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

GrainGrasp: Dexterous Grasp Generation with Fine-grained Contact Guidance

Fuqiang Zhao[1], Dzmitry Tsetserukou[2] and Qian Liu[1]*

***Abstract*— One goal of dexterous robotic grasping is to allow robots to handle objects with the same level of flexibility and adaptability as humans. However, it remains a challenging task to generate an optimal grasping strategy for dexterous hands, especially when it comes to delicate manipulation and accurate adjustment the desired grasping poses for objects of varying shapes and sizes. In this paper, we propose a novel dexterous grasp generation scheme called *GrainGrasp* that provides fine-grained contact guidance for each fingertip. In particular, we employ a generative model to predict separate contact maps for each fingertip on the object point cloud, effectively capturing the specifics of finger-object interactions. In addition, we develop a new dexterous grasping optimization algorithm that solely relies on the point cloud as input, eliminating the necessity for complete mesh information of the object. By leveraging the contact maps of different fingertips, the proposed optimization algorithm can generate precise and determinable strategies for human-like object grasping. Experimental results confirm the efficiency of the proposed scheme. Our code is available at *https://github.com/wmtlab/GrainGrasp*.**

I. Introduction

The past decades have witnessed the rapid development of high-fidelity humanoid hands in computer graphics, e.g. MANO [1], as well as real-world dexterous robotic hands, e.g. Shadow Hand [2]. These advancements have enabled realistic hand pose imitation and dexterous manipulation, with widespread applications in Virtual Reality [3], [4], Human-Computer Interactions [5] and robotics [6]-[10].

The advances of digital and robotic hands have promoted dexterous grasping as an increasingly important research direction. Recent studies focus on obtaining diverse and high-quality grasps, including data-driven approaches, optimization algorithms, as well as combinations of these techniques. A typical procedure is to first utilize a Deep Learning algorithm to predict the contact maps from objects in datasets, then employ an optimization algorithm or a reinforcement learning approach to guide the hand to match the predicted contact maps and achieve grasping. Unfortunately, we should admit that the predicted the contact maps of this typical procedure can only provide a coarse approximation of ideal contacts. The ambiguity in the contact map makes it difficult to achieve stable grasps for complex objects. In order to achieve better grasping, it is necessary to adjust the position and force of each individual finger on the dexterous hand with fine-grained contact maps. However, existing grasp generation methods generally produce grasps without detailed finger adjustments, resulting in unsatisfactory performance on complex tasks. Therefore, realizing fine granularity over each finger becomes an essential task for high-quality grasp generation.

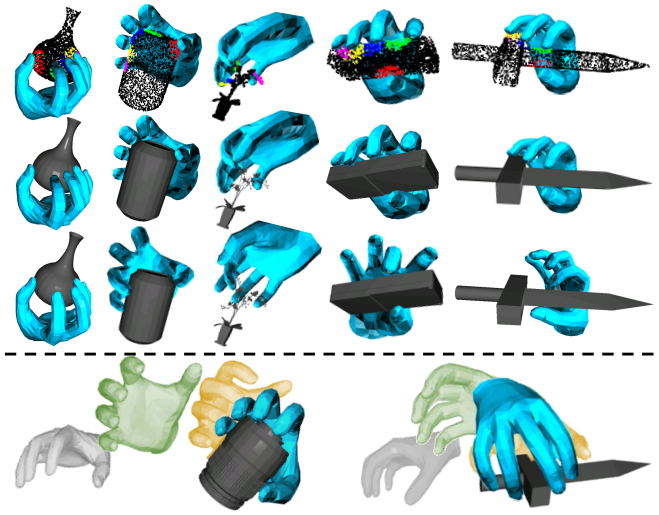

Fig. 1: Examples of grasping results and grasping processes generated by the GrainGrasp. Top: The grasp results are displayed with point cloud and mesh representations in the first two rows. The colored points in the point cloud indicate the contact map of each finger. The third row illustrates the determinable results achieved by controlling the contribution of each finger. Bottom: The grasping process is represented in four steps: gray → green → yellow → blue. Throughout the process, the hand first adjusts its direction, then quickly approaches the object, which presents human-like characteristics of object grasping.

On the other hand, due to the lack of explicit modeling between contact maps and executed grasps, even with fine-grained contact maps, it is not guaranteed to achieve human-desired grasping positions. This limitation poses a challenge to the effectiveness of using contact map prediction as a reliable supervision signal. Although methods like ContactGrasp [9] set specific contact locations for the thumb, controlling the thumb alone is insufficient for dexterous manipulation, which requires coordinated motion of multiple fingers. The key to resolving this problem lies in establishing a direct mapping and conducting joint optimization of finger positions, incorporating appropriate contact constraints.

To address the above-mentioned limitations, we reformulate the whole-hand grasping task by focusing on managing the individual contribution and coordination of each finger, rather than regulating the entire hand. In order to obtain a fine-grained contact map for grasping, we propose to predict a distinct contact map for each fingertip. This allows us

This work was supported in part by the National Science Foundation of China(Grant No.62071083), and in part by the Dalian Science and Technology Innovation Foundation (No. 2022JJ12GX014).

