MSGeoGrasp: A 6-DoF Grasp Detection Based on Multi-Scale Feature and Geometric Enhancement

Yunpeng Guo[1], Jing Zhang[1,2*], Jie Su[1*], Junzheng Yang[2], Tianchi Zhang[3*]
[1] School of Information Science and Engineering, University of Jinan, Jinan 250022, China
[2] Yantai Institute of Science and Technology, Yantai 26560, China
[3] School of Information Science and Engineering, Chongqing Jiaotong University, Chongqing 400074, China
Email: 840292452@qq.com, ise_zhangjing@ujn.edu.cn, ise_sujie@ujn.edu.cn, zhangtianchi@cqjtu.edu.cn

***Abstract*—In recent years, with the increasing complexity of robotic manipulation tasks, 6-DoF grasp detection has become a key technology in robotic grasping. EconomicGrasp introduces an efficient economic supervision strategy, significantly reducing computational cost while improving grasp detection accuracy. However, when dealing with large-scale variations and severe occlusions, the method still suffers from limitations in fixed-scale neighborhood modeling and geometric representation. To address these issues, we propose MSGeoGrasp, a 6-DoF grasp detection method based on multi-scale feature modeling and geometric enhancement. We introduce multi-scale cylindrical local feature modeling to enhance scale adaptability, and employ geometric feature enhancement to improve geometric representation. Experimental results on GraspNet-1Billion demonstrate that MSGeoGrasp achieves superior performance while introducing only a negligible increase in model parameters.**

***Index Terms*—point cloud, 6-DoF grasp detection, multi-scale feature modeling, geometric feature enhancement**

I. Introduction

6-DoF grasp detection has been widely applied to robots and robotic arms for handling various tasks in industrial, service [1], and household [2] scenarios [3]. The core objective is to predict executable grasp poses from a single RGB-D image and evaluate and rank the predicted grasp candidates [4]. Compared with 2D grasping, 6-DoF grasping enables pose prediction in 3D space and exhibits greater flexibility in complex scenarios [5]. However, in practical applications, the model is required to generate stable and reliable grasp candidates while maintaining strong generalization capability under challenging conditions such as occlusions and clutter [6]. Therefore, improving prediction accuracy and generalization has become a central problem in this field.

In recent years, with the development of large-scale datasets and deep learning, 6-DoF grasp detection has achieved significant performance improvements [7]. GSNet [6] leverages dense supervision from the GraspNet-1Billion [8] dataset to achieve high grasping accuracy. However, such dense supervision models incur substantial training costs and often suffer from slow convergence. EconomicGrasp [9] introduces an efficient supervision strategy that reduces training cost while improving grasp accuracy. Nevertheless, the method still exhibits limitations in complex environments. Specifically, its fixed-scale cylindrical neighborhood struggles to handle small-scale objects, often introducing background and irrelevant points, which degrades feature extraction accuracy. In addition, the model mainly relies on implicit feature learning from point clouds, making it difficult to capture geometric information, thereby affecting grasp pose generation.

To address these issues, we propose MSGeoGrasp, a 6-DoF grasp detection network based on multi-scale feature modeling and geometric enhancement. Specifically, based on the EconomicGrasp framework, we introduce two key improvements. First, a multi-scale cylindrical local feature modeling module is proposed, which constructs cylindrical neighborhoods with different radii to extract features at multiple scales and fuses them via an attention mechanism to enhance scale adaptability. Second, a geometric feature enhancement module is introduced, which extracts geometric information including normals, curvature, and density via covariance decomposition and integrates them with the original features through a lightweight mapping to improve geometric representation.

The main contributions of this paper are summarized as follows:

- We propose a multi-scale cylindrical local feature modeling module to enhance scale adaptability.
- We propose a geometric feature enhancement module that integrates geometric features to improve geometric representation.
- We build the MSGeoGrasp network upon the EconomicGrasp framework to achieve effective fusion of multi-scale and geometric features.
- Extensive comparative experiments and ablation studies are conducted on the large-scale 6-DoF grasp detection dataset GraspNet-1Billion to validate the effectiveness of the proposed MSGeoGrasp network.

