Efficient Heatmap-Guided 6-Dof Grasp Detection in Cluttered Scenes

Siang Chen[1,2,4,†], Wei Tang[3,4,†], Pengwei Xie[1,4], Wenming Yang[3,4], Guijin Wang[1,2,4,*]

arXiv:2403.18546v2 [cs.RO] 14 May 2024

***Abstract*—Fast and robust object grasping in clutter is a crucial component of robotics. Most current works resort to the whole observed point cloud for 6-Dof grasp generation, ignoring the guidance information excavated from global semantics, thus limiting high-quality grasp generation and real-time performance. In this work, we show that the widely used heatmaps are underestimated in the efficiency of 6-Dof grasp generation. Therefore, we propose an effective local grasp generator combined with grasp heatmaps as guidance, which infers in a global-to-local semantic-to-point way. Specifically, Gaussian encoding and the grid-based strategy are applied to predict grasp heatmaps as guidance to aggregate local points into graspable regions and provide global semantic information. Further, a novel non-uniform anchor sampling mechanism is designed to improve grasp accuracy and diversity. Benefiting from the high-efficiency encoding in the image space and focusing on points in local graspable regions, our framework can perform high-quality grasp detection in real-time and achieve state-of-the-art results. In addition, real robot experiments demonstrate the effectiveness of our method with a success rate of 94% and a clutter completion rate of 100%. Our code is available at https://github.com/THU-VCLab/HGGD.**

***Index Terms*—Deep Learning in Grasping and Manipulation; RGB-D Perception; Grasping**

I. INTRODUCTION

OBJECT grasping is a critical component of robotics in manufacturing, service, medical assistance, etc. Despite its vital importance, fast and accurate grasping is still challenging for robots. Recent advances in deep learning have enabled data-driven methods to generalize to unseen objects. Representative methods [1]–[3] generate grasp configurations as oriented grasp rectangles by adopting pixel-wise heatmaps to represent planar grasps, achieving good performance in simple scenarios with high-efficiency. However, such a representation forces the gripper perpendicular to the camera plane, limiting the applications.

Recently, considerable attention has been paid to 6-Dof grasping, enabling robots to grasp from arbitrary directions. [4], [5] adopt a sampling-evaluation strategy which is extremely time-consuming. [6], [7] directly regress grasp attributes from the extracted per-point features. However, these methods tend to generate unreliable grasps since ignoring the local contextual geometric information. To ameliorate this issue, recent methods [8], [9] encode locally aggregated features to generate or fine-tune grasp poses. Despite the impressive performances, these approaches still struggle to perform high-quality 6-Dof grasp detection in real-time.

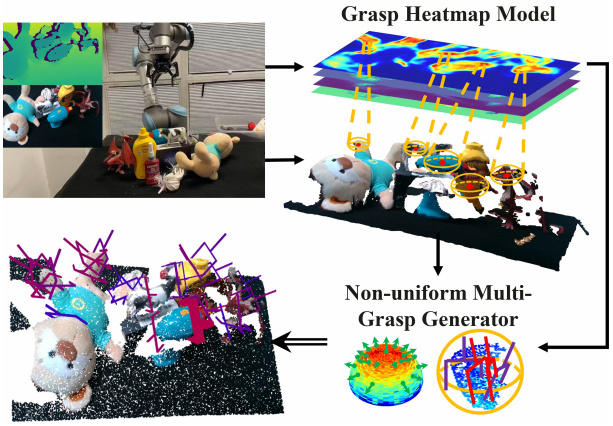

Fig. 1. The key insight of our method is generating the grasp heatmaps as guidance for regional geometric feature mining and further grasp pose generation via a novel local grasp generator.

Accordingly, motivated by the satisfying performance brought by the widely used heatmaps in object detection [10], human pose estimation [11] and planar grasping [1], this paper extends heatmaps for high-quality 6-Dof grasp generation with high-efficiency through a carefully designed local grasp generator. We propose a novel efficient 6-Dof grasp detection framework in cluttered scenes. As shown in Fig. 1, our key insight is constructing the grasp heatmaps as guidance to aggregate local points into graspable regions for further grasp pose generation. Thus, it can distinguish the graspable regions and significantly reduce the input size. Specifically, Gaussian encoding and the grid-based strategy are introduced to predict grasp confidence and attribute heatmaps robustly and efficiently. Furthermore, we design a local grasp generator combined with a novel non-uniform anchor sampling mechanism to precisely estimate spatial rotations in local aggregated regions, enabling anchor sampling with minimal fitting error between anchors and the actual grasp rotation distributions. Notably, our pipeline can fully leverage the global semantic and local geometrical representations benefiting from our semantic-to-point feature fusion.

In summary, our primary contributions are as follows:

- We propose a novel global-to-local semantic-to-point 6-Dof grasp detection framework, achieving state-of-the-

[1]Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.
[2]Shanghai AI Laboratory, Shanghai 200232, China.
[3]Department of Electronic Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518071, China.
[4]Beijing National Research Center for Information Science and Technology (BNRist), China.
[†]These authors contributed equally.
*Correspondence: wangguijin@tsinghua.edu.cn.

art performance in real-time through a low-cost training procedure.

- Grasp attribute heatmaps are predicted by the proposed Gaussian encoding and the grid-based strategy, significantly improving the encoding efficiency and reducing the local input size for grasp generation.
- A local grasp generator with a novel non-uniform anchor sampling mechanism is designed to generate dense grasps precisely, and an extra local semantic-to-point feature fusion makes grasp generation more robust.

II. RELATED WORKS

Existing grasp detection methods can be roughly divided into model-based and model-free. The model-based methods [12]–[15] transfer the task to object pose estimation and project grasps from the pre-prepared database. In contrast, model-free methods like [1], [16], [17] consider the problem as grasp rectangle detection in images, still limited in some scenarios since the grippers can only be perpendicular to the camera plane.

