nature machine intelligence

Supplementary information https://doi.org/10.1038/s42256-026-01292-y

Learning contact representations in real-world clutter for universal robotic grasping

In the format provided by the authors and unedited

Supplementary Information Contents

Supplementary Notes

Supplementary Figures

Supplementary Tables

Legends for Supplementary Videos

Supplementary Note 1: Training strategies of contact estimation

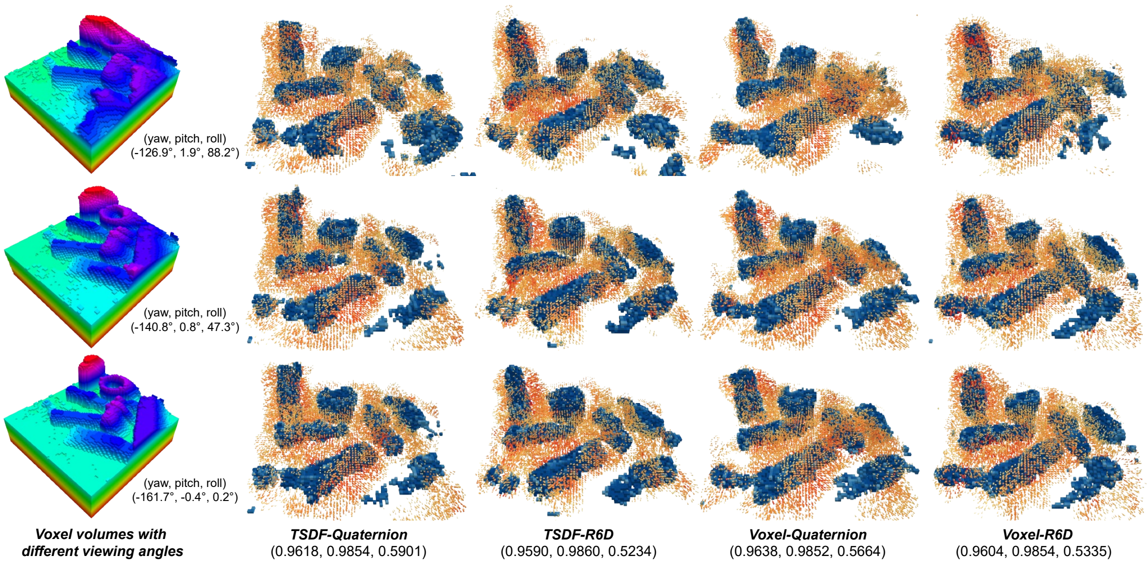
We implemented the 3D U-Net in Pytorch and optimized it using the Adam optimizer with a learning rate of 1×10^{-3}. We leveraged the same data-split strategy of GraspNet-1Billion to train and evaluate the neural network. Specifically, the contact dataset $\{(V, S_{contact}, S_{centroid})\}^{12160}$ can be split into 100 training scenes and 90 test scenes. Because the contact dataset is created from score-based grasp labels, it lacks non-contact labels for gradient-based network training. To balance the label types, we exclusively collected N_{tc} contact-feature samples with a ratio a_c of non-contact samples. Specifically, the training data of one scene contains $(1 - a_c) \cdot N_{tc}$ contact features randomly sampled from $S_{contact}$ and $a_c \cdot N_{tc}$ samples of non-contact labels $((u, v, w), (r_0, 0))$, where $r_0 = [0, 0, 0, 1]$ in quaternion expression. The voxel coordinates (u, v, w) of the non-contact samples were randomly determined inside V apart from $S_{contact}$. Similarly, the contact-centroid training data comprises $(1 - a_e) \cdot N_{te}$ samples from $S_{centroid}$ and $a_e \cdot N_{te}$ non-centroid samples of $((u, v, w), 0)$, where a_e and N_{te} are the ratio of non-centroid labels and the number of total centroid samples, respectively. Note that the numbers of contact poses and contact centroids (in one scene) have the relation: $N_{vc} = (3 \sim 5) \cdot N_{ve}$. Thus, we similarly set $N_{tc} = 4 \cdot N_{te}$ in the training phase. Empirically, higher non-contact ratios a_c introduce more distinct effects of smoothing the contact-feature estimates, and N_{tc} slightly varies the training efficiency. Additionally, both a_c and N_{tc} have little influence on the estimation accuracy. In physical experiments, we determined $a_c = 0.20$ and $N_{tc} = 3000$. Because the contact centroid set $S_{centroid}$ reveals distinct clustering boundaries, the effect of tuning a_e is not evident in contact-centroid estimation and we set $a_e = 0.40$. With the sampled contact features, the 3D U-Net is trained for 64 epochs with an observation batch of 4.

Supplementary Note 2: Investigation of observation type and feature representation

Different patterns of contact features lead to varied grasp optimization in terms of optimizing effectiveness and grasping performance. We investigated the superior representation of visual observation and contact orientation adopted in contact-feature estimation. Regarding the estimation input, the TSDF and the Voxel are considered potential observation formats. For representing contact orientation, Quaternion and R6D are selected as the optional types in this evaluation. Four groups of 3D U-Net are separately trained on data of different observation-orientation types (i.e., TSDF-Quaternion, TSDF-R6D, Voxel-Quaternion, and Voxel-R6D). Supplementary Fig. 1 depicts the estimation performance of contact features in four data-type pairs. Different rows present the feature distributions under observations of varied camera perspectives.