[1]Fuqiang Zhao and Qian Liu are with the the Department of Computer Science and Technology, Dalian University of Technology, China. Emails: fuqiangzh@mail.dlut.edu.cn, qianliu@dlut.edu.cn.

[2]Dzmitry Tsetserukou is with the Skolkovo Institute of Science and Technology, 121205 Moscow, Russia. Email: D.Tsetserukou@skoltech.ru.

*Corresponding author: Qian Liu.

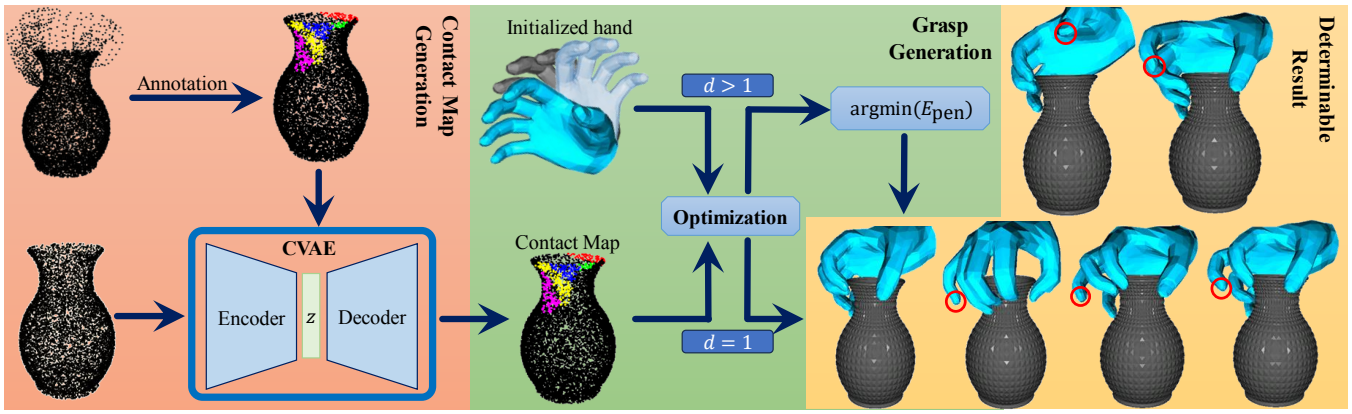

Fig. 2: Pipeline of our method. **1)** We automatically annotate point cloud data and train a CVAE model to generate individual contact maps. **2)** We utilize the object point cloud and contact maps to optimize the initialized hand pose. If the number of directions d rotates more than once (i.e., $d > 1$), the final grasping result is obtained by minimizing E_{pen} under the condition that $E_{\text{pen}} > 0$. **3)** We generate determinable grasping results by adjusting individual finger contributions.

to capture the details of finger-object contacts, instead of predicting a holistic contact map or only a small number of contact points as that of existing methods. Inspired by previous work [10], [11], we adopt generative models for this research.

Furthermore, we propose the **DCoG** (Directional Consistency optimization for Grasping) to comprehensively model the intricate relationship between the contact map and the grasping determination. DCoG emphasizes the consistency in the physically appropriate direction of each finger between the finger and the object surface during the grasping process. This way, DCoG can reveal the role of each finger in the grasping task and its coordinated relationship with other fingers. As a result, it provides a fine-grained guidance to the grasping behavior of each finger, enabling advanced grasping generation. Examples of grasping results and the corresponding grasping processes generated by the proposed scheme can be found in Fig. 1.

In summary, the proposed scheme consists of two key components: the Contact Map Generation and the Grasp Generation modules, as shown in Fig. 2. We further summarize the contributions of this research as follows.

1) We propose a new method to predict distinct contact maps for individual fingertips. This method offers enhanced precision in achieving a stable grasp by individually adjusting the contact of each finger, rather than making direct adjustments to the entire hand. Consequently, it can provide superior guidance for achieving grasping that closely resembles human capabilities.

2) Our proposed optimization method, known as DCoG, allows for interpretable and determinable robotic grasping using contact maps. This contact-based grasping algorithm enables explainable and intuitive robot manipulation.

II. Related Work

A. Dexterous Grasp Synthesis

The synthesis of dexterous grasps poses significant challenges due to the numerous degrees of freedom and complexity of modeling grasp interactions.

Analytical methods typically focus on optimizing the grasp pose to achieve force closure with consideration of physical constraints [12]-[15]. Recently, Liu *et al*. [16] developed a differentiable force closure estimator that enables the generation of varied and physically stable grasps for robotic hands with arbitrary structures. Building upon this method, Wang *et al*. [17] refined the initial hand pose, and contributed a new grasp dataset called DexGraspNet. Similarly, Li *et al*. [10] collected a dataset of grasps from different hands using the method described in [16].