II. Related Work

A. 6-DoF Grasp Detection

Existing 6-DoF grasp detection methods are typically based on point cloud representations, where local features are extracted to generate grasp candidates and predict grasp poses. Early methods directly predict grasp poses from point clouds. PointNetGPD [10] introduces PointNet [11] style point clouds for local feature extraction. The large-scale GraspNet-1Billion [8] dataset and its benchmark method [12] provide a unified evaluation protocol for 6-DoF grasp detection.

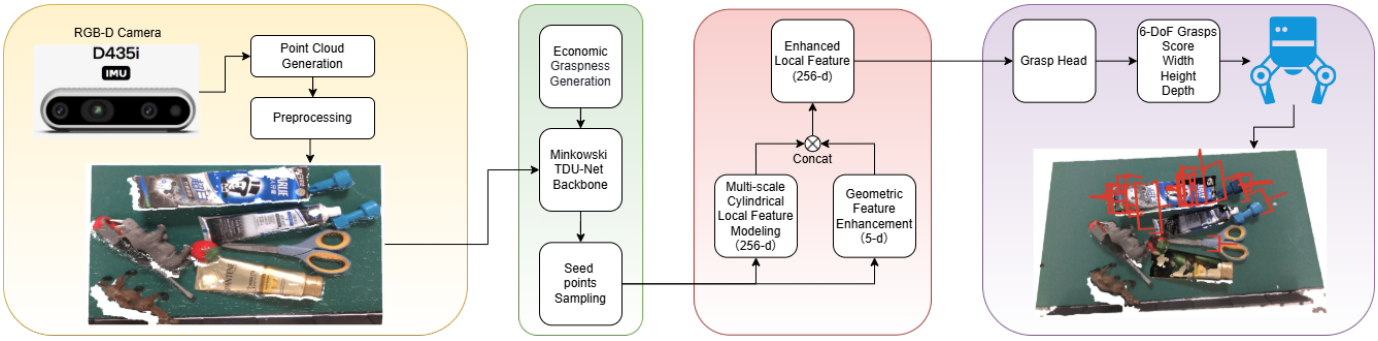

Fig. 1: Overall framework of MSGeoGrasp with multi-scale and geometric enhancement.

AnyGrasp [13] achieves efficient and robust grasp detection through spatio-temporal modeling of grasp features. Based on dense supervision, GSNet [6] introduces graspness to predict graspable regions, significantly improving performance in cluttered scenes. However, such methods typically rely on dense supervision and have high training costs. To address this issue, EconomicGrasp [9] proposes an efficient supervision strategy that maintains good performance while reducing training cost through grasp label pruning and related strategies.

B. Multi-scale Feature Modeling for 6-DoF Grasp Detection

In 6-DoF grasp detection, local feature modeling has an important impact on grasp pose generation. In real-world grasping scenarios, object scales vary significantly, and single-scale neighborhoods struggle to capture both detailed and global features of objects. KPConv [14] uses deformable convolution kernels on point clouds to extract features at different scales. TSB [15] addresses the scale imbalance problem through multi-scale modeling and scale-balanced learning. Wang et al. [16] enhance grasp perception for small-scale objects by introducing multi-scale cylindrical grouping and balanced sampling strategies.

C. Geometric Feature Modeling for 6-DoF Grasp Detection

Geometric feature modeling has an important impact on grasp stability and pose generation in 6-DoF grasp detection. Lu et al. [17] construct grasp confidence by integrating multiple physical constraints. FineGrasp [18] introduces surface normals as priors to improve grasping performance for small-scale objects. FlexLoG [19] proposes a locally guided grasping framework with multiple guidance strategies, improving both grasping performance and generalization ability.

III. Method

In this section, we present a detailed description of the proposed MSGeoGrasp 6-DoF grasp detection method. Section III-A provides an overview of the MSGeoGrasp framework; Section III-B introduces the multi-scale cylindrical local feature modeling module; and Section III-C further describes the design and implementation of the geometric feature enhancement module.

A. MSGeoGrasp Framework Overview

MSGeoGrasp is a 6-DoF grasp detection method based on multi-scale feature modeling and geometric enhancement. MSGeoGrasp is built upon the economic supervision framework of EconomicGrasp [9] and enhances the feature extraction stage of grasp candidates to address the limitation of single-scale cylindrical local feature modeling.

