

Rethinking 6-Dof Grasp Detection: A Flexible Framework for High-Quality Grasping

Pengwei Xie^{1,†}, Siang Chen^{1,3,†}, Wei Tang^{2,†}, Dingchang Hu¹, Wenming Yang², Guijin Wang^{1,3,*}

Abstract—**Robotic grasping** is a primitive skill for complex tasks and is fundamental to intelligence. For general 6-Dof grasping, most previous methods directly extract scene-level semantic or geometric information, while few of them consider the suitability for various downstream applications, such as target-oriented grasping. Addressing this issue, we rethink 6-Dof grasp detection from a grasp-centric view and propose a versatile grasp framework capable of handling both scene-level and target-oriented grasping. Our framework, *FlexLoG*, is composed of a *Flexible Guidance Module* and a *Local Grasp Model*. Specifically, the *Flexible Guidance Module* is compatible with both global (e.g., grasp heatmap) and local (e.g., visual grounding) guidance, enabling the generation of high-quality grasps across various tasks. The *Local Grasp Model* focuses on object-agnostic regional points and predicts grasps locally and intently. Experiment results reveal that our framework achieves over 18% and 23% improvement on unseen splits of the GraspNet-1Billion Dataset. Furthermore, real-world robotic tests in three distinct settings yield a 95% success rate.

I. INTRODUCTION

Robotic grasping is fundamental to various complex robotic manipulation tasks across various fields, including manufacturing, service industries, and medical assistance. Despite its critical importance, the efficiency and quality of grasp detection across diverse scenarios remain unsatisfactory, particularly for target-oriented grasping.

Recent advances in deep learning have enabled data-driven methods for generating grasps of objects without 3D models. Representative methods [1], [2] generate planar grasps similar to rotated object detection and achieve good performance in simple scenarios with high efficiency. However, this representation inherently constrains the gripper to be perpendicular to the camera plane, limiting its applicability in specific contexts.

Accordingly, 6-Dof grasping has drawn extensive attention because of its broader applications. Pioneer methods employ a sampling-evaluation strategy [3], [4] or a direct regression paradigm [5], [6] to predict grasp attributes, which are either time-consuming or inaccurate. Advanced works [7]–[10] excavate local features based on global information, implicitly split the grasp generation into two stages: *where*

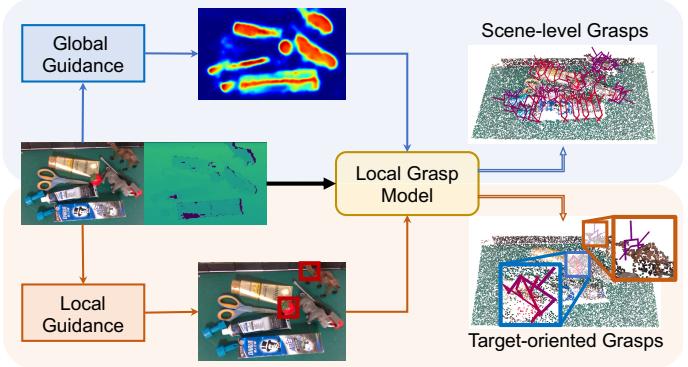


Fig. 1. Our framework can be flexibly integrated with global or local guidance methods for different application scenarios. Global guidance, such as grasp heatmap, can be utilized to generate scene-level grasps. Local guidance, such as object detection, can be utilized to generate target-oriented grasps.

and *how* [9]. The *where* stage encodes the global information to provide scene-level guidance helping to locate grasps, and then the *how* stage generates refined grasps based on it. Despite the impressive performance, these approaches are flawed in two ways. Firstly, they are limited to integrating only with scene-level guidance in their *where* stage, without considering the suitability for different downstream applications, such as target-oriented grasping aiming to grasp specific objects or parts. For target-oriented grasping, [11]–[13] first generate scene-level grasps and apply grasp filtering. These approaches introduce unnecessary computational load and risk interference from irrelevant objects, potentially leading to low-quality grasps or even no grasp in the targeted area. Secondly, the quality of their generated grasps is still unsatisfying, especially for unseen objects.

Different from these scene-level methods, we rethink the 6-Dof grasp detection problem in a grasp-centric view and turn our attention from the scene-level to the region-level. We propose a novel flexible framework capable of handling both scene-level and target-oriented grasping. Concretely, our framework, named *FlexLoG*, is composed of a *Flexible Guidance Module* and a *Local Grasp Model*. The *Flexible Guidance Module* is compatible with both global and local guidance, aggregating multiple local regions. Subsequently, the *Local Grasp Model* processes the regional data points and efficiently predicts grasps within each region.

There are several highlights to reformulate this problem in a local grasp-centric view. Firstly, as shown in Fig. 1, our framework can be guided by either global or local guidance methods (e.g., scene-level heatmap [10], object-level

¹The Department of Electronic Engineering, Tsinghua University, Beijing 100084, China.

²The Department of Electronic Engineering, Shenzhen International Graduate School, Tsinghua University, Shenzhen 518071, China.

³Shanghai AI Laboratory, Shanghai 200232, China.

[†]Equal Contribution.