Recently, 6-Dof grasping has been widely researched due to the flexibility that the gripper can approach the object from arbitrary directions. [4], [5] propose the sample-evaluation strategy to select high-quality grasps from a large number of grasp proposals. [7] directly regresses grasps from per-point features extracted by PointNet++. Different from these methods, our method designs a global-to-local semantic-to-point framework to detect grasps, avoiding time-consuming grasp candidates sampling and significantly increasing the quality of generated grasps.

Based on the large-scale grasp dataset and benchmark [18], [19] utilize RGB images to generate pixel-wise orientation heatmaps and filters unreasonable grasps with point clouds. Similarly, [20] proposes point-wise graspness to represent the possibility of grasp locations and approach directions. Unlike these methods, we predict heatmaps as guidance to aggregate local points into graspable regions and only focus on these regions for further grasp pose generation.

To fully exploit the local contextual features, [9] refines grasp poses locally, and [8] presents a network based on grasp regions and gripper closing areas to obtain local shape information of grasps. More targeted than them, our method constructs heatmaps to distinguish graspable regions and perform grasp orientation prediction and center refinement on these regions, fully leveraging the fusion of semantic and geometrical representations and facilitating real-time detection performance. Besides, to precisely predict the spatial rotation angle or the approach view, [8], [9], [18] introduce the anchor-based strategy, achieving higher rotation prediction accuracy [21] than directly regressing. Unlike sampling pre-defined anchors from the surface of a unit sphere or using Fibonacci lattices [22], we design a novel non-uniform anchor sampling mechanism to predict the spatial rotation.

III. PROBLEM STATEMENT

Given a monocular RGBD image $\chi \in R^{H\times W\times 4}$ and corresponding camera intrinsics $\mathbf{c}$, our framework aims to efficiently learn parallel-jaw grasp configurations $\mathbf{G}$ in a cluttered scene. Equivalent to former definitions of the grasp in [5], [9], [18], to better fit proposed global-to-local semantic-to-point framework, one grasp pose $\boldsymbol{g}$ is defined as:

$$\boldsymbol{g} = (u, v, \theta, w, d, \gamma, \beta).$$

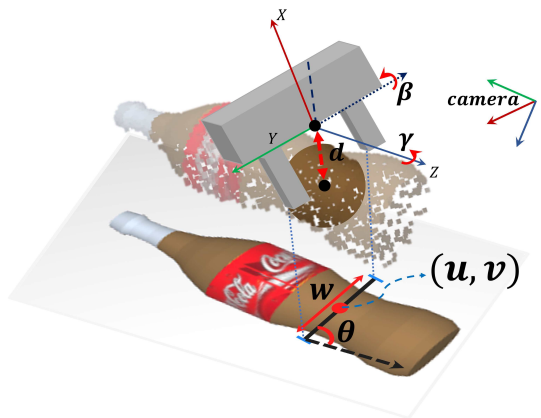

Fig. 2. Proposed grasp representation as $(u, v, \theta, w, d, \gamma, \beta)$.

As Shown in Fig. 2, rather than directly locating grasp centers in the 3-dimensional coordinate, we adopt a 2-dimensional tuple (u, v) to represent the grasp center in the image plane. (u, v, θ, w) with another fixed grasp height parameter h compose an oriented rectangle, representing a 4-Dof grasp similar to [23]. $\gamma, \beta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ are grasp Euler angles among axis-z and axis-y in the gripper coordinate, while d denotes the depth offset from the grasp center to corresponding point cloud surface point. Refer to appendices for more details of our data preparation.

IV. METHOD

A. Overview

We aim to efficiently generate high-quality and abundant grasps $\mathbf{G}$ in a novel global-to-local semantic-to-point way, with monocular RGBD images and camera intrinsics $\mathbf{c}$. As illustrated in Fig. 3, instead of directly processing the observed point cloud, our method inputs RGBD images to encode grasp heatmaps as graspable region guidance efficiently. With the critical heatmap guidance, only these regions' semantic and geometrical representations are extracted and fused. Then a novel local grasp generator enables **HGGD** (**Heatmap-Guided 6-Dof Grasp Detection**) to detect grasps with high-quality and diversity in real-time. Our model comprises two sub-modules: **Grasp Heatmap Model** (**GHM**) and **Non-uniform Multi-Grasp Generator** (**NMG**).

GHM preprocesses input RGBD images, extracts semantic features with an efficient CNN, and further generates four grasp heatmaps as guidance. Gaussian encoding is applied to encode the grasp ground truths, assisting in locating graspable areas more precisely. The grid-based strategy transforms the discontinuous pixel-wise regression into the prediction based

