2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Joint Segmentation and Grasp Pose Detection with Multi-Modal Feature Fusion Network

Xiaozheng Liu[1], Yunzhou Zhang[1*], He Cao[2], Dexing Shan[1], Jiaqi Zhao[1]

Abstract— **Efficient grasp pose detection is essential for robotic manipulation in cluttered scenes. However, most methods only utilize point clouds or images for prediction, ignoring the advantages of different features. In this paper, we present a multi-modal fusion network for joint segmentation and grasp pose detection. We design a point cloud and image co-guided feature fusion module that can be used to fuse features and adaptively estimate the importance of the point-pixel feature pairs. Moreover, we develop a seed point sampling algorithm that simultaneously considers the distance, semantics and attention scores. For selected seed points, we adopt a local feature aggregation module to fully utilize the local spatial features in the grasp region. Experimental results on the GraspNet-1Billion Dataset show that our network outperforms several state-of-the-art methods. We also conduct real robot grasping experiments to demonstrate the effectiveness of our approach.**

I. INTRODUCTION

Grasp pose detection is the basis of robotic manipulation that has promising applications. It has been proven challenging due to cluttered environments, sensor noise, and occlusions. Traditional methods [1], [2] use 3D models for physical analysis to generate grasp poses. However, it is infeasible and impractical to assume that all objects in new environments have been modeled. Data-driven methods aim to solve general grasp problems and generalize to unknown objects, including 2D planar grasping and 6-DoF grasp detection. The 2D planar grasping methods [3]–[7] use convolutional neural networks (CNNs) to extract image features and then predict 2D grasp rectangles on feature maps, greatly simplifying the grasping modeling. However, rectangles-based methods only allow the robot to approach objects from top to bottom, which is limited in application scenarios and less adaptable to complex environments.

Recently, 6-DoF grasp detection [8]–[15] has drawn much attention and developed rapidly. They utilize a PointNet-like [16] network to extract geometric features and then predict 6-DoF grasp poses around the seed points. 6-DoF grasp poses allow the robot to approach objects from any direction and adapt well to complex environments. However, we found that these approaches still have several potential drawbacks. a) Only geometric features are utilized in the prediction process. Semantic and geometric features are two complementary data. RGB images contain rich color information that can help robots quickly obtain the semantics of objects. Point clouds contain rich geometric information that can help robots understand the physical structure of different objects. However, existing 6-DoF grasp detection methods only use point clouds to extract geometric features and ignore the importance of RGB images. b) Traditional sampling algorithm contains many negative samples. The uniform sampling or farthest point sampling uses distance information to generate seed points, including some background and inappropriate grasp points. These negative samples are not helpful for the prediction and consume computing resources.

To solve these problems, we propose a multi-modal fusion network for joint segmentation and grasp pose detection. Specifically, we propose a point cloud and image co-guided feature fusion module to fuse the geometric and color information. ResNet [17] and PointNet++ [16] are used to extract the image and geometric features, respectively. Then, the co-guided fusion module performs feature fusion and highlights critical point-pixel feature pairs. The fused features are first used for foreground point segmentation, which is not sensitive to categories and can accurately distinguish tables and objects. Furthermore, we meticulously design a seed point sampling algorithm that combines segmentation results and attention scores. As a result, we can generate more positive seed points with high grasp confidence. The grasp pose is generated based on the grasp region around the seed point, and it is worth exploring to effectively utilize the local features of the grasp region. Therefore, we utilize a local feature aggregation module to obtain better local spatial features. The aggregated features are fed into the grasp pose predictor to generate final 6-DoF grasp poses. Experiments on the GraspNet-1Billion Dataset [11] and real robot systems demonstrate the effectiveness of our method.

To summarize, our main contributions are as follows:

- We propose a point cloud and image co-guided fusion module to obtain multi-modal features and adaptively estimate the importance of the point-pixel feature pairs.
- We propose a simple but effective sampling algorithm and feature aggregation module, which can generate high-quality seed points and fully use their local spatial features.
- We integrate these modules into a new grasping framework that outperforms several state-of-the-art methods on the GraspNet-1Billion Benchmark, especially for similar and novel scenes.

*The corresponding author of this paper.