^{*}Correspondence: wangguijin@tsinghua.edu.cn

This work was partly supported by the Special Foundations for the Development of Strategic Emerging Industries of Shenzhen(Nos.CJGJZD20210408092804011&JSGG20211108092812020).

detection [14]), and generates high-quality scene-level grasps or target-oriented grasps efficiently according to demands. Secondly, the grasp-centric representation makes the network capable of learning grasp-related and object-agnostic geometric features, resulting in a sizeable 23% improvement for novel objects of GraspNet-1Billion [15]. Additionally, our pipeline exhibits greater flexibility and superior generalization capabilities in novel scenarios compared to other scene-level approaches, making it more adaptable and effective in a broader range of applications. Finally, we deploy our method to the robot and conduct real-world experiments under three distinct settings: Cluttered, Randomly Arranged, and Click-and-Grasp, achieving an impressive average success rate of 95%.

II. RELATED WORKS

A. Scene-level Grasping

Scene-level grasping means generating grasps for the entire scene in a target-agnostic manner. Some methods directly extract scene-level semantic or geometric information. [5], [6] utilize PointNet++ [16] to encode scene points and then directly regress grasp attributes. Recent methods in grasp generation predominantly adopt a two-stage approach: *where* and *how*.

Chen et al. [10] leverage ResNet [17] to generate heatmaps, guiding the identification of grasping locations in image space. Similarly, works like [7], [9], [15], [18], [19] utilize PointNet++ as an encoder to extract global per-point features, assigning each point a “graspness” or confidence score. Recognizing the importance of local information for effective grasping, these methods typically employ techniques like ball query or cylinder grouping to segment graspable regions around potential grasp centers. Chen et al. [10] further augment this process by integrating points within the graspable regions with corresponding semantic features. A vanilla PointNet [20] is then used to extract both geometric and semantic features. In contrast, other studies [7]–[9], [15], [18], [19] employ Multilayer Perceptrons (MLP) to encode deeper features based on the global per-point features extracted by PointNet++. Based on these extracted features, various mechanisms are developed to predict different grasp attributes tailoring to the specific grasp representation.

In contrast to the aforementioned scene-level methods that generate scene-wide grasps, we reformulate the 6-Dof grasp detection problem in a grasp-centric view. FlexLoG, our innovative framework, stands as the first to predict high-quality grasps solely based on local information. This focus on local data significantly improves the framework’s ability to generalize to unseen objects, a benefit stemming from the object-agnostic characteristics of the regional points.

Besides, other methods such as Ten et al. [21], and Liang et al. [4] follow a sample-evaluation strategy and also sample grasps locally. Compared to our streamlined approach, their reliance on complex, hand-crafted features is more time-consuming and results in lower grasp quality.

B. Target-Oriented Grasping

Recently, several approaches [11]–[13] focus on grasping specific objects in cluttered by integrating an additional segmentation branch. Such a strategy, though practical, often results in excess computational load and cannot consistently yield high-quality grasps for the intended target. [22] utilizes a visual grounding algorithm for object detection, using bounding box filters to obtain object-level point clouds. These clouds are then processed by a network [23] trained on scene-level data, which can result in suboptimal grasping due to a domain mismatch with the training data.

Different from them, our framework, which is trained on regional grasp-centric data, can directly generate grasps on the target parts and seamlessly integrate various guidance methods, including object detection [24], [25] and instance segmentation [26], [27]. This approach allows for direct processing of local point clouds rather than the whole scene, eliminating redundant computations. The object-agnostic nature of our method ensures higher grasp quality on target objects. Besides, our framework can readily incorporate part affordance concepts [28], [29] to facilitate grasping specific object parts.

III. PROBLEM FORMULATION

Previous approaches generate scene-level grasps [1], [10], [15], processing the whole scene information directly. Different from them, we reformulate this problem to region-level from a grasp-centric view. Assuming multiple grouped regions have been obtained from different guidance methods, we aim to predict grasps within each region, which can be formulated as

$$G_p = \Phi(\mathcal{T}(f_i | i = 1, \dots, K)),$$

where $f_i \in R^{N \times 3}$ is the i -th grouped region which is cropped from the scene point cloud and centered at (x_p, y_p, z_p) in the camera frame, as shown in Fig. 2(A). $\mathcal{T}(\cdot)$ means canonical transformation (transform points to

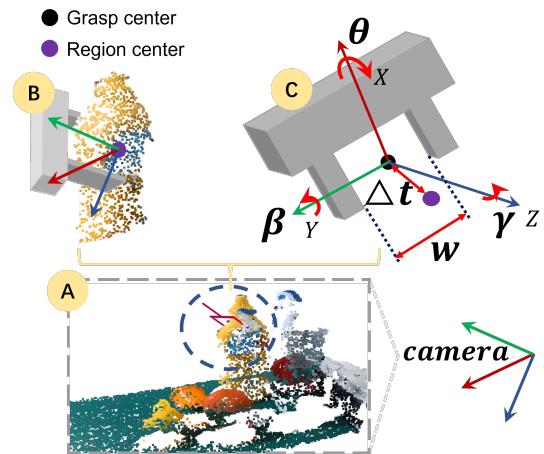


Fig. 2. A: The region is cropped from the scene point cloud in the camera frame. B: The local neighbor points are transformed to the local region frame. C: The regional grasp representation as $(\theta, \gamma, \beta, w, \Delta x, \Delta y, \Delta z)$.

