GenDexGrasp: Generalizable Dexterous Grasping

Puhao Li[†1,2], Tengyu Liu[†1], Yuyang Li[1,2], Yiran Geng[1,3], Yixin Zhu[3], Yaodong Yang[1,3], Siyuan Huang[‡1]

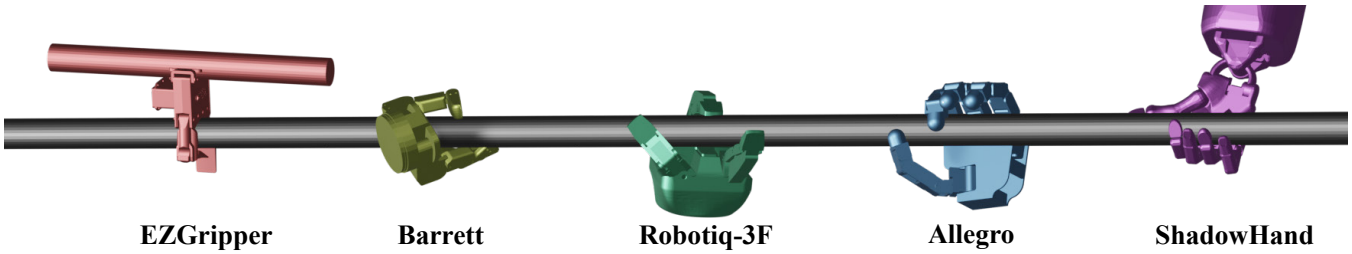

Fig. 1: **The proposed GenDexGrasp synthesizes and generalizes versatile dexterous grasps across arbitrary robot hands.**

***Abstract*— Generating dexterous grasping has been a long-standing and challenging robotic task. Despite recent progress, existing methods primarily suffer from two issues. First, most prior arts focus on a specific type of robot hand, lacking *generalizable* capability of handling unseen ones. Second, prior arts oftentimes fail to *rapidly* generate *diverse* grasps with a *high success rate*. To jointly tackle these challenges with a unified solution, we propose GenDexGrasp, a novel hand-agnostic grasping algorithm for generalizable grasping. GenDexGrasp is trained on our proposed large-scale multi-hand grasping dataset MultiDex synthesized with force closure optimization. By leveraging the contact map as a hand-agnostic intermediate representation, GenDexGrasp efficiently generates diverse and plausible grasping poses with a high success rate and can transfer among diverse multi-fingered robotic hands. Compared with previous methods, GenDexGrasp achieves a three-way trade-off among success rate, inference speed, and diversity.**

I. Introduction

Humans' ability to grasp is astonishingly versatile. In addition to the full grasp with five fingers, humans can efficiently *generalize* grasps with two or three fingers when some fingers are occupied and imagine *diverse* grasping poses given a new type of hand we have never seen, all happened *rapidly* with *a high success rate*. These criteria are in stark contrast to most prior robot grasping methods, which primarily focus on specific end-effectors, requiring redundant efforts to learn the grasp model for every new robotic hand. On top of this challenge, prior methods often have difficulties generating diverse hand poses for unseen scenarios in a short period, further widening the gap between robot and human capabilities. Hence, these deficiencies necessitate a generalizable grasping algorithm, efficiently handling arbitrary hands and allowing fast prototyping for new robots.

Fundamentally, the most significant challenge in generalizable dexterous grasping [1–7] is to find an *efficient* and *transferable* representation for diverse grasp. The *de facto* representation, joint angles, is unsuitable for its dependency on the structure definition: two similar robotic hands could have contrasting joint angles if their joints are defined differently. Existing works use contact points [8–10], contact maps [11, 12], and approach vectors [13] as the representations, and execute the desired grasps with complex solvers. A simple yet effective representation is still in need.

In this paper, we denote **generalizable dexterous grasping** as the problem of generating grasping poses for unseen hands. We evaluate generalizable grasping in three aspects:

- **Speed:** Hand-agnostic methods adopt inefficient sampling strategies [8, 11, 12], which leads to extremely slow grasp generation, ranging from 5 minutes to 40 minutes.
- **Diversity:** Hand-aware methods [9, 10, 13] rely on deterministic solvers, either as a policy for direct execution or predicted contact points for inverse kinematics, resulting in identical grasping poses for the same object-hand pair.
- **Generalizability:** Hand-aware methods [9, 10, 13] also rely on hand descriptors trained on two- and three-finger robotic hands, which hinders their generalizability to new hands that are drastically different from the trained ones.

To achieve a three-way trade-off among the above aspects and alleviate the aforementioned issues, we devise GenDexGrasp for generalizable dexterous grasping. Inspired by Brahmbhatt *et al*. [11], we first generate a hand-agnostic contact map for the given object using a conditional variational autoencoder [14]. Next, we optimize the hand pose to match the generated contact map. Finally, the grasping pose is further refined in a physics simulation to ensure a physically plausible contact. GenDexGrasp provides generalizability by reducing assumptions about hand structures and achieves fast inference with an improved contact map and an efficient optimization scheme, resulting in diverse grasp generation by a variational generative model with random initialization.