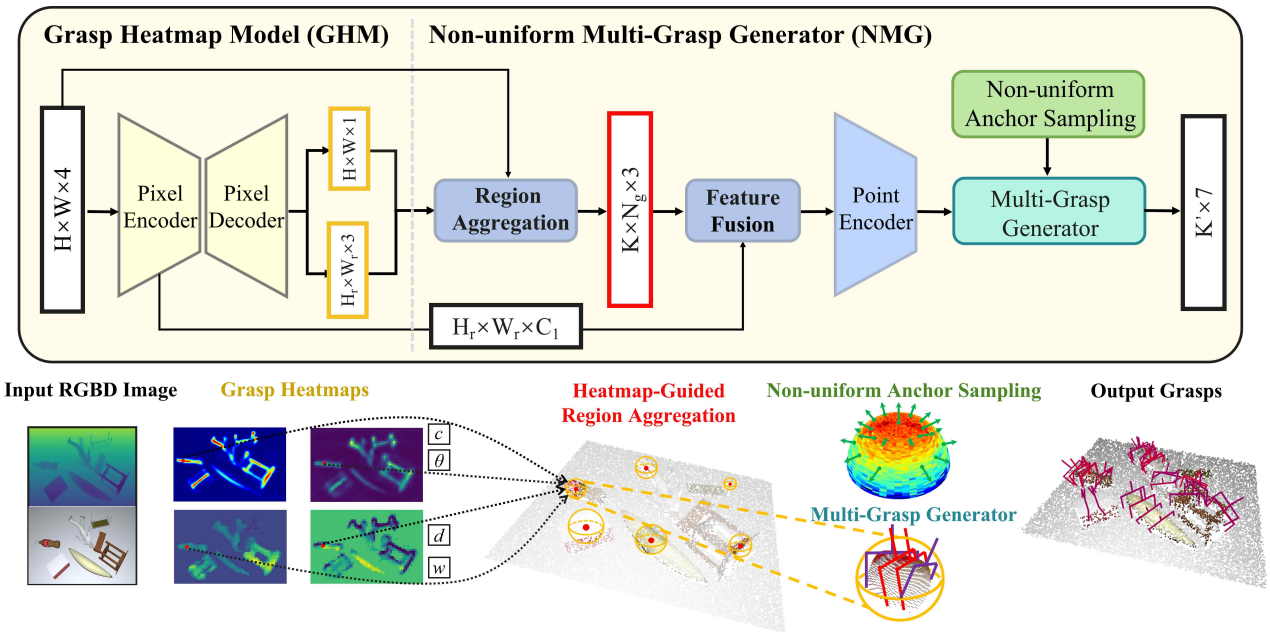

Fig. 3. The architecture of HGGD. Taking a monocular RGBD image as input, GHM generates grasp confidence heatmap Q_c and grided attributes heatmaps (Q_θ, Q_w, Q_d). Then NMG transfers the depth image to the point cloud through camera intrinsics **c** for region aggregation under the guidance of heatmaps. Feature fusion and the point encoder extract regional features fused with semantic information from GHM. Finally, a multi-grasp generator combined with a novel non-uniform anchor sampling mechanism utilizes the fusion features to output the grasps.

on neighborhood similarity, which makes heatmap generation more robust for each local region.

NMG utilizes the heatmaps generated in GHM as guidance to aggregate local points into graspable regions and detects grasps through a light-weighted point encoder in each region. The proposed non-uniform anchor sampling mechanism adopted in NMG improves grasp quality by fitting the ground truth distribution better. Furthermore, a novel semantic-to-point feature fusion module is applied to detect grasps more robustly.

B. Grasp Heatmap Model

GHM is an encoder-decoder model containing two output branches, the confidence branch aiming at constructing grasp confidence heatmap Q_c and the attribute branch aiming at generating attribute heatmaps (Q_θ, Q_w, Q_d). Inspired by [24], [25], we apply Gaussian encoding and the grid-based strategy to decouple this task on account of different peculiarities between heatmaps. As is shown in Fig. 4, the ground truth 6-Dof grasps are projected to the image plane and encoded to heatmaps $(\hat{Q}_c, \hat{Q}_\theta, \hat{Q}_w, \hat{Q}_d)$.

The Gaussian encoding strategy adopts a 2D Gaussian kernel to encode the projected grasp ground truth centers before training. This approach effectively highlights grasp centers without ignoring nearby pixels, because pixels nearby will also serve as helpful guidance for further grasp detection. The value of pixel (u, v) in the confidence heatmap for training can be calculated by

$$q = \exp\left(-\frac{(u-u_0)^2 + (v-v_0)^2}{2\sigma_g^2}\right), \tag{1}$$

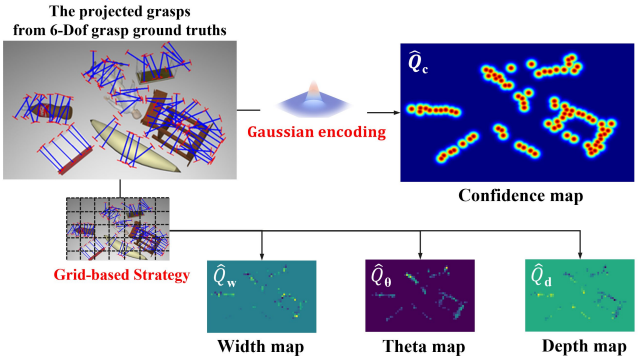

Fig. 4. Visualization of how the ground truth 6-Dof grasps are projected. Grasp confidence heatmap $\hat{Q}_c$ and attribute heatmaps $(\hat{Q}_\theta, \hat{Q}_w, \hat{Q}_d)$ are encoded with Gaussian kernel and grids, respectively.

where (u_0, v_0) denotes the center point of a grasp ground truth, and σ_g is the standard deviation which depends on the width of each grasp. Supervised by $\hat{Q}_c$, the confidence branch applies pixel-wise classification to predict the Q_c.

The proposed grid-based strategy encodes and predicts grasp attributes (θ, w, d) within a specific local grid instead of direct pixel-wise regression. Grasp attributes usually have high similarity in these areas due to the similar geometric structure. Thus, by fully exploiting the similarity of adjacent grasps, more robust grasp attribute prediction can be accomplished. Concretely, the full-scale image is divided into $H_r \times W_r$ grid cells with side length r. Based on the oriented anchor box mechanism in [25], [26], for each grid cell, k_a multiple oriented anchors are introduced with uniformly sampled angles. Therefore, the ground truth θ can be assigned to the nearest anchor. We obtain the number distribution of anchors in each grid, and a sigmoid function is applied afterward to

acquire $\hat{Q}_\theta$. Besides, we calculate the average normalized w, d in grids to obtain ground truth attribute heatmaps $(\hat{Q}_w, \hat{Q}_d)$. Supervised by these heatmaps, in a patch-wise manner, the attribute branch predicts Q_θ through the combination of anchor classification and offset regression, and estimates (Q_w, Q_d) via direct regression.