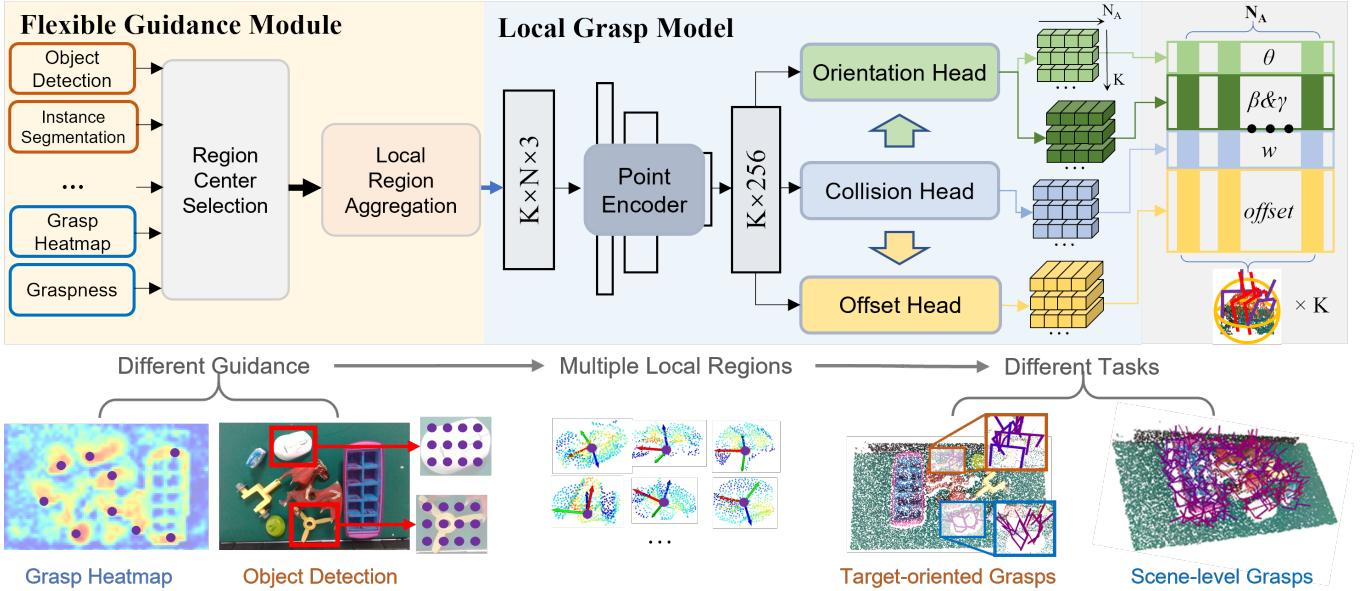


Fig. 3. The architecture of FlexLoG. Taking a monocular observation image as input, the Flexible Guidance Module (FGM) utilizes different guidance methods (e.g., grasp heatmap for global guidance and object detection for local guidance) to identify potential graspable areas and sample points as regional centers. These points are then clustered into multiple local regions. The Local Grasp (LoG) Model then extracts geometric features and predicts grasps. Depending on the guidance method used in the FGM, the output is either scene-level or target-oriented grasps.

local region frame from camera frame in this paper). And $\Phi(\cdot)$ denotes the Local Grasp Model. To better fit the task that predicts grasps within a region, one region-level grasp $g_p \in G_p$ is centered at (x_p, y_p, z_p) and defined as:

$$g_p = (\Delta x, \Delta y, \Delta z, \theta, \gamma, \beta, w).$$

As shown in Fig. 2(C), $(\theta, \gamma, \beta) \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ are grasp Euler angles in the gripper frame and w denotes the grasp width. Given that the guidance for grasp centers is often imprecise, there tends to be a misalignment between the actual grasp center and the regional center. To address this, we introduce an offset, $\Delta t = (\Delta x, \Delta y, \Delta z) \in R^3$, which quantifies the deviation. Once g_p has been determined, the final grasp g can be represented as:

$$g = (x_p + \Delta x, y_p + \Delta y, z_p + \Delta z, \theta, \gamma, \beta, w).$$

IV. METHOD

A. Overview

Our flexible 6-Dof grasp detection framework, **FlexLoG**, consists of two main components: a **Flexible Guidance Module** (FGM) and a **Local Grasp Model** (LoG). As is shown in Fig. 3, FGM leverages global or local guidance methods to sample potential grasp points. These points, designated as regional centers, are subsequently clustered with nearby points to form multiple distinct regions. LoG then extracts local geometric features from these regions and predicts grasps aligned with corresponding regional centers by employing three specially designed heads. The framework's region-level, grasp-centric methodology facilitates the integration of various guidance methods within the FGM and significantly enhances grasp quality, especially in unseen scenarios. LoG's

proficiency in identifying grasp-related and object-agnostic geometric features from regional points is a critical factor in this enhanced performance.

B. Flexible Guidance Module

The **Flexible Guidance Module** (FGM) aims to identify and aggregate local regions with high graspability from the entire scene or specific targets. According to different downstream tasks, we categorize the guidance methods into two types: **scene-level** and **target-oriented** guidance. As depicted in Fig. 3, for **scene-level** grasping, we mainly adopt two effective guidance methods, heatmap [10] and graspness [9], to generate high-quality scene-level grasps. Both of them serve to indicate point-wise graspability and facilitate the sampling of points with high confidence, which are then used as regional centers for further aggregation.

Regarding **target-oriented** grasping, local guidance methods such as object detection [24] and semantic segmentation [26] can be utilized to sample regional centers from the predicted bounding boxes or segmentation masks. Furthermore, our FlexLoG framework is also compatible with other versatile guidance methods, such as part affordance [28], [29], and even user-driven pixel selection through mouse clicks.

