2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

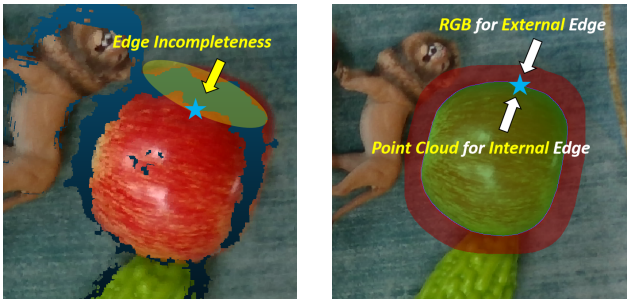
EdgeGrasp: Enhancing Edge Perception for 7-DoF Grasping Pose Estimation in Cluttered Scenes

Junning Qiu[1,2,3] Fei Wang[1] Yu Guo[1,*] Yonggen Ling[2] Minglei Lu[2,*]

***Abstract*— Estimating 7-DoF grasping poses (6-DoF with gripper width) in cluttered scenes is a critical challenge for robotic manipulation. In such environments, object edges often contain many promising grasp candidates, but relying solely on incomplete single-view point cloud to infer them is difficult. While neural networks excel at learning edge features from RGB images, simply combining these with point clouds often fails to generalize to novel scenes. To address these challenges, we propose EdgeGrasp, which enhances edge perception by allowing each modality to contribute to the most suitable edge information source for improving grasping performance. The internal edge features are extracted through voxel-based sparse 3D convolution on the aggregated point cloud from the edge interior, ensuring a rich geometric representation while mitigating incompleteness at the edge. For external edge and junction, vision foundation model is employed to extract local zero-shot semantic features, capturing fine-grained details and improving cross-object generalization. Finally, edge spatial attention fuses these features into edge-enhanced features by encoding edge distance for estimating 7-DoF grasping poses. Experimental results demonstrate our method's effectiveness, achieving state-of-the-art performance on the Graspnet-1Billion benchmark. Real-world robotic experiments further validate its practical applicability.**

I. Introduction

Estimating grasping poses in cluttered environments represents a critical challenge for robotic intelligence, focusing on the precise grasping of objects by adjusting the gripper. Recent advancements have shifted from traditional techniques that estimate the 6D pose of objects [1], [2], [3], followed by grasping the object at predefined locations, to direct estimation of grasping poses. This new approach [4], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], known as 7-DoF grasping pose estimation, includes both the gripper width and the 6D pose of the gripper. This evolution allows robots to grasp previously unseen objects, eliminating the need for predefined grasping locations and even object models. Point clouds, as a direct and effective representation of 3D geometry, have become a common input form in robotic tasks [26], [27]. State-of-the-art 7-DoF grasping pose estimation methods [16], [28], [25] also use point clouds as input, employing techniques such as PointNet++ [29], Transformer [17], voxel-based 3D convolutions [16], [24], [23] and graph neural networks [22] for feature extraction to estimate 7-DoF grasping poses. This learning framework has proven effective and is widely adopted by most methods in the field.



(a) (b)

Fig. 1. Illustration of the limitations of a single RGBD sensor. The blue star icons represent the edge points to be estimated. (a) The point cloud is incomplete at the object edges, making grasp pose estimation difficult. (b) Our method's strategy allows each modality to contribute to the most suitable edge information source, with the green region representing the internal edge and the red region representing the external edge.

Many grasp candidates along object's edges exhibit great force-closure metric [30], [4], a common and effective approach for evaluating 7-DoF grasping poses that calculates the minimum friction coefficient required for force closure between the gripper and the object. A lower friction coefficient, therefore, indicates better grasping performance [14]. When using a single RGBD sensor to capture the scene's point cloud, the object's edges typically lie at its boundaries, often corresponding to inflection points in its shape. However, point clouds captured by a single RGBD sensor are often incomplete at these edges as shown in Figure 1(a), making it challenging to effectively perceive the edges, leading to suboptimal performance in these regions, especially in grasping tasks where accurate edge information is critical [31], [32]. Neural networks excel at learning edge features from RGB images, but simply and directly fusing them with point clouds does not yield significant improvements, as these features often fail to generalize well to novel scenes.