The overall framework of MSGeoGrasp is illustrated in Fig. 1. It follows the standard two-stage pipeline of 6-DoF grasp detection. First, the input point cloud is preprocessed, and grasp candidate points are sampled. Then, local neighborhoods are constructed around each candidate point for subsequent feature extraction and grasp pose prediction. In the local feature modeling stage, a multi-scale cylindrical local feature modeling module and a geometric feature enhancement module are introduced. Finally, the enhanced features are fed into the grasp head for pose regression and quality evaluation, producing executable 6-DoF grasp poses.

B. Multi-scale Cylindrical Local Feature Modeling

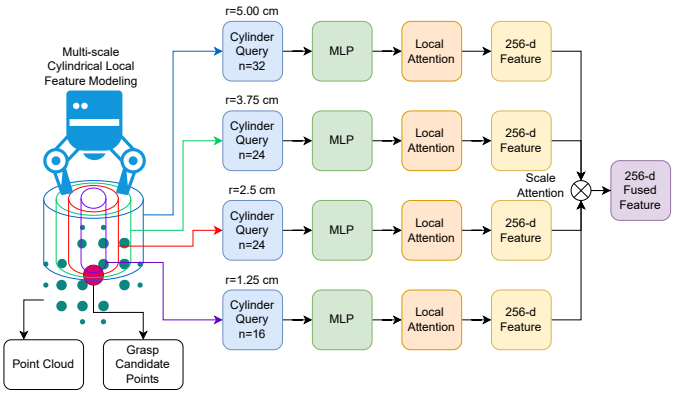

Fig. 2: Illustration of the multi-scale cylindrical local feature modeling module.

MSGeoGrasp introduces a multi-scale cylindrical local feature modeling module at grasp candidate points to enhance the model's perception of objects with different scales. This module models local point cloud structures using K cylindrical neighborhoods at different scales, enabling the model to

capture both local details and global information of objects. An illustration of the module is shown in Fig. 2.

Specifically, for a grasp candidate point p_c, K cylindrical neighborhoods $\{N_k\}_{k=1}^{K}$ with different scales are constructed in the grasp-view-aligned coordinate system. Each cylindrical neighborhood is defined by its radius r_k, height range $[h_{\min}, h_{\max}]$, and the number of sampled points n_k, and can be formulated as:

$$N_k = \{p_i \,|\, \|(p_i - p_c)_\perp\| < r_k,\ h_{\min} < h_i < h_{\max}\}. \quad (1)$$

Here, $(p_i - p_c)_\perp$ denotes the projection of $(p_i - p_c)$ onto the grasp plane, such that only contact points near the grasp center plane are considered. h_i represents the coordinate of point p_i along the local height axis, which constrains the reachable height range of the gripper.

In our implementation, we adopt $K = 4$ scales. The corresponding cylindrical radii r_k are set to $\{0.0125, 0.025, 0.0375, 0.05\}\,\mathrm{m}$, with the numbers of sampled points n_k set to $\{16, 24, 24, 32\}$, respectively. For all scales, the height range is uniformly set to $[-0.02, 0.04]\,\mathrm{m}$.

For each cylindrical scale, the local point set N_k is first selected from the point cloud according to the cylindrical constraint defined in Eq. (1), centered at the grasp candidate point in the view-aligned coordinate system. The relative coordinates of the points are then concatenated with their original features and fed into a MLP for point-wise encoding. Subsequently, a local self-attention module is introduced to enhance the correlations among points within the neighborhood. Finally, a pooling operation is applied to obtain the local feature representation at the current scale. In this way, each grasp candidate point is associated with a set of multi-scale local feature vectors $f_k \in \mathbb{R}^{256}$, which characterize object information at different scales.

During the scale fusion stage, an attention mechanism is employed to adaptively weight multi-scale features. For each scale feature f_k, a shared lightweight scale-attention function is first used to compute its corresponding score. These scores are then normalized across the scale dimension using a softmax operation to obtain the attention weights α_k. Finally, the fused multi-scale feature $f_{\text{ms}} \in \mathbb{R}^{256}$ is computed as a weighted sum of the scale-specific features:

$$\alpha_k = \frac{\exp\left(g(f_k)\right)}{\sum_{i=1}^{K} \exp\left(g(f_i)\right)}, \quad (2)$$

$$f_{\text{ms}} = \sum_{k=1}^{K} \alpha_k f_k. \quad (3)$$

This scale fusion mechanism allows the model to dynamically adjust the contribution of features from different scales according to object characteristics, enabling more flexible and effective multi-scale feature integration.

