2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Learning Realistic and Reasonable Grasps for Anthropomorphic Hand in Cluttered Scenes

Haonan Duan[1,2,*], Yiming Li[1,2,*], Daheng Li[1,2], Wei Wei[1,2], Yayu Huang[1,2], Peng Wang[1,2,3,✉]

***Abstract*— Grasping is one of the most fundamental skills for humans to interact with objects. However, it remains a challenging problem for anthropomorphic hands, due to the lack of object affordance understanding and high-dimensional grasp planning. In this work, we propose an anthropomorphic hand grasping framework to learn realistic and reasonable grasps in cluttered scenes, which tackles the problem in three items: 1) graspable point segmentation; 2) hand grasp generation and 3) grasp optimization. Specifically, our method generates high-quality hand grasps efficiently without complete object models by learning graspable points, associated grasp configurations from observed point cloud in a parallel manner and optimizing predicted grasps based on hand-object contacts. Simulation experiments show that our model generates physical plausible grasps for the anthropomorphic hand effectively with over 70% success rate. Real-world experiments demonstrate that the model trained in simulation performs satisfactorily in real-world scenarios for unseen objects.**

I. Introduction

Robotic grasping is one of the most fundamental and long-lasting problems in the robotics community. While grasping with parallel-jaw grippers has been well investigated and widely applied in industrial and domestic services [1], [2], dexterous grasping with an anthropomorphic hand in complex scenes remains an open challenge. With the concept and trend of Embodied AI [3], and the increasing expectation of integrating humanoid robots into human life and production, substituting parallel-jaw gripper with anthropomorphic hand is a necessary step of accomplishing that long-cherished wish.

Traditionally, anthropomorphic hand grasping is tackled by analytical methods with commonly used metrics: force closure [4] and ϵ-quality metric [5]. However, such methods are inevitably computationally expensive [6] based on the assumption of being aware of object physical and geometric information.Although early data-driven methods solve the shortcomings to a certain extent through generic grasping, it is still limited by this assumption [7]–[10].

This work was supported in part by the National Natural Science Foundation of China under Grants (91748131, 62006229 and 61771471), in part by the Strategic Priority Research Program of Chinese Academy of Science under Grant XDB32050106, and in part by the InnoHK Project.

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China. (E-mail: {duanhaonan2021, lidaheng2020, wei.wei2018, huangyayu2021}@ia.ac.cn, ymli.cn@gmail.com)

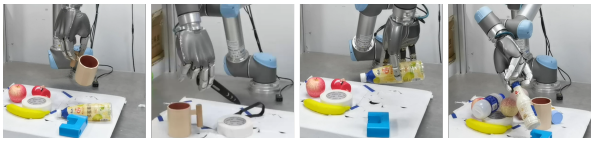
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China.

[3] Centre for Artificial Intelligence and Robotics, Hong Kong Institutue of Science and Innovation, Chinese Academy of Sciences, Hong Kong, China.

* Authors contribute equally.

✉ Corresponding author: peng_wang@ia.ac.cn

(a)



(b)

Fig. 1: Inspired by human hand grasping, we design an anthropomorphic hand grasping framework for cluttered scenes based on object affordance to generate realistic and reasonable grasps for anthropomorphic hand. (a). Human hand grasping. (b) Anthropomorphic hand grasping.

Recent work pays attention to generating hand grasps from partial observation. [11], [12] deal with grasping for single object. However, very limited work focuses on anthropomorphic hand grasping in cluttered scenes, which needs to handle higher collision probabilities and lower approaching space under occluded situations. One can reconstruct the full scene model from partial observations [13]. Nevertheless, it severely restricts the diversity and generalizability. [14] proposes to directly predict collision-free grasps in cluttered scenes with different grasp types. Above methods remain a considerable gap between the generated grasps and human grasping manners without taking into account object affordance. Inspired by the work of hand-object interactions [8], [15] and object affordance understanding [16], [17], we present a method to learn object affordance as well as grasp configurations in cluttered scenes under partial observation to improve realism and reasonableness (shown in Fig. 1).