To address the above issues, we propose EdgeGrasp, which improves grasping performance by enhancing edge perception. Specifically, it assigns the extraction of internal edge features to the point cloud and external edge features to the RGB image, allowing each modality to contribute to the most suitable edge information source as shown in Figure 1(b). We input RGBD data into the Internal-External Edge Feature Encoder, which separately encodes the point cloud and RGB

[1]State Key Laboratory of Human-Machine Hybrid Augmented Intelligence, Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China {wfx, yu.guo}@stu.xjtu.edu.cn
[2]Tencent Robotics X, China {rolandling, mingleilu}@tencent.com
[3]EngineAI, China qiujn@engineai.com.cn
*Yu Guo and Minglei Lu are corresponding authors.

image as different edge feature source. The encoder mitigates the interference caused by incomplete point clouds at the edges by voxelizing the region inside the edges of the point cloud and aggregating its geometric features using sparse 3D convolutions [33]. For external edge and junction, the encoder extracts local zero-shot semantic features [34], [35] using foundation vision models like DINOv2 [36], capturing fine-grained local details at edge junctions and extracting features that extend outward from the edge, contributing to preserving generalization across different objects. Additionally, to fully integrate the features mentioned above and form edge-enhanced features, we propose a edge spatial encoding attention mechanism. This mechanism guides the fusion of these features using residuals of the coordinates and edges encoding, resulting in edge-enhanced features for 7-DoF grasp pose estimation.

In conclusion, the key contributions of this work can be outlined as follows:

1) We propose EdgeGrasp, a novel RGBD-based framework that enhances edge perception and improves grasp performance through a strategy that clearly distinguishes the sources of edge features.
2) We propose edge spatial encoding attention mechanism for guiding the fusion of different edge feature sources.
3) Our method attained state-of-the-art performance on benchmark GraspNet-1Billion dataset[14] and demonstrated its practicality in real-world robotic experiments across diverse scenes.

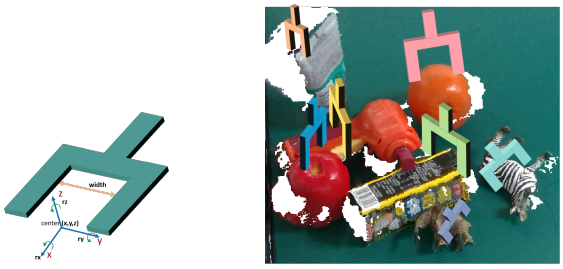
II. Related Work

A. 7-DoF grasping pose estimation based on point cloud

Conventional 6-DoF grasping methods begin by estimating an object's 6-DoF pose [1], [2], [3], followed by the determination of the final 6-DoF grasping pose based on precalculated poses for known objects. These methods, however, struggle with objects that lack predefined models in new environments. To enable grasping to work on novel objects, 2D planar grasping mothods [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49] that represent grasping pose by an rectangle are proposed. But the grasping of this kind of method can only approach the object in a top-down direction. In response, newer methods have been developed that directly estimate 7-DoF grasping pose for parallel-jaw grippers [14], [15], [16], [17], [18], [19], [21], [20], [28], [50], [51], [23], [22]. This advancement bypasses the need for 3D models of the objects to be grasped. Point cloud is an intuitive and effective representation of 3D data, which is widely used in the field of robotics [26], [27], including 7-DoF grasping. As a landmark work, GraspNet-1Billion[14] provides a large-scale grasping pose estimation benchmark and a PointNet++based [29] baseline. Subsequently, many works emerged to improve the performance of grasping pose estimation based on point cloud. In addition to continuing the use of PointNet++ from Grasp-1Billion as in [52], [19], voxel-based 3D convolutions [16], Transformer [17], and graph neural networks [22] have also been employed. In addition, there have been recent efforts to introduce self-supervision into this field [25]. Unlike the aforementioned works, we address the problem of hindered edge perception caused by the incomplete edge point cloud from a single RGBD sensor, improving grasp performance by distinguishing the sources of edge features.

B. Robotic Grasping based on RGBD

RGB images are commonly used in 2D planar grasping mothods, with many methods directly concatenating RGBD images as input. Although there are still some attempts to introduce RGB in 7-DoF grasping [15], [19], most methods still retain the point cloud learning framework. RGBMatters [15] uses RGB solely to predict the orientation of objects, reducing the reliance on high-quality depth maps. SymmetryGrasp [19] utilizes RGB images to identify the precise parts of objects. Unlike the aforementioned works, we leverage foundation vision models [36] to extract local zero-shot semantic features from RGB images to capture fine-grained local details at the edge junctions.