C. Geometric Feature Enhancement

After multi-scale feature fusion, MSGeoGrasp further introduces a geometric feature enhancement module to strengthen the model's ability to represent object geometric shapes. Geometric features are computed based on covariance analysis of point sets within multi-scale cylindrical neighborhoods, extracting normal, curvature, and density information, which is then jointly modeled with the fused multi-scale features. All geometric statistics are computed in the grasp-view-aligned coordinate system to ensure spatial consistency with the multi-scale features.

Specifically, for each grasp candidate point and its corresponding cylindrical neighborhood at scale k, the local point set within the neighborhood is first collected and transformed into a relative coordinate system centered at the candidate point, yielding the point set $\{\tilde{p}_i\}_{i=1}^{N_k}$. This process can be expressed as:

$$\bar{p} = \frac{1}{N_k} \sum_{i=1}^{N_k} \tilde{p}_i, \quad (4)$$

$$\hat{p}_i = \tilde{p}_i - \bar{p}. \quad (5)$$

Subsequently, the covariance matrix of the local point cloud is constructed as:

$$C = \frac{1}{N_k} \sum_{i=1}^{N_k} \hat{p}_i \hat{p}_i^\top + \epsilon I, \quad (6)$$

where ϵ is a numerical stability term and I denotes the identity matrix. Eigenvalue decomposition is then applied to the covariance matrix C, yielding the eigenvalues and their corresponding eigenvectors:

$$\lambda_1 \le \lambda_2 \le \lambda_3. \quad (7)$$

Based on these eigenvalues and eigenvectors, the following geometric features are constructed. The normal feature $\mathbf{n}$ describes the local surface orientation around the grasp candidate point. The eigenvector corresponding to the smallest eigenvalue λ_1 is regarded as the normal direction $\mathbf{n}$ of the local surface. The curvature feature c is used to characterize the flatness and edge properties of the local surface region:

$$c = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \lambda_3}. \quad (8)$$

The density feature d is defined using a geometric dispersion measure to describe the spatial distribution intensity of the local point cloud:

$$d = \log\left(1 + \frac{\lambda_1 + \lambda_2 + \lambda_3}{r_k^2}\right), \quad (9)$$

where r_k denotes the radius of the cylindrical neighborhood at scale k. The geometric feature vector at scale k is defined as:

$$\mathbf{g}_k = [n_x,\, n_y,\, n_z,\, c,\, d] \in \mathbb{R}^5. \quad (10)$$

Across multiple scales, geometric feature vectors from different scales are aggregated by average pooling to obtain the final geometric enhancement feature $\mathbf{g} \in \mathbb{R}^5$. The geometric feature $\mathbf{g}$ is then concatenated with the fused multi-scale feature f_{ms} to form a feature vector $f \in \mathbb{R}^{261}$, which is subsequently projected back to the original feature dimension through a lightweight linear mapping, yielding the final geometry-enhanced feature $f_{\text{geo}} \in \mathbb{R}^{256}$.

TABLE I: Performance comparison on GraspNet-1Billion under the RealSense setting. CD denotes collision detection. The best results without CD are highlighted in bold.

methods	seen			similar			novel			Average
	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	
GPD [20]	22.87	28.53	12.84	21.33	27.83	9.64	8.24	8.89	2.67	17.48
PointNetGPD [10]	25.96	33.01	15.37	22.68	29.15	10.76	9.23	9.89	2.74	19.29
GraspNet [8]	27.56	33.43	16.95	26.11	34.19	14.23	10.55	11.25	3.98	21.41
TransGrasp [21]	39.81	47.54	36.42	29.32	34.80	25.19	13.83	17.11	7.67	27.65
GraNet [22]	43.33	52.56	34.04	39.98	48.66	32.00	14.90	18.66	7.76	32.74
TSB [15]	58.95	68.18	54.88	52.97	63.24	46.99	22.63	28.53	12.00	44.85
GSNet [6]	65.70	76.25	61.08	53.75	65.04	45.97	23.98	29.93	14.05	47.81
EconomicGrasp [9]	68.21	79.60	63.54	61.19	73.60	53.77	25.48	31.46	13.85	51.63
MSGeoGrasp	**72.93**	**82.62**	**70.32**	**62.86**	**73.72**	**57.64**	**27.09**	**33.62**	**15.43**	**54.29**
MSGeoGrasp+CD	74.02	84.03	71.04	63.90	75.07	58.31	27.44	34.05	15.49	55.12