By adopting the ideas in previous work [12], [14], our single-shot grasp proposal network consists of three components: 1) Graspable point segmentation module for detecting graspness points and affordance area; 2) Hand grasp generation module for predicting coarse 6-DoF wrist pose and joint angles of the anthropomorphic hand; 3) Grasp optimization module for refining precise grasp configurations based on hand-object contact. Different from previous work [12], [13] that utilizes object completion technologies to reconstruct the scene, our method learns object affordance implicitly, which reduces interference caused by cascaded modules. To train the network, we collect 179 household objects and label their affordance areas based on human grasping manner. A large-scale grasp synthetic dataset is

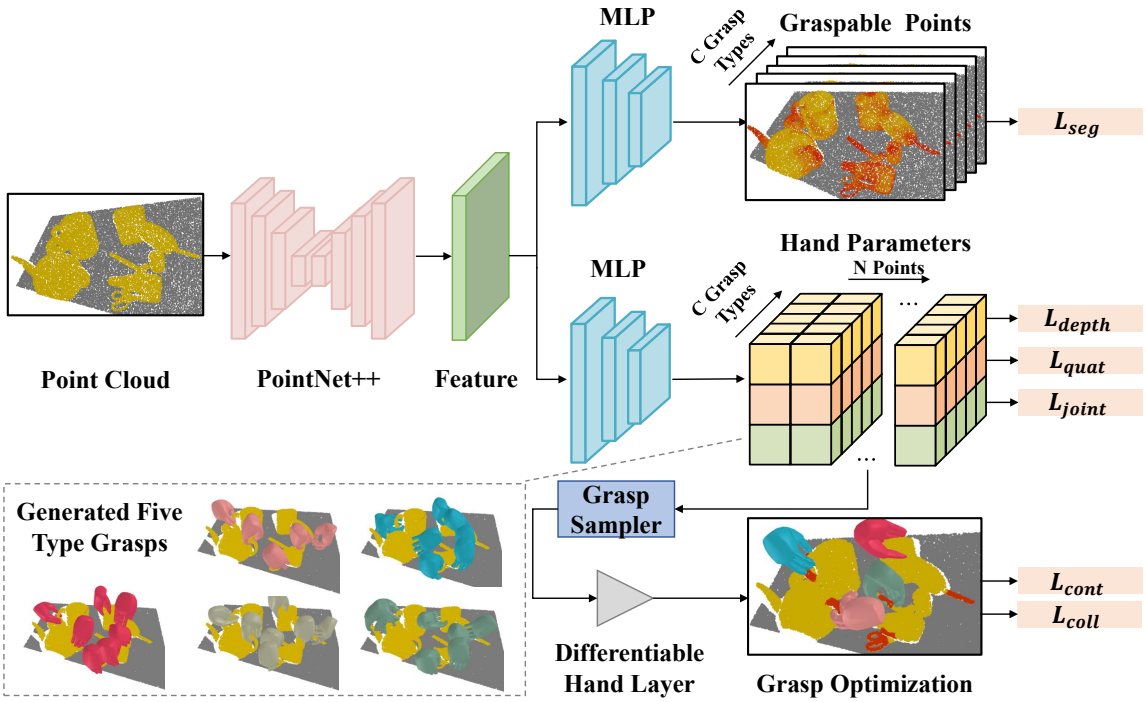


Fig. 2: Framework of the proposed anthropomorphic hand grasping network for generating realistic and reasonable grasps. Given the observed point cloud, our model extracts hierarchical point features and predicts graspable points as well as associated grasp configurations for each grasp type. Generated grasps are sampled and a differentiable hand layer is also followed to reconstruct the hand and optimize grasp affordances based on hand-object contacts.

built in simulation that contains over 5K cluttered scenes and 10M grasp annotations. The effectiveness, robustness and generalizability of our method are demonstrated by simulation and real robot experiments.

Overall, our contributions can be summarized as follows:

- Present an anthropomorphic hand grasping framework that jointly learns graspable points and grasp configurations to generate high-quality grasps in cluttered scenes without complete object models.
- Segment graspable points based on both graspness and affordance area, and optimize grasps based on hand-object contact to guarantee realism and reasonableness.
- Conduct experiments in simulation and on the real robot to demonstrate the method not only reasonably generates high-quality grasps, but also performs well for affordance-aware grasping in realistic cluttered scene.

II. Related Work

Anthropomorphic Hand Grasp Planning. The topic of anthropomorphic hand grasping is defined as an analytical problem based on the optimization criteria [4], [5] in the early work. Such methods mainly depend on metrics by requiring precise object models [6] without computational cost consideration. Emerging data-driven grasp synthesis is divided into object-aware and object-agnostic [18]. The former still assumes the explicit object geometric information is known. [7], [8] build the grasp dataset for anthropomorphic hands on the captured images or videos. [9], [10] generate coarse grasp postures from a sampler and optimize the hand-object contact to obtain reasonable results. On the contrary, object-agnostic methods predict the feasible grasp configurations based on partial observation. [11] utilizes multi-view depth images to predict the human-like grasp directly. [12] generates diverse grasp poses on the object after single-view point cloud completion. [13] applies the same techniques on cluttered scenes. [14] predicts dense grasp configurations with different grasp types on the single-view point cloud.