(a) (b)

Fig. 2. (a) Schematic of the 7-DoF grasping pose. It contains 3 translation components, 3 rotation angles and a gripper opening width (b) Schematic of 7-DoF grasping poses in cluttered scenes. A series of grasping poses in the current cluttered scene, where each gripper has the same 7 degrees of freedom as in (a).

III. Problem Statement

Given the RGB-D image of the scene $I = (C, D)$, where C, represented as $\in \mathbb{R}^{H\times W\times 3}$, denotes the RGB image, and D, denoted by $\in \mathbb{R}^{H\times W}$, signifies the depth image. The goal of 7-DoF grasping pose estimation is to find the grasping configuration $\mathbf{P} = \{\mathbf{p}_k \in \mathbb{R}^7\}_{k=1}^{N}$ to manipulate objects, which consists of 6-DoF pose with parallel-jaw gripper opening distance. Specifically, where $\mathbf{p}_k = (r_x, r_y, r_z, x, y, z, w)$ consists of the rotation r_x, r_y, r_z, the translation x, y, z, opening distance w of the parallel-jaw gripper as illustrated in the Figure 2.

IV. Proposed Approach

A. Overview

The overall approach of our work is illustrated in Figure 3. First, we extract internal edge features from the point cloud within the edge and external edge features from the RGB of the outer edge and junctions using the Internal-External Edge Feature Encoder. Then, to effectively fuse these features and enhance grasping, we introduce an edge spatial encoding attention mechanism, which leverages residuals of the coordinates and edges encoding to guide the fusion process.

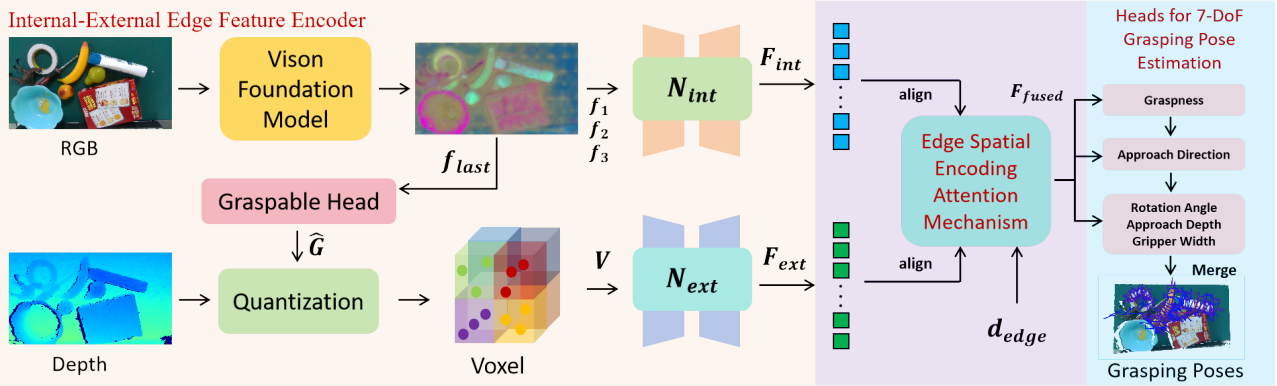


Fig. 3. Overview of our method including Internal-External Edge Feature Encoder, Edge Spatial Encoding Attention Mechanism and Heads for 7-DoF Grasping Pose Estimation.

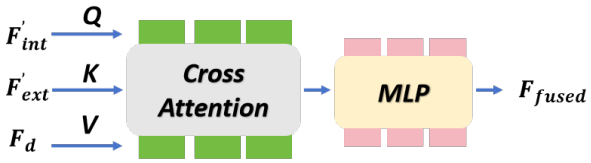


Fig. 4. Details of Edge Spatial Encoding Attention Mechanism. F_d is derived from the position encoding related to the edge distance, which is used to guide the fusion of features from different sources.

Finally, the fused features are used for 7-DoF grasping pose estimation in cluttered scenes.

B. Internal-External Edge Feature Encoder

For the input pair (C, D), we first utilize the foundation vision model DINOv2 [36] to extract local zero-shot semantic features [34], [35] from C. The features from the last layer of DINOv2 are then fed into the Graspable Head, implemented with ResNet, to predict a graspable map with confidence values in the range $[0, 1]$, thereby determining the object edge. This can be represented as:

$$\hat{G} = Graspable\ Head(f_{last})$$

where $\hat{G}$ is the predicted graspable map with confidence values in the range $[0, 1]$.