Numerical metrics are established to quantify the performance of different estimation schemes in estimating $\{\widehat{P}_c, \widehat{\Phi}_c, \widehat{P}_e\}$. We refer to $\{P_c, \Phi_c, P_e\}$ as their ground truth, which can be generated from $\{S_{contact}, S_{centroid}\}$ by voxel assignment. Specifically, to evaluate the regression performance of contact certainty scores $\widehat{P}_c$, the Kullback-Leibler (KL) divergence between P_c and $\widehat{P}_c$ is calculated and normalized to $(0, 1]$ with a negative exponential: $\mathcal{E}_{P_c} = \exp(-D_{\mathrm{KL}}(p_c \| \widehat{p}_c))$, where $\mathcal{E}_{P_c}$ is calculated voxel-wise and averaged over all voxels of $p_c \in P_c$ and $\widehat{p}_c \in \widehat{P}_c$. Metric $\mathcal{E}_{P_c}$ closer to 1 indicates a smaller divergence between P_c and $\widehat{P}_c$. Similarly, the performance of contact-centroid estimation $\widehat{P}_e$ is measured by $\mathcal{E}_{P_e}$: $\mathcal{E}_{P_e} = \exp(-D_{\mathrm{KL}}(p_e \| \widehat{p}_e))$. Regarding the estimate of contact orientation $\widehat{\Phi}_c$, we constructed the metric as $\mathcal{E}_{\mathrm{Angle}} = \mathrm{Vct}(\widehat{\phi}_c) \cdot \mathrm{Vct}(\phi_c)$, where $\mathcal{E}_{\mathrm{Angle}}$ computes the cosine similarity of two unit vectors $\mathrm{Vct}(\cdot)$ rotated by $\widehat{\phi}_c \in \widehat{\Phi}_c$ and $\phi_c \in \Phi_c$. Because Φ_c only contains valid orientations as labeled in $S_{contact}$, we obtain $\mathcal{E}_{\mathrm{Angle}}$ by weight-averaging all voxel-wise cosine similarities with their contact certainty scores $p_c \in P_c$.

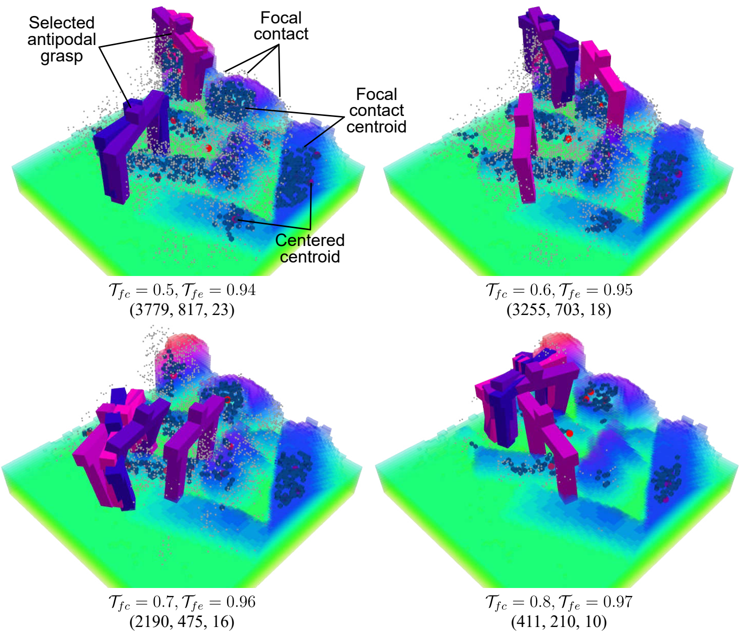
Concerning the four potential data types, we calculate their average $(\mathcal{E}_{P_c}, \mathcal{E}_{P_e}, \mathcal{E}_{\mathrm{Angle}})$ on 64 camera viewing poses of 90 test scenes (see the bottom of Supplementary Fig. 1). The results of $\mathcal{E}_{\mathrm{Angle}}$ indicate that Quaternion, which possesses fewer parameters for representing orientation, significantly excels R6D in contact-orientation estimation (e.g., 0.5901 (Quaternion) versus 0.5234 (R6D) in TSDF observations). Regarding observation types, TSDF and Voxel are both capable of facilitating the certainty-score estimation (measured by $\mathcal{E}_{P_c}$ and $\mathcal{E}_{P_e}$), while TSDF gains better orientation estimation than Voxel (0.6901 versus 0.5664 in Quaternion type). Referring to Supplementary Fig. 1, the contact features estimated from TSDF exhibit distinct boundaries between objects, while the features estimated from Voxel are smoother and more consistent across the scenes of different viewing angles. Despite these distinctions, all estimated contact features hold merits under significantly varied viewing angles.



Supplementary Fig. 1 Estimation results of different types of observation and contact orientation. Different rows demonstrate the distributions of estimated features under observations with different viewing angles. The bottom line indicates $(\mathcal{E}_{P_c}, \mathcal{E}_{P_e}, \mathcal{E}_{\mathrm{Angle}})$ of different ObservationType-OrientationType pairs.

Supplementary Note 3: Parameter investigation of feature refinement

In contact-feature refinement, we introduced $\mathcal{T}_{fc}$ and $\mathcal{T}_{fe}$ for selecting focal features from the estimated $\{\widehat{P}_c, \widehat{\Phi}_c, \widehat{P}_e\}$. We apply grid search to determine them within $\mathcal{T}_{fc} \in [0.5, 0.6, 0.7, 0.8]$ and $\mathcal{T}_{fe} \in [0.80, 0.90, 0.94, 0.95, 0.96, 0.97]$ by conducting dataset-based antipodal grasp generation. To achieve a rapid search of 4×6 threshold pairs, we adopt ten *Seen* test scenes in GraspNet-1Billion (each scene contains 256 observations in different viewing angles) for grasp generation. For each tested annotation, one volume observation (Voxel or TSDF, with a grid size of $80 \times 80 \times 80$ and a spatial dimension of 0.4 m $\times$ 0.4 m $\times$ 0.4 m) is encoded, resembling the dataset generation process. The trained 3D U-Nets estimated contact features from the encoded volumes. With different pairs of thresholds $(\mathcal{T}_{fc}, \mathcal{T}_{fe})$, we select the focal feature sets S_{fc} and S_{fe} following the contact-refinement procedure. We generated antipodal grasps using the parallel-jaw gripper in Fig. 2b as the hand model H, where $D_{Depth} = 0.06$ m, $D_{Height} = 0.02$ m, and $D_{Width} \leq 0.12$ m. We set $\gamma_j = 0.2$ and $N_{batch} = 500$ to initialize the hand batch in grasp optimization. As stated in the established optimization strategy, the initialized grasp poses are optimized by 200 iterations. The top 50 grasps of 500 optimized grasp candidates are scored with the analytical force-closure [17] under different friction coefficients μ. We adopted AP_μ and **AP** metrics [26] to assess the generated grasps. AP_μ indicates the average $Precision@k$ for k ranging from 1 to 50 once given a specific μ, where $Precision@k$ measures the precision of top-k grasps. **AP** denotes the averaged AP_μ, where μ ranges from 0.2 to 1.0 along with an interval of 0.2.