Previous methods like [1], [2], [19] encode grasps as pixel-wise rectangles, suffering from two defects. First, they fail to highlight the importance of the most considerable grasping probability at the center point [3]. Second, ground truth attribute heatmaps $(\hat{Q}_\theta, \hat{Q}_w, \hat{Q}_d)$ are not smooth as the confidence heatmap $\hat{Q}_c$ due to the relatively dense grasp annotations in cluttered scenes. In contrast, the designed GHM can highlight grasp centers and predict more robust grasp attributes, especially in cluttered scenes.

C. Non-uniform Multi-Grasp Generator

NMG takes the heatmaps and the scene point cloud as input and aggregates multiple graspable local areas efficiently under the guidance of the heatmaps. Subsequently, utilizing the grasp attributes in each grid, NMG predicts the remaining grasp rotation attributes and refines the former generated ones by the local features to generate multiple grasps. The proposed non-uniform anchor sampling mechanism improves the grasp quality and the novel semantic-to-point feature fusion contributes to the robustness of the detected grasps. According to different functions, the overall structure of NMG can be split into two parts, Heatmap-Guided Region Aggregation, and Non-uniform Multi-Grasp Generator.

*1) **Heatmap-Guided Region Aggregation**:* The first part of NMG processes heatmaps and the point cloud into useful local features, consisting of two steps: the region aggregation and the feature fusion.

Firstly, the region aggregation aggregates local points into graspable regions under the guidance of heatmaps from GHM for the subsequent multi-grasp generator. Concretely, the grasp confidence heatmap is downsampled with bilinear interpolation to $H_r \times W_r$, which is the same shape as the attribute heatmaps. Then top k_{center} grids with the highest predicted confidence are selected, containing k_{center} local peaks in total as regional centers. This grid-based selection suppresses center density to reduce the aggregation of duplicate areas. During training, k_{center} is set to a larger number to ensure most graspable local regions are extracted. During inference, it is convenient to adjust k_{center} to achieve grasp detection with different coverage rates.

Afterward, we transform the pixel centers (u, v) with corresponding depth d to point centers (x, y, z) in 3D space using camera intrinsics $\mathbf{c}$. As Fig. 3 suggests, a ball query [27] is utilized to crop the points within a sphere with a radius of the predicted grasp width w for each center. In each local ball region, N_g points are sampled by farthest point sampling to reduce the complexity of subsequent calculations while maintaining the local geometric information to the maximum extent.

Additionally, pixel-wise features extracted in GHM contain rich semantic information, which can reasonably supplement

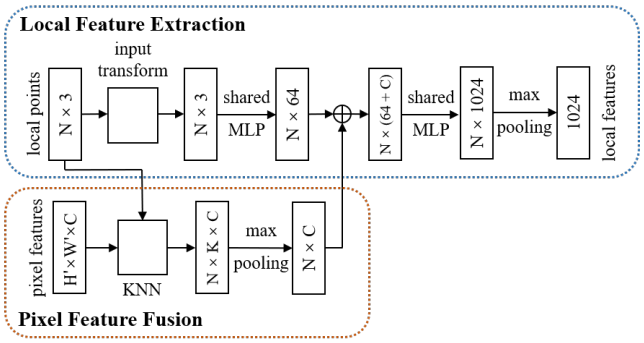

Fig. 5. The pipeline of local region feature extraction with semantic-to-point feature fusion.

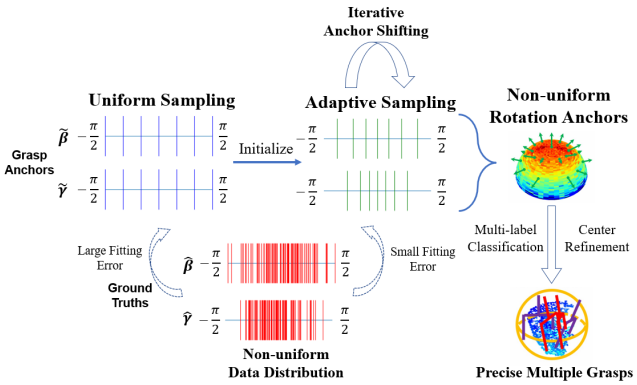

Fig. 6. Visual illustration for the procedure of the anchor shifting algorithm and the multi-grasp generation.

local point clouds. Therefore, we design a new light-weighted PointNet-based [27] network with semantic-to-point feature fusion for local feature extraction. The overall process is illustrated above in Fig. 5, where pixel features are grouped via a KNN operation to each local point. Then we integrate the pooled features with point features by point-wise concatenation for further feature extraction. By conducting KNN grouping and combining shared MLP (Multi-Layer Perceptrons) with max-pooling, we consider both local geometric and semantic information in the following grasp generator, increasing the grasp robustness, especially when point cloud input is unreliable.

*2) **Non-uniform Multi-Grasp Generator**:* The second part of NMG adopts anchor-based methods to detect grasps by the local features, consisting of the non-uniform anchor sampling and the multi-grasp generator. As Fig. 6 suggests, we generate non-uniform rotation anchors via iterative anchor shifting and conduct multi-label classification in each local region to generate multiple grasps. Notably, center refinement is applied to alleviate the errors resulted from the first stage and acquire more precise results.