[1]Xiaozheng Liu, Yunzhou Zhang, Dexing Shan, Jiaqi Zhao, are with College of Information Science and Engineering, Northeastern University, Shenyang, China (Email: zhangyunzhou@mail.neu.edu.cn).

[2]He Cao are with the Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China.

Supported by National Natural Science Foundation of China (No. 61973066) , Major Science and Technology Projects of Liaoning Province(No.2021JH1/10400049), Fundation of Key Laboratory of Equipment Reliability(No.WD2C20205500306), Fundation of Key Laboratory of Aerospace System Simulation(No.6142002200301).

II. RELATED WORK

A. Multi-Modal Feature Fusion

Contact-GraspNet [18] proposes a cascading approach that first segments objects from images and then generates grasp poses based on the corresponding point clouds. RGB Matters [19] predicts the orientation of the gripper from images and then calculates other grasp parameters based on point clouds. However, cascading methods need extra 2D annotations, and the 2D detector bounds their performance. The work done in [20] and [5] takes the depth image as an extra channel and then co-encodes it with the RGB image. However, compared with depth image, point clouds are more intuitive and contain more geometric information. FFB6D [21] designs a bidirectional fusion module that allows sufficient information fusion in the feature extraction process. However, it has problems with the problematic alignment of feature pairs in the middle fusion stage, which might generate improper feature matching. DenseFusion [22] extracts image features and point features individually and then fuses them in the final stage. Although DenseFusion ensures strict feature alignment, it only performs simple feature concatenation and does not fully fuse multi-modal information. Therefore, it is necessary to perform efficient multi-modal feature fusion while ensuring feature alignment.

B. Grasp Pose Detection

Jiang [3] firstly proposes the Cornell Dataset and searching-based algorithm to acquire high-confidence 2D grasp poses. GQ-CNN [23] generates 2D grasp poses and evaluates them by neural networks. Chu [24] proposes a grasp proposal network to predict 2D grasp poses on multi-object images. Recently, 6-DoF grasp pose detection has attracted much attention due to the maturity of RGB-D sensors. GPD [9] and PointNetGPD [8] propose a sampling-evaluation method that uses a neural network to evaluate the candidate grasp poses and select the best one. S^4G [10] uses PointNet++ to extract the geometric features and directly regress 6-DoF grasp poses based on point clouds. GraspNet-1Billion [11] proposes a large-scale object grasping dataset and designs a cascading network to predict grasp poses. A multi-task learning network with simultaneous instance segmentation and collision detection is proposed by SSCL [14]. GSNet [13] proposes a plug-and-play 'graspness' model that outperforms previous methods by a large margin. To extract long-distance contexts, TransGrasp [25] designs a multi-scale geometry encoding module. These methods only utilize single-modality information without considering the advantages of multi-modal information.

III. PROBLEM STATEMENT

Given a scene point cloud, the task of grasp pose detection is to predict suitable grasp poses. Similar to previous work [11], we define the grasp pose $\boldsymbol{P}$ as

$$\boldsymbol{P} = \begin{bmatrix} \boldsymbol{r} & \boldsymbol{t} & w \end{bmatrix} \tag{1}$$

where $\boldsymbol{r} = (r_x, r_y, r_z) \in \mathbb{R}^3$, $\boldsymbol{t} = (x, y, z) \in \mathbb{R}^3$ represent the orientation and translation of the two-finger gripper. $w \in \mathbb{R}$ represents the proper grasp width of the gripper. Fig. 1 shows the grasp pose representation.

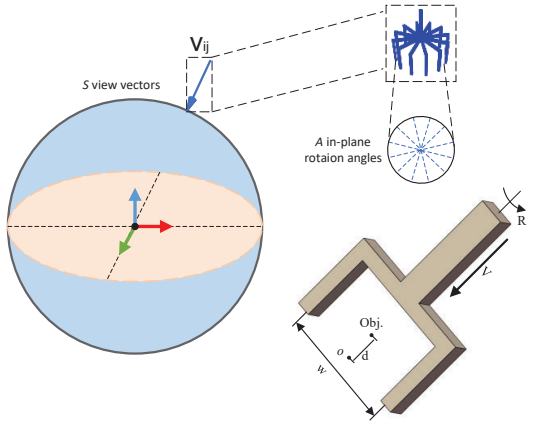

Fig. 1. Grasp representation. o represents the origin of the gripper. Object coordinates (obj.) represents the object point, the orientation of the gripper is decoupling into approaching vector V and in-plane rotation R, w represents the grasp width, d represents the grasp depth. V_{ij} represents the approaching vector of j^{th} virtual view of i^{th} graspable point.