With the development of Deep Learning and grasp datasets [17]-[22], data-driven methods have gained immense popularity. These methods utilize Deep Neural Networks to learn from datasets, enabling direct predictions of grasping parameters or critical information for grasping, such as contact points and contact maps. GraspCVAE [11] and GraspGlow [23] generated corresponding grasping parameters for 3D objects. These methods generally employed test-time adaptation (TTA) to refine the generated poses. UniGrasp [24] and EfficientGrasp [25] utilized a neural network called PSSN to select contact points. ContactNet [11], [23] and DeepContact [26] were trained to estimate contact maps. Additionally, other data-driven methods also showed promising results. One example is the training of Grasping Field [27] as an implicit function which used the signed distances [28] of 3D points to infer contact regions. Another example is CGF [7], which trained a Conditional Variational Auto-Encoder (CVAE) [29] framework to generate continuous trajectories.

B. Contact-based Approaches

In robotic grasping, contact information visually represents the physical interactions between the robot hand and the target object. Numerous research studies utilize this information to enhance grasping capabilities.

UniGrasp [24] predicted contact points on the object and adjusted the gripper trajectory using inverse kinematics. Wu *et al*. [30] employed the CVAE to predict finger placements, which were used to initialize a Bilevel Optimization for grasp configuration. ContactOpt [26] trained the DeepContact model by performing multiple random perturbations on hand-object grasps using the ContactPose dataset [22] to generate the target contact map. S^2Contact [31] utilized visual and geometric consistency constraints to generate pseudo-labels for semi-supervised learning and employed a graph-

(a) (b) (c)

Fig. 3: (a)&(b) Five fingertips and fingers are presented in different colors for distinction. (c) Our initial hand pose exhibits a slight flexion at the interphalangeal joint compared to a fully flat hand.

based network to infer the contact map. ContactGrasp [9] sampled random grasps using GraspIt! [32] and ranked them based on their consistency with a thumb-aligned contact map to synthesize functional grasps using DART [33]. Jiang *et al*. [11] proposed a self-supervised approach to improve grasping. It updated the hand pose based on the difference between the contact map predicted by ContactNet and the contact map estimated by the distance between the hand and the object during inference. Similar to the approach presented in [11], UniDexGrasp [23] incorporates information from the predicted contact map generated by ContactNet. This information is used to optimize the grasping pose during the intermediate step of the grasping process. GenDexGrasp [10] trained a generative model CVAE to generate a contact map for the object point cloud, then used an optimization algorithm to fit the hand to the generated contact map. Mandikal *et al*. [34] modeled the contact map as the object's affordance and treated the affordance learning problem as a segmentation task to predict binary per-pixel labels on the contact map.

Inspired by the aforementioned methods, we approach the generation of contact maps as a point cloud segmentation task. In this study, we generate individual contact maps for each fingertip on the object point cloud. By predicting contact maps per fingertip, we can independently optimize the contact location and force for each finger. Consequently, our proposed approach offers fine-grained grasping guidance for a dexterous hand.

III. Contact Map Generation

The development of an efficient separate contact map generator involves a two-step process. We first automatically annotate the ObMan dataset [20] and then use it to train a CVAE model. This CVAE model can generate separate contact maps for each finger when given an input of object point cloud.

A. Dataset Annotation

Unlike previous algorithms [9], [26] that directly utilize thermal contacts from the ContactDB [21] and Contact-Pose [22] datasets, our method enables automatic annotation of contact maps on the Grasp dataset solely relies on hand-object point cloud data. In this research, we adopt the ObMan dataset [20] and conduct all experiments based on the MANO hand model, which can be described by $\theta_H = (\theta, t, R)$, with θ, t, R indicating the pose, translation, and rotation, respectively. The pose θ is represented as a 45-dimensional Principal Component Analysis (PCA) manifold. It should be noted that the MANO model also requires shape parameters to describe the hand model. In this paper, we use the default shape parameter, which remains unchanged during the optimization process.

First, we manually delineate the fingertip region of the hand using the Blender software [35]. As shown in Fig. 3(a), the five fingertips are delineated in different colors, indicating a specific fingertip encountering the object on the annotated contact map. *Second*, for points within the delineated regions representing the five fingertips, we compute the K nearest points on the object point cloud. An issue may arise during this step, where some points may belong to contact maps of multiple fingertips simultaneously. To resolve this problem, we should determine which fingertip region is closest to each point and assign it accordingly in the contact map. *Finally*, we classify each point in the object point cloud into six categories: contacted by the thumb, index, middle, ring, pinky fingertip, or uncontacted, respectively.

Since we only use object point clouds, we can only rely on Euclidean distance instead of the aligned distance adopted by Li *et al*. [10]. Consequently, some false contact maps may be generated when annotating thin objects. Therefore, we cannot have a large K. On the other hand, a smaller K value can lead to a reduction in the overall size of annotated contact maps, which can cause difficulty in model prediction. Therefore, we encourage the utilization of a more appropriate K value. Based on our experiments, it seems that setting the value of K to 50 for a point cloud of 3000 points resulted in satisfactory point labeling performance.