Meanwhile, to extract the external edge features, we employ a learnable network, denoted as N_{ext}, that outputs the external edge feature map F_{ext}. The intermediate-layer features $\{f_1, f_2, f_3\}$, where $f_1 \subseteq \mathbb{R}^{192\times\frac{H}{4}\times\frac{W}{4}}$, $f_2 \subseteq \mathbb{R}^{384\times\frac{H}{8}\times\frac{W}{8}}$, and $f_3 \subseteq \mathbb{R}^{768\times\frac{H}{16}\times\frac{W}{16}}$, are first aligned to the size of f_1. This alignment is achieved using learnable upsampling CNNs, U_2 and U_3, for f_2 and f_3, respectively. After upsampling, the features are concatenated along the channel dimension. The concatenated feature map is then passed through a convolutional layer to fuse the information. Following fusion, the resulting feature map is upsampled to match the original image size using another learnable upsampling operation. The extracted external edge features are represented as:

$$F_{\text{ext}} = N_{\text{ext}}(f_1, f_2, f_3) \tag{1}$$

Next, based on the Graspable Head prediction, we only quantize the points inside the object's edge region to avoid interference caused by incomplete point cloud data at the boundaries. This quantization process can be expressed as follows:

$$V = \{v = q(x_i, y_i, z_i) \mid \hat{G}(x_i, y_i, z_i) > S_1\} \tag{2}$$

where q is the quantization function from the Minkowski Engine [33], which maps the point clouds' coordinates (x_i, y_i, z_i) from depth map D to the nearest voxel center. In this process, $\hat{G}(x_i, y_i, z_i)$ refers to the confidence score of the point (x_i, y_i, z_i) predicted by the Graspable Head, and S_1 is the threshold above which the points are called **graspable points** P_g. After quantization, we apply 3D voxel convolution to aggregate the features from this region, resulting in the internal edge features, denoted as F_{int}. The process of extracting the internal edge features can be represented as:

$$F_{\text{int}} = \mathbf{N}_{\text{int}}(V) \tag{3}$$

where V is the set of quantized points inside the object's edge region, and the 3D ResUNet network N_{int} aggregates these features to produce the internal edge feature F_{int}.

C. Edge Spatial Encoding Attention Mechanism

To integrate the different sources of edge information, F_{int} and F_{ext}, we design a edge spatial encoding attention mechanism. The spatial relationship between the points to be grasped and edges is used as fusion clues, which guides the fusion of the two into edge-enhanced features for grasping as shown in Figure 4. First, the internal edge features F_{int} and external edge features F_{ext} are aligned in the graspable point space. The internal edge features F_{int} are aligned by inverse mapping through the voxelization process, resulting in F'_{int}, while the external edge features F_{ext} are aligned based on the graspable map, where points are selected using the threshold S_2, resulting in F'_{ext}. The d_{edge} denotes the distance from each grasping point (in the UV plane) to the nearest edge determined by $\hat{G}$ and graspable points' coordinates, computed in pixel space and normalized by the image width and height to avoid the influence of incomplete edge point clouds. To incorporate spatial information, we compute the

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 [47]	15.48	21.84	10.25	13.26	18.37	4.62	5.52	5.93	1.86
Chu et al. [46]	15.97	23.66	10.80	15.41	20.21	7.06	7.64	8.69	2.52
GPD [4]	22.87	28.53	12.84	21.33	27.83	9.64	8.24	8.89	2.67
PointnetGPD [5]	25.96	33.01	15.37	22.68	29.15	10.76	9.23	9.89	2.74
Graspnet-baseline [14]	27.56	33.43	16.95	26.11	34.18	14.23	10.55	11.25	3.98
RGBMatter [15]	27.98	33.47	17.75	27.23	36.34	15.60	12.25	12.45	5.62
SSCL [52]	36.55	47.22	19.24	28.36	36.11	10.85	14.01	16.56	4.82
Transgrasp [53]	39.81	47.54	36.42	29.32	34.80	25.19	13.83	17.11	7.67
GraNet [22]	43.33	52.56	34.03	39.98	48.66	32.00	14.90	18.66	7.76
HGGD [24]	59.36	-	-	51.20	-	-	22.17	-	-
Scale-Balanced [28]	58.95	68.18	54.88	52.97	63.24	46.99	22.63	28.53	12.00
Graspness [16]	66.83	79.98	59.23	58.41	72.18	48.19	25.15	31.03	13.08
GraspContrast [25]	67.22	-	-	55.91	-	-	25.79	-	-
Graspness w. RGB	68.61	80.16	63.08	58.64	71.54	49.21	26.17	32.79	14.62
Graspness w. ResNet	69.81	81.99	63.90	56.60	69.79	45.84	27.05	33.84	15.11
Ours	**73.39**	**84.53**	**69.21**	**62.08**	**74.81**	**54.03**	**27.54**	**34.56**	**15.98**

