in: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004, pp. 2149–2154, https://doi.org/10.1109/IROS.2004.1389727.
[38] W. Chen, Y. Xu, Z. Chen, P. Zeng, R. Dang, R. Chen, J. Xu, Bidirectional sim-to-real transfer for gelsight tactile sensors with cyclegan, IEEE Robot. Autom. Lett. 7 (2022) 6187–6194, https://doi.org/10.1109/LRA.2022.3167064.
[39] Y. Lin, J. Lloyd, A. Church, N.F. Lepora, Tactile gym 2.0: sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch, IEEE Robot. Autom. Lett. 7 (2022) 10754–10761, https://doi.org/10.1109/LRA.2022.3195195.
[40] Y. Liu, H. Xu, D. Liu, L. Wang, A digital twin-based sim-to-real transfer for deep reinforcement learning-enabled industrial robot grasping, Robot. Comput. Integr. Manuf. 78 (2022), 102365, https://doi.org/10.1016/j.rcim.2022.102365.
[41] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, S. Levine, OmniTact: a multi-directional high-resolution touch sensor, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2020, pp. 618–624, https://doi.org/10.1109/ICRA40945.2020.9196712.
[42] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2242–2251, https://doi.org/10.1109/ICCV.2017.244.
[43] P. Isola, J.Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134, https://doi.org/10.1109/CVPR.2017.632.
[44] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778, https://doi.org/10.1109/CVPR.2016.90.
[45] A. ten Pas, R. Platt, Using geometry to detect grasp poses in 3d point clouds, in: A. Bicchi, W. Burgard (Eds.), Robotics Research. Springer Proceedings in Advanced Robotics. 2, Springer, Cham, 2018, pp. 307–324, https://doi.org/10.1007/978-3-319-51532-8_19.
[46] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, A.Y. Ng, ROS: an open-source robot operating system, in: Proceedings of the IEEE International Conference on Robotics and Automation Workshop on Open Source Software, 2009.
[47] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, A. M. Dollar, Yale-CMU-Berkeley dataset for robotic manipulation research, Int. J. Robot. Res. 36 (2017) 261–268, https://doi.org/10.1177/0278364917700714.
[48] A. Singh, J. Sha, K.S. Narayan, T. Achim, P. Abbeel, BigBIRD: a large-scale 3D database of object instances, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2014, pp. 509–516, https://doi.org/10.1109/ICRA.2014.6906903.
[49] A. Kasper, Z. Xue, R. Dillmann, The KIT object models database: an object model database for object recognition, localization and manipulation in service robotics, Int. J. Robot. Res. 31 (2012) 927–934, https://doi.org/10.1177/0278364912445831.
[50] I.-M. Chen, J.W. Burdick, Finding antipodal point grasps on irregularly shaped objects, IEEE Trans. Robot. 9 (1993) 507–512, https://doi.org/10.1109/70.246063.
[51] MVTec Software GmbH: HALCON Vision Software - Version 20.11.2.0, https://www.mvtec.com/products/halcon, 2022 (accessed 15 November 2022).
[52] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2794–2802, https://doi.org/10.1109/ACCESS.2022.3158343.