With the local regions aggregated, only the 2D rotation (γ, β) has to be determined. Because both angles are continuous in $[-\frac{\pi}{2}, \frac{\pi}{2}]$, anchor-based methods prove to achieve better localization accuracy than direct regression [21]. Most former works predefine approach vectors uniformly on a sphere surface for grasp rotation prediction [8], [18], while

the actual distribution proves to be uneven. Hence, there is an inevitable trade-off between rotation prediction accuracy and time efficiency, which means denser anchors provide better accuracy but slower speed. Unlike methods before, we present a novel anchor-shifting algorithm applied in the training process, gradually shifting our anchors to minimize the fitting error between anchors and the acquired grasp rotation distributions, successfully attaining higher performance with fewer anchors. To simplify the problem, we consider γ and β equally and take γ as the example. The goal of the anchor shifting for the cumulative K grasps and k_r defined anchors $\tilde{\boldsymbol{\gamma}}$ can be formulated below:

$$\tilde{\boldsymbol{\gamma}}^*, \mathbf{B}_\gamma^* = \underset{\tilde{\boldsymbol{\gamma}}, \mathbf{B}_\gamma}{\arg\min} \left\| \mathbf{B}_\gamma^T \tilde{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}} \right\|_2^2, \tag{2}$$

in which our grasp anchors are defined as $\tilde{\boldsymbol{\gamma}} \in [-\frac{\pi}{2}, \frac{\pi}{2}]^{k_r \times 1}$, and $\mathbf{B}_\gamma \in \{0,1\}^{k_r \times K}$ represents one-hot encodings of the nearest anchor's index for each grasp ground truth. $\hat{\boldsymbol{\gamma}} \in [-\frac{\pi}{2}, \frac{\pi}{2}]^{K \times 1}$ is the rotation angles of the cumulative K selected grasp ground truths during the training procedure.

For dynamically and efficiently shifting our anchors during training, we solve the above problems via a block coordinate descent algorithm:

- Get updated anchor encodings $\mathbf{B}_\gamma^*$ according to $\tilde{\boldsymbol{\gamma}}$ by following simple comparison,

$$\mathbf{B}_\gamma^{(i,j)} = \begin{cases} 1 & \text{if } \underset{k \in \{1,\dots,k_r\}}{\arg\min} \left\| \hat{\boldsymbol{\gamma}}^{(j)} - \tilde{\boldsymbol{\gamma}}^{(k)} \right\| = i, \\ 0 & \text{else.} \end{cases} \tag{3}$$

- Fix $\mathbf{B}_\gamma$ and update $\tilde{\boldsymbol{\gamma}}$: Eq.(2) reduces to a linear regression problem which can be solved with the least square method as

$$\tilde{\boldsymbol{\gamma}}^* = \left(\mathbf{B}_\gamma \mathbf{B}_\gamma^T\right)^{-1} \mathbf{B}_\gamma \hat{\boldsymbol{\gamma}}. \tag{4}$$

We conduct anchor shifting during the whole training procedure, represented as the python-style pseudocode in Algorithm 1.

Algorithm 1 Non-uniform anchor shifting during training

Parameters: $\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\beta}} \in [-\frac{\pi}{2}, \frac{\pi}{2}]^{k_r \times 1}$ - current anchors
K - grasp number threshold, T - shifting iterations
Python-Style Pseudocode:

```
Grasps = list()
while training do
    G = GetGraspGroudTruthsInEachRegion()
    Grasps.extend(G)
    if len(Grasps) > K then
        γ̂, β̂ = Grasps.γ, Grasps.β
        for t = 1 → T do
            Get B_γ,t, B_β,t with γ̃_t-1, β̃_t-1 per Eq.(3)
            Update γ̃_t, β̃_t with B_γ,t, B_β,t per Eq.(4)
        end for
        Grasps.clear()
    end if
end while
```

At the same time with anchor shifting, supervised by local grasp ground truths, our multi-grasp generator takes the region aggregated features as input and combines anchors of the two angles to form a higher dimensional (k_r^2-class) multi-label classification problem. Then an MLP is utilized to generate the multi-label classification results and form multiple grasps in each local region. It is nonnegligible that the errors of the first stage, especially the errors of center localization, might affect the performance of the grasp generator. Thus, our multi-grasp generator not only predicts grasp rotation attributes but also refines grasp centers estimated in the first stage by regressing 3-dimensional center offsets for each anchor.

D. Loss & Implementation Details

*1) **Loss:*** The overall training objective of HGGD is the weighted sum of the heatmap losses and the anchor losses, formulated as:

$$L = \underbrace{L_{Q_c} + a \times L_{cls} + b \times L_{reg}}_{heatmap\ losses} + \underbrace{L_{anchor} + c \times L_{offset}}_{anchor\ losses},$$

where L_{Q_c} represents the pixel-wise cross-entropy loss between the predicted grasp confidence Q_c and the encoded ground truth $\hat{Q}_c$. We present a penalty-reduced focal loss [28] similar to [10] to match the Gaussian-based heatmap. Another focal loss L_{cls} is employed to supervise the multi-label classification [29] for θ learning. Moreover, a masked Smooth $L1$ loss L_{reg} is adopted for all the regression problems in GHM. L_{anchor} represents the local grasp rotation anchor classification loss calculated using focal loss, and L_{offset} is a Smooth $L1$ loss adopted to predicted grasp center offsets of different rotation candidates.

*2) **Implementation Details:*** We adopt ResNet-34 [30] as the backbone to build our pixel encoder for heatmap generation in GHM. The channel number is reduced to 128 for better inference efficiency. With extra skip connections, the pixel decoder is composed of stacking transposed convolution layers to upsample the feature map and fit the output shape as the heatmaps. We set the input image resolution to 640×360, the oriented anchor number k_a to 6, and the grid size r to 8. In NMG, during training, we set $k_{center} = 128$ to cover areas as much as possible, while during the inference, k_{center} is set to 32 and 48 for two datasets respectively. Anchor shifting iteration number T is set to 1 to get gentler anchor value movement. We aggregate local points with $N_g = 512$ for each region and generate multiple grasps with $k_r = 7$. End-to-end joint training is adopted.