TABLE I

EVALUATION RESULTS OF REALSENSE CAMERA DATA ON THE GRASPNET-1BILLION BENCHMARK COMPARED WITH OTHER METHODS, '-' INDICATES THAT THE RESULTS ARE NOT AVAILABLE.

Method		Seen	Similar	Novel
IEFE	ESEA	AP	AP	AP
×	×	66.83	58.41	25.15
w.o.Srategy	×	68.35(+1.52)	55.37(-2.54)	24.94(-0.21)
w.Srategy	×	71.67(+4.84)	61.74(+3.33)	26.30(+1.15)
w.Srategy	✓	73.39(+6.56)	62.08(+3.67)	27.54(+2.39)

TABLE II

ABLATION STUDY RESULTS OF KEY COMPONENTS.

positional encoding $\mathrm{PE}(d_{\text{edge}})$ for each distance d_{edge}, and then concatenate it with the distance:

$$\mathrm{PE}(d_{\text{edge}}) = \left[\sin\left(d_{\text{edge}} \cdot 2^k \cdot \pi\right), \cos\left(d_{\text{edge}} \cdot 2^k \cdot \pi\right)\right] \quad (4)$$

where $k \in \{0, 1, \cdots, 9\}$. The concatenation of the distance and positional encoding is then passed through a Multi-Layer Perceptron (MLP) to obtain the feature V:

$$F_d = MLP(concat(d_{\text{edge}}, \mathrm{PE}(d_{\text{edge}}))) \quad (5)$$

Next, a Cross-Attention mechanism [54] is used to integrate the internal edge features F'_{int} and the external edge features F'_{ext}. The mechanism ensures that the most relevant information from both feature sets is combined based on the spatial relationship of the grasping points to the edges. The edge-enhanced feature F_{fused} for each grasping point is then obtained through an MLP that processes the fused features, allowing for non-linear transformations to refine the feature representation for grasping pose estimation:

$$F_{\text{fused}} = MLP(CrossAttention(F'_{\text{int}}, F'_{\text{ext}}, F_d)) \quad (6)$$

D. Heads for 7-DoF Grasping Pose Estimation

Here, we have obtained the graspable point with the edge-enhanced feature F_{fused}. Next, this feature is used to estimate the 7-DoF grasping pose at P_g, which is decomposed into graspness estimation, approach direction, rotation angle, gripper width, and approach depth, following the process in [16]. We define graspness and use a graspness head to predict it. P_g values greater than the threshold S_2 are filtered and denoted as P'_g. Next, we use the Approach head to predict the predefined viewpoints classification [14] for these P'_g. Except for the gripper width, all other aspects are treated as classification problems, consistent with previous work [16]. The above heads are all composed of MLPs.

E. Loss Function

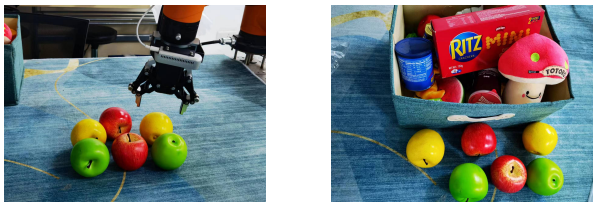
The loss function is as follow:

$$L = \lambda_1 L_{graspness} + \lambda_2 L_{view} + \lambda_3 L_{pos} + \lambda_4 L_{width} \quad (7)$$

$L_{graspness}$ is for point-wise graspness metric. L_{view} is for view-wise graspness metric. L_{pos} is for grasping scores of different preset rotations and depths. L_{width} is for gripper opening width. We use smooth-l_1 as the loss function for all tasks. λ_1, λ_2, λ_3, λ_4 are the weights to balance the multi-task learning loss. The rest of the details are the same as [16].