To obtain regional points in the graspable areas, we employ the ball query method in conjunction with Furthest Point Sampling (FPS) [20]. This technique isolates a specified number of points within a spherical area from the scene's point cloud, effectively clustering them into local regions abundant with geometric features. As demonstrated in Fig. 2, we then transform these points $f_i \in R^{N \times 3}, i = 1, \dots, K$, from the camera frame to the local region frame, resulting in

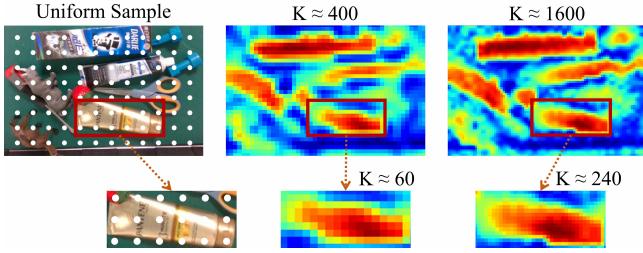


Fig. 4. Local grasp scores can be spliced to form a grasp heatmap, illustrating the graspability. As the number of sampled centers K increases, so does the heatmap’s resolution, leading to a more accurate depiction of the graspable areas.

a set of regional points $\mathcal{T}(f_i|i = 1, \dots, K)$, where K denotes the number of regions (i.e., sampled regional centers).

Addressing the challenge of generating scene-level grasps without external guidance, we employ a heuristic approach to analyze graspable areas in the scene. As shown in Fig. 4, in this scenario, we treat the entire scene as the whole graspable area, which is actually a special case of global guidance. We start by uniformly sampling pixels in a 2D mesh grid across the scene with a fixed grid size. Each grid center correlates to a specific point in 3D space, which is then utilized as a center for local region aggregation. Based on the aggregated regions, LoG can generate local grasps efficiently. Subsequently, the highest grasp scores within these regions are spliced to form a grasp heatmap, serving as a visual representation of the potential distribution of graspable areas. Notably, the mesh grid size is adjustable, allowing for a tailored balance between the quality of grasp prediction and time efficiency.

C. Local Grasp Model

From a grasp-centric view, each grasp is fundamentally influenced by the information of its surrounding region. To this end, we develop the **Local Grasp Model** (LoG), a robust network designed for extracting local geometric features and predicting multiple regional grasps. While existing studies [7], [9], [15], [18] utilize feature extractors for local geometry, they predominantly rely on global features from PointNet++. The more recent HGGD [10] is more similar to our proposed framework but still requires necessary pixel-wise grasp attribute prediction from global images. Our LoG, however, is the first model capable of predicting high-quality grasps solely utilizing local information, represented as $\mathbf{g}_p = (\theta, \gamma, \beta, w, \Delta x, \Delta y, \Delta z)$. This advanced capability aligns seamlessly with the functionalities of our FGM.

LoG processes the transformed local points $\mathcal{T}(f_i|i = 1, \dots, K)$ and predicts grasps for each region. As illustrated in Fig. 5, our proposed point encoder, inspired by PointMLP [30], is adept at local geometric feature extraction and achieves a better balance between efficiency and precision. With the local features extracted, LoG employs three specialized heads – **Collision Head**, **Orientation Head**, and **Offset Head** – all based on Multilayer Perceptron (MLP) architecture, to predict

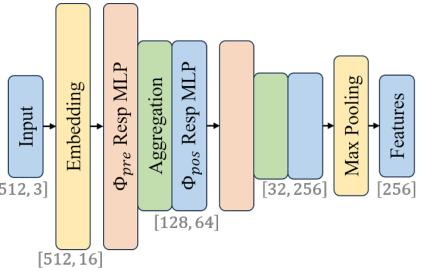


Fig. 5. The proposed light-weighted PointMLP-based encoder structure of Local Grasp Model.

various grasp attributes: w , (θ, β, γ) , and $(\Delta x, \Delta y, \Delta z)$. These attributes collectively form the 6-Dof grasps $\{\mathbf{g}_{p(i,j)}|i = 1, \dots, K, j = 1, \dots, N_A\}$ within each region, where N_A denotes the number of grasp anchors.

In the **Collision Head** of LoG, informed by a predefined range, grasp width prediction is approached as a regression problem to avoid collisions with nearby objects. These predicted widths, encapsulating vital local collision information, are subsequently utilized in the Offset and Orientation Heads. For the **Offset Head**, our approach mirrors that of HGGD, regressing the normalized grasp location offset along three axes, thereby refining grasp placements within each region. In the **Orientation Head**, we first predict the grasp in-plane rotation angle θ , a critical step due to its role as a precursor for spatial rotation angles (β, γ) in our Euler angle setting. The angle range $\theta \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ is discretized into $k_\theta = 6$ bins. We employ a combined approach of bin classification and residual regression within each bin to derive the final θ value. Furthermore, the predicted θ values are incorporated as contextual data for (β, γ) predictions. Following the non-uniform anchor sampling strategy [10], our strategy views spatial rotation prediction as a multi-label classification task. With 7 anchors each for (β, γ) , we generate up to $N_A = 49$ possible grasps per region and preserve those with the highest scores.

Compared with scene-level methods [7], [9], [10], [15], [18], our LoG exclusively focuses on regional data and local geometric feature extraction, which significantly enhances grasp detection quality. A vital aspect of our strategy involves the adoption of regional point clouds, which frequently lack complete object shape information, as the network input. This compels the network to adapt to learning object-agnostic features, consequently leading to significantly enhanced generalization, especially in unseen scenarios.