Grasp in Clutter. Most research centers on dealing with single object placed on planar surface [1], [19], especially for high dimensional grippers [9], [11], [12]. However, such planners designed for single-object grasping cannot be directly transferred to cluttered scenes due to their weak ability to handle the increased probabilities of collisions, the incremental areas of occlusions and the limited space of approaching [20], [21]. Some work propose to predict 6-DoF grasp configurations in clutter with parallel-jaw grippers [2], [22]. Yet these methods gain desirable performance, grasping in clutters using an anthropomorphic hand is still an open problem. [13] generates several grasp configurations through a multi-stage method. [14] tackles the task in end-to-end fashion.

Grasp Affordance. Affordance is the physical and geometric property of objects, which refers to the part of the objects with the high probability to prosper grasps based on human grasping habits [18]. Previous work mainly guided by hand-object interactions [8], [15]. Some literature propose grasp affordance dataset for anthropomorphic hand [8], [15] and grasp synthesis method [17]. Grasping with affordance in clutter remains a challenge because of the declining success rate of affordance detection resulting from object occlusions.

III. Problem Statement

This work concentrates on anthropomorphic hand grasping in cluttered scenes, which requires understanding object affordance to generate physically plausible and collision-free grasp configurations. More formally, given the observed point cloud $\mathcal{P}$ as input, our model $\mathcal{M}$ produces grasp configurations $\mathcal{G}$:

$$\mathcal{M} : \mathcal{P} \rightarrow \mathcal{G} \tag{1}$$

Each grasp configuration $\boldsymbol{g} \in \mathcal{G}$ is represented by a hand wrist pose $\boldsymbol{p}$, hand joint angles $\boldsymbol{\theta}$ and a predefined grasp type $\boldsymbol{c}$, *i. e.* $\boldsymbol{g} = \{\boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{c}\}$. Hand wrist pose $\boldsymbol{p} = (\boldsymbol{t} \in \mathbb{R}^3, \boldsymbol{q} \in \mathbb{R}^4)$ is given in $\mathbb{SE}(3)$, including the translation and the orientation. $\boldsymbol{\theta} \in \mathbb{R}^{20}$ denotes 20-DoF hand joint angles of our HIT-DLR II hand. Grasp type $\boldsymbol{c}$ in grasp type sets $\boldsymbol{C}$ is defined by $[\boldsymbol{\theta}_{\boldsymbol{c}}^{open}, \boldsymbol{\theta}_{\boldsymbol{c}}^{close}]$, which respectively represents joint boundaries of the anthropomorphic hand when it opens and closes.

IV. Dataset Generation

To train and test our proposed method, we construct a large-scale grasp synthetic dataset for the used HIT-DLR II anthropomorphic hand.

A. Single-Object Grasp Generation

Object. We collect 179 household objects with various shapes and categories. All objects are rigid, watertight and sharing the same coefficients (*e.g.* density and friction).

Grasp Type. Based on the work in human grasp type classification [23] and mechanical properties of HIT-DLR II hand, five selected grasp types are shown in Fig. 3(a).

Grasp Sampling. Approach-based scheme [24] are adopted for sampling grasp with moving towards the opposite direction of the point normal. Dense grasp candidates are generated by sampling 15 uniform depths (at $5mm$ intervals) and 24 in-plane rotation angles (at 15-degree intervals) on 512 sampled points with normals on each object surface.

Grasp Simulation. We annotate grasp labels in MuJoCo [25] physical simulator. The grasp simulation consists of three steps: 1) Objects are stationary while the HIT-DLR II hand executes a grasp attempt with sampled pose and type. The hand closes from $\boldsymbol{\theta}_{\boldsymbol{c}}^{open}$ to $\boldsymbol{\theta}_{\boldsymbol{c}}^{close}$ until all fingers contact the object or reach maximum joint angles. 2) All fingers maintain the grasp while gravity is present till the simulator reaches a stable state or the object falls from the hand. 3) Slight shaking of the hand is performed to filter out unstable grasps. The distribution of the number of generated grasps for each grasp type is shown in Tab. I. Examples of successful grasps are shown in Fig. 3(b).

TABLE I: Distribution of the number of grasps.

Grasp Type	No. of Valid Grasps (Million)	Proportion(%)
Parallel Extension	4.35	24.8
Pen Pinch	2.41	13.7
Palmar Pinch	0.81	4.6
Precision Sphere	4.91	27.9
Medium Wrap	5.10	29.0

Affordance Annotation. 103 out of 179 objects with clear usage properties (*i.e.* electric drill, clamp, and hammer) are selected and manually labeled with the affordance 3D bounding box through LabelCloud [26] (shown in Fig. 3(c)). Rest 76 objects are annotated with *unknown* labels.