TABLE II: Performance comparison on GraspNet-1Billion under the Kinect setting. CD denotes collision detection. The best results without CD are highlighted in bold.

methods	seen			similar			novel			Average
	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	AP	$AP_{0.8}$	$AP_{0.4}$	
GPD [20]	24.38	30.16	13.46	23.18	28.64	11.32	9.58	10.14	3.16	19.05
PointNetGPD [10]	27.59	34.21	17.83	24.38	30.84	12.83	10.66	11.24	3.21	20.88
GraspNet [8]	29.88	36.19	19.31	27.84	33.19	16.62	11.51	12.92	3.56	23.08
TransGrasp [21]	35.97	41.69	31.86	29.71	35.67	24.19	11.41	14.42	5.87	25.70
GraNet [22]	41.38	49.84	33.86	35.29	43.15	26.89	11.57	14.31	5.24	29.41
TSB [15]	39.42	58.38	43.66	41.49	50.29	34.60	15.35	19.18	7.98	32.09
GSNet [6]	61.19	71.46	56.04	47.39	56.78	40.43	19.01	23.73	10.60	42.53
EconomicGrasp [9]	62.59	73.89	55.99	**51.73**	**62.70**	43.45	**19.54**	**24.24**	11.12	44.62
MSGeoGrasp	**64.69**	**74.05**	**60.79**	50.71	60.27	**43.58**	18.94	23.52	**11.15**	**44.78**
MSGeoGrasp+CD	66.41	76.18	62.03	52.50	62.56	44.80	19.90	24.72	11.60	46.27

IV. EXPERIMENTS

In this section, we conduct a comprehensive experimental evaluation of the proposed MSGeoGrasp 6-DoF grasp detection method. Section IV-A describes the experimental setup in detail; Section IV-B compares MSGeoGrasp with representative methods; Section IV-C presents ablation studies to analyze the contribution of each component; and Section IV-D provides qualitative visualizations of grasping results.

A. Experimental Setup

a) Dataset: We evaluate our method on the large-scale 6-DoF grasp detection dataset GraspNet-1Billion [8]. This dataset contains cluttered real-world scenes captured by two popular RGB-D sensors, RealSense and Kinect, across 190 scenes. It comprises 97,280 RGB-D images and over one billion grasp poses. Following the official training and testing split, the first 100 scenes are used for training, while the remaining 90 scenes are used for testing. The test set is further divided into three categories based on object characteristics: Seen, Similar, and Novel, each containing 30 scenes.

b) Baselines: We compare MSGeoGrasp with eight representative 6-DoF grasp detection methods, including GPD [20], PointNetGPD [10], GraspNet [8], TransGrasp [21], GraNet [22], TSB [15], GSNet [6], and EconomicGrasp [9]. All experiments follow the official evaluation protocol of the GraspNet-1Billion dataset, using average precision (AP) as the evaluation metric. The AP score is computed by averaging AP_μ over multiple friction coefficient thresholds $\mu \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$.

c) Implementation Details: Our model is implemented using the PyTorch framework, and both training and evaluation are conducted on a single Nvidia RTX 5070 Ti GPU. The Adam optimizer is adopted for training, with a cosine learning rate decay schedule. During the training stage of the multi-scale cylindrical local feature modeling module, the batch size is set to 4 with an initial learning rate of 0.001, and the model is trained for 10 epochs. Subsequently, the geometric feature enhancement module is fine-tuned based on the previous stage, using a batch size of 2 with an initial learning rate of 0.0003 for 8 epochs.

B. Comparing with Representative Methods

Table I and Table II report the quantitative comparison results of MSGeoGrasp against the eight representative methods on the GraspNet-1Billion dataset under the RealSense and Kinect settings, respectively, where the best results are highlighted in bold. MSGeoGrasp achieves competitive performance under both camera settings, with especially strong results on the RealSense setup.