To address contact ambiguities (especially for thin-shell objects) during grasp optimization, we devise an aligned distance to compute the distance between surface point and hand, which helps to represent accurate contact maps for grasp generation. Specifically, the traditional Euclidean distance would mistakenly label both sides of a thin shell as contact points when the contact is on one side, whereas the aligned distance considers directional alignment to the surface normal of the contact point and rectifies the errors.

To learn the hand-agnostic contact maps, we collect a large-scale multi-hand dataset, MultiDex, using force closure optimization [8]. MultiDex contains 436,000 diverse grasping poses for 5 hands and 58 household objects.

† Puhao Li and Tengyu Liu contributed equally to this paper.
‡ Corresponding email: syhuang@bigai.ai.
[1] Beijing Institute of General Artificial Intelligence (BIGAI).
[2] Tsinghua University. [3] Peking University.
Website: https://github.com/tengyu-liu/GenDexGrasp.

We summarize our contributions as follows:

1) We propose GenDexGrasp, a versatile generalizable grasping algorithm. GenDexGrasp achieves a three-way trade-off among speed, diversity, and generalizability to unseen hands. In experiments, we demonstrate that GenDexGrasp is significantly faster than existing hand-agnostic methods and generates more diversified grasping poses than prior hand-aware methods. Our method also achieves strong generalizability, comparable to existing hand-agnostic methods.
2) We devise an aligned distance for properly measuring the distance between the object's surface point and hand. We represent a contact map with the aligned distance, which significantly increases the grasp success rate, especially for thin-shell objects. The ablation analysis in Tab. II shows the efficacy of such a design.
3) We collect and open-source a large-scale synthetic dataset, MultiDex, for generalizable grasping with 5 robotic hands, 58 household objects, and 436,000 diverse grasping poses. MultiDex is by far the largest multi-hand grasp dataset with diverse hand structures.

II. Related Work

A. Generalizable Dexterous Grasping

Existing solutions to generalizable grasping fall into two categories: hand-aware and hand-agnostic. The hand-aware methods are limited by the diversity of generated poses, whereas the hand-agnostic methods are oftentimes too slow for various tasks. Below, we review both methods in detail.

Hand-aware approaches [9, 10, 13] learn a data-driven representation of the hand structure and use a neural network to predict an intermediate goal, which is further used to generate the final grasp. For instance, UniGrasp [9] and EfficientGrasp [10] extract the gripper's PointNet [15] features in various poses and use a PSSN network to predict the contact points of the desired grasp. As a result, contact points are used as the inverse kinematics's goal, which generates the grasping pose. Similarly, AdaGrasp [13] adopts 3D convolutional neural networks to extract gripper features, ranks all possible poses from which the gripper should approach the object, and executes the best grasp with a planner. However, all hand-aware methods train and evaluate the gripper encoders only with two- and three-finger grippers, hindering their ability to generalize to unseen grippers or handle unseen scenarios. Critically, these methods solve the final grasp deterministically, yielding similar grasping poses.

Hand-agnostic methods rely on carefully designed sampling strategies [8, 11, 12]. For instance, ContactGrasp [11] leverages the classic grasp planner in *GraspIt!* [16] to match a selected contact map, and Liu *et al*. [8] and Turpin *et al*. [12] sample hand-centric contact points/forces and update the hand pose to minimize the difference between desired contacts and actual ones. All these methods adopt stochastic sampling strategies that are extremely slow to overcome the local minima in the landscape of objective functions. As a result, existing hand-agnostic methods take minutes to generate a new grasp, impractical for real-world applications.

B. Contact Map

Contact map has been an essential component in modern grasp generation and reconstruction. Initialized by GraspIt! [16] and optimized by DART [17], ContactGrasp [11] uses thumb-aligned contact maps from ContactDB [18] to retarget grasps to different hands. ContactOpt [19, 20] uses estimated contact map to improve hand-object interaction reconstruction. NeuralGrasp [21] retrieves grasping poses by finding the nearest neighbors in the latent space projections of contact maps. Wu *et al*. [7] samples contact points on object surfaces and uses inverse kinematics to solve the grasping pose. Mandikal *et al*. [22] treats contact maps as object affordance and learns an RL policy that manipulates the object based on the contact maps. DFC [8] simultaneously updates hand-centric contact points and hand poses to sample diverse and physically stable grasping from a manually designed Gibbs distribution. GraspCVAE [4] and Grasp'D [12] use contact maps to improve grasp synthesis: GraspCVAE generates a grasping pose and refines the pose w.r.t. an estimated contact map, whereas Grasp'D generates and refines the expected contact forces while updating the grasping pose. IBS-Grasp [23] learns a grasping policy that takes an interaction bisector surface, a generalized contact map, as the observed state. Compared to prior methods, the proposed GenDexGrasp differs by treating the contact map as the *transferable* and *intermediate* representation for *hand-agnostic* grasping. We use a less restrictive contact map and a more efficient optimization method for faster and more diversified grasp generation; see detailed in Sec. IV-A.

C. Grasp Datasets

3D dexterous grasping poses are notoriously expensive to collect due to the complexity of hand structures. The industrial standard method of collecting a grasping pose is through kinesthetic demonstration [24], wherein a human operator manually moves a physical robot towards a grasping pose. While researchers could collect high-quality demonstrations with kinesthetic demonstrations, it is considered too expensive for large-scale datasets. To tackle this challenge, researchers devised various low-cost data collection methods.