IV. METHODS

Fig. 2 illustrates our multi-modal feature fusion network for joint foreground point segmentation and grasp pose detection. The network prediction process starts with foreground point segmentation on the fused features, followed by grasp pose prediction based on the selected seed points. Our proposed method consists of a point cloud and image co-guided feature fusion module, an effective seed point sampling algorithm, and a local feature aggregation module. More details are introduced as follows.

A. Co-Guided Feature Fusion Network

Feature Extraction. Point clouds and images have different data formats, making it challenging to fuse them correctly. A straightforward fusion approach is concatenating the XYZ channels with the RGB channels and then extracting information from the concatenated data. However, this fusion method is gradually replaced by other better fusion strategies [21], [22]. Inspired by these fusion methods, we design a novel fusion module. The point cloud stream and the image stream are used to extract geometric and color features separately. Given an aligned RGB-D image, we first convert the depth map into 3D point clouds using the camera intrinsic. Then, the point cloud stream utilizes PointNet++ [16] to extract point-wise geometric features. We adopt the segmentation version of PointNet++, which contains four set abstraction (SA) layers and four feature propagation (FP) layers. The SA layer downsamples the point cloud and extracts local features, and the FP layer upsamples the point features to restore the point cloud to its original size.

Simultaneously, the image stream utilizes ResNet [17] to extract pixel-wise image features. Our ResNet contains four convolutional blocks and deconvolutional layers. Each convolution block shrinks the feature map by half to expand the perceptual field and extract image features, and the

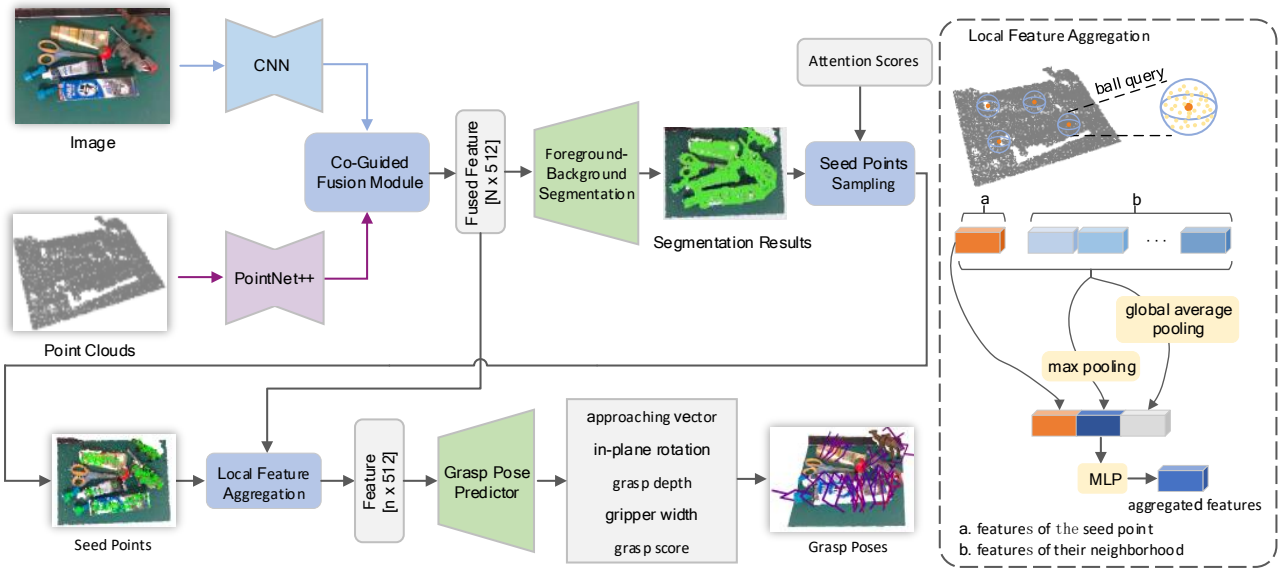

Fig. 2. The structure of our multi-modal fusion network. In the final stage of the backbone network, the co-guided fusion module fuses the extracted geometric and color features. Then, the segmentation network uses the fused features for foreground-background segmentation. Next, we perform seed point sampling, where the attention scores come from the co-guided fusion module. Our designed local feature aggregation module is shown on the right. Finally, we use the grasp pose predictor to generate grasp poses for each seed point.

deconvolution layer is used to recover the original image size. Finally, we obtain point-wise geometric features and pixel-wise image features, represented by F_{point} and F_{image}.