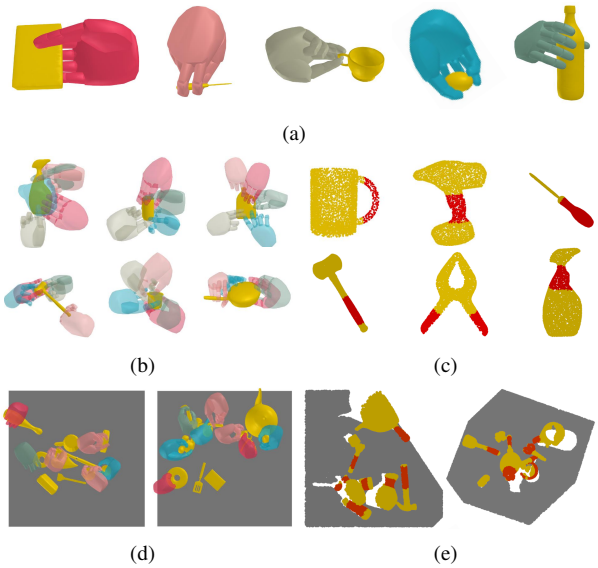


Fig. 3: Illustrations of our synthetic dataset. (a) Selected five grasp types. Left to right: parallel extension, pen pinch, palmar pinch, precision sphere and medium wrap. (b) Generated grasps of single objects. (c) Affordance area of single objects. (d) Scene grasps. (e) Scene affordance grasp areas.

B. Scene Grasp Generation

Cluttered Scene Construction. We use BlenderProc [27] to generate 5K cluttered scenes and 20K images for model training. For each scene, 10 objects are sampled with random poses and fell on a table to construct a clutter. 4 RGB-D images are rendered from random views to obtain single-view scene captures.

Scene Grasp Transformation. We apply the same transformation of scene objects to object-centric grasps and utilize an offline collision detection module to filter invalid grasps that collide with the scene (shown in Fig. 3(d).).

Point Cloud Labelling. Scene point clouds are generated from RGB-D images according to camera intrinsics. To map scene point clouds with grasp labels, we adopt a KD-Tree algorithm for each grasp to search the nearby points among point clouds with query radius $r = 0.005$m. Scene points in a group will be broadcast with the same label as the grasp point and affordance area (shown in Fig. 3(e)).

V. Method

The overall pipeline is illustrated in Fig. 2. Our model utilizes a multi-task learning framework to predict graspable points and grasp configurations in a parallel manner, as the high correlation of these two properties.

A. Network Overview

Previous work [14] explores predicting dense grasps in cluttered scenes for anthropomorphic hand, yet lacks the ability to generate feasible grasps due to insufficient understanding of object properties. For our method, the object affordance area and hand-object contacts are taken into consideration to promote realistic and reasonable grasps.

We utilize PointNet++ [28] with multi-scale group (MSG) as backbone network for point cloud features encoding. Multi-layer perceptrons (MLPs) are followed for graspable point segmentation and hand grasp generation. Five predefined grasp types are adopted to reduce the high-dimensional search space of the anthropomorphic hand. For each grasp type, the network predicts point-wise grasp parameters. A differentiable layer designed on the forward kinematics derivation of HIT-DLR II hand is employed to reconstruct selected high-score grasps and optimize hand-object contacts.

B. Graspable Point Segmentation

We consider graspable points in two aspects: graspness and affordance areas. Graspness represents which part can be grasped to avoid falling, while affordance areas mean which part is compatible with human grasp manner.

For affordance area, each observed scene point is labelled with *positive*, *negative* or *unknown*. For graspness, the three annotations are labeled five times for each grasp type. During training, we supervise the positive and negative points through two-class cross-entropy loss, while points with unknown annotations are ignored:

$$\mathcal{L}_{seg} = \sum_{i \in P_f} \mathcal{F}_{cls}(y_i^f, \widehat{y_i^f}) + \sum_{i \in P_g, c \in C} \mathcal{F}_{cls}(y_{i,c}^g, \widehat{y_{i,c}^g}), \tag{2}$$

where $\mathcal{F}_{cls}$ denotes weighted cross-entropy loss. $y_i^f, \widehat{y_i^f}$ are affordance label and predicted mask for point i, while $y_{i,c}^g, \widehat{y_{i,c}^g}$ are graspness label and associated prediction for grasp type c on point i, respectively. P_f and P_g are two point sets with known labels (positive or negative) in terms of affordance areas and graspness. To handle the imbalance of graspable points, the weighted of positive and negative points are set to (5.0, 1.0) for the affordance and (10.0, 1.0) for graspness. We also randomly sample 5% table points as negative points to decrease the interference caused by the table plane.