The straightforward idea is to replace kinesthetic demonstration with a motion capture system. Recent works have leveraged optical [25–27] and visual [20, 28–30] MoCap systems to collect human demonstrations. Another stream of work collects the contact map on objects by capturing the heat residual on the object surfaces after each human demonstration and using the contact map as a proxy for physical grasping hand pose [18, 20]. Despite the differences in data collection pipelines, these prior arts collect human demonstrations within a limited setting, between pick-up and use. Such settings fail to cover the long-tail and complex nature of human grasping poses as depicted in the grasping taxonomy [31] and grasp landscape [8]. As a result, the collected grasping poses are similar to each other and can be represented by a few principal components [32, 33]. We observe the same problem in programmatically generated datasets [34–38] using GraspIt! [16].

Fig. 2: **Exemplar grasps of different hands and objects from the proposed synthesized dataset.** From top to bottom: EZGripper, Barrett, Robotiq-3F, Allegro, and ShadowHand. From left to right: alarm clock, apple, binocular, and meat can.

III. Dataset Collection

To learn a versatile and hand-agnostic contact map generator, the grasp dataset ought to contain diverse grasping poses and corresponding contact maps for different objects and robotic hands with various morphologies.

A. Grasp Pose Synthesis

Inspired by Liu *et al.* [8], we synthesized a large-scale grasping dataset by minimizing a differentiable force closure estimator DFC, a hand prior energy E_{n}, and a penetration energy E_{p}. We use the qpos q_H to represent the kinematics pose of a robotic hand H, denoted as

$$q_H = \{q_{\text{global}} \in \mathbb{R}^6, q_{\text{joint}} \in \mathbb{R}^N\}, \tag{1}$$

where q_{global} is the rotation and translation of the root link, and q_{joint} describes the rotation angles of the revolute joints.

We selected 58 daily objects from the YCB dataset [39] and ContactDB [18], together with 5 robotic hands (EZGripper, Barrett Hand, Robotiq-3F, Allegro, and Shadowhand) ranging from two to five fingers. We split our dataset into 48 training objects and 10 test objects. We show a random subset of the collected dataset in Fig. 2.

Given an object O, a kinematics model of a robotic hand H with pose q_H and surface $\mathcal{H}$, and a group of n hand-centric contact points $X \subset \mathcal{H}$, we define the differentiable force closure estimator DFC as:

$$\text{DFC} = Gc, \tag{2}$$

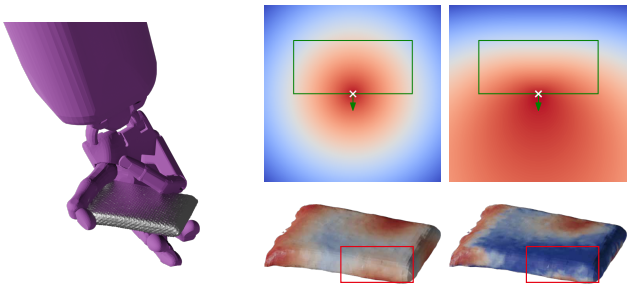

Fig. 3: **Comparison between aligned and euclidean distances on thin shell objects.** Given an exemplar grasp (left), we show both distances from the object to hand surfaces in 3D; red regions denote shorter distances and blue longer. An illustration of both distances is also shown in 2D (top middle and top right); the green rectangle, white cross, and green arrow represent a rectangular object, the point of interest, and the surface normal n_o at the point, respectively. The Euclidean distance (top middle) labels the upper edge of the object as close to the point of interest, whereas the aligned distance (top right) is geometry-aware. The 3D aligned distances of the exemplar grasp (bottom right) correctly reflect non-contact areas in the highlighted area, where the finger contacts the opposite side of the thin object. The Euclidean distances fail to distinguish contacts on one side from contacts on the other side.

where $c \in \mathbb{R}^{3n\times 1}$ is the object surface normal on the contact points X, and

$$G = \begin{bmatrix} I_{3\times3} & I_{3\times3} & \dots & I_{3\times3} \\ \lfloor x_1 \rfloor_\times & \lfloor x_2 \rfloor_\times & \dots & \lfloor x_n \rfloor_\times \end{bmatrix}, \tag{3}$$

$$\lfloor x_i \rfloor_\times = \begin{bmatrix} 0 & -x_i^{(3)} & x_i^{(2)} \\ x_i^{(3)} & 0 & -x_i^{(1)} \\ -x_i^{(2)} & x_i^{(1)} & 0 \end{bmatrix}. \tag{4}$$

DFC describes the total wrench when each contact point applies equal forces, and friction forces are neglectable. As established in Liu *et al.* [8], DFC is a strong estimator of the classical force closure metric.

Next, we define the prior and penetration energy as

$$E_{\text{p}}(q_H, O) = \sum_{x\in\mathcal{H}} \text{R}(-\delta(x, O)) \tag{5}$$

$$E_{\text{n}}(q_H) = \|\text{R}(q_H - q_{H\uparrow}) + \text{R}(q_{H\downarrow} - q_H)\|_2, \tag{6}$$

where $q_{H\uparrow}$ and $q_{H\downarrow}$ are the upper and lower limits of the robotic hand parameters, respectively. $\delta(x, O)$ gives the signed distance from x to O, where the distance is positive if x is outside O and is negative if inside.