Co-Guided Fusion Module. One thing worth noting is how to find the correct correspondence between points and pixels. The aligned RGB-D image uses camera intrinsic to generate XYZ map [21], which forms a strict alignment. Therefore, the XYZ map can help point features get the corresponding image features, forming strict geometric-color feature pairs.

Furthermore, efficient fusion is difficult since different points have different importance. Points not in the region of interest may interfere with the prediction results. Instead, points with high grasp confidence will improve the predicted accuracy. Our designed co-guided feature fusion module is shown in Fig. 3, which can capture the critical information of each data modality. Specifically, we utilize the attention mechanism to suppress the interference and highlight important feature pairs. For each geometric feature and its color feature, we use fully connected layers to compress their channels and concatenate them together. We then use a fully connected layer to fuse the concatenated information, forming the basis for the point cloud and image co-guidance. To obtain attention scores, we normalize the features using the max pooling and sigmoid functions. Finally, we reweight the original feature pairs using attention scores and concatenate them together again to get the final fused features,

$$\begin{aligned} \text{AttenS} &= \sigma\left(\text{Max}\left(\text{MLP}\left(F_{point} \oplus F_{image}\right)\right)\right) \\ F_{fusion} &= \text{MLP}\left(\left(\text{AttenS} \otimes F_{point}\right) \oplus \left(\text{AttenS} \otimes F_{image}\right)\right) \end{aligned} \tag{2}$$

where AttenS represents the attention scores that indicates the importance of each feature pair, F_{fusion} represents final fused features.

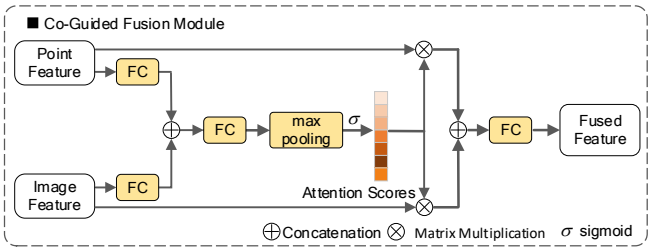

Fig. 3. Illustration of the co-guided fusion module. The attention scores will be used to reweight the extracted features and select seed points.

B. Seed Point Sampling and Local Feature Aggregation

After obtaining the fused features, we first perform a foreground-background segmentation task. There are many unseen object categories in unstructured environments. Foreground point segmentation is insensitive to object categories and can help the robot to obtain point-wise semantic labels, including tables and objects. The modeling for segmentation is implemented with a multi-layer perception (MLP) network. Specifically, taking the fused features $N \times 512$ as input, the MLP outputs a classification score $N \times 2$. The focal loss [26] is applied to supervise the segmentation task,

$$L_{seg}^{A} = -\alpha\left(1-c\right)^{\gamma}\log c \tag{3}$$

where α = 0.25 and γ = 2.0 are default hyper-parameters. c represents the probability of the point belonging to each category.

The Farthest Point Sampling (FPS) or uniform sampling is usually used to generate seed points covering the entire space as possible. However, these methods only utilize the distance information between points and do not fully utilize the attribute features, such as the semantic information and the importance of points. Therefore, we propose a simple but

effective sampling algorithm that simultaneously considers the distance, semantics, and significance, as shown in Algorithm 1. First, the FPS is used to sample λn candidate points in the foreground region, which allows the candidate points to cover the whole foreground. Then, we use the attention scores from the co-guided module to rank the candidate points. Next, the top n points of the attention scores among the candidate points are selected as seed points. As a results, we chose the seed points with better attributes and considered multi-modal information.

Algorithm 1: Seed point sampling

Input: Attention scores S, Foreground points K, Number of sampling points n.
Export: Candidate point set $C = \{\}$, Seed point set $P = \{\}$.
if $K < \lambda n$ **then**
  Randomly sample K until $K = \lambda n$
  Candidate point $C = K$
else
  Candidate point $C = K$
Use FPS for candidate points, sampling λn
Sort C_i by Attention scores S
for $i \leftarrow 0$ **to** n **do**
  $P_i = C_i$
Output: Seed point set P.