During testing, the network predicts point-wise graspable probabilities, the softmax function is adopted to calculate scores in terms of graspness and affordance. The final grasp score $s_{i,c}$ for point i and grasp type c is:

$$s_{i,c} = \mu s_i^f + (1 - \mu) s_{i,c}^g, \tag{3}$$

where s_i^f denotes the affordance grasping score for point i and $s_{i,c}^g$ is the graspness score for point i under grasp type c. μ is a coefficient to balance the two items.

C. Hand Grasp Generation

As mentioned in Section III, a grasp configuration g is represented by hand wrist pose $\boldsymbol{p}$, joint angle $\boldsymbol{\theta}$ and grasp type $\boldsymbol{c}$. A hand wrist pose $\boldsymbol{p} = (\boldsymbol{t}, \boldsymbol{q})$ in $\mathbb{SE}(3)$ can be simplified by the 3D coordinates of graspable point i, a 3D rotation $\boldsymbol{q}$ and a grasp depth $\boldsymbol{d}$ [14] via approach-based grasp sampling. We adopt smooth $\mathcal{L}_1$ loss to regress grasp depth. For optimizing 3D rotations, we utilize a distance loss to measure the related angle between predicted quaternions $\hat{q}$ and ground truth q. The loss function is formulated as follows:

$$\mathcal{L}_{depth}^{i,c} = \lambda_1 \mathcal{F}_{\mathcal{SL}_1}(d_{i,c}, \widehat{d_{i,c}}) \tag{4}$$

$$\mathcal{L}_{quat}^{i,c} = \lambda_2(-\mathcal{D}_q(q_{i,c}, \widehat{q_{i,c}}))), \tag{5}$$

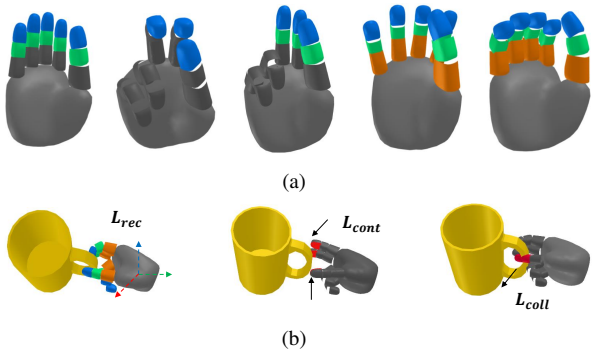


Fig. 4: (a) Contact part of each grasp type. Points on the inside of the contact finger are used to compute the contact loss. (b) Illustrations of three losses for generating a hand mesh based on predicted grasp configurations.

where λ_1 and λ_2 are two constants to balance different items. $\mathcal{F}_{\mathcal{SL}_1}$ donates smooth $\mathcal{L}_1$ loss, $\mathcal{D}_q(q_{i,c}, \widehat{q_{i,c}}) = (2\langle q_{i,c}, \widehat{q_{i,c}} \rangle^2 - 1)$ represents the cosine distance between two quaternions, and $\langle \cdot, \cdot \rangle$ donates a dot product. $d_{i,c}, \widehat{d_{i,c}}, q_{i,c}, \widehat{q_{i,c}}$ are ground truth and predict values of grasp depth and 3D rotation for point i at grasp type c, respectively.

Finally, we utilize $\mathcal{L}_2$ loss to regress finger joint angles:

$$\mathcal{L}_{joint}^{i,c} = \lambda_3 ||N(\theta_{i,c}) - N(\widehat{\theta}_{i,c})||_2^2, \tag{6}$$

where $\theta_{i,c}$ and $\widehat{\theta}_{i,c}$ are labels and predict values of joint angles for point i at grasp type c. $N(\cdot)$ is a normalization function. λ_3 is a constant. The hand reconstruction loss is a sum of the three individual loss functions:

$$\mathcal{L}_{rec} = \sum_{i \in P_{pos}, c \in C} (\mathcal{L}_{depth}^{i,c} + \mathcal{L}_{quat}^{i,c} + \mathcal{L}_{joint}^{i,c}), \tag{7}$$

where P_{pos} is a set of points with positive graspness labels. During testing, each scene point predicts associated hand configurations for five grasp types. All of the generated grasps are ranked by grasp score $s_{i,c}$ and we use a pose-NMS [2] algorithm to select top K grasps for grasp optimization.