Generating valid grasps requires finding the optimal set of contact points $X \subset \mathcal{H}$ that minimize $E = \text{DFC} + E_{\text{n}} + E_{\text{p}}$. For computational efficiency, we sample $X \subset \mathcal{H}$ from a set of rectangular contact regions predefined for each robotic hand. This strategy allows us to update the contact point positions via a gradient-based optimizer and improve sample efficiency. We use the DeepSDF [40, 41] to approximate the signed distance and surface normal of an object.

We use a Metropolis-adjusted Langevin algorithm (MALA) [8] to simultaneously sample the grasping poses and contact points. We run the MALAalgorithm on an NVIDIA A100 80GB with a batch size of 1024 for each hand-object pair and obtain 436,000 valid grasping poses. It takes about 1,400 GPU hours to synthesize the entire dataset.

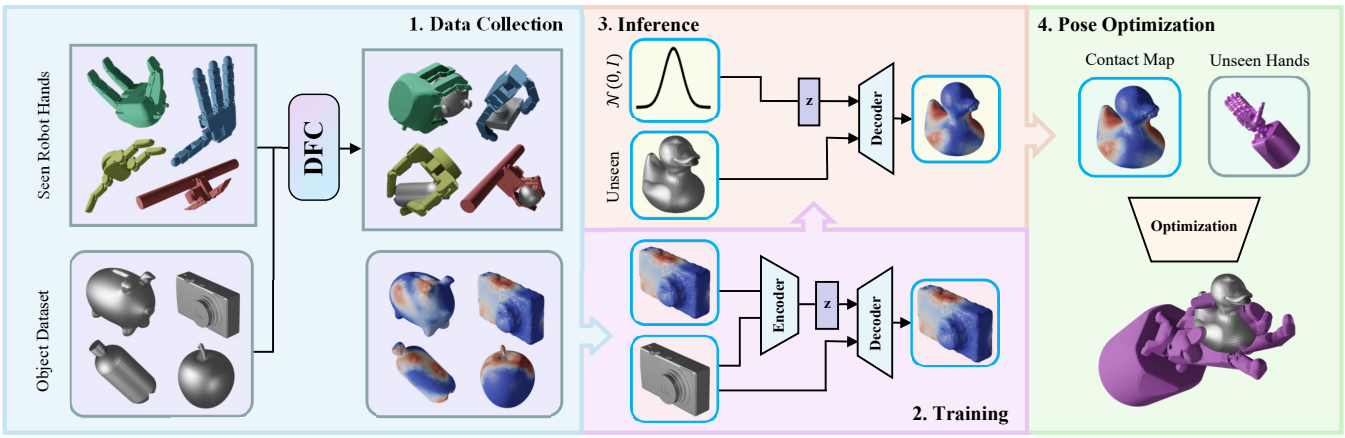

Fig. 4: **An overview of the GenDexGrasp pipeline.** We first collect a large-scale synthetic dataset for multiple hands with DFC. Then, we train a CVAE to generate hand-agnostic contact maps for unseen objects. We finally optimize grasping poses for unseen hands using the generated contact maps.

B. Contact Map Synthesis

Given the grasping poses, we first compute the object-centric contact map Ω as a set of normalized distances from each object surface point to the hand surface. Instead of using Euclidean distance, we propose an aligned distance to measure the distance between the object's surface point and the hand surface. Given the object O and the hand H with optimized grasp pose q_H, we define $\mathcal{O}$ as the surface of O and $\mathcal{H}$ as the surface of H. The aligned distance $\mathcal{D}$ between an object surface point $v_o \in \mathcal{O}$ and $\mathcal{H}$ is defined as:

$$\mathcal{D}(v_o, \mathcal{H}) = \min_{v_h \in \mathcal{H}} e^{\gamma(1-\langle v_o - v_h, n_o \rangle)} \sqrt{\|v_o - v_h\|_2}, \quad (7)$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product of two normalized vectors, and n_o denotes the object surface normal at v_o. γ is a scaling factor; we empirically set it to 1. The aligned distance considers directional alignment with the object's surface normal on the contact point and reduces contact ambiguities on thin-shell objects. Fig. 3 shows that our aligned distance correctly distinguishes contacts from different sides of a thin shell, whereas the Euclidean distance mistakenly labels both sides as contact regions.

Next, we compute the contact value $\mathcal{C}(v_o, \mathcal{H})$ on each object surface point v_o following Jiang *et al.* [4]:

$$\mathcal{C}(v_o, \mathcal{H}) = 1 - 2\Big(\text{Sigmoid}\big(\mathcal{D}(v_o, \mathcal{H})\big) - 0.5\Big), \quad (8)$$

where $\mathcal{C}(v_o, \mathcal{H}) \in (0, 1]$ is 1 if v_o is in contact with $\mathcal{H}$, and is 0 if it is far away. $\mathcal{C} \leqslant 1$ since $\mathcal{D}$ is non-negative.