For clearer illustration, we denote the subset involving the five contacted categories as C. Then, we employ the notations $\mathbf{T}$, $\mathbf{H}$, and $\mathbf{O}$ to symbolize the point cloud sets corresponding to fingertips, full hands, and objects, respectively. We concurrently manually delineated the finger region, as shown in Fig. 3(b). We denote these labeled point cloud sets as $\mathbf{F}$, which will be used in the proposed optimization algorithm in Sec. IV-A.

B. CVAE Synopsis

Inspired by [10], [24], we regard the contact map generating problem as a segmentation task to predict multiclass pointwise labels. Meanwhile, we adopt the CVAE [29] to generate the separate contact maps of fingertips.

CVAE consists of an encoding stage and a decoding stage. For the **encoding** stage, we first utilize the PointNet encoder [36] to extract pointwise features from the object point cloud. In parallel, an embedding layer is employed to obtain embedded category features from the corresponding contact map. Then, we concatenate the two features and apply Multi-Layer Perceptrons (MLP), followed by a max pooling layer, to obtain the latent distribution $\mathcal{N}(\mu, \sigma^2)$. For the **decoding** stage, we sample the latent code $z \sim \mathcal{N}(\mu, \sigma^2)$ and replicate it n times, where n indicates the number of points in the point cloud. The replicated latent code is concatenated with the pointwise features and passed through an MLP to generate the contact map.

For the training of the CVAE model, we leverage two types of supervised loss functions: a reconstruction loss that penalizes errors in reconstructing the contact map and a regularization loss that encourages the latent space to have desirable properties.

TABLE I: Comparison Experiments

Methods	**Volume**$(cm^3)\downarrow$	**Depth**$(cm)\downarrow$	**Displacement**$(cm)\downarrow$	**Succ. Rate**$(\%.)\uparrow$	**Perc. Scores**$\uparrow$
GT [20]	2.20	0.66	**0.75**	**52.28**	3.51
Contactopt [26]	3.25	0.68	0.91	18.52	-
GF [27]	2.38	0.94	0.89	20.60	2.93
GA (w/o TTA) [11]	2.99	0.77	0.87	33.78	-
GA (w/ TTA) [11]	3.65	0.75	**0.80**	35.61	3.43
Ours (only opt.)	**1.48**	0.57	0.82	**45.51**	**3.66**
Ours (complete)	1.79	**0.48**	0.84	43.10	3.41

GT represents results from the dataset. GF represents results from the Grasping Field method. GA represents results from the GraspTTA method.

The reconstruction loss consists of a multiclass cross-entropy loss $\mathcal{L}_{\text{cross}}$ and a multiclass dice loss $\mathcal{L}_{\text{dice}}$ [37], which enables the network to generate more accurate contact map from the input point cloud. We use a KL loss $\mathcal{L}_{\text{KD}}$ as the regularization loss. The $\mathcal{L}_{\text{KD}}$ is defined by maximizing the KL-Divergence between the latent code distribution $\mathcal{N}(\mu, \sigma^2)$ and a standard Gaussian distribution $\mathcal{N}(0, 1)$, which encourages the latent code distribution learned by the encoder to be close to the prior standard Gaussian distribution. The complete loss function to train the CVAE can be expressed as:

$$\mathcal{L} = \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{dice}}\mathcal{L}_{\text{dice}} + \lambda_{\text{KD}}\mathcal{L}_{\text{KD}} \tag{1}$$

where λ_{cross}, λ_{dice} and λ_{KD} represent the weights of corresponding losses. Their values are set as 0.5, 0.9 and 0.01, respectively.

IV. Proposed Grasping Optimization

When a human grasps an object, the fingertips apply force in the direction perpendicular to the surface to ensure a stable grip. To mimic the human grasping process, we propose Directional Consistency optimization for Grasping (DCoG). Given an object point cloud and its corresponding contact map, DCoG guides the grasping process to simultaneously optimize both hand direction and grasp distance through well-designed energy terms.

A. Optimization Method

The DCoG consists of multiple energy terms to achieve deterministic grasping results.

Contact-based Distance Energy: One key objective of grasping is to ensure the actual contact between the hand and the object. To achieve this, we propose a contact-based distance energy E_{dis}. This energy function aims to guide the hand to a suitable position by minimizing the squared Euclidean distance between the fingertips' points and the object contact map points, considering they belong to the same category. Mathematically, this energy can be expressed as:

$$E_{\text{dis}} = \sum_{c}^{C}\sum_{i}^{|\mathbf{T}^c|} \text{Drop}\left(\min_{j}\left\|\mathbf{T}_i^c - \mathbf{O}_j^c\right\|_2^2, p\right) \tag{2}$$

where $\mathbf{T}^c$ and $\mathbf{O}^c$ denote the subsets of the point cloud associated with category c within $\mathbf{T}$ and $\mathbf{O}$, respectively. $\mathbf{T}^c$ is manually delineated, while $\mathbf{O}^c$ is generated by the CVAE model mentioned in Sec. III-B. To reduce the risk of getting trapped in a local optima, we employ the $\text{Drop}(\cdot, p)$ operation. It increases the randomness of the optimization process by introducing a probability parameter, denoted as p, to reset certain calculation results to zero.