D. Grasp Optimization

A differentiable HIT-DLR II hand layer is designed to reconstruct the anthropomorphic hand. The layer takes batch hand configurations $(\boldsymbol{t}, \boldsymbol{q}, \boldsymbol{\theta})$ as input and outputs corresponding vertices, normals and faces of hand mesh. The hand parameters are optimized by minimizing the distance between hand vertices and object surfaces.

We first introduce a contact loss $\mathcal{L}_{cont}$ to model hand-object contacts naturally. For each grasp type, contact part is defined to represent which part of the hand should touch the object (shown in Fig. 4(a)), and calculate the unsigned distance between scene points and contact part. If the distance is less than a fixed threshold $\mathcal{T}$, we take the scene point as the contact point and optimize the contact loss as follows:

$$\mathcal{D}_k(\mathcal{P}_i^s) = \min_j ||\mathcal{V}_j^t - \mathcal{P}_i^s||_2^2, \tag{8}$$

$$\mathcal{L}_{cont} = \lambda_4 \sum_k \sum_i f(\mathcal{D}_k(\mathcal{P}_i^s)) \ for\ all\ \mathcal{D}_k(\mathcal{P}_i^s) \leq \mathcal{T} \tag{9}$$

where $\mathcal{D}_k(\mathcal{P}_i^s)$ represents the unsigned distance between k-th hand vertex and i-th scene point. $\mathcal{V}_j^i$ is the j-th hand

vertex located in the touched part. $f(\mathcal{D}(\cdot)) = 1 - 2 \cdot \text{Sigmoid}((2\mathcal{D}(\cdot)) - 0.5)$ is a normalize function scaled the distance to $[0, 1]$. Threshold $\mathcal{T}$ is set to 5mm. The hand object contact loss encourages hand vertices to be as close as possible to contact points on objects.

To alleviate surface intersection between the scene and reconstructed hand mesh, we also propose a collision loss $\mathcal{L}_{coll}$ that explicitly learns collision-free grasps:

$$\mathcal{L}_{coll} = \lambda_5 \sum_k \sum_j - \min(\mathcal{D}_s(\mathcal{V}_j^k), 0), \tag{10}$$

where $\mathcal{D}_s(\mathcal{V}_j^k)$ is the signed distance between j-th hand vertex of k-th grasp and the scene. The collision objective is to penalize the negative sum of signed distances of the hand mesh to the scene.

During training, we sample N_o points from the surface of the complete object model for computing hand-object contact. To encourage the anthropomorphic hand to achieve realistic grasping, N_f points located in affordance areas are additionally sampled when calculating contact loss. Fig. 4(b) illustrates three losses for generating a hand mesh.

The total grasp loss is a sum of above loss functions $\mathcal{L}_{grasp} = \mathcal{L}_{seg} + \mathcal{L}_{rec} + \mathcal{L}_{cont} + \mathcal{L}_{coll}$.

VI. Experiments

In this section, we first introduce implementation details and evaluation metrics of our proposed method, then we conduct a series of experiments both in MuJoCo simulation and real-world platform with a UR5 robot arm and a HIT-DLR II anthropomorphic hand.

A. Implementation Details

For each scene, we randomly sample 40K points from observed single-view point cloud as input for network. The network is trained for 100 epochs with an Adam optimizer. The learning rate is set to 0.01 at the start and decreased by a factor of 2 for every 10 epochs with a batch size of 8. Hyper-parameters of $\mu, \lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, K, N_o, N_f$ are respectively set to 0.1, 10, 1, 10, 5, 1, 50, 8096 and 2048.

B. Evaluation Metrics

Four metrics regarding previous work on robot grasping [29] and hand grasp generation [12] are utilized: 1) **Interpenetration (Int.)** to measure the penetration between hand and scene; 2) **Success rate (SR)** to measure the quality of generated grasps through robotic experiments; 3) **Completion rate (CR)** to measure how many objects are successfully grasped in a scene and 4) **Time cost (TC)** to measure the time efficiency.

C. Simulation Experiments

We first select 18 objects absent in the training dataset to build cluttered scenes, then use a depth camera to capture observed images to generate single-view point cloud. We select two grasp configurations with the highest scores for each object to execute. Successful grasp attempts should lift the object over $20cm$ within 1000 simulation steps.

Grasp quality and efficiency. We compare the grasping performance with two baseline methods (GraspIt! [6] and HGC-Net [14]). Quantitative results is shown in Tab. II. Our method greatly surpasses the baseline in terms of grasping success rate and completion rate, but also inevitably increases the interpenetration between hand and scene to form a stable force closure. Compared with GraspIt!, an analytical method that requires complete object models and involves a considerable amount of time for extensive sampling and computation, our data-driven approach with one-stage network architecture design boosts grasp prediction and naturally reduces time cost. Compared with HGC-Net [14], our network focuses more on hand-object contacts with better grasping performance. Qualitative results of are shown in Fig. 5.