Supplementary Fig. 2 Distribution of contact features under different focal contact threshold $\mathcal{T}_{fc}$ and focal centroid threshold $\mathcal{T}_{fe}$. The bottom lines of the sub-figures present (N_{fc}, N_{fe}, N_{ee}) of the visualized scenes. Five representative antipodal grasps are depicted, where brighter magenta colors indicate better grasps.

Supplementary Table 1 tabulates the resultant **AP** of antipodal grasp generation under different estimation inputs (TSDF and Voxel) and selection thresholds ($\mathcal{T}_{fc}$ and $\mathcal{T}_{fe}$). The examples of filtered contact features and generated grasps are further shown in Supplementary Fig. 2. As $\mathcal{T}_{fc}$ increases, N_{fc} gradually decreases from thousands to hundreds and leads to unimodal change in **AP** (reaching the best performance at $\mathcal{T}_{fc} = 0.7$). Similar trends occur on $\mathcal{T}_{fe}$ and (N_{fe}/N_{ee}), with the best **AP** achieved at $\mathcal{T}_{fe} = 0.94 \sim 0.96$. Within the same interval of $\mathcal{T}_{fe}$, N_{ee} falls between 12 and 16, which is about $1.5 \sim 2$ times the number of objects placed in tested scenes. This statistical result characterized the preferred number of centered centroids for each object in general cluttered grasping tasks. Regarding the type of scene observation, Voxel generally gains superior performance over TSDF by approximately 3 in **AP**, attributing to more consistent feature distributions under varied visual conditions. In the validation, we set the hand batch $N_{batch} = 500$ and optimized the hand configurations by 200 iterations. According to Supplementary Table 1, it generally takes less than two seconds for every 100 iterations. Additionally, T_{time} rises about 0.12 seconds as N_{batch} increases 100 hands. $\mathcal{T}_{fc}$ significantly affects T_{time} with its resultant quantities of focal contacts N_{fc}, attributed to the time-consuming contact pairing process. In comparison, $\mathcal{T}_{fe}$ only affects the spatial distribution of grasp candidates related to the hand initialization and the contact-centroid optimization, resulting in minor perturbation (less than 0.01 seconds) in the optimizing time. Based on the studies, $\mathcal{T}_{fc}$ and $\mathcal{T}_{fe}$ are set as 0.7 and 0.95 in grasp generation. Note that other optimization parameters are tentatively set in the validation of Supplementary Table 1.

Supplementary Table 1 Performance of grasp generation under different feature-selection parameters and volumetric types. The observation type, focal contact threshold $\mathcal{T}_{fc}$, and focal centroid threshold $\mathcal{T}_{fe}$ are assessed under the metrics of optimization time and grasping **AP**. The left column presents $\mathcal{T}_{fe}(N_{fe}/N_{ee})$, and the row T_{time} denotes the optimization time for every 100 optimizing iterations when $N_{batch} = 500$.

Volumetric type: TSDF

$\mathcal{T}_{fc}$	0.5	0.6	0.7	0.8
N_{fc}	3097.1	2637.0	1624.8	253.9
T_{time} (seconds)	2.29	2.16	1.83	1.68
0.80 (822.2/20.4)	20.30	20.58	20.34	16.58
0.90 (693.3/19.2)	21.12	21.17	21.27	17.34
0.94 (550.1/17.2)	21.94	21.90	22.21	17.71
0.95 (480.9/15.9)	22.11	22.05	22.27	17.72
0.96 (376.6/13.8)	22.09	22.17	22.07	17.74
0.97 (241.2/10.4)	19.44	19.24	19.37	15.53

Volumetric type: Voxel

$\mathcal{T}_{fc}$	0.5	0.6	0.7	0.8
N_{fc}	2614.9	2173.1	1391.6	205.1
T_{time} (seconds)	2.26	1.89	1.71	1.58
0.80 (639.2/18.8)	23.22	23.75	23.75	20.00
0.90 (537.6/17.2)	23.96	24.55	24.30	21.04
0.94 (449.7/15.2)	24.99	25.09	25.44	21.54
0.95 (395.5/13.9)	25.49	25.46	25.76	22.29
0.96 (309.9/11.8)	25.31	25.33	25.75	21.86
0.97(201.1/9.1)	23.93	23.92	24.18	20.27

Supplementary Note 4: Strategies of hand initialization

Proper initial hand configurations are crucial, as they shorten the time to reach optimization convergence and result in higher-quality grasp outcomes. Specifically, initializing joints at extended finger positions generally yields more feasible optimized hand poses, as a larger opening width reduces finger penetration into the target object during the early optimization stages. A similar consideration applies to the initial translation: poses should be initialized above scene objects to mitigate initial penetration and value fluctuation in the objective functions during the early stage of optimization. Specifically, some optimizing hand configurations can get stuck in local minimums due to the high complexity of the constructed objectives and coarsely distributed contact features. To alleviate this issue, we imposed stochastic initialization on $\{g_H\}^{N_{batch}}$ to sufficiently explore the possible grasp configurations before gradient descent. Specifically, the hand joints g_{joint} are randomly sampled within a range of $[j_{low}, j_{low} + \gamma_j \cdot (j_{up} - j_{low}))$. In the defined range, $\gamma_j \in (0, 1]$ is set to encourage the initial robotic hand to open its fingers widely. To initialize hand orientation, we first assigned an approaching direction $\vec{v}_{sa}$ for the adopted robotic hand, along which it commonly approach and grip objects. Besides, an intended reaching direction $\vec{v}_{hg}$ was determined for a given grasping scene. For instance, in the case of top-down grasping, $\vec{v}_{hg}$ can be perpendicular to the desktop where the objects are placed. For general tasks, $\vec{v}_{hg}$ can be along the camera-viewing perspective (resulting in less obscured data for optimization) or random orientations. With $\vec{v}_{sa}$ and $\vec{v}_{hg}$, the initial hand rotation $g_{rotation}$ was randomized such that the angle α between these two vectors satisfies $\alpha(\vec{v}_{sa}, \vec{v}_{hg}) < (\pi/6)$. It is also preferable to incorporate a hand-approaching objective $\mathcal{O}_{ab} = b_{ab} \cdot \|\vec{v}_{sa} - \vec{v}_{hg}\|$ into $\mathcal{O}_{overall}$ to maintain the desired grasping directions. After initializing g_{joint} and $g_{rotation}$, the hand translation $g_{translation}$ is determined by shifting the centered candidate of H (defined as $\bar{c}_h = (1/N_{ch}) \cdot \sum^{C_H} c_h$) away from a randomly chosen centered centroid of S_{ee} along $-\vec{v}_{sa}$ for a distance of 0.05 m. These adjustable configurations provide the flexibility to alter the proposed optimization framework for tailored deployment. Specifically, under varying scene complexities and constraints imposed by accessible robotic hardware (such as 4-DoF SCARA robots or other manipulators with limited reachable space), practitioners can correspondingly regulate or expand the variety of grasp poses by actively modifying the permissible ranges of initial hand orientations and robotic joints, thereby aligning the optimization process with the practical requirements of the deployment environment.