D. Regional Grasp-centric Dataset Generation

To effectively train the Local Grasp Model, we generate a new dataset containing regional point clouds and local grasp labels derived from existing scene-level annotations in [15]. The dataset creation involves selecting potential grasp centers and cropping local regions around them. Intuitively, sampling grasp center candidates randomly across the entire space is straightforward but proved inefficient, yielding a high proportion of invalid data. Alternatively, sampling centers

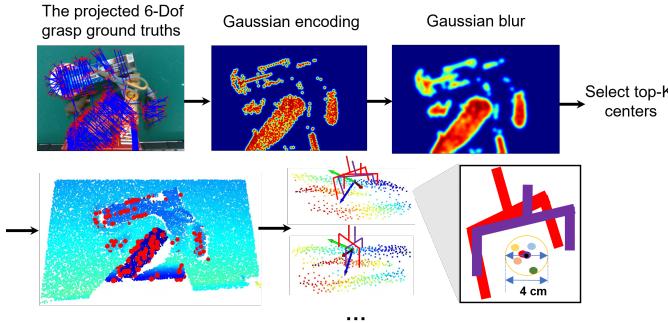


Fig. 6. The pipeline of region-level grasp-centric data generation. Grasp centers are sampled using the Gaussian-based strategy. Then, local neighbor points around each center are cropped as regions. Only the grasp labels within a radius of 2 cm from the region center are preserved.

directly from grasp label projections, as suggested in [10], leads to suboptimal results during inference due to domain shift caused by heatmap prediction deviation.

Addressing these challenges, we implement a Gaussian-based strategy for center selection, which introduces a small proportion of invalid data, naturally incorporating noise into the training process. As shown in Fig. 6, this process begins by projecting 6-Dof grasp ground truths to 4-Dof planar grasps. Each planar grasp center is encoded using a Gaussian kernel. Subsequently, we apply another Gaussian filter to create a blurred heatmap, depicting the distribution of graspable areas. Centers are then selected from this heatmap through grid-based sampling and converted into 3D points using corresponding depth values. For each center, a sphere with a radius between 6 and 12 cm – informed by the gripper width – is defined, and the ball query method [20] is applied to extract relevant points.

Label processing involves preserving only those grasp labels within a 2 cm Euclidean distance from the corresponding region centers, reinforcing the grasp-centric nature of our framework and dataset. This criterion ensures that LoG focuses on generating grasps near regional centers, which is critical for target-oriented grasping. Points and labels are then transformed from the camera frame to the local frame, normalizing the data distribution.

Through the above steps, we construct a substantial dataset of millions of grasp-centric and object-agnostic regional data, providing a basis for the LoG training.

E. Implementation Details

As depicted in Section. IV-C, our FlexLoG incorporates classification and regression components. Thus, the overall loss, which is equal to the regional grasp prediction loss L , could then be formulated with different loss terms as

$$L = a \times (L_{\theta}^{cls} + L_{\theta}^{reg}) + b \times L_w + c \times L_{offset} + d \times L_{anchor},$$

where L_{θ}^{cls} represents the cross-entropy loss for the bin classification of θ . A Smooth L1 loss L_{θ}^{reg} is adopted for the residual regression within one bin. Another two Smooth L1 loss L_w and L_{offset} are employed to supervise the regression of local grasp width and center offset. L_{anchor} represents the

multi-label classification loss calculated using focal loss [31] for the non-uniform anchor sampling strategy.

For FGM, we directly adopt the pretrained heatmap checkpoint [10] and the reimplemented graspsness model [9] as our scene-level guidance. With specific guidance, we aggregate $K = 48$ local regions by default. As for the uniform sampling without guidance, we uniformly sample points in the 2D mesh grid with grid size 12 pixels by default. However, in addition to the default sampling settings, considering the trade-offs between inference time and detection performance, it is practicable to adjust the center number of local regions in different scenarios.

For LoG, similar to the local region settings in [10], we aggregate local point clouds with $N = 512$ points via Furthest Point Sampling (FPS) for each region. Notably, the same anchor shifting strategy is employed to generate more accurate grasps.

V. EXPERIMENT

To thoroughly evaluate the performance of our proposed framework, we conduct experiments both on the dataset and the real robot platform.

A. Performance Evaluation

1) **Scene-Level Grasping:** To better compare overall performance with other methods, we firstly evaluate them in the scene-level grasping situation. GraspNet-1Billion dataset [15] is widely used in grasp detection as a standard evaluation platform for the task of general robotic grasping, containing RGBD images captured in the real world from 190 cluttered scenes and more than 1 billion grasp annotations. We utilize the method in Section. IV-D to generate the regional grasp-centric data based on the GraspNet-1Billion dataset. Around **6.5M** local regions are obtained from the training split to train the proposed LoG and compare its performance with other methods.

Following previous works, the **Average Precision (AP)** [15] evaluation metric is adopted, which adopts the same average precision calculation metric in object detection task for the force-closure scores [6] of the top 50 grasps after non-maximum suppression.