We then perform local feature aggregation for each selected seed point. We use ball query to find points within the grasp region and utilize max pooling and global average pooling to obtain the most representative features and global features, respectively. Next, these features are concatenated and fed into the MLP network to obtain richer local spatial features.

C. Grasp Pose Predictor

Based on the aggregated features, we predict the final 7-DoF grasp pose. Earlier research [14], [19] has demonstrated that it is challenging to regress the rotation matrix or quaternion directly. As a result, we decouple the orientation into approaching direction V and in-plane rotation R, and estimating these parameters is treated as a multi-class classification problem, as shown in Fig. 1. For each grasp point, we pre-define S approaching views in the sphere space and A in-plane rotation angles, totaling $S \times A$ classes of orientations. The approaching network [11] is used to predict the approaching vectors and graspable mask,

$$L^B\left(c_i, s_{ij}\right) = \frac{1}{N_{cls}} \sum_i L_{cls}\left(c_i, c_i^*\right) + \lambda_1 \frac{1}{N_{reg}} \sum_i \sum_j c_i^* 1\left(\left|v_{ij}, v_{ij}^*\right| < 5^\circ\right) L_{reg}\left(s_{ij}, s_{ij}^*\right) \tag{4}$$

where c_i and s_{ij} denote the predicted graspable score and confidence score for viewpoint j of point i. And correspondingly, c_i^* and s_{ij}^* denote the ground-truths. Indicator function $\mathbf{1}(\cdot)$ constrains the grasping loss on graspable points.

In addition, we utilize operation networks [11] to predict the remaining in-plane rotation R, grasp depth D and grasp width W, and grasp score S,

$$L^C\left(R_{ij}, S_{ij}, W_{ij}\right) = \sum_{d=1}^{K} \left(\frac{1}{N_{cls}} \sum_{ij} L_{cls}^d\left(R_{ij}, R_{ij}^*\right) + \lambda_2 \frac{1}{N_{reg}} \sum_{ij} L_{reg}^d\left(S_{ij}, S_{ij}^*\right) + \lambda_3 \frac{1}{N_{reg}} \sum_{ij} L_{reg}^d\left(W_{ij}, W_{ij}^*\right) \right) \tag{5}$$

where R_{ij} denotes the binned rotation degrees, S_{ij}, W_{ij} and d denote the grasp confidence scores, grasp width and grasp depth respectively. L^d means loss for the d^{th} binned distance.

V. EXPERIMENTS

A. Experiments on Dataset

1) Dataset and Evaluation Metric: The general object grasping dataset of GraspNet-1Billion contains 97,280 aligned RGB-D images, 100 training scenes and 90 testing scenes, over one billion grasp poses. The test scenes are divided into seen, similar and novel scenes that can verify the model's generalization. These images are captured in real environments. The widely used average *Precision@k* [11] is adopted as our evaluation metric, which evaluates the precision of top-k ranked grasp poses. Given a coefficient of friction $\mu \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$, AP_μ represents the average *Precision@k* for k ranges from 1 to 50.

2) Implementation Details: We randomly sample N = 20000 points to extract geometric features from point clouds. For corresponding RGB images, we perform pre-processing operations including resizing and data augmentation. The output size of the fused features is $N \times 512$. In the seed points sampling module, the number of seed points n is set to 2048, λ is set to 2 to ensure that the number of candidate points is greater than the number of seed points. S approaching views and A in-plane rotation angles are same as in [11] , grasp depth $d \in \{0.01, 0.02, 0.03, 0.04\}$. In loss functions, we set $\lambda_1, \lambda_2, \lambda_3$ = 0.5, 1.0, 0.2. We implemented our network with PyTorch on Nvidia RTX 3090 GPUs and the adam optimizer [27] is used to adjust model parameters.