Supplementary Note 5: Optimization time

Throughout extensive evaluations, SpaHybGen can be computation-intensive and time-consuming for general-hand grasp planning. It is critical to balance its efficiency and performance in different physical tasks. To investigate this problem, we characterize the relationship between the generation time and the resultant minimum objectives of seven robotic hands in Supplementary Table 2. To obtain the statistics, the cluttered scenes presented in Fig. 3c (i.e., five observations for each robotic hand in a clear-out grasping sequence, each containing different numbers of remaining objects) are utilized to rerun the SpaHybGen pipeline in three redesigned iteration schemes listed in Supplementary Table 2. The grasp generation is repeated ten times for each of the five observations, resulting in fifty runs for every robotic hand and iteration scheme, which are summarized in aspects of averaged generation time and averaged minimal objective. To facilitate the comparison across different robotic hands, we scaled their minimal objectives within $[0, 1]$.

From statistics, the generation time is primarily affected by the number of finger joints, which determines the computation complexity of kinematic chain in gradient descent. Notably, the scene complexity (e.g., the number of remaining objects) has less influence on the generation time because the number of focal contact features remains similar under different scene complexities. Our physical experiments adopted an optimizing iteration scheme of [200, 40], resulting in desirable grasping performance with the lowest objectives in Supplementary Table 2. Nevertheless, competitive objectives can also be achieved by reducing the overall iterations from 200 to 60 (especially for robotic hands with fewer joints), which can significantly shorten the ultimate computation time. Similarly, as shown in Supplementary Fig. 6, it is feasible to reduce the optimizing time by enabling the 40 iterations of penetration check in advance.

Supplementary Table 2 Generation time and grasping objective under different robotic hands and optimizing iteration schemes. Three optimizing schemes indicate [*Overall iterations*, *Iterations with collision avoidance*]. The numerical results present "*Averaged generation time* (seconds) / *Averaged minimal objectives* (scaled)", with n = 50 trials.

Robotic hand		**Iteration scheme**		
(Finger/Joint/Contact)		[200, 40]	[100, 40]	[60, 60]
FinRay-2F	(2/2/18)	1.69/0.34	0.91/0.36	0.65/0.40
Robotiq-2F	(2/2/18)	1.51/0.18	0.81/0.19	0.59/0.24
Robotiq-3F	(3/11/30)	3.68/0.86	1.97/0.95	1.41/1.00
ModelO-3F	(3/5/27)	2.22/0.51	1.20/0.51	0.86/0.52
FinRay-4F	(4/4/20)	2.07/0.60	1.10/0.60	0.78/0.67
Brunel Hand	(5/9/40)	3.45/0.43	1.88/0.54	1.39/0.71
LEAP Hand	(4/16/30)	4.33/0.20	2.29/0.25	1.58/0.31

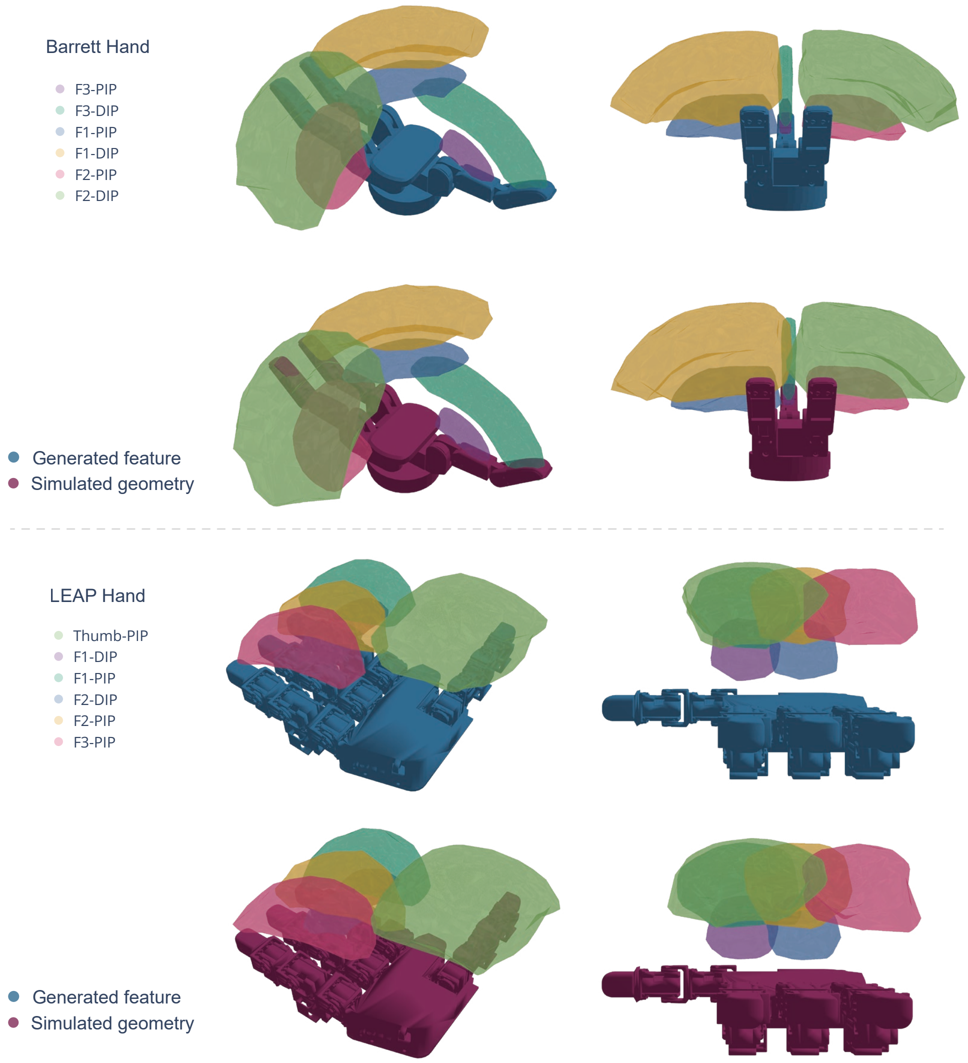
Supplementary Note 6: Grasp diversity and contact pattern of multi-finger dexterous hands