V. DATASET & EVALUATION METRICS

A. Dataset

Grasp datasets can be roughly divided into real and synthetic according to the type of observations. GraspNet-1Billion [18] builds a large-scale grasp dataset in which the observations are captured in the real world. Although more than 1 billion grasp annotations are provided, it is still limited in the scale of objects, which are only 88 with similar shapes.

On the other hand, simulating observations provides a more scalable alternative [31]. Thus, similar to [7], [8], we build a large-scale simulated grasp dataset based on ACRONYM

TABLE I
RESULTS ON TS-ACRONYM DATASET

Method	CR ↑	CFR ↑	AS ↑	Time[1,2] (ms) ↓
GPD (3 channels) [4]	-	69.3 %	0.408	20342
GPD (12 channels)	-	72.9 %	0.412	19756
PointNetGPD [5]	-	74.4 %	0.434	10212
S4g [7]	0.177	83.2 %	0.618	432
REGNet [8]	0.296	94.3 %	0.662	441
HGGD	**0.503**	**98.2 %**	**0.686**	**28**

[1] Including network inference time and post-processing time.
[2] Evaluated with AMD 5600x CPU and single NVIDIA RTX 3060Ti GPU.

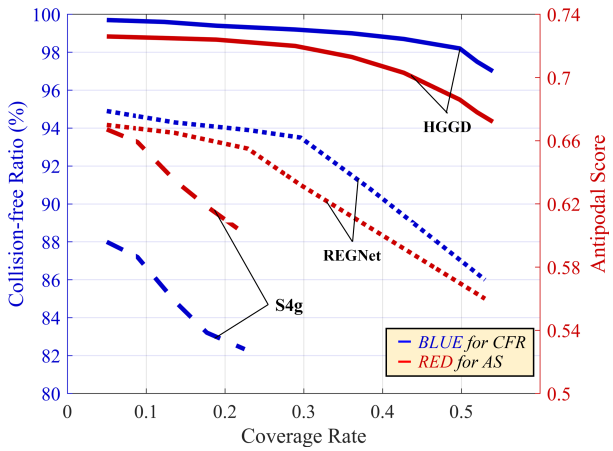

Fig. 7. (CFR, CR) and (AS, CR) curves. The red lines represent the Collision-Free Ratio and the blue lines represent the Antipodal Score. When CR increases, compared with baselines, HGGD still remains a relatively high grasp quality.

[31], containing mesh-based grasps generated with a physics simulator. We manually select 300 objects with various typical geometries from ShapeNetSem [32] with supplementary textures for dataset construction. Then, we construct 500 cluttered scenes, each containing 6-8 objects, render 50 synthetic RGBD images from random perspectives for each scene, and project collision-free grasp labels to the camera frame. Additionally, we calculate each grasp a quality score to describe the force closure property following [7]. We call this generated dataset TS-ACRONYM, which tells textured and scored. Besides, different from the [18], [31], TS-ACRONYM contains compact scene-level grasp labels instead of dense object-level grasp labels, which avoids extra data transformation during training and reduces storage consumption.

B. Evaluation Metrics

1) For TS-ACRONYM: Following S4g [33] and REGNet [8], we use **Collision-Free Ratio (CFR)** and **Antipodal Score (AS)** to evaluate the quality of each generated grasp. CFR describes the possibility of not colliding with the scene, and AS describes the force closure property. In addition to grasp quality, grasp diversity is important to achieve successful executable grasps. 6-Dof Graspnet [33] introduces **Coverage Rate (CR)**, which describes the diversity of the grasps and measures how well the generated grasps cover all ground truths. Nevertheless, CR defined in [33] only considers the distance between grasps, while the grasp rotation is also rather crucial. Thus, according to the rotation distance metrics elaborated in [34], in our work, grasp label $\hat{\boldsymbol{g}}$ is regarded as covered if predicted grasp $\boldsymbol{g}$ satisfies the following two conditions:

- $|\mathbf{t}_{\boldsymbol{g}} - \mathbf{t}_{\hat{\boldsymbol{g}}}| \leq 2\ cm$
- $1 - |\mathbf{q}_{\boldsymbol{g}} \cdot \mathbf{q}_{\hat{\boldsymbol{g}}}| \leq 0.1$

where $\mathbf{t}$ and $\mathbf{q}$ denote translations and quaternions of the grasps. Following [33], we also draw the curves of (CFR, CR) and (AS, CR) by sampling different grasp numbers for detailed analysis and evaluation.

2) For GraspNet-1Billion: GraspNet-1Billion proposes the **Average Precision (AP)** as an evaluation metric, which calculates the friction coefficient of the top 50 grasp poses by force-closure metric after non-maximum suppression.

VI. EXPERIMENT

We evaluate our framework both in simulation, including synthetic dataset (TS-ACRONYM), real dataset (GraspNet-1Billion), and in real robot grasping.

A. Performance Evaluation

Firstly, we compare HGGD with state-of-the-arts on TS-ACRONYM, including GPD [4], PointNetGPD [5], S4g [7] and REGNet [8]. As illustrated in Table I, HGGD significantly outperforms other methods on CR, AS and CFR metrics, which indicates that HGGD can generate more precise and dense grasps. Meanwhile, Fig. 7 suggests that HGGD has a more acceptable descending on grasp quality when CR increases. As for time efficiency, owing to its light-weighted architecture and avoiding redundant points processing, HGGD can detect grasps at a real-time speed, around 28 ms, much faster than any previous works.

Our model trained on the GraspNet-1Billion dataset is evaluated by AP and compared with other methods. As illustrated in Table II, on the premise of ensuring the method's efficiency, our approach achieves nonnegligible AP performance gains on all the seen, similar and novel dataset split compared with REGNet, demonstrating the effectiveness of HGGD. With comparable performance with Graspness [20], HGGD is nearly three times faster and can run in real-time and performs better in difficult scenarios ($AP_{0.4}$, which means lower friction factor) and unseen scenes.