Finally, we define the contact map $\Omega(\mathcal{O}, \mathcal{H})$ as

$$\Omega(\mathcal{O}, \mathcal{H}) = \{\mathcal{C}(v_o, \mathcal{H})\}_{v_o \in \mathcal{O}}. \quad (9)$$

IV. GenDexGrasp

Given an object O and the kinematics model of an arbitrary robotic hand H with N joints, our goal is to generate a dexterous, diverse, and physically stable grasp pose q_H.

A. Generate Hand-Agnostic Contact Maps

Generating q_H directly for unseen H is challenging due to the sparsity of the observed hands and the non-linearity between q_H and hand geometry. Inspired by Brahmbhatt *et al.* [11], we adopt the object-centric contact map as a hand-agnostic intermediate representation of a grasp. Instead of directly generating q_H, we first learn a generative model that generates a contact map over the object surface. We then fit the hand to the generated map.

We adopt CVAE [14] to generate the hand-agnostic contact map. Given the point cloud of an input object and the corresponding pointwise contact values $\mathcal{C}$, we use a PointNet [15] encoder to extract the latent distribution $\mathcal{N}(\mu, \sigma)$ and sample the latent code $z \sim \mathcal{N}(\mu, \sigma)$. When decoding, we extract the object point features with another PointNet, concatenate z to the per-point features, and use a shared-weight MLP to generate a contact value $\hat{\mathcal{C}}(v_o)$ for each $v_o \in \mathcal{O}$, which forms the predicted contact map $\hat{\Omega}(\mathcal{O}) = \{\hat{\mathcal{C}}(v_o)\}_{v_o \in \mathcal{O}}$.

We learn the generative model by maximizing the log-likelihood of $p_{\theta,\varphi}(\Omega \mid O)$, where θ and ϕ are the learnable parameters of the encoder and decoder, respectively. According to Sohn *et al.* [14], we equivalently maximize the ELBO:

$$\begin{aligned} \log p_{\theta,\varphi}(\Omega \mid O) \geqslant & \mathbb{E}_{z \sim Z}[\log p_\varphi(\Omega \mid z, O)] \\ & - D_{KL}(p_\theta(z \mid \Omega, O) \,||\, p_Z(z)), \end{aligned} \quad (10)$$

where Z is the prior distribution of the latent space; we treat Z as the standard normal distribution $\mathcal{N}(0, I)$.

We leverage a reconstruction loss to approximate the expectation term of ELBO:

$$\mathbb{E}_{z \sim Z}[\log p_\varphi(\Omega \mid z, O)] = \frac{1}{N_o} \sum_{i=0}^{N_o - 1} \|\hat{\Omega}^i - \Omega^i\|_2, \quad (11)$$

where N_o is the number of examples. Ω^i and $\hat{\Omega}^i$ denote the expected and generated contact map of the i-th example, respectively.

Of note, since the generated contact map is empirically more ambiguous than the ground-truth contact map, we sharpen the generated contact map with

$$\hat{\hat{\Omega}} = \begin{cases} \hat{\Omega} & \text{if } \hat{\Omega} < 0.5 \\ 1 & \text{otherwise.} \end{cases} \quad (12)$$

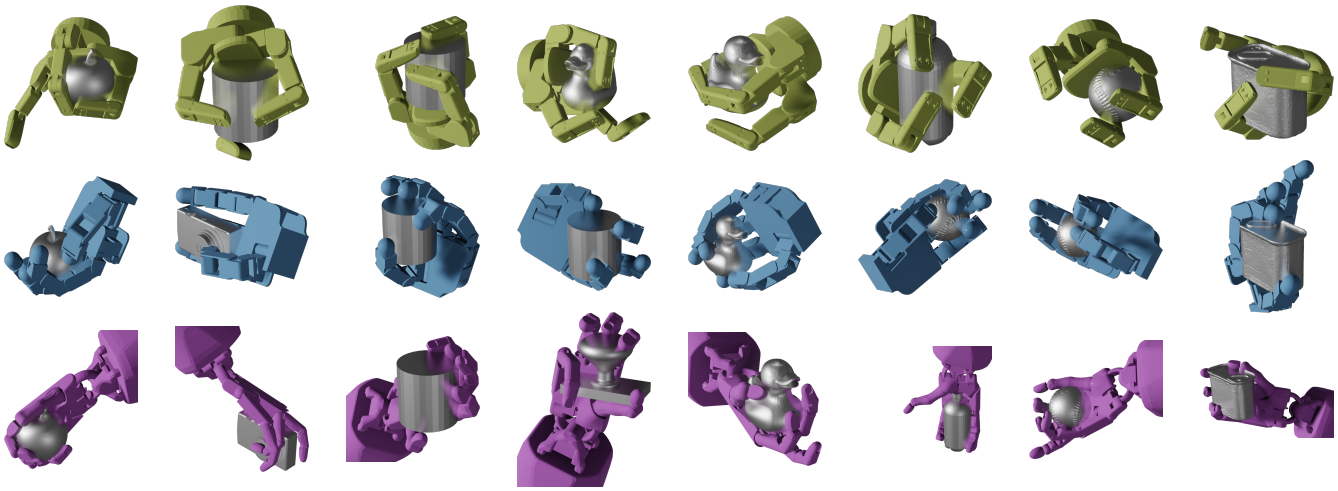

Fig. 5: **Examples of the generated grasping poses for unseen hands and objects.** From top to bottom: Barrett, Allegro, and ShadowHand.