3) Results, Analysis, and Comparison Experiments: We compare previous methods on the GraspNet-1Billion, and the results are shown in Table I. We achieve 1.2 **AP** and 1.33 **AP** gains on the similar and novel scenes, respectively. We believe that these two kinds of information can complement each other. Color information can help robots distinguish objects with similar geometric structures. Simultaneously, geometric information can help robots distinguish objects with similar appearances. Our method fully uses multi-modal information to facilitate foreground-background segmentation, which is category-insensitive and can better distinguish foreground points. Then, with the help of segmentation results and multi-modal features, we can sample more positive points and thus improve the accuracy of grasp pose detection,

TABLE I

GRASPING RESULTS ON GRASPNET-1BILLION DATASET.

Methods	Seen			Similar			Novel		
	AP	$AP_{0.8}$	$AP_{0.4}$	**AP**	$AP_{0.8}$	$AP_{0.4}$	**AP**	$AP_{0.8}$	$AP_{0.4}$
GG-CNN [4] (RSS 2018)	15.48	21.84	10.25	13.26	18.37	4.62	5.52	5.93	1.86
GPD [9] (IJRR 2017)	22.87	28.53	12.84	21.33	27.83	9.64	8.24	8.89	2.67
PointNetGPD [8] (ICRA 2019)	25.96	33.01	15.37	22.68	29.15	10.76	9.23	9.89	2.74
GraspNet [11] (CVPR2020)	27.56	33.43	16.95	26.11	34.18	14.23	10.55	11.25	3.98
RGB Matters [19] (ICRA 2021)	27.98	33.47	17.75	27.23	36.34	15.60	12.25	12.45	5.62
SSCL [14] (IROS 2021)	36.55	47.22	19.24	28.36	36.11	10.85	14.01	16.56	4.82
TransGrasp [25] (ICRA 2022)	**39.81**	**47.54**	**36.42**	29.32	34.80	**25.19**	13.83	17.11	**7.67**
ours	36.29	44.51	29.73	**30.52**	**36.57**	23.36	**15.34**	**18.24**	6.85

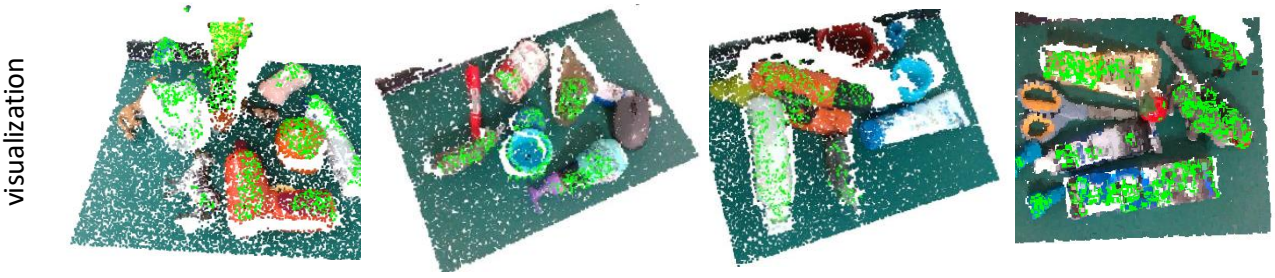

Fig. 4. Some visualization examples on GraspNet-1Billion Dataset. We visualize the selected seed points, and the generated grasp poses are concentrated in the seed point region.

TABLE II

COMPARISON EXPERIMENTS ON FUSION MODULES.

Fusion method	AP	$AP_{0.8}$	$AP_{0.4}$
DenseFusion	35.25	43.36	29.07
Ours	**36.29**	**44.51**	**29.73**

TABLE III

COMPARISON EXPERIMENTS ON SEED POINTS SAMPLING.

Method	AP	$AP_{0.8}$	$AP_{0.4}$
FPS	34.82	40.63	27.79
Ours	**36.29**	**44.51**	**29.73**

especially for similar and novel objects. Fig. 4 shows some visualization examples. It is obvious from the figure that the seed points are concentrated in places with high grasp confidence, such as the center of easily grasped objects.

We conduct comparison experiments on our framework. Firstly, we replace the fusion structure to verify the effectiveness of our co-guided fusion module. The fusion structures most relevant to us are FFB6D [21] and DenseFusion [22], which are fused in the middle and final stages, respectively. The image size is gradually reduced during the encoding process, which causes the point-pixel relationship cannot be strictly aligned at the middle fusion stages. This phenomenon might make FFB6D form improper feature matching and interfere with the prediction results. DenseFusion ensures strict feature alignment but only performs simple feature concatenation. Instead, our co-guided fusion module can both guarantee strict feature alignment and adaptively estimate the importance of each point to enhance or suppress the original feature pairs. The results are shown in Table II, and our co-guided fusion method is 1.04 **AP** higher than DenseFusion. In Table III, we study the effect of the seed point sampling algorithm. Compared with FPS, our sampling algorithm considers more attribute features of candidate points and has a significant improvement.