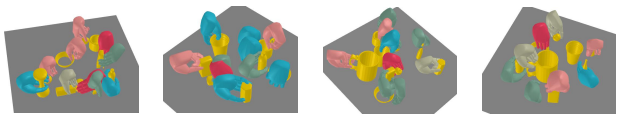


Fig. 5: Results for generated grasps of our proposed method.

TABLE II: Experiments results in Simulation.

Methods	Int.(cm^{-6})↓	SR(%)↑	CR (%)↑	TC (s)↓
GraspIt! [6]	**1.43**	54.7	51.8	40
HGC-Net [14]	6.64	65.9	73.8	**0.25**
Ours	9.26	**71.2**	**78.1**	0.27

Effectiveness of grasp generation. Tab. III demonstrates the effectiveness of proposed method. We compare the grasp performance on different grasp types. It shows that the larger contact area between hand and object can improve the stability of grasping. The grasp optimization (G.O.) module gains 11.8% and 9.7% improvements in terms of success rate and completion rate. In addition, a larger μ (μ=0.5) indicates the model focuses less on graspness, which lead to a decrease of grasping performance in SR and CR.

Performance of affordance grasping. Fig. 6 shows representative examples of affordance grasping, indicates the ability of our model to grasp the functional area of an object with clear usage (such as the handle of tools or mugs).

Fig. 6: Examples of affordance grasping in simulation.

D. Real-world Experiments

We capture the depth image with an Ensenso N35 camera mounted on the top of the robot. Over 27 novel objects absent in the training dataset with various shapes and sizes are selected to evaluate the effectiveness (Fig. 7(b)). The experiment settings are: 1) 5-15 Objects are randomly selected and placed within a $45cm \times 45cm$ square area to construct a cluttered scene; 2) The camera captures the scene from backward at a 60-degree viewpoint; 3) Execute the grasp with the highest score and use MoveIt! [30] for motion planning; 4) A grasp attempt is classified as a successful grasp if the robot can pick the object and place it to predefined home position without dropping; 5) The anthropomorphic

TABLE III: Grasping performance of five grasp types in simulation.

	Grasp Type	Parallel Extension	Pen Pinch	Palmar Pinch	Precision Sphere	Medium Wrap
w/o G.O. $\mu = 0.1$	Int. (cm^3)↓	6.40	**5.44**	5.52	7.06	6.85
	SR (%)↑	68.1	29.2	45.7	**79.6**	69.7
	CR (%)↑	78.7	31.6	57.8	**89.5**	82.6
with G.O. $\mu = 0.1$	Int. (cm^3)↓	10.0	8.28	9.74	9.81	**7.99**
	SR (%)↑	79.1	49.8	58.3	**85.5**	78.4
	CR (%)↑	85.3	57.8	67.6	**91.3**	86.7
with G.O. $\mu = 0.5$	Int. (cm^3)↓	11.1	8.49	9.08	8.33	**7.44**
	SR (%)↑	74.4	43.6	52.8	**80.3**	72.9
	CR (%)↑	78.2	49.5	58.7	**84.7**	79.4

(a) (b)

Fig. 7: (a) Robot system. (b) Representative objects for real-world experiments. Some objects are with clear usage properties (top), while some objects are easily deformed (Bottom).

hand attempts several grasps until all of objects are grasped or encounters three consecutive failures. The experiment is repeated 3 times.

TABLE IV: Grasping performance on real robot platform.

Methods	SR(%)	CR(%)	SR (FG) (%)	CR (FG) (%)
HGC-Net	57.8	63.2	21.9	24.8
Ours $\mu = 0.1$	**70.1**	**79.2**	29.6	32.3
Ours $\mu = 0.5$	61.5	65.3	**49.2**	**56.3**

Quantitative results. Tab. IV shows grasping results in terms of SR and CR of real-world experiments. We introduce SR(FG) and CR(FG) to evaluate success rate and completion rate of affordance grasping. In this case, only affordance grasping can be regarded as a successful grasping, which is determined by two users. Our methods trained in simulation generalizes well in real-world scenarios. The decrease in SR and CR as μ increases is intuitive, because the network excessively tend to generate grasps in the affordance area.

Qualitative results. Fig. 8 shows our method can grasp unseen household objects with various sizes and shapes. The third column shows affordance grasping (clamp, mug and knife) and the last column shows deformable objects grasping (milk carton and rubber toy), demonstrates the method can generate high-quality realistic and reasonable grasps.

Fig. 8: Visualization results of real-world experiments.