B. Grasp Optimization

Given the generated contact map $\hat{\hat{\Omega}}$ on object O, we optimize the grasping pose q_H for hand H. We initialize the optimization by randomly rotating the root link of the hand and translating the hand towards the back of its palm direction. We set the translation distance to the radius of the minimum enclosing sphere of the object.

We compute $\mathcal{H}$ by differentiable forward kinematics and obtain the current contact map $\dot{\Omega}$. We compute the optimization objective E as

$$E(q_H, \hat{\hat{\Omega}}, O) = E_{\text{c}}(q_H, \hat{\hat{\Omega}}) + E_{\text{p}}(q_H, O) + E_{\text{n}}(q_H), \quad (13)$$

where E_{c} is the MSE between the goal contact map $\hat{\hat{\Omega}}$ and the current contact map $\dot{\Omega}$. E_{p} and E_{n} describe the penetration between hand and object and if the hand pose is valid, respectively, described in Eqs. (5) and (6).

Since the computation of the objective function is fully differentiable, we use the Adam optimizer to minimize E by updating q_H. We run a batch of 32 parallel optimizations to keep the best result to avoid bad local minima.

C. Implementation Details

We optimize the CVAE for hand-agnostic contact maps using the Adam optimizer with a learning rate of $1\text{e}{-4}$. Other Adam hyperparameters are left at default values. We train the CVAE for 36 epochs, which takes roughly 20 minutes on an NVIDIA 3090Ti GPU. The grasp optimizer Adam uses a learning rate of $5\text{e}{-3}$.

V. Experiment

We quantitatively evaluate GenDexGrasp in terms of success rate, diversity, and inference speed.

Success Rate: We test if a grasp is successful in the Isaac Gym environment [42] by applying an external acceleration to the object and measuring the movement of the object. We test each grasp by applying a consistent 0.5ms^{-2} acceleration at the object for 1 second or 60 simulation steps and evaluate if the object moves more than 2cm after the simulation. We repeat this process for each grasp six times with acceleration along $\pm xyz$ directions. A grasp fails if it fails one of the six tests. Since generative methods usually exhibit minor errors that result in floatation and penetration near contact points, we apply a contact-aware refinement to the generated examples of all compared methods. Specifically, we first construct a target pose by moving the links close enough to the object (within 5mm) towards the object's direction. Next, we update q_H with one step of gradient descent of step size 0.01 to minimize the difference between the current and the target pose. Finally, we track the updated pose with a positional controller provided by the Isaac Gym.

Diversity: We measure the diversity of the generated grasps as the standard deviation of the joint angles of the generated grasps that pass the simulation test.

Inference Speed: We measure the time it takes for the entire inference pipeline to run.

TABLE I: Comparative Experiments

Methods	**Gen.**	**Succ.**(%)	**Div.**(rad.)	**Speed**(sec.)
DFC [8]	✓	**79.53**	**0.344**	>1,800
GC (w/o TTA) [4]	✗	19.38	**0.340**	**0.012**
GC (w/ TTA) [4]	✗	22.03	**0.355**	43.233
UniG.(top-1) [9]	✓	**80.00**	0.000	9.331
UniG.(top-8) [9]	✓	50.00	0.167	9.331
UniG.(top-32) [9]	✓	48.44	0.202	9.331
Ours	✓	**77.19**	0.207	16.415

We compare GenDexGrasp with DFC [8], GraspCVAE [4] (GC), and UniGrasp [9] (UniG.) in Tab. I. The columns represent method names, whether the method is generalizable, success rate, diversity, and inference speed. We evaluate all methods with the test split of the ShadowHand data in MultiDex. We trained our method with the training split of EZGripper, Robotiq-3F, Barrett, and Allegro. Since GraspCVAE is designed for one specific hand structure, we train GraspCVAE on the training split of the ShadowHand data and keep the result before and after test-time adaptation (TTA). We evaluate UniGrasp with its pretrained weights.

Of note, since the UniGrasp model only produces three contact points, we align them to the thumb, index, and middle finger of the ShadowHand for inverse kinematics. In addition,