For the FGM, we adopt uniform sampling (without any guidance), point-wise graspsness guidance, and heatmap guidance to aggregate local regions and conduct scene-level grasp detection. As illustrated in Table I, FlexLoG makes significant gains and achieves the new state-of-the-art on the GraspNet-1Billion benchmark. It is worth noting that our heatmap-guided approach achieves **10.4/9.83** and **5.73/3.89** performance gains on the similar and novel splits compared to the previous state-of-the-art method Graspness, which demonstrates that our region-level and grasp-centric method has a more robust generalization to unseen scenarios. Furthermore, FlexLoG based on local region grasp generation is robust to different guidance inputs and manages to achieve high-quality grasp detection results even without any guidance, which indicates FlexLoG's potential for other target-oriented grasp scenarios.

TABLE I
RESULTS ON GRASPNET DATASET, SHOWING APs ON REALSENSE/KINECT SPLIT AND METHOD EFFICIENCY

| Method | Seen \uparrow | Similar \uparrow | Novel \uparrow | Average \uparrow | FPS \uparrow |
|---------------------------------------|-----------------------------|-----------------------------|-----------------------------|-----------------------------|----------------|
| GPD [3] | 22.87 / 24.38 | 21.33 / 23.18 | 8.24 / 9.58 | 17.48 / 19.05 | - |
| RGB Matters [32] | 27.98 / 32.08 | 27.23 / 30.40 | 12.25 / 13.08 | 22.49 / 25.19 | 2.3 |
| REGNet [7] | 37.00 / 37.76 | 27.73 / 28.69 | 10.35 / 10.86 | 25.03 / 25.77 | 2.2 |
| TransGrasp [18] | 39.81 / 35.97 | 29.32 / 29.71 | 13.83 / 11.41 | 27.65 / 25.70 | - |
| Graspness [9] | 67.12 / 63.50 | 54.81 / 49.18 | 24.31 / 19.78 | 48.75 / 44.15 | \sim 10 |
| Scale Balanced Grasp [33] | 63.83 / - | 58.46 / - | 24.63 / - | 48.97 / - | - |
| HGGD [10] | 59.36 / 60.26 | 51.20 / 48.59 | 22.17 / 18.43 | 44.24 / 42.43 | 28 |
| FlexLoG (<i>Uniform Sampling</i>) | 68.45 / <u>64.05</u> | 61.61 / <u>55.72</u> | 27.94 / 21.59 | 52.67 / <u>47.12</u> | 10 |
| FlexLoG (<i>Graspness Guidance</i>) | <u>69.23</u> / 62.86 | <u>62.40</u> / 53.22 | <u>29.63</u> / 25.15 | <u>53.75</u> / 47.08 | 9.0 |
| FlexLoG (<i>Heatmap Guidance</i>) | 72.81 / 69.44 | 65.21 / 59.01 | 30.04 / <u>23.67</u> | 56.02 / 50.67 | <u>26</u> |

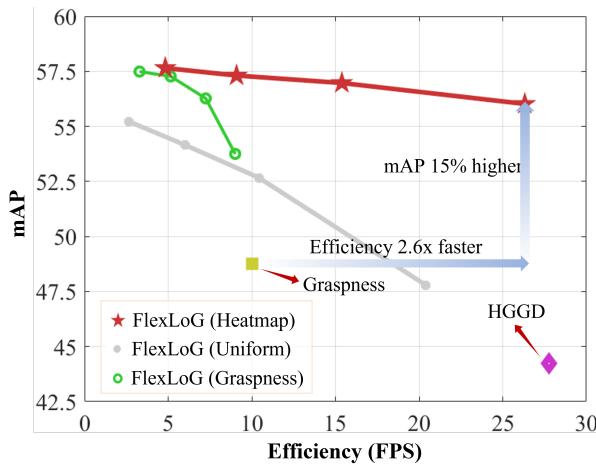


Fig. 7. The performance curve (mAP on all test scenes) on GraspNet dataset. Our framework with heatmap guidance outperforms previous state-of-the-art method Graspness with 2.6 times faster speed and 15% higher mAP.

Besides, we draw curves of (inference efficiency (FPS) - mAP) by sampling different numbers of candidate region centers to conduct detailed analysis and evaluation. With more candidate centers, methods detect grasps with higher mAPs and object coverage rates but at a slower speed. As shown in Fig. 7, our method with heatmap guidance achieves state-of-the-art grasp detection results at different speeds (center numbers). Even without any guidance, our method can still outperform other methods by a large margin. Thus, the experiment results prove that FlexLoG is a flexible framework for high-quality and real-time grasping.

2) **Target-Oriented Grasping:** To further evaluate the performance of our LoG for target-oriented grasping scenarios, we modified the evaluation metric from **AP** to **Target-Oriented Average Precision (TOAP)** and adjusted the evaluation process following the same procedure of TOGNet [34].

Specifically, for fair comparisons, we use the ground truth object segmentation mask from the GraspNet-1Billion dataset to randomly select one object, not fully occluded by others, as the target for each RGB-D image. Since there are

512 RGB-D images with different cameras and viewpoints for each scene, and approximately 10 objects per scene, all objects can be selected as targets. During evaluation, we compute the distance from the detected grasp centers to the target object mesh model, retaining only the grasps on the target. Finally, due to reduced grasp diversity, we compute force-closure scores [6] for the **top 10** grasps after non-maximum suppression and use scores under different friction coefficients to derive **TOAP**.

The evaluation results are shown in Table II. Notably, our LoG model achieves significantly higher results than other baseline methods and delivers comparable target-oriented grasping performance to TOGNet. However, unlike TOGNet, which utilizes both RGB and XYZ features, our LoG model uses only local XYZ features, making it more convenient for certain applications.