The constructed contact features combine all feasible contacts available in the entire scene as one single contact-feature map, which facilitates a dense, smooth, scene-level contact distribution for synthesizing possible diverse multi-contact grasps using dexterous robotic hands. In this regard, despite that the contact features are decomposed from antipodal grasps, the combination of these dense generated contact instances provides sufficiently rich potential contacts for generating dexterous grasp configurations. To investigate the degree of hand dexterity synthesized with contact features, we designed a comparative ablation study between the proposed features and simulated object geometry. Specifically, we leveraged the fine-grained object geometry of GraspNet-1Billion to generate the surface normals and voxel occupancy of the simulated cluttered scene (noted as Simulated geometry). We created geometric contact alternatives around the cluttered objects within a distance threshold, where their directions align with the surface normals estimated on the simulated object models. The geometric contact-certainty scores are assigned according to the distances from geometric contact alternatives to the object surfaces. Geometric centroid alternatives are obtained through object occupancy in the simulated spatial space. In comparison with the proposed contact features, the simulated geometric alternatives are denser, smoother (as calculated from precise object surfaces), and only correlated with local geometry of separate objects without encoding grasp properties and scene-related information. Moreover, in contrast with the harsh sensing conditions adopted in this work, the simulated alternatives are only accessible with known object prior of fine-grained scanned geometries, resembling the common data routines adopted in prior simulation-heavy methods in Table 1. The geometric alternatives are leveraged to reveal the original hand dexterity of multi-finger robotic hands during grasp optimization (similar to simulation-based methods), serving as a comparative baseline for the generated contact features.

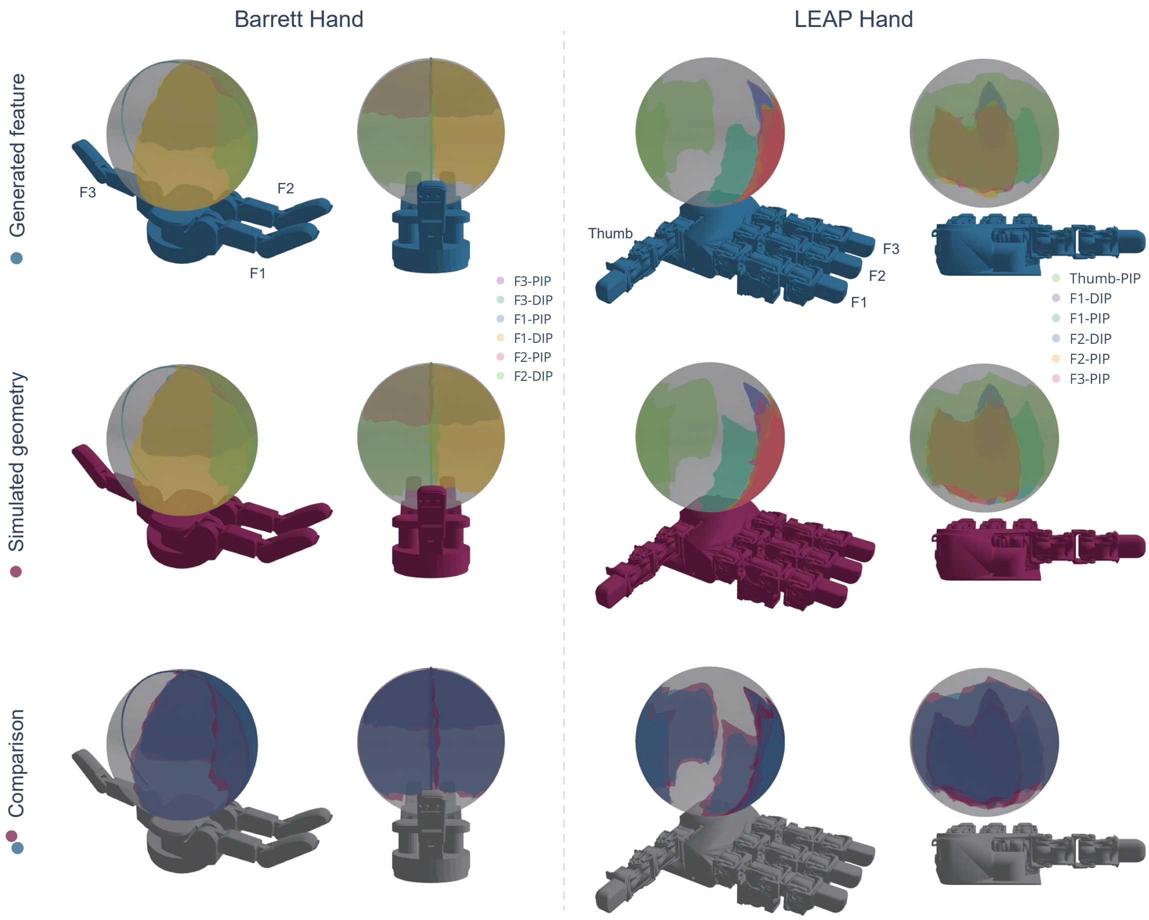
Using Barrett Hand and LEAP Hand, we conducted grasp optimization on the *Seen* test set of GraspNet (30 cluttered scenes) with contact features and geometric alternatives. The grasp synthesis is repeated 48 times for each scene, with hand batch set to 48 and optimizing iterations of 200. By selecting the top-three grasps for every grasp synthesis, there are 4320 selected generated grasps for each of data types, all of which are jointly mapped to hand coordinates to characterize general grasp patterns. To provide a granular analysis of all top-ranked grasps, Supplementary Table 3 tabulates their average engaged contact regions, contact candidates, and robotic fingers. In the study, six potential contact candidates C_H (i.e., contact pads) are annotated to both robotic hands, as shown in Supplementary Fig. 3. Notably, among six contact candidates, the generated grasps of Barrett Hand averagely made 3.44 contacts with objects, distributing on 2.42 robotic fingers. For LEAP Hand, the average engaged robotic fingers reach 3.29 out of four fingers, indicating that most grasps are effectively synthesized with three and four fingers. To reveal the resultant hand dexterity under two different data types, we resort to the diversity of the generated contact positions and contact orientations. Supplementary Fig. 3 comparatively visualizes the reachable volumes of the engaged contact regions that are optimized by the contact features or simulated geometric alternatives. The spatial volumes spanned by the grasp group generated from contact features, i.e., engaged volumes of blue hand models in Supplementary Fig. 3, closely resemble and match that of simulated geometry (magenta hand models). This implies that the feature density of the proposed contact dataset sufficiently supports grasp optimization to explore the inherent hand reachability without evident positional constraints. Moreover, as another critical indication of hand dexterity, Supplementary Fig. 4 depicts the variety of contact directions to reveal the engaged contact patterns of all generated grasps. As evidently shown in the first-row spheres, the generated contact-orientation maps of dexterous fingers (e.g., F1 and F2 of Barrett Hand and the thumb of LEAP Hand) cover a wide range of orientations, revealing distinct non-antipodal contact patterns when collectively synthesized with other fingers. Additionally, as shown in the bottom-row comparative depiction, the explored contact-orientation ranges generated from contact features closely align with that of simulated geometric ones. Taking together with the contact-engagement statistics, the aligned similarities of both reachable volumes and contact maps prove that the dense distribution of contact features sufficiently preserves the feasible grasp primitives and dexterity of multi-finger robotic hands, while maintaining robustly obtainable in clutter scenes and encoding further grasp-level information.