Directional Consistency Energy: In order to enable precise force applied to fingers during grasping, we introduce two energy terms to ensure the directional consistency: E_{dct}, defined over the fingertip, and E_{dcf}, defined over the finger. E_{dct} and E_{dcf} are defined as the squared Euclidean distance between the unit surface normal vector of a point in the set $\mathbf{T}^c(\mathbf{F}^c)$ and the unit directional vector pointing towards the nearest point within the set $\mathbf{O}^c(\mathbf{O})$. $\mathbf{F}^c$ denote the subsets of the point cloud associated with category c within $\mathbf{F}$, which is also manually delineated, similar to $\mathbf{T}^c$. Since the mesh of the hand can be obtained directly, the surface normal vectors of all points in $\mathbf{T}$ and $\mathbf{F}$ can be simply calculated. Let $N(x)$ denote the unit surface normal vector at point x, and let $u(x) = \frac{x}{\|x\|_2}$ indicate the unit normalization of x. Thus, E_{dct}, E_{dcf} and the final directional consistency energy E_{dc} are defined as

$$E_{\text{dct}} = \sum_{c}^{C}\sum_{i}^{|\mathbf{T}^c|} \text{Drop}\left(\left\|N\left(\mathbf{T}_i^c\right) - u\left(\mathbf{O}_k^c - \mathbf{T}_i^c\right)\right\|_2^2, p\right) \tag{3}$$

$$E_{\text{dcf}} = \sum_{c}^{C}\sum_{i}^{|\mathbf{F}^c|} \text{Drop}\left(\left\|N\left(\mathbf{F}_i^c\right) - u\left(\mathbf{O}_{\dot{k}} - \mathbf{F}_i^c\right)\right\|_2^2, p\right) \tag{4}$$

$$E_{\text{dc}} = \lambda_{\text{dct}}E_{\text{dct}} + \lambda_{\text{dcf}}E_{\text{dcf}} \tag{5}$$

where k and $\dot{k}$ can be calculated via $k = \text{argmin}_j \left\|\mathbf{T}_i^c - \mathbf{O}_j^c\right\|$ and $\dot{k} = \text{argmin}_j \left\|\mathbf{F}_i^c - \mathbf{O}_j\right\|$, respectively.

Adaptive Weight Adjustment Strategy: A potential challenge in the optimization process is determining the weight coordination of the two energy terms in Eq. (2) and Eq. (5). To tackle this task, we propose an adaptive weight adjustment strategy. In the initial optimization phase, we assign a lower weight to E_{dis} and a higher weight to E_{dc}, guiding the optimization algorithm to enhance the hand's grasping orientation. As the number of iterations increases, the weight of E_{dis} progressively decreases, while the weight of E_{dc} increases. This adaptive weight adjustment promotes a more natural grasping process.

The non-convex nature of the dexterous grasping task presents challenges for the optimization algorithm in achieving a global optimal solution. Inspired by [38], we train a binary classification network based on the PointNet ar-

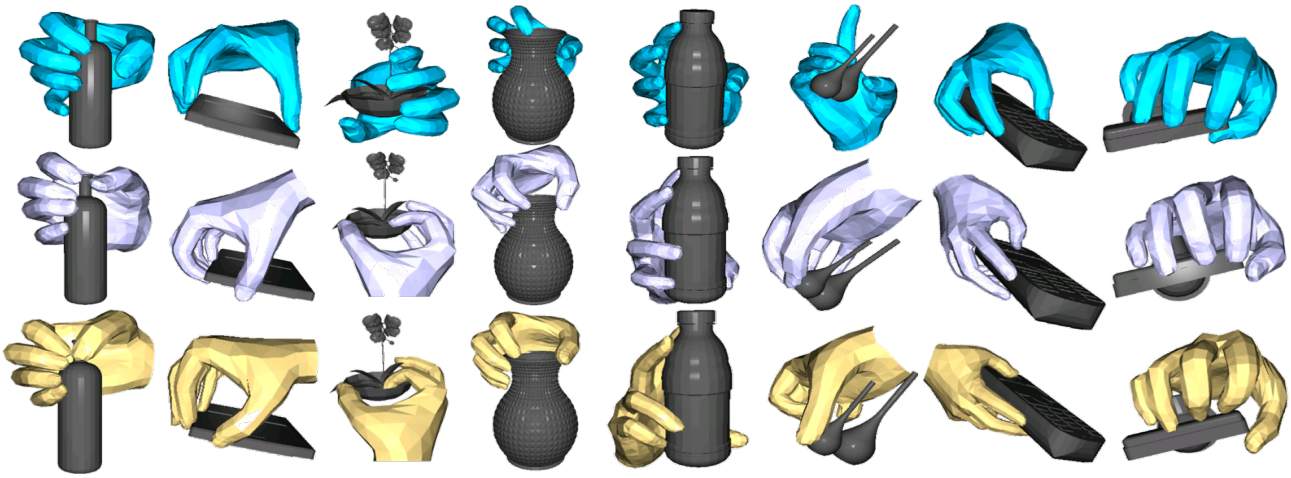

Fig. 4: Qualitative experimental results. The *top* shows grasps generated by our complete method, the *middle* displays grasps obtained with the optimization-only method, and the *bottom* represents the Ground Truth.