As is shown in Fig. 8, the visualization results also prove that HGGD can predict denser and higher-quality grasps than REGNet, benefiting from the effectiveness of the proposed global-to-local and semantic-to-point scheme.

B. Ablation Studies

To objectively analyze the role of each module in our method, we build the baseline framework with a random center selection strategy, single-label classification on uniformly sampled anchors and no center refinement for generated grasps. Then we apply the proposed modules to the baseline in order, and experiments show the effectiveness of our method.

TABLE II
DETAILED RESULTS ON GRASPNET DATASET, SHOWING APS ON REALSENSE/KINECT SPLIT AND METHOD TIME USAGE

Method	Seen			Similar			Novel			Time[1]
	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	/ms
GPD [4]	22.87/24.38	28.53/30.16	12.84/13.46	21.33/23.18	27.83/28.64	9.64/11.32	8.24/9.58	8.89/10.14	2.67/3.16	-
PointnetGPD [5]	25.96/27.59	33.01/34.21	15.37/17.83	22.68/24.38	29.15/30.84	10.76/12.83	9.23/10.66	9.89/11.24	2.74/3.21	-
GraspNet-1B [18]	27.56/29.88	33.43/36.19	16.95/19.31	26.11/27.84	34.18/33.19	14.23/16.62	10.55/11.51	11.25/12.92	3.98/3.56	296
RGB Matters [19]	27.98/32.08	33.47/39.46	17.75/20.85	27.23/30.40	36.34/37.87	15.60/18.72	12.25/13.08	12.45/13.79	5.62/6.01	440
REGNet [8]	37.00/37.76	- / -	- / -	27.73/28.69	- / -	- / -	10.35/10.86	- / -	- / -	452
TransGrasp [35]	39.81/35.97	47.54/41.69	36.42/31.86	29.32/29.71	34.80/35.67	25.19/24.19	13.83/11.41	17.11/14.42	7.67/5.84	-
GSNet [20]	**67.12/63.50**	**78.46/74.54**	60.90/**58.11**	**54.81/49.18**	**66.72/59.27**	**46.17/41.89**	24.31/**19.78**	**30.52/24.60**	14.23/11.17	∼100[2]
HGGD	64.45/61.17	72.81/69.82	**61.16**/56.52	53.59/47.02	64.12/56.78	45.91/38.86	**24.59**/19.37	30.46/23.95	**15.58/12.14**	**36**

"-": Result Unavailable.
[1] Evaluated with AMD 5600x CPU and single NVIDIA RTX 3060Ti GPU.
[2] Reported in [20] on NVIDIA RTX 1080Ti GPU since the code is not available.

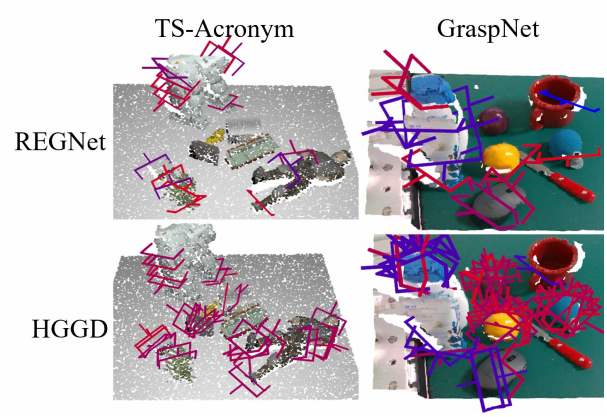

Fig. 8. Qualitative results on TS-Acronym and GraspNet-1Billion datasets. Grasps are color-coded based on their test (antipodal/force-closure) scores in RGB space, with red indicating better quality and blue indicating lower quality.

TABLE III
ABLATION ANALYSIS OF EACH MODULE

TS-ACRONYM	CR ↑	CFR ↑	AS ↑
baseline	0.144	59.7 %	0.338
+ heatmap guidance	0.450	96.9 %	0.656
+ center refinement	0.467	97.5 %	0.669
+ non-uniform anchor	0.481	97.8 %	0.679
+ multi-label classification	0.498	**98.2** %	**0.686**
+ feature fusion	**0.503**	**98.2 %**	**0.686**

As illustrated in Table III, the baseline method shows abysmal performance on each evaluation metric as expected. When the heatmap guidance is offered for local region aggregation, all performance metrics increase significantly, which proves that heatmap guidance is crucial for the grasp detection pipeline to excavate the graspable regions. Ablation experiment for the center refinement module in NMG proves that local geometric features can remarkably increase grasp location precision and reduce the influence of prediction errors in GHM. Then, by adapting anchors during training, HGGD successfully reduces the anchor fitting error and improves detected grasp quality, especially in CR and AS. Beneficial from the non-uniform anchor sampling mechanism, extending traditional single-label classification to the multi-label one is straightforward, enabling HGGD to detect more than one potential graspable rotation in each local region. Multi-label classification successfully improves the grasp quality and diversity of HGGD to a higher level.

TABLE IV
ABLATION ANALYSIS OF METHOD ROBUSTNESS

TS-ACRONYM with extra noise	CR ↑	CFR ↑	AS ↑
REGNet	0.159	92.5 %	0.629
HGGD w/o feature fusion	0.464	97.5 %	0.636
HGGD	**0.469**	**97.9 %**	**0.653**

TABLE V
RESULTS OF ROBOTICS EXPERIMENTS

Scene	Object	Success	Attempt
1	9	9	10
2	8	8	8
3	10	10	11
4	8	8	9
5	9	9	10
6	8	8	8
7	10	10	10
Success Rate[1]	62 / 66 = **94%**		
Completion Rate[2]	7 / 7 = **100%**		

[1] The sum of **Attempt** dividing the sum of **Success**.
[2] The total scene number dividing the successfully cleared scene number.