3) **Ablation Studies:** To further analyze the role of the proposed LoG, we design ablation studies on the local grasp model backbone. We mainly discuss the performance of the point encoder with different architectures. As illustrated in Table III, firstly, we change our encoder to PointNet [20] as the same as [10], which shows a faster speed but an unignorable performance drop. Then, we explore the influence of the PointMLP structure. *Wider PointMLP* means double the network width (embedding dimension), and *Deeper PointMLP* means double the network depth (network grouping and MLP layer number). Experiments show that a wider network slightly improves grasp detection performance but with a much slower speed. Moreover, a deeper and slower network cannot bring any increase in model performance. In conclusion, the overall results show the efficiency of our designed light-weighted PointMLP encoder for local grasp detection.

B. Real Robot Experiments

We also perform real-world robot grasping experiments using a UR-5e robot equipped with a Robotiq 2-finger parallel-jaw gripper. To capture single-view RGBD images, we employ a Realsense-D435i camera. Our experiment procedure follows that of previous studies, but we quantitatively

TABLE II
TARGET-ORIENTED APs ON GRASPNET DATASET (REALSENSE/KINECT)

| Method | Seen | Similar | Novel | Mean |
|---------------------------|----------------------|----------------------|----------------------|----------------------|
| GraspNet-baseline [15] | 22.64 / 13.28 | 20.63 / 12.67 | 8.35 / 3.85 | 17.21 / 9.93 |
| Scale Balanced Grasp [33] | 38.04 / - | 33.94 / - | 15.97 / - | 29.32 / - |
| HGGD [10] | 38.91 / 34.82 | 34.70 / 28.77 | 16.73 / 12.02 | 30.11 / 25.20 |
| Graspness [9] | 44.80 / 34.81 | 37.67 / 29.61 | 18.06 / 12.82 | 33.51 / 25.75 |
| TOGNet [34] | 51.84 / 49.60 | 46.62 / 40.03 | 23.74 / 19.58 | 40.73 / 36.40 |
| LoG | 50.57 / 44.67 | 44.59 / 39.37 | 22.59 / 16.04 | 39.25 / 33.36 |

"-": Result unavailable

TABLE III
MODEL ABLATION EXPERIMENTS, SHOWING APs ON
REALSENSE/KINECT SPLIT AND METHOD EFFICIENCY

| Method | Average \uparrow | FPS \uparrow |
|-----------------|-----------------------------|----------------|
| Ours | <u>56.02</u> / 50.67 | <u>26</u> |
| PointNet | 52.65 / 47.98 | 30 |
| Wider PointMLP | 56.32 / 51.86 | 21 |
| Deeper PointMLP | 54.48 / <u>51.02</u> | 20 |

TABLE IV

RESULTS OF ROBOTICS EXPERIMENTS

| Scene | Success Rate ¹ | Completion Rate ² |
|-----------------|---------------------------|------------------------------|
| Cluttered | 44 / 46 = 96 % | 6 / 6 = 100 % |
| Random Arranged | 48 / 51 = 94 % | 6 / 6 = 100 % |
| Click-and-Grasp | 48 / 51 = 94 % | 6 / 6 = 100 % |
| Total | 140 / 148 = 95 % | 18 / 18 = 100 % |

¹ Success / Attempt = Success Rate

² Cleared Scene / Total Scene = Completion Rate

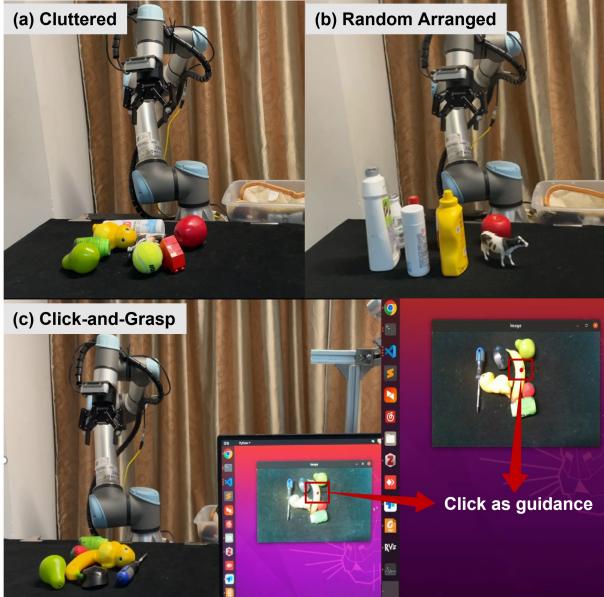


Fig. 8. Three settings of real-world robot experiments. (a) **Cluttered**: the poses of objects are randomized by shaking them in a box. (b) **Random Arranged**: arrange objects randomly and simulate the state of them in daily life. (c) **Click-and-Grasp**: the points of click are used as local guidance for target-oriented grasping.

evaluate our algorithms in three different settings: **Cluttered**, **Random Arranged**, and **Click-and-Grasp**. We assemble a collection of 30 objects with diverse shapes and sizes commonly encountered in everyday life, 15 of which have never been seen in the GraspNet-1Billion dataset. Subsequently, we randomly select and arrange 7 to 9 of these objects in each scene, placing them on the table in various orientations.

As illustrated in Fig. 8, to evaluate FlexLoG’s performance of *scene-level* grasping, we design the **Cluttered** and **Ran-**

dom Arrange settings. To evaluate the performance of *target-oriented* grasping, we demonstrate a **Click-and-Grasp** setting for simulating local guidance.