Supplementary Table 3 Statistics of engaged contact patterns for multi-finger robotic hands. The contact area represents the proportion of engaged area given all predefined regions of contact candidates for Barrett Hand and LEAP Hand. The averaged numbers of engaged pads and fingers are collected depending on their associated contact area (n = 4320 grasps).

Hand type	Barrett Hand		LEAP Hand	
Data type	Simulated geometry	Contact feature	Simulated geometry	Contact feature
Contacted area	0.549/1	0.591/1	0.725/1	0.777/1
Engaged pads	3.32/6	3.44/6	4.09/6	4.04/6
Engaged fingers	2.53/3	2.42/3	3.54/4	3.29/4



Supplementary Fig. 3 Contact reachable volumes for multi-finger robotic hands under contact features and simulated geometric features. For both robotic hands, the reached ranges of six contact candidates are visualized given the 4320 generated grasps. The collected contact positions are smoothed by averaging 5% outliers with their nearest neighbors, which are then used to compute concave hulls as the robust reachable volumes. The finger terminology is annotated in Supplementary Fig. 4 for clarity.

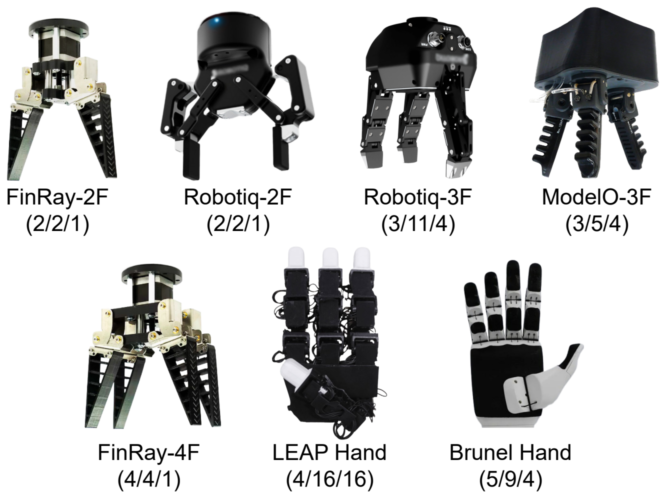


Supplementary Fig. 4 Contact orientation maps for multi-finger robotic hands under contact features and simulated geometric features. The contact orientations of 4320 generated grasps are projected to a standard sphere according to their relative directions in the hand frame. To obtain robust map regions, the projected points are smoothed by averaging 5% outliers with their nearest neighbors, which are used to compute concave regions as the orientation maps.

Supplementary Note 7: Benchmark setup for three physical assessments

Task 1: comparative semi-cluttered grasping

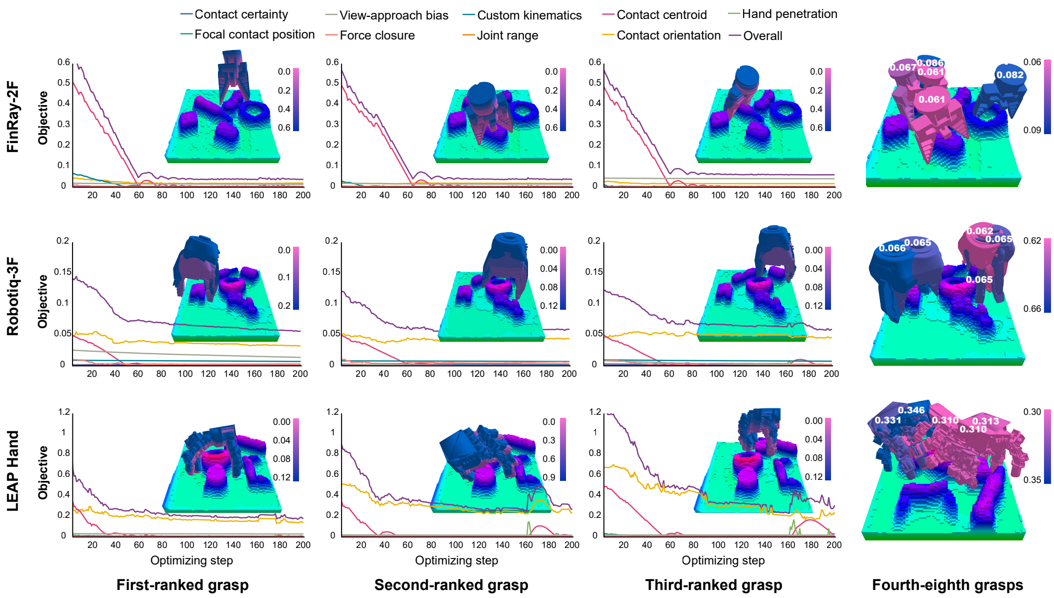
The hardware in Fig. 3a is adopted to perform general-hand assessment. In the semi-cluttered grasping, all five objects shown in the figure were randomly poured into the tray. An RGB-D camera (Azure Kinect) was attached above the tray to capture depth images for encoding voxel volumes. A robot arm (UR-5e) mounted on the workbench navigated a robotic hand for grasping objects from a tray. All hardware devices were coordinated by a computer (running Ubuntu 18.04 and Robot Operating System) equipped with an Intel Core i7-11700 CPU and NVIDIA GeForce 3060-Ti graphics card. For hand optimization, 48 (N_{batch}) grasp candidates underwent 200 optimization iterations. The objective of penetration $\mathcal{O}_{pn}$ was activated in the last 40 iterations. The following procedure was repeated to clear objects until the tray is empty. First, five objects (one object for single-object tests in Supplementary Table 4) were randomly placed inside the picking region. On obtaining the best grasp proposal from SpaHybGen, the robotic hand was actuated to the pre-grasp configuration, and the robot arm navigated the gripper to the desired grasp pose. Then, all fingers were simultaneously closed until they gained stable contacts. If contact equilibrium is formed between the gripper and the object, the arm-gripper system lifted the grasped object and placed it into the dropping container. One grasp trial is successful when an object is picked from the tray and dropped to the targeted place. For data in Fig. 4 and Supplementary Table 4, the *grasping success rate* is computed as *(number of successful grasps) / (number of grasp trials)*. For both single-object and semi-cluttered scenes tested in Supplementary Table 4, the *number of grasp trials* corresponds to the total grasp execution times needed for clearing 25 items (five objects repeated over five experimental runs).