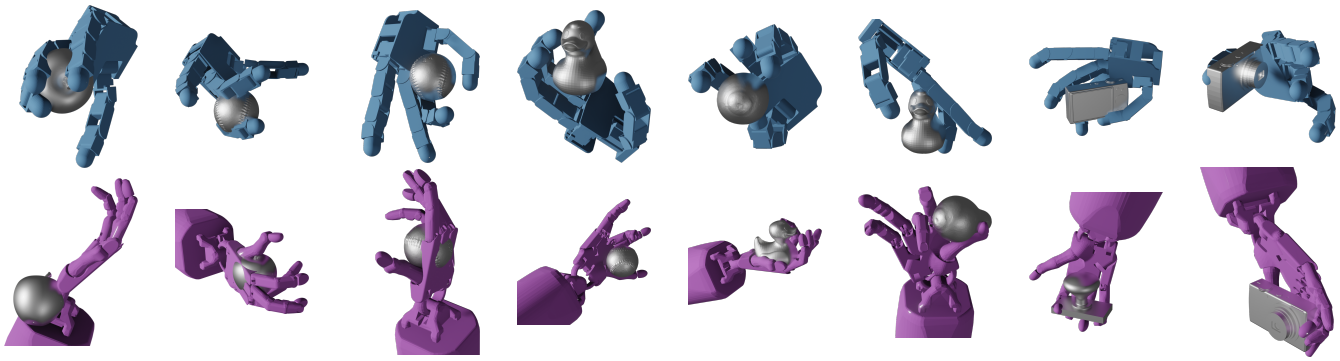

Fig. 6: **Failure cases with Allegro (top) and ShadowHand (bottom).** The last two columns show artifacts caused by contact ambiguities when using Euclidean distances instead of aligned distances.

UniGrasp yields zero diversity since it produces the top-1 contact point selection for each object. We include top-8, top-32, and top-64 contact point selections to evaluate its diversity. We observe that DFC achieves the best success rate and diversity but is overwhelmingly slow. GraspCVAE can generate diverse grasping poses but suffers from a low success rate and cannot generalize to unseen hands. We attribute the low success rate to our dataset's large diversity of grasping poses. The original GraspCVAE was trained on HO3D [28], where grasp poses are similar since six principal components can summarize most grasping poses. UniGrasp can generalize to unseen hands and achieve a high success rate. However, it fails to balance success rate and diversity.

Our method achieves a slightly lower success rate than DFC and UniGrasp top-1 but can generate diverse grasping poses in a short period of time, achieving a good three-way trade-off among quality, diversity, and speed.

We examine the efficacy of the proposed aligned distance in Tab. II. Specifically, we evaluate the success rate and diversity of the full model (full) and the full model with Euclidean distance contact maps (-align). The experiment is repeated on EZGripper, Barrett, and ShadowHand to show efficacy across hands. In all three cases, we observe that using the Euclidean distance lowers the success rate significantly while improving the diversity slightly. Such differences meet our expectations, as contact maps based on Euclidean distances are more ambiguous than those based on aligned distances. During the evaluation, such ambiguities bring more uncertainties, which are treated as diversities using our current metrics. We also observe that the model performs worse on the EZGripper due to the ambiguities in aligning two-finger grippers to multi-finger contact maps.

TABLE II: Ablation Study - Contact

Methods	**Succ. Rate**(%)	**Diversity**(rad.)
Full (EZGripper)	**38.59**	0.248
-align (EZGripper)	29.53	**0.312**
Full (Barrett)	**70.31**	0.267
-align (Barrett)	52.19	**0.349**
Full (ShadowHand)	**77.19**	0.207
-align (ShadowHand)	58.91	**0.237**

We further compare the performances of GenDexGrasp on seen and unseen hands in Tab. III. We train two versions of GenDexGrasp for each hand. The in-domain version is trained on all five hands and evaluated on the selected hand. The out-of-domain version is trained on all four hands except the selected hand and evaluated on the selected hand. Our result shows that our method is robust in out-of-domain scenarios for various hand structures.

TABLE III: Ablation Study - Generalization

Robots	**Domain**	**Succ. Rate**(%)	**Diversity**(rad.)
Ezgripper	in	**43.44**	0.238
Ezgripper	out	38.59	**0.248**
Barrett	in	**71.72**	**0.281**
Barrett	out	70.31	0.267
Shadowhand	in	77.03	**0.211**
Shadowhand	out	**77.19**	0.207

The qualitative results in Fig. 5 show the diversity and quality of grasps generated by GenDexGrasp. The generated grasps cover diverse grasping types that include wraps, pinches, tripods, quadpods, hooks, *etc*. We also show failure cases in Fig. 6, where the first six columns show failures from our full model, and the last two columns show failures specific to the *-align* ablation version. The most common failure types are penetrations and floatations caused by imperfect optimization. We observe an interesting failure case in the first example in the bottom row, where the algorithm tries to grasp the apple by squeezing it between the palm and the base. While the example fails to pass the simulation test, it shows the level of diversity that our method provides.

VI. Conclusion

This paper introduces GenDexGrasp, a versatile dexterous grasping method that can generalize to unseen hands. By leveraging the contact map representation as the intermediate representation, a novel aligned distance for measuring hand-to-point distance, and a novel grasping algorithm, GenDexGrasp can generate diverse and high-quality grasping poses in reasonable inference time. The quantitative experiment suggests that our method is the first generalizable grasping algorithm to properly balance among quality, diversity, and speed. In addition, we contribute MultiDex, a large-scale synthetic dexterous grasping dataset. MultiDex features diverse grasping poses, a wide range of household objects, and five robotic hands with diverse kinematic structures.

REFERENCES

[1] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto, "Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation," *arXiv preprint arXiv:2203.13251*, 2022.

[2] A. Nagabandi, K. Konolige, S. Levine, and V. Kumar, "Deep dynamics models for learning dexterous manipulation," in *Conference on Robot Learning (CoRL)*, 2020.

[3] P. Mandikal and K. Grauman, "Dexvip: Learning dexterous grasping with human hand pose priors from video," in *Conference on Robot Learning (CoRL)*, 2022.