Under the RealSense setting, MSGeoGrasp improves the average AP by 2.67% compared to EconomicGrasp, with the most significant gain of 4.72% observed in the Seen category. Improvements of 1.67% and 1.61% are achieved in the Similar and Novel categories, respectively. These results demonstrate that the proposed method can effectively improve grasp detection performance under high-quality point cloud conditions.

Under the Kinect setting, due to higher noise levels and increased sparsity in the point clouds, the overall performance

(a) EconomicGrasp (Scene 110) (b) EconomicGrasp (Scene 149) (c) EconomicGrasp (Scene 176)

(d) MSGeoGrasp (Scene 110) (e) MSGeoGrasp (Scene 149) (f) MSGeoGrasp (Scene 176)

Fig. 3: 6-DoF grasp pose visualization.

of all methods decreases. MSGeoGrasp achieves a 0.16% improvement in average AP over EconomicGrasp, with a notable gain of 2.10% in the Seen category, while slight decreases are observed in the Similar and Novel categories. These results indicate that the performance of the proposed method is affected under low-quality point cloud conditions, but MSGeoGrasp still maintains competitive performance in most scenarios.

C. Ablation Studies

We conduct systematic ablation studies on the GraspNet-1Billion dataset to evaluate the effectiveness of the multi-scale cylindrical local feature modeling and geometric feature enhancement modules. The results are presented in Table III.

TABLE III: Ablation study of MSGeoGrasp on GraspNet-1Billion under the RealSense/Kinect settings.

MS	Geo	Seen	Similar	Novel
		68.21/62.59	61.19/51.73	25.48/19.54
✓		71.48/**65.37**	62.13/**52.25**	27.16/**19.55**
	✓	70.07/65.08	60.91/51.21	25.24/19.35
✓	✓	**72.93**/64.69	**62.86**/50.71	**27.09**/18.94

The ablation results demonstrate that introducing the multi-scale cylindrical local feature modeling module leads to consistent performance improvements under both the RealSense and Kinect settings. Specifically, under the RealSense setting, the average AP increases from 51.62 to 53.59, while under the Kinect setting, it improves from 44.62 to 45.72. These results indicate that the proposed module effectively enhances the model's adaptability to object scale variations and remains effective even under low-quality point cloud conditions. When introducing the geometric feature enhancement module alone, improvements are observed in the average AP and the Seen category under both settings, while slight decreases occur in the Similar and Novel categories. This may be because geometric features are more sensitive to noise and distribution variations in point clouds.

We further evaluate the combined effect of the two modules. Under the RealSense setting, the two modules exhibit strong complementarity, with the average AP further improving from 53.59 to 54.29, and gains observed across all three categories. In contrast, under the Kinect setting, the combined improvement is limited. Although the multi-scale cylindrical local feature modeling module improves performance, further introducing geometric enhancement results in an overall performance drop. This phenomenon may be attributed to the higher noise and lower density of Kinect point clouds, which affects the reliability of geometric feature estimation.

D. Efficiency Analysis

Table IV presents the inference time comparison between MSGeoGrasp and EconomicGrasp. Here, MS denotes the multi-scale cylindrical local feature modeling module, and Geo denotes the geometric feature enhancement module. Although MSGeoGrasp introduces two additional modules at the feature extraction stage, the overall number of model parameters increases by only 7.18%, from 15.76M to 16.89M.

TABLE IV: Inference time comparison on GraspNet-1Billion under the RealSense/Kinect settings.

Methods	Inference Time(s)		
	Seen	Similar	Novel
EconomicGrasp	0.2795/0.2671	0.2819/0.2646	0.2827/0.2645
MSGeoGrasp	0.2887/0.2814	0.2885/0.2813	0.2865/0.2813

During inference, MSGeoGrasp achieves nearly the same runtime efficiency as EconomicGrasp, with inference time increasing by only 2.31% under the RealSense setting and 5.99% under the Kinect setting. Although multi-scale cylindrical local

feature modeling introduces additional computational overhead, this computation can be highly parallelized. As a result, in practical evaluations, the inference time of MSGeoGrasp is only slightly higher than that of the original model, while still meeting the requirements for real-time applications.