chitecture which utilizes a binary cross-entropy loss function, taking the **H** and **O** as inputs, while predicting the success outcome of grasping. To avoid mismatching, we also incorporate the binary cross-entropy loss function as an energy term, denoted as $E_{\text{net}} = -\log p_{\text{net}}$, for network supervision in the optimization process. Here, p_{net} represents the probability of being successfully grasped according to the network prediction. Since the network has learned the underlying principles of grasping, it gains an advantage in the optimization algorithm by effectively avoiding local optimal solutions. Another energy term considered is the penetration energy. As our method relies solely on point clouds, direct computation of the signed distance is not feasible. Consequently, we adopt E_{pen} as proposed by Jiang *et al*. [11].

Finally, we formulate the total energy as:

$$E = i_{\text{c}}\lambda_{\text{dis}}E_{\text{dis}} + (i_{\text{s}} - i_{\text{c}})\lambda_{\text{dc}}E_{\text{dc}} + \lambda_{\text{net}}E_{\text{net}} + \lambda_{\text{pen}}E_{\text{pen}} \quad (6)$$

where i_{c} denotes the current iteration index in the optimization process, and i_{s} denotes the total number of iterations. These two parameters correspond to the adaptive weight adjustment strategy.

In this research, the parameters in Eq. (6) are set as: $i_{\text{s}} = 300, \lambda_{\text{dis}} = 0.5, \lambda_{\text{dct}} = 0.8, \lambda_{\text{dcf}} = 0.6, \lambda_{\text{dc}} = 1.0, \lambda_{\text{net}} = 0.6, \lambda_{\text{pen}} = 10$.

B. Initialization Strategy

Inspired by [17], we adopt a similar hand pose preparing for grasping, as illustrated in Fig. 3(c). Then, we place the hand to a random distance from the contact map, aligning with the middle fingertip to ensure that it is away from the object. However, in real-world grasps, it is generally impractical to achieve a direct orientation of the palm facing the object. To address this, we introduce diversity by rotating the palm in various directions during initialization. The final grasping result is determined based on the minimum of E_{pen}, as incorrectly grasped results often exhibit a significant area of penetration with the object.

C. Analysis of Determinable Generation

Different grasping tasks require distinct finger contributions. Our method leverages individualized contact for each finger, employing highly interpretable energy terms. This enables us to achieve a customizable selection of individual finger contributions, which is in particular realized by adjusting the weight of each category when calculating E_{dis}. Such a strategy has proven to be effective in various scenarios, enabling the algorithm to determine which fingers should be given priority when grasping objects.

V. Experiment

We conduct both quantitative and qualitative evaluations on the proposed approach, comparing it with other grasping methods. Additionally, we perform ablation experiments to illustrate the individual contributions of different modules and demonstrate the synergistic effect achieved through their combination.

A. Quantitative Evaluations

Penetration: We evaluate both the penetration volume and maximum penetration depth for all test samples. A larger penetration indicates a deeper intrusion of the hand into the object. Conversely, if the penetration volume or penetration depth is 0, it may infer that the hand has no contact with the object.

Physical Displacement: We conduct a quantitative assessment of grasp stability using a physics simulator, following the methodology introduced in [11], [39]. In this evaluation, we place the hand and object within the simulator, then it calculates contact forces based on the penetration volume, and simulates the displacement of the object under these forces. A larger displacement suggests that the grasp is unable to apply sufficient forces to stably hold the object.

Success Rate: When a grasp involves a minimal penetration, it may result in a significant physical displacement if the hand cannot exert an appropriate grip force. Conversely, if a grasp exhibits considerable physical displacement, it may be due to excessive penetration between the hand and the object. Therefore, we define the success rate of a grasp by considering a trade-off between these two factors. In particular, a grasp is deemed successful if the penetration volume is less than 5 cm^3 and the object displacement is smaller than 2 cm. Otherwise, we categorize the grasp as a

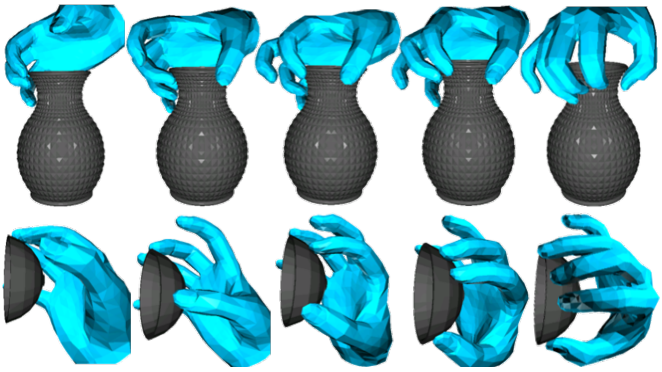

Fig. 5: Determinable results. By sequentially setting the contribution of each finger to zero, we can obtain five determinable grasps.

failure. We should admit that this definition of success rate may be stringent, emphasizing the granularity of the grasp.