Notably, when the point cloud is unreliable, it is difficult for point-cloud-only methods to mine adequate information for grasp detection. In this circumstance, the semantic information from GHM is helpful due to its containing shape information of objects. To verify this, we apply an extra considerable Gaussian noise to the input point cloud and depth image, significantly reducing the performance of methods without feature fusion, which is quite evident in Table IV. Applying feature fusion in NMG can better recover helpful information

from the corrupted local point cloud for grasp detection, avoiding more significant performance loss by fully leveraging the semantic and geometrical representations.

C. Training Efficiency

Benefiting from the proposed global-to-local framework, our method can infer in real-time and be trained efficiently with few key grasp ground truths after non-maximum suppression. Average ~ 500 grasps per scene ($\sim 2\%$ of all labels for GraspNet-1Billion dataset) are sufficient for HGGD to achieve state-of-art performance. Based on its efficient framework and refined dataset organization, HGGD can be fully trained on a single RTX 3090 GPU in 4 hours.

D. Real Robot Experiments

We also conduct real robot grasping experiments in cluttered scenes on UR-5e with a Robotiq 2-finger parallel-jaw gripper. Realsense-D435i is used to acquire single-view RGBD images. Similar to the experiment procedure of previous works [7], [8], [35], we prepare 20 objects of various shapes and sizes used in daily life, then randomly select 8-10 of them for each scene and place them on the table in random poses. The robot performs grasp detection and executes grasping for each scene until no grasp is detected or 15 attempts are tried. We adopt the **Success Rate** and **Completion Rate** to evaluate the performance.

Results in Table.V reports the performance of HGGD, indicating that HGGD can generalize to the real world to generate high-quality grasps efficiently. Some failures occurred when grasping transparent objects like glass because of the unreliable depth image captured.

VII. CONCLUSION

This paper presents a novel end-to-end 6-Dof grasp poses detection framework in cluttered scenes. Through the global-to-local and semantic-to-point scheme, HGGD achieves state-of-the-art performance in two representative datasets, much faster than all previous methods. However, our framework is single-view-based and open-loop, inhibiting applications in more complex scenarios. In the future, we intend to utilize it for closed-loop grasp detection, endowing robots to adjust predicted grasp poses while approaching the target object and reacting to the changing environment tactfully.

REFERENCES

[1] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[2] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 9626–9633.

[3] S. Yu, D.-H. Zhai, Y. Xia, H. Wu, and J. Liao, "Se-resunet: A novel robotic grasp detection method," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5238–5245, 2022.

[4] 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.

[5] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 3629–3635.

[6] P. Ni, W. Zhang, X. Zhu, and Q. Cao, "Pointnet++ grasping: Learning an end-to-end spatial grasp generation algorithm from sparse point clouds," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 3619–3625.

[7] 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," in *Conference on robot learning*, 2020, pp. 53–65.

[8] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "Regnet: Region-based grasp network for end-to-end grasp detection in point clouds," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 474–13 480.

[9] W. Wei, Y. Luo, F. Li, G. Xu, J. Zhong, W. Li, and P. Wang, "Gpr: Grasp pose refinement network for cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4295–4302.

[10] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.

[11] D. Osokin, "Real-time 2d multi-person pose estimation on cpu: Lightweight openpose," *arXiv preprint arXiv:1811.12004*, 2018.

[12] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[13] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *2017 IEEE international conference on robotics and automation (ICRA)*, 2017, pp. 1386–1383.

[14] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3343–3352.

[15] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun, "Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 632–11 641.

[16] 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.

[17] S. Wang, X. Jiang, J. Zhao, X. Wang, W. Zhou, and Y. Liu, "Efficient fully convolution neural network for generating pixel wise robotic grasps with high resolution images," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, 2019, pp. 474–480.

[18] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[19] M. Gou, H.-S. Fang, Z. Zhu, S. Xu, C. Wang, and C. Lu, "Rgb matters: Learning 7-dof grasp poses on monocular rgbd images," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 459–13 466.

[20] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, "Graspness discovery in clutters for fast and accurate grasp detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.

[21] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, 2023.

[22] Á. González, "Measurement of areas on a sphere using fibonacci and latitude–longitude lattices," *Mathematical Geosciences*, vol. 42, no. 1, pp. 49–64, 2010.

[23] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *2011 IEEE International conference on robotics and automation*, 2011, pp. 3304–3311.

[24] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: Keypoint triplets for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6569–6578.

[25] X. Zhou, X. Lan, H. Zhang, Z. Tian, Y. Zhang, and N. Zheng, "Fully convolutional grasp detection network with oriented anchor box," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 7223–7230.

[26] A. Depierre, E. Dellandréa, and L. Chen, "Scoring graspability based on grasp regression for better grasp prediction," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 4370–4376.

[27] 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.

[28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[29] M.-L. Zhang and Z.-H. Zhou, "A review on multi-label learning algorithms," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 8, pp. 1819–1837, 2013.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[31] C. Eppner, A. Mousavian, and D. Fox, "Acronym: A large-scale grasp dataset based on simulation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 6222–6227.

[32] M. Savva, A. X. Chang, and P. Hanrahan, "Semantically-enriched 3d models for common-sense knowledge," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 24–31.

[33] 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.

[34] D. Q. Huynh, "Metrics for 3d rotations: Comparison and analysis," *Journal of Mathematical Imaging and Vision*, vol. 35, no. 2, 2009.

[35] Z. Liu, Z. Chen, S. Xie, and W.-S. Zheng, "Transgrasp: A multi-scale hierarchical point transformer for 7-dof grasp detection," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 1533–1539.