V. EXPERIMENTS

In this section, we conduct experiments to compare our method with SOTA methods. We first describe the evaluation metrics for 7-DoF grasping pose estimation and the implementation details of our pipeline. Next, we conduct main experiments, ablation experiments, quantitative analysis, and qualitative analysis on the benchmark GraspNet-1Billion [14] dataset. Finally, real robot experiments across multiple scenes and targets verify the effectiveness of improving grasping performance by enhancing edge perception.



(a) (b)

Fig. 5. (a) Physical robotic arm AUBO-i5 with realsense D435 depth camera andDH AG-95 parallel-jaw gripper. (b) Various objects we used for experiments.

A. Experimental Setup

1) Dataset and Evaluation Metrics: We conduct experiments on the benchmark GraspNet-1Billion dataset [14] and use its metrics to evaluate our results. The metric uses non-maximum suppression and evaluates the top 50 predictions, using the average accuracy over a range of assumed different friction coefficients ($\mu \in \{0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$) AP_μ, and the average value of AP_μ is AP. The GraspNet-1Billion [14] dataset contains 190 scenes, of which 90 scenes are used for testing and are divided into the same, similar and novel, and each scene has 256 RGBD images. For robot experiments, we evaluate success rate (SR) and completion rate (CR) same as [17], [20]. Denote N_{sa}, N_{ta}, and N_{no} represent the number of successful attempts, the total number of attempts, and the number of graspable relevant objects in the scene.

$$SR = \frac{N_{sa}}{N_{ta}}, \quad CR = \frac{N_{sa}}{N_{no}} \tag{8}$$

2) Implementation Details: In the training phase, we use the masks of all graspable objects to segment them, retaining the weights of DINOv2 (ViT-B version) while training the other modules, which is completed on an RTX8000. The model is trained for 15 epochs on a dataset collected with a Realsense camera, using a batch size of 4. The learning rate is multiplied by 0.95 at each epoch. Adam is used as the optimizer, with an initial learning rate of 1e-3 for all ablation studies [55]. In the test phase, farthest point sampling (FPS) is uniformly applied to select 1,024 points for evaluation, as per the benchmark [14], without any post-processing, such as collision detection. The above process unifies the number of point clouds to 25,000 through random downsampling. Other parameters include the threshold values S_1 set to 0.98, and S_2 set to 0.1. The number of channels in F_{ext} and F_{int} is 256. The weight coefficients for the losses in section IV.E are λ_1, λ_2, λ_3, and λ_4 set to 10, 100, 15, and 10, respectively.

B. Main Results

We compare our proposed method with different methods on the GraspNet-1Billion benchmark, achieving state-of-the-art performance as shown in Table I. In terms of grasping performance, compared with the previous state-of-the-art method GraspContrast [25] without instance label, our pipeline not only achieves performance gains of +6.17/+6.17/+1.75 AP on the seen, similar, and novel test sets, respectively. Additionally, we modified Graspness with two extra versions, where we directly supplemented the point cloud with the original RGB and RGB features extracted by ResNet, to validate the effectiveness of our use of multimodal information based on the code here[1]. The results show that our strategy of allowing each modality to contribute to the most suitable edge information source more effectively utilizes the RGBD information.

Method	Seen	Similar	Novel
	AP	AP	AP
Baseline	41.37	35.89	16.18
Ours	**47.60**	**40.20**	**18.76**

TABLE III

QUANTITATIVE ANALYSIS OF THE IMPACT OF ENHANCING EDGE PERCEPTION FOR 7-DOF GRASPING POSE ESTIMATION

C. Ablation Study

To analyze the effectiveness of our two key components: Internal-External Edge Feature Encoder and Edge Spatial Encoding Attention Mechanism, we designed a series of ablation studies, as shown in Table II. "IEFE" stands for "Enhanced Edge Local Features" and "ESEA" stands for "Edge Spatial Encoding Attention Mechanism".

1) Srategy of distinguishing the sources of edge features: To verify the effectiveness of the Strategy of distinguishing the sources of edge features, we conducted experiments on the Internal-External Edge Feature Encoder as shown in Table II, where the rows with "*ESEA* is ×". Here, *w.Srategy* refers to the strategy that separately encodes the point cloud and RGB image as different edge feature sources, while *w.o.Srategy* refers to the strategy where the features extracted from both modalities are directly added without distinguishing. The experiments show that simply using features extracted from a foundation vision model may hinder generalization for both "similar" and "novel" cases (-2.54,-0.21)AP, while our strategy leads to better generalization and improved performance for the RGBD modalities (+4.84,+3.33,+1.15)AP for seen, similar and novel.