E. Qualitative Results

Fig. 3 presents qualitative comparisons of grasp poses generated by MSGeoGrasp and EconomicGrasp. One scene is selected from each of the Seen, Similar, and Novel categories in the GraspNet-1Billion dataset, and the top 30 grasp poses with the highest scores are visualized. Compared with EconomicGrasp, MSGeoGrasp produces more concentrated grasp poses that are better aligned with the object geometry. For objects with complex geometric structures, MSGeoGrasp is still able to generate valid grasp poses, demonstrating its superior performance in complex geometric scenarios.

V. Conclusion

In this paper, we propose MSGeoGrasp, a 6-DoF grasp detection method based on multi-scale feature modeling and geometric enhancement. Built upon the economic supervision framework of EconomicGrasp, MSGeoGrasp improves the perception of objects with varying scales and complex geometric structures by introducing a multi-scale cylindrical local feature modeling module and a geometric feature enhancement module. Experimental results on the GraspNet-1Billion dataset demonstrate that MSGeoGrasp achieves competitive performance under both RealSense and Kinect sensor settings. Future work will explore adaptive strategies for integrating multi-scale and geometric features under low-quality point cloud conditions, with the goal of further improving robustness and generalization across different sensors.

Acknowledgment

This research was funded by 2022–2025, the National Natural Science Foundation of China under Grant No. 52171310.

References

[1] H. Zhou, X. Wang, W. Au, H. Kang, and C. Chen, "Intelligent robots for fruit harvesting: Recent developments and future challenges," *Precision Agriculture*, vol. 23, no. 5, pp. 1856–1907, 2022.

[2] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation using low-cost whole-body teleoperation," in *8th Annual Conference on Robot Learning*, 2024.

[3] H. Zhang, J. Tang, S. Sun, and X. Lan, "Robotic grasping from classical to modern: A survey," *arXiv preprint arXiv:2202.03631*, 2022.

[4] Z. Xie, X. Liang, and C. Roberto, "Learning-based robotic grasping: A review," *Frontiers in Robotics and AI*, vol. 10, p. 1038658, 2023.

[5] C. Yuanwei, M. H. M. Zaman, and M. F. Ibrahim, "A review on six degrees of freedom (6d) pose estimation for robotic applications," *IEEE Access*, 2024.

[6] 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.

[7] A. D. Vuong, M. N. Vu, H. Le, B. Huang, H. T. T. Binh, T. Vo, A. Kugi, and A. Nguyen, "Grasp-anything: Large-scale grasp dataset from foundation models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 14 030–14 037.

[8] 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.

[9] X.-M. Wu, J.-F. Cai, J.-J. Jiang, D. Zheng, Y.-L. Wei, and W.-S. Zheng, "An economic framework for 6-dof grasp detection," in *European Conference on Computer Vision*. Springer, 2024, pp. 357–375.

[10] 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, pp. 3629–3635.

[11] 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.

[12] H.-S. Fang, M. Gou, C. Wang, and C. Lu, "Robust grasping across diverse sensor qualities: The graspnet-1billion dataset," *The International Journal of Robotics Research*, vol. 42, no. 12, pp. 1094–1103, 2023.

[13] H.-S. Fang, C. Wang, H. Fang, M. Gou, J. Liu, H. Yan, W. Liu, Y. Xie, and C. Lu, "Anygrasp: Robust and efficient grasp perception in spatial and temporal domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023.

[14] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[15] 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.

[16] H. Wang, Y. Zhang, Y. Wang, and J. Li, "6-dof grasp detection in clutter with enhanced receptive field and graspable balance sampling," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 3000–3007.

[17] 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, pp. 8238–8244.

[18] Y. Du, M. Zhao, T. Lin, Y. Jin, C. Huang, and Z. Su, "Finegrasp: Towards robust grasping for delicate objects," *arXiv preprint arXiv:2507.05978*, 2025.

[19] P. Xie, S. Chen, W. Tang, K. Yang, and G. Wang, "Rethinking 6-dof grasp detection: A flexible framework for high-quality grasping," *Pattern Recognition*, vol. 170, p. 112088, 2026.

[20] 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.

[21] 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, pp. 1533–1539.

[22] H. Wang, W. Niu, and C. Zhuang, "Granet: A multi-level graph network for 6-dof grasp pose generation in cluttered scenes," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 937–943.