As mentioned earlier, to mitigate the impact of faulty grasps on our experimental results, we calculate the average of penetration and physical displacement across all successful grasps.

Perceptual Scores: We conduct a perceptual study to evaluate the naturalness of the generated grasps following [27] and [11]. In contrast to their approaches, we recruit 15 participants who are able to observe the complete view of the grasps in a 3D window, in contrast to only a few perspectives as that of [27] and [11]. Each participant is asked to score 20 randomly generated grasp results by different method shown in Tab. I, where ContactOpt is not included due to low success rate, while the GA (w/o TTA) is excluded due to an obvious worse performance compared with GA (w/ TTA).

The experimental results presented in Tab. I exhibit that our success rate and perceptual scores surpass those of other methods. As ContactOpt requires complete MANO parameters as input, we utilize the generated parameters from our method for ContactOpt. Due to the potential instability of contact maps produced by the generation model, we conduct separate evaluations using the optimization-only method and the complete method. In the case of the optimization-only method, the contact maps are annotated using our annotation method presented in Sec. III-A.

We observe better results when using only the optimization method, primarily because the annotated contact maps obtained are more realistic. This highlights the effectiveness of our proposed optimization method. It also indicates that as the model's capability to generate accurate contact maps improves, the proposed approach will be more likely to achieve better grasp results.

B. Qualitative Evaluations

In Fig. 4, we present visuals of grasps generated by our proposed complete method, the optimization-only method, and the Ground Truth (GT). The optimization-only method utilizes GT-annotated contact maps, resulting in generated grasps that closely resemble the GT. In contrast, the complete method autonomously generates contact maps using the CVAE, leading to differences between the generated grasps and the GT. Through qualitative experimental results, we demonstrate that the proposed method is capable of producing stable grasps. In addition, the determinable results are shown in Fig. 5, highlighting the deterministic capabilities of our method in generating grasps. By adjusting the contributions of different fingers, we generate grasp poses with determinable outcomes. We believe that this controllability in grasping will enhance the adaptability of the grasp algorithm, thereby enabling it to emulate the versatility and adaptability observed in human hand grasping.

TABLE II: Ablation Study - Energy Terms

Energy	**Volume**(cm^3)	**disp.**$(cm)\downarrow$	**Succ. R.**$(\%)\uparrow$
w/o E_{dct}	2.10	0.89	40.80
w/o E_{dcf}	2.55	0.85	39.67
w/o E_{dc}	-	-	7.34
w/o E_{net}	**1.48**	0.93	43.52
complete E	**1.48**	**0.82**	**45.51**

C. Ablation Study

The ablation study aims to evaluate the proposed energy terms: E_{dct}, E_{dcf}, E_{dc}, and E_{net}.

The results of the ablation experiments in Tab. II indicate that each energy term plays a distinct role. The removal of E_{dc} results in a significantly reduced number of successful grasps, making other evaluation metrics meaningless in its absence. E_{dct} and E_{dcf} primarily contribute to increasing the success rate of grasps, while E_{net}, learned from the dataset, emphasizes the close interaction between the hand and the object, thereby enhancing the overall quality of grasps.

D. Methodological Analysis

Advantages: Our experimental results demonstrate the enhanced grasping granularity achieved by the proposed method. Notably, the proposed approach effectively preserves small object displacements in the simulation environment while minimizing penetration volume. Moreover, the proposed method only requires sampled point clouds of objects as input. In addition, our method exhibits a high degree of hand control, facilitating easy adjustment of each finger's contribution during the process of object grasping. This remarkable flexibility not only highlights the adaptability of our grasping method but also establishes a robust groundwork for future extensions and enhancements.

Limitations: We also investigate the scalability of our proposed method. We observe that if the CVAE has not been trained on objects that have similar shapes to the input object, it may face challenges in generating accurate contact maps, leading to grasp failures. This indicates the dependency of our method on the generation capability of the CVAE model. To effectively overcome this limitation, comprehensive training on a large-scale dataset becomes necessary.

VI. Conclusion

In this study, we first revealed the limitations of existing methods in grasping generation. In order to tackle these challenges, we proposed a new scheme that utilized a CVAE model to predict individual contact maps for each finger. We also developed a new optimization method, DCoG, to guide the generation of grasps. Our approach **GrainGrasp** demonstrated promising performance in terms of granularity

in grasping. In particular, we can determine the contribution of each finger to the final grasp result, which was not presented in previous methods. We should point out that the performance of the proposed scheme relied on the generation capability of CVAE during the contact map generation process. We leave this for future research.

REFERENCES

[1] 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.

[2] "Shadowrobot," 2005. [Online]. Available: https://www.shadowrobot.com/dexterous-hand-series/

[3] S. Han, P.-c. Wu, Y. Zhang, B. Liu, L. Zhang, Z. Wang, W. Si, P. Zhang, Y. Cai, T. Hodan *et al.*, "Umetrack: Unified multi-view end-to-end hand tracking for vr," in *SIGGRAPH Asia 2022 Conference Papers*, 2022, pp. 1–9.