VII. Conclusion

In this work, we propose an anthropomorphic hand grasping method based on object affordance that can learn realistic and reasonable anthropomorphic hand grasps in cluttered scenes. Our model learns graspable points in terms of graspness and affordance grasping and hand configurations in a parallel manner, following with a grasp optimization module based on hand-object contact. To train the model, we build a large-scale grasp synthetic dataset in simulation for our HIT-DLR II hand. Simulation experiments demonstrate that our method is significantly superior to baseline methods in both grasping performance and efficiency. Real-world experiments also show that our method performs well for unseen in real-world scenarios, which is able to generate physical plausible and reasonable anthropomorphic hand grasps. In future work, we will (1) generalize our method to universal human-like grasping with the different robotic hands and (2) explore in-hand manipulation for stable object grasping and tool using.

References

[1] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[2] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," *arXiv preprint arXiv:2108.02425*, 2021.

[3] R. Pfeifer and F. Iida, "Embodied artificial intelligence: Trends and challenges," *Lecture notes in computer science*, pp. 1–26, 2004.

[4] V. Nguyen, "Constructing force-closure grasps," *The International Journal of Robotics Research (IJRR)*, 1988.

[5] C. Ferrari and J. F. Canny, "Planning optimal grasps," in *IEEE International Conference on Robotics and Automation (ICRA)*, 1992.

[6] A. T. Miller and P. K. Allen, "Graspit! A versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, 2004.

[7] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "Grab: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision (ECCV)*, 2020.

[8] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays, "Contactpose: A dataset of grasps with object contact and hand pose," in *European Conference on Computer Vision (ECCV)*, 2020.

[9] J. Lundell, E. Corona, T. N. Le, F. Verdoja, P. Weinzaepfel, G. Rogez, F. Moreno-Noguer, and V. Kyrki, "Multi-fingan: Generative coarse-to-fine sampling of multi-finger grasps," in *IEEE International conference on robotics and automation (ICRA)*, 2021.

[10] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[11] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Deep differentiable grasp planner for high-dof grippers," in *Robotics: Science and Systems (RSS)*, 2020.

[12] W. Wei, D. Li, P. Wang, Y. Li, W. Li, Y. Luo, and J. Zhong, "Dvgg: Deep variational grasp generation for dextrous manipulation," *IEEE Robotics and Automation Letters*, 2022.

[13] J. Lundell, F. Verdoja, and V. Kyrki, "Ddgc: Generative deep dexterous grasping in clutter," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6899–6906, 2021.

[14] Y. Li, W. Wei, D. Li, P. Wang, W. Li, and J. Zhong, "Hgc-net: Deep anthropomorphic hand grasping in clutter," in *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.

[15] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from rgb-d videos," *The International journal of robotics research*, vol. 32, no. 8, pp. 951–970, 2013.

[17] T. Zhu, R. Wu, X. Lin, and Y. Sun, "Toward human-like grasp: Dexterous grasping via semantic representation of object-hand," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 741–15 751.

[18] H. Duan, P. Wang, Y. Huang, G. Xu, W. Wei, and X. Shen, "Robotics dexterous grasping: The methods based on point cloud and deep learning," *Frontiers in Neurorobotics*, 2021.

[19] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[20] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, "6-dof grasping for target-driven object manipulation in clutter," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 6232–6238.

[21] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, "S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes." PMLR, 2020, pp. 53–65.

[22] W. Wei, Y. Luo, F. Li, G. Xu, J. Zhong, W. Li, and P. Wang, "Gpr: Grasp pose refinement network for cluttered scenes," *arXiv preprint arXiv:2105.08502*, 2021.

[23] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[24] C. Eppner, A. Mousavian, and D. Fox, "A billion ways to grasp: An evaluation of grasp sampling schemes on a dense, physics-based grasp data set," *arXiv preprint arXiv:1912.05604*, 2019.

[25] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[26] C. Sager, P. Zschech, and N. Kuhl, "labelCloud: A lightweight labeling tool for domain-agnostic 3d object detection in point clouds," *Computer-Aided Design and Applications*, vol. 19, no. 6, pp. 1191–1206, mar 2022. [Online]. Available: http://cad-journal.net/files/vol_19/CAD_19(6)_2022_1191-1206.pdf

[27] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, 2019.

[28] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *arXiv preprint arXiv:1706.02413*, 2017.

[29] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

[30] D. Coleman, I. Sucan, S. Chitta, and N. Correll, "Reducing the barrier to entry of complex robotic software: a moveit! case study," *Journal of Software Engineering in Robotics, Special issue on Best Practice in Robot Software Development*, 2014.