TABLE IV

RESULTS OF SINGLE OBJECT GRASPING EXPERIMENT.

Object	Attempt	Success	SR
Banana	10	10	100%
Screwdriver	10	7	70%
Toothpaste	10	8	80%
Peach	10	6	60%
Tape	10	10	100%
Average	10	8.2	82%

TABLE V

RESULTS OF CLUTTERED SCENES GRASPING EXPERIMENT.

Scene	Number	Attempt	SR
Scene1	4	5	80%
Scene2	3	3	100%
Scene3	3	5	60%
Scene4	5	7	71.4%
Scene5	4	7	57.1%
Average	3.8	5.4	70.4%

B. Experiments on Real Robot

We also conduct real robot grasping experiments for single objects and cluttered scenes. The grasping experiments are performed on an RM-65 robot with an Intel Realsense camera, as shown in Fig. 5. Our pipeline runs on an ubuntu 18.04 operating system with an NVIDIA RTX 2060 GPU. The MoveIt! is used for trajectory planning and control.

1) Single object scenes: We collect 15 different objects and select 5 of them for single-object grasping experiments. We randomly place these objects on the table and conduct 10 grasping experiments. Table IV shows the success rate.

2) Cluttered scenes: Grasping in cluttered scenes are challenging for robots and we conduct experiments in five

Fig. 5. Real-world robotic grasping experimental setup.

different scenes. Similar to single object grasping, we randomly select some objects for each scene and place them randomly. The grasp pose with the highest grasp score is selected and executed until all objects are cleared. The success rate are shown in Table V.

VI. CONCLUSION

We propose a point cloud and image co-guided feature fusion network for joint segmentation and grasp pose detection. Compared with single-modality methods, our method not only obtains multi-modal features but also highlights key feature pairs. Moreover, we design a seed point sampling algorithm and a local feature aggregation module, which can generate more positive seed points and obtain better local spatial features. Experiments on the public GraspNet-1Billion Dataset and real robots prove the effectiveness of our method.

Limitation. Extracting both point cloud and image features simultaneously requires a significant amount of GPU computing resources. In future work, we expect to design a lightweight fusion structure.

REFERENCES

[1] H. Dang and P. K. Allen, "Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 1311–1317.

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

[3] 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*. IEEE, 2011, pp. 3304–3311.

[4] 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.

[5] 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, pp. 9626–9633.

[6] R. Araki, T. Onishi, T. Hirakawa, T. Yamashita, and H. Fujiyoshi, "Mt-dssd: Deconvolutional single shot detector using multi task learning for object detection, segmentation, and grasping detection," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 487–10 493.

[7] H. Zhang, X. Lan, S. Bai, X. Zhou, Z. Tian, and N. Zheng, "Roi-based robotic grasp detection for object overlapping scenes," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 4768–4775.

[8] 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.

[9] 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.

[10] 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*. PMLR, 2020, pp. 53–65.

[11] 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.

[12] 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, pp. 4295–4302.

[13] 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.

[14] Y. Li, T. Kong, R. Chu, Y. Li, P. Wang, and L. Li, "Simultaneous semantic and collision learning for 6-dof grasp pose estimation," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3571–3578.

[15] 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, pp. 13 474–13 480.

[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] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 438–13 444.

[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)*. IEEE, 2021, pp. 13 459–13 466.

[20] H. Zhu, Y. Li, F. Bai, W. Chen, X. Li, J. Ma, C. S. Teo, P. Y. Tao, and W. Lin, "Grasping detection network with uncertainty estimation for confidence-driven semi-supervised domain adaptation," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9608–9613.

[21] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3003–3013.

[22] 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.

[23] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *arXiv preprint arXiv:1703.09312*, 2017.

[24] F.-J. Chu, R. Xu, and P. A. Vela, "Real-world multiobject, multigrasp detection," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3355–3362, 2018.

[25] 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.

[26] 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.

[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.