Supplementary Fig. 5 Seven robotic hands used in the physical tests. The numbers below each robotic hand indicate the quantities of *(Finger/Joint/Actuator)* of their hand models.

Supplementary Table 4 Grasping statistics of seven robotic hands in semi-cluttered scenes (visualized in Fig. 3d). Values indicate *(total grasp trials) / (grasping success rate)*, with n = total grasp trials.

Robotic hand (Finger/Joint/Actuator)		FinRay-2F (2/2/1)	Robotiq-2F (2/2/1)	Robotiq-3F (3/11/4)	ModelO-3F (3/5/4)	FinRay-4F (4/4/1)	Brunel Hand (5/9/4)	LEAP Hand (4/16/16)
AdaGrasp	Single	36/69.4%	30/83.3%	28/89.3%	37/67.6%	46/54.3%	N/A	N/A
	Clutter	36/63.9%	27/**92.6%**	34/61.8%	32/78.1%	38/65.8%	N/A	N/A
	Overall	72/66.7%	57/87.7%	62/74.2%	69/72.5%	84/59.5%	N/A	N/A
HybridGen	Single	30/83.3%	27/92.6%	28/89.3%	26/**96.2%**	37/67.6%	N/A	N/A
	Clutter	30/83.3%	30/83.3%	31/67.7%	24/87.5%	31/54.8%	N/A	N/A
	Overall	60/83.3%	57/87.7%	59/78.0%	50/92.0%	68/61.8%	N/A	N/A
SpaHybGen	Single	25/**100.0%**	26/**96.2%**	25/**100.0%**	26/**96.2%**	25/**100.0%**	26/**96.2%**	25/**100.0%**
	Clutter	27/**92.6%**	27/**92.6%**	26/**96.2%**	25/**100.0%**	26/**96.2%**	26/**96.2%**	26/**96.2%**
	Overall	52/**96.2%**	53/**94.3%**	51/**98.0%**	51/**98.0%**	51/**98.0%**	52/**96.2%**	51/**98.0%**



Supplementary Fig. 6 Optimization curves for FinRay-2F, Robotiq-3F, and LEAP Hand. The fourth to eighth ranked grasps are visualized alongside their final $\mathcal{O}_{overall}$ values. The related scenes and successfully executed top-ranked grasps are demonstrated in 3b.

Task 2: dynamic grasp servo in densely-cluttered scenes

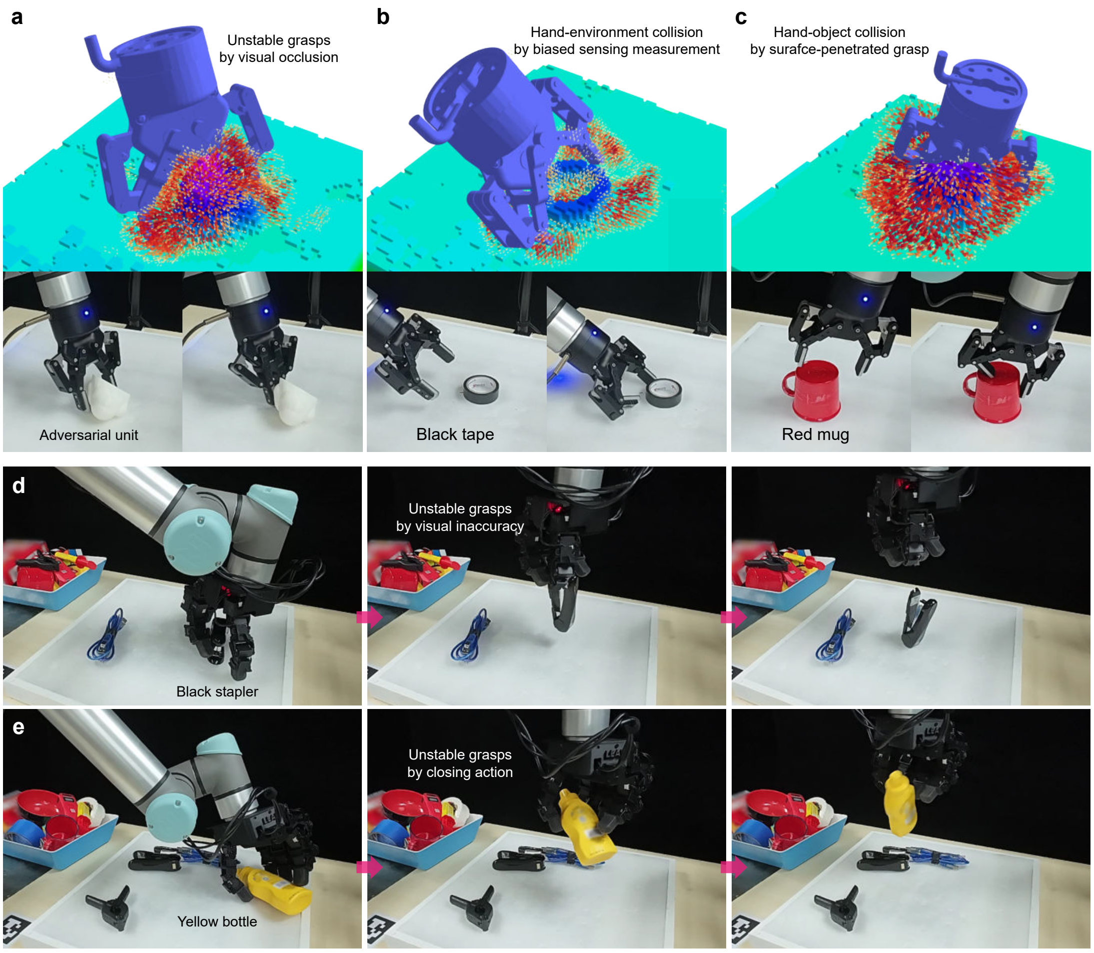
In highly cluttered scenes, visual occlusion becomes inevitable and severe when using a single, fixed visual sensor. Voxel volumes generated from such highly occluded depth observations frequently contain misleading volumetric voxels that are falsely occupied due to being unseen, ultimately leading to inferior contact features during the estimation process. Consequently, static grasping with a fixed camera configuration cannot reach the maximal potential of SpaHybGen in these environments. To enable more effective grasping in dense clutter, we implemented a strategy of online dynamic grasp updates during the robotic hand's approach phase toward targeted objects. To facilitate observation of the workspace from different, less occluded views during the approach, we installed a miniaturized depth camera board (Pico Flexx) on the palm of the robotic hands (see the bottom-left inset of Fig. 4b). This setup provided real-time depth capture from a closer vantage point, replacing the static Azure Kinect camera used in previous experiments. We conducted this dynamic grasping test with two multi-finger robotic hands: the Robotiq-3F and the LEAP Hand.