Specifically, for the **Cluttered** setting, following [4], we shake a box with all the objects in it, ensuring that the poses of all the objects are randomized. For the **Random Arranged** setting, we consider the general states of objects in daily life, such as a bottle standing upright. Then we arrange and organize these objects randomly on a tabletop. For the **Click-and-Grasp** setting, to evaluate the target-oriented grasping performance of LoG, we randomly select one object, click at a specific prehensile part, and use the points in the local region to generate grasps. Please see the supplementary video for the demonstration. We adopt the **Success Rate** and **Completion Rate** to evaluate the performance. Note that for the Click-and-Grasp setting, success is contingent upon successfully grasping the selected object.

It is worth noting that, in the **Cluttered** and **Random Arranged** settings, we employ a naive uniform sampling strategy instead of any other global guidance models. Table IV reports that our method achieves an average grasp success rate of 95% across 18 test scenarios, with a 100% scene completion rate. It indicates that the FlexLoG can generalize to the real world and generate high-quality grasps efficiently. Some failures are observed in situations where objects have smooth surfaces (e.g., the bottle) and the gripper slides over the object, or sometimes the gripper collides with other objects in the clutter.

VI. CONCLUSION

In this paper, we rethink the 6-Dof grasp detection problem and propose a flexible 6-Dof grasp framework, FlexLoG. Through a novel grasp-centric view, the designed Local Grasp Model can be integrated with either global or local

guidance methods for scene-level grasping or target-oriented grasping. Our framework achieves state-of-the-art performance on GraspNet-1Billion Dataset. Besides, the quantitative real-world robot grasping experiments demonstrate the effectiveness of our method. However, the LoG uses only point cloud, which is prone to generate low-quality grasps when the point cloud is of poor quality (e.g., transparent or reflective objects). In the future, we consider adding more semantic information to enhance robustness.

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)*. IEEE, 2020.
- [3] 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.
- [4] 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)*. IEEE, 2019.
- [5] 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)*. IEEE, 2020.
- [6] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, “S4g: Amodal single-view single-shot sc (3) grasp detection in cluttered scenes,” in *Conference on robot learning*. PMLR, 2020.
- [7] 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)*. IEEE, 2021.
- [8] 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)*. IEEE, 2021.
- [9] C. Wang, H.-S. Fang, M. Gou, H. Fang, J. Gao, and C. Lu, “Graspness discovery in clutter for fast and accurate grasp detection,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 964–15 973.
- [10] S. Chen, W. Tang, P. Xie, W. Yang, and G. Wang, “Efficient heatmap-guided 6-dof grasp detection in cluttered scenes,” *IEEE Robotics and Automation Letters*, vol. 8, no. 8, pp. 4895–4902, 2023.
- [11] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, “6-dof grasping for target-driven object manipulation in clutter,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020.
- [12] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, “Contact-grasnet: Efficient 6-dof grasp generation in cluttered scenes,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021.
- [13] Z. Liu, Z. Wang, S. Huang, J. Zhou, and J. Lu, “Ge-grasp: Efficient target-oriented grasping in dense clutter,” in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 1388–1395.
- [14] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in neural information processing systems*, vol. 28, 2015.
- [15] 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.
- [16] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, “Pointnet++: Deep hierarchical feature learning on point sets in a metric space,” *Advances in neural information processing systems*, vol. 30, 2017.
- [17] 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.
- [18] 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)*. IEEE, 2022.
- [19] X. Liu, Y. Zhang, H. Cao, D. Shan, and J. Zhao, “Joint segmentation and grasp pose detection with multi-modal feature fusion network,” in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 1751–1756.
- [20] 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.
- [21] A. Ten Pas and R. Platt, “Using geometry to detect grasp poses in 3d point clouds,” *Robotics Research: Volume 1*, pp. 307–324, 2018.
- [22] Y. Lu, Y. Fan, B. Deng, F. Liu, Y. Li, and S. Wang, “Vi-grasp: a 6-dof interactive grasp policy for language-oriented objects in cluttered indoor scenes,” *arXiv preprint arXiv:2308.00640*, 2023.
- [23] Y. Lu, B. Deng, Z. Wang, P. Zhi, Y. Li, and S. Wang, “Hybrid physical metric for 6-dof grasp pose detection,” in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022.
- [24] S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, et al., “Grounding dino: Marrying dino with grounded pre-training for open-set object detection,” *arXiv preprint arXiv:2303.05499*, 2023.
- [25] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, “Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7464–7475.
- [26] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al., “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.
- [27] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.
- [28] S. Deng, X. Xu, C. Wu, K. Chen, and K. Jia, “3d affordancenet: A benchmark for visual object affordance understanding,” in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1778–1787.
- [29] K. Mo, L. J. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, “Where2act: From pixels to actions for articulated 3d objects,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 6813–6823.
- [30] X. Ma, C. Qin, H. You, H. Ran, and Y. Fu, “Rethinking network design and local geometry in point cloud: A simple residual mlp framework,” *arXiv preprint arXiv:2202.07123*, 2022.
- [31] 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.
- [32] 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)*. IEEE, 2021.
- [33] H. Ma and D. Huang, “Towards scale balanced 6-dof grasp detection in cluttered scenes,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2004–2013.
- [34] P. Xie, S. Chen, D. Hu, Y. Dai, K. Yang, and G. Wang, “Target-oriented object grasping via multimodal human guidance,” *arXiv preprint arXiv:2408.11138*, 2024.