[4] M. Höll, M. Oberweger, C. Arth, and V. Lepetit, "Efficient physics-based implementation for realistic hand-object interaction in virtual reality," in *2018 IEEE conference on virtual reality and 3D user interfaces (VR)*. IEEE, 2018, pp. 175–182.

[5] T. Xue, W. Wang, J. Ma, W. Liu, Z. Pan, and M. Han, "Progress and prospects of multimodal fusion methods in physical human–robot interaction: A review," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10 355–10 370, 2020.

[6] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*. PMLR, 2021, pp. 1602–1611.

[7] J. Ye, J. Wang, B. Huang, Y. Qin, and X. Wang, "Learning continuous grasping function with a dexterous hand from human demonstrations," *IEEE Robotics and Automation Letters*, vol. 8, no. 5, pp. 2882–2889, 2023.

[8] Y. Qin, H. Su, and X. Wang, "From one hand to multiple hands: Imitation learning for dexterous manipulation from single-camera teleoperation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 873–10 881, 2022.

[9] S. Brahmbhatt, A. Handa, J. Hays, and D. Fox, "Contactgrasp: Functional multi-finger grasp synthesis from contact," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 2386–2393.

[10] P. Li, T. Liu, Y. Li, Y. Geng, Y. Zhu, Y. Yang, and S. Huang, "Gendexgrasp: Generalizable dexterous grasping," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 8068–8074.

[11] H. Jiang, S. Liu, J. Wang, and X. Wang, "Hand-object contact consistency reasoning for human grasps generation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 107–11 116.

[12] J.-W. Li, H. Liu, and H.-G. Cai, "On computing three-finger force-closure grasps of 2-d and 3-d objects," *IEEE Transactions on Robotics and Automation*, vol. 19, no. 1, pp. 155–161, 2003.

[13] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 326–336, 2012.

[14] C. Rosales, R. Suárez, M. Gabiccini, and A. Bicchi, "On the synthesis of feasible and prehensile robotic grasps," in *2012 IEEE international conference on robotics and automation*. IEEE, 2012, pp. 550–556.

[15] D. Prattichizzo, M. Malvezzi, M. Gabiccini, and A. Bicchi, "On the manipulability ellipsoids of underactuated robotic hands with compliance," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 337–346, 2012.

[16] 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*, vol. 7, no. 1, pp. 470–477, 2021.

[17] R. Wang, J. Zhang, J. Chen, Y. Xu, P. Li, T. Liu, and H. Wang, "Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 359–11 366.

[18] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9044–9053.

[19] S. Hampali, M. Rad, M. Oberweger, and V. Lepetit, "Honnotate: A method for 3d annotation of hand and object poses," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3196–3206.

[20] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 807–11 816.

[21] S. Brahmbhatt, C. Ham, C. C. Kemp, and J. Hays, "Contactdb: Analyzing and predicting grasp contact via thermal imaging," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8709–8719.

[22] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 361–378.

[23] Y. Xu, W. Wan, J. Zhang, H. Liu, Z. Shan, H. Shen, R. Wang, H. Geng, Y. Weng, J. Chen *et al.*, "Unidexgrasp: Universal robotic dexterous grasping via learning diverse proposal generation and goal-conditioned policy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4737–4746.

[24] 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*, vol. 5, no. 2, pp. 2286–2293, 2020.

[25] 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*, vol. 7, no. 4, pp. 8619–8626, 2022.

[26] P. Grady, C. Tang, C. D. Twigg, M. Vo, S. Brahmbhatt, and C. C. Kemp, "Contactopt: Optimizing contact to improve grasps," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 1471–1481.

[27] K. Karunratanakul, J. Yang, Y. Zhang, M. J. Black, K. Muandet, and S. Tang, "Grasping field: Learning implicit representations for human grasps," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 333–344.

[28] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[29] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," *Advances in neural information processing systems*, vol. 28, 2015.

[30] 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.

[31] T. H. E. Tse, Z. Zhang, K. I. Kim, A. Leonardis, F. Zheng, and H. J. Chang, "S 2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning," in *European Conference on Computer Vision*. Springer, 2022, pp. 568–584.

[32] 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.

[33] T. Schmidt, R. A. Newcombe, and D. Fox, "Dart: Dense articulated real-time tracking." in *Robotics: Science and systems*, vol. 2, no. 1. Berkeley, CA, 2014, pp. 1–9.

[34] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," in *2021 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2021, pp. 6169–6176.

[35] Blender Online Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Blender Institute, Amsterdam, 2023. [Online]. Available: http://www.blender.org

[36] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.

[37] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *2016 fourth international conference on 3D vision (3DV)*. Ieee, 2016, pp. 565–571.

[38] 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.

[39] D. Tzionas, L. Ballan, A. Srikantha, P. Aponte, M. Pollefeys, and J. Gall, "Capturing hands in action using discriminative salient points and physics simulation," *International Journal of Computer Vision*, vol. 118, pp. 172–193, 2016.