To create highly cluttered scenes, we randomly chose 12 to 16 items from each group of four object clusters (see Supplementary Fig. 7) and piled them up in the tray (repeated five times). A focused grasping metric *FGM* is defined as *(number of successful grasps) / (number of grasped objects)* to quantitatively characterize the level of multi-object undesirable grasping. The grasping process began by positioning the robotic hand in a predefined initializing pose, enabling the capture of the first depth frame via the palm camera. Then, a preliminary grasp pose was generated by SpaHybGen utilizing the first depth observation (chosen from 96 grasp candidates optimized by 48 iterations). To start dynamic grasp update, the robotic hand was navigated to a starting pose located at 0.15 m offset along the reverse approaching direction of the preliminary grasp pose (see Fig. 4b). Then, a second grasp-refinement program iteratively optimized the single hand candidate (initialized by the preliminary grasp pose) using newly captured depth images. Meanwhile, the robotic hand approached the latest optimal grasp pose with incrementally shorter offsets before reaching the interacting pose (0.05 m offset, the sensing limit of the camera board, see Fig. 4b). Then, the robotic hand was directly navigated to the final optimal grasp and gripped the target. One grasp is deemed successful execution when it grips objects from the tray to the targeted place. Because the second grasp-refinement optimization is well initialized with the preliminary grasp, it can reach new convergence within three iterations on the latest depth observation, operating at $15 \sim 20$ Hz. To reveal the superiority of dynamic grasp update, we also conducted static grasping on the same cluttered scenes by running full grasp optimization on the first depth observation. The static grasping generation optimized 96 grasp candidates for 128 iterations, which is close to the number of overall optimizing iterations in dynamic grasping.

Supplementary Fig. 7 Diverse items used in densely-cluttered dynamic grasp servo. The experiments include 64 objects clustered into four groups, including fabric toys, artificial vegetables, adversarial units, and household items.

Task 3: multi-gripper simultaneous sorting

We equipped Robotiq-2F, ModelO-3F, and Robotiq-3F robotic hands on three robot arms (two UR-3e and one UR-5e) in the same workbench (see Fig. 5b). One fixed Azure RGB-D camera captured depth observations for the grasping system. Six objects with different sizes were collected to facilitate a size-specific sorting task. The Robotiq-3F gripper has a wider grasp opening of 0.155 m than the Robotiq-2F and ModelO-3F grippers (the maximal grasp widths of both grippers are 0.085 m). Thus, considering object sizes, the carton and blue ball are suitable for the Robotiq-3F gripper, while other objects can be grasped by the ModelO-3F and Robotiq-2F grippers. Three robotic hands were coordinated within the same grasp planning pipeline shown in Fig. 5a. In the simultaneous grasping loop, the camera captured one depth image for encoding one voxel volume, which was then sent to three individual grasp optimization nodes established for three adopted robotic hands. Each optimization node returned the generated grasp configurations for its targeted gripper. Given optimized grasp candidates of multiple robotic hands, we manually selected one solution according to the status of the cluttered scene. Other setting and parameters of grasp generation are identical to semi-cluttered grasping.



Supplementary Fig. 8 Case analysis of failure modes for SpaHybGen in physical grasping. **a**, Unstable grasps may be formed by inferior contacts due to visual occlusion from single-viewed observations. **b**, Biased depth measurement leaded to inaccurate table heights, causing collision between the robotic hand and environment. **c**, The size of the object is close to the maximal span of the hand opening width, causing possible collisions between hand and object due to errors of hardware calibration/execution, inferior navigation planning, or slight hand penetration. **d**, The reflective and black surfaces of objects significantly degrade depth measurement, leading to wrong contact estimation and unstable grasps. **e**, The closing action from optimal grasps may cause gripping failure.

Supplementary Video 1

Semi-cluttered grasping experiments of general robotic hands. The video demonstrates extensive assessments on seven robotic hands, ranging from two to five fingers.

Supplementary Video 2

Densely-cluttered dynamic grasp update. The scenes are piled up with four groups of 64 diverse objects, which are cleared using the Robotiq-3F and LEAP Hand.

Supplementary Video 3

Multi-hand simultaneous sorting experiment. The demonstration adopted three distinct robotic hands for sorting a cluttered scene.