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

[5] I. Radosavovic, X. Wang, L. Pinto, and J. Malik, "State-only imitation learning for dexterous manipulation," in *International Conference on Intelligent Robots and Systems (IROS)*, 2021.

[6] M. Kokic, D. Kragic, and J. Bohg, "Learning task-oriented grasping from human activity datasets," *IEEE Robotics and Automation Letters (RA-L)*, vol. 5, no. 2, pp. 3352–3359, 2020.

[7] A. Wu, M. Guo, and C. K. Liu, "Learning diverse and physically feasible dexterous grasps with generative model and bilevel optimization," *arXiv preprint arXiv:2207.00195*, 2022.

[8] T. Liu, Z. Liu, Z. Jiao, Y. Zhu, and S.-C. Zhu, "Synthesizing diverse and physically stable grasps with arbitrary hand structures using differentiable force closure estimator," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 1, pp. 470–477, 2021.

[9] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg, "Unigrasp: Learning a unified model to grasp with multifingered robotic hands," *IEEE Robotics and Automation Letters (RA-L)*, 2020.

[10] K. Li, N. Baron, X. Zhang, and N. Rojas, "Efficientgrasp: A unified data-efficient learning to grasp method for multi-fingered robot hands," *IEEE Robotics and Automation Letters (RA-L)*, vol. 7, no. 4, pp. 8619–8626, 2022.

[11] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *International Conference on Intelligent Robots and Systems (IROS)*, 2019.

[12] D. Turpin, L. Wang, E. Heiden, Y.-C. Chen, M. Macklin, S. Tsogkas, S. Dickinson, and A. Garg, "Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands," in *European Conference on Computer Vision (ECCV)*, 2022.

[13] Z. Xu, B. Qi, S. Agrawal, and S. Song, "Adagrasp: Learning an adaptive gripper-aware grasping policy," in *International Conference on Robotics and Automation (ICRA)*, 2021.

[14] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[15] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *IEEE Robotics & Automation Magazine*, vol. 11, no. 4, pp. 110–122, 2004.

[17] T. Schmidt, R. A. Newcombe, and D. Fox, "Dart: Dense articulated real-time tracking." in *Robotics: Science and Systems (RSS)*, 2014.

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

[19] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[20] 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.

[21] N. Khargonkar, N. Song, Z. Xu, B. Prabhakaran, and Y. Xiang, "Neuralgrasps: Learning implicit representations for grasps of multiple robotic hands," *arXiv preprint arXiv:2207.02959*, 2022.

[22] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *International Conference on Robotics and Automation (ICRA)*, 2021.

[23] Q. She, R. Hu, J. Xu, M. Liu, K. Xu, and H. Huang, "Learning high-dof reaching-and-grasping via dynamic representation of gripper-object interaction," *arXiv preprint arXiv:2204.13998*, 2022.

[24] T. Eiband and D. Lee, "Identification of common force-based robot skills from the human and robot perspective," in *International Conference on Humanoid Robots (Humanoids)*, 2021.

[25] 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.

[26] O. Taheri, V. Choutas, M. J. Black, and D. Tzionas, "Goal: Generating 4d whole-body motion for hand-object grasping," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[27] Z. Fan, O. Taheri, D. Tzionas, M. Kocabas, M. Kaufmann, M. J. Black, and O. Hilliges, "Articulated objects in free-form hand interaction," *arXiv preprint arXiv:2204.13662*, 2022.

[28] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[29] S. Hampali, S. D. Sarkar, M. Rad, and V. Lepetit, "Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[30] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[31] 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.

[32] J. Romero, D. Tzionas, and M. J. Black, "Embodied hands: modeling and capturing hands and bodies together," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 6, pp. 1–17, 2017.

[33] M. Ciocarlie, C. Goldfeder, and P. Allen, "Dimensionality reduction for hand-independent dexterous robotic grasping," in *International Conference on Intelligent Robots and Systems (IROS)*, 2007.

[34] C. Goldfeder, M. Ciocarlie, H. Dang, and P. K. Allen, "The columbia grasp database," in *International Conference on Robotics and Automation (ICRA)*, 2009.

[35] 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 *International Conference on Robotics and Automation (ICRA)*, 2021.

[36] J. Lundell, F. Verdoja, and V. Kyrki, "Ddhc: Generative deep dexterous grasping in clutter," *IEEE Robotics and Automation Letters (RA-L)*, vol. 6, no. 4, pp. 6899–6906, 2021.

[37] Y. Hasson, G. Varol, D. Tzionas, I. Kalevatykh, M. J. Black, I. Laptev, and C. Schmid, "Learning joint reconstruction of hands and manipulated objects," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[38] 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.

[39] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *International Journal of Robotics Research (IJRR)*, vol. 36, no. 3, pp. 261–268, 2017.

[40] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] T. Davies, D. Nowrouzezahrai, and A. Jacobson, "Overfit neural networks as a compact shape representation," *arXiv preprint arXiv:2009.09808*, 2020.

[42] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa *et al.*, "Isaac gym: High performance gpu-based physics simulation for robot learning," *arXiv preprint arXiv:2108.10470*, 2021.