2) Edge Spatial Encoding Attention for Fusion: To effectively fuse different edge feature sources into edge-enhanced features, we designed the Edge Spatial Encoding Attention mechanism and performed an ablation study as shown in Table II, where the rows with "*IEFE* is *w.Strategy*". Experimental results demonstrate that the introduction of this mechanism (ESEA is ✓) effectively utilizes the distance difference between grasp points and edges as a cue to fuse the internal and external edge features. This results in an improvement of (+1.72/+0.34/+1.24) AP compared to the direct summation fusion strategy (ESEA is ×).

D. Quantitative Analysis

To further demonstrate how the proposed method enhances edge perception by allowing each modality to contribute to the most suitable edge information source for 7-DoF grasp pose estimation, we compare our method with the baseline

[1]https://github.com/rhett-chen/graspness_implementation

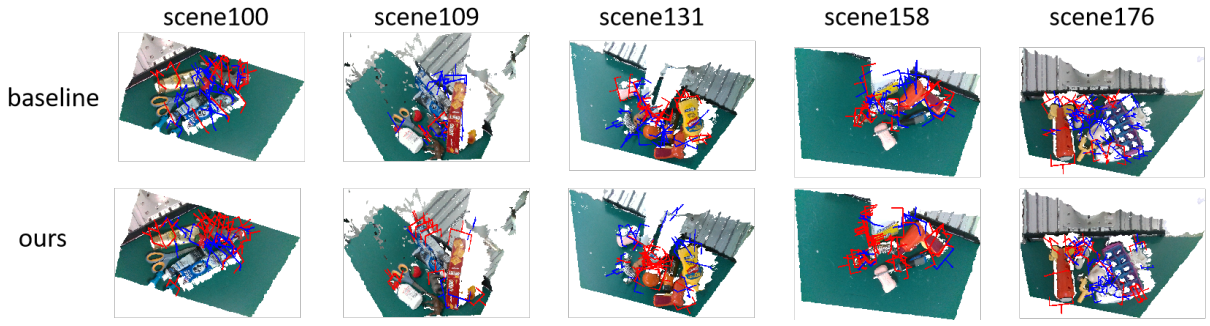


Fig. 6. Qualitative Analysis. Red represents valid grasps, while blue represents invalid grasps. Compared to the baseline method, our approach enhances edge perception, effectively reducing invalid grasps of object edges.

where the proposed strategy of distinguishing the sources of edge features and Edge Spatial Encoding Attention for Fusion are removed, as shown in Table III. Using the FPS algorithm, we retain a sufficient number of 8192 points per scene and evaluate the AP of object edge points solely based on the real values of the mask. Experimental results show that our method can enhance the pose estimation performance of object edges. The AP gains of object edges on the seen, similar, and novel test sets are (+6.23/+4.31/+2.58). Additionally, as shown in Table II, we can conclude that the performance improvement achieved by our method is largely due to the enhanced grasping at the edge.

Scene	Number of objects	Attempt	SR	CR
scene1	6	6	100%	100%
scene2	6	6	100%	100%
scene3	10	10	100%	100%
scene4	5	6	83.33%	100%
scene5	6	6	100%	100%
Average	7	7.33	95.50%	100%

TABLE IV

EXPERIMENTAL RESULTS OF CLUTTERED SCENE GRASPING ON REAL ROBOT. SR IS SUCCESS RATE. CR IS COMPLETION RATE.

E. Qualitative Analysis

We further present some qualitative results in Figure 6 to illustrate the relationship between enhanced edge perception and 7-DoF grasp pose estimation. We observe that by enhancing edge perception, our proposed method reduces invalid grasps (such as collisions or empty grasps) of object edges, which in turn improves grasping and leads to an overall increase in the grasp pose score of the scene.

F. Experiments on Robotic in the Real World

We conducted grasping experiments in cluttered scenes on a real robot. As shown in Figure 5, we mounted an Intel Realsense D435 depth camera on an AUBO i5 robotic arm and performed the experiments using a DH AG-95 parallel gripper. The robotic arm's planning algorithm is provided by the AUBO manufacturer. The final executed grasp pose is determined based on the task setup, which includes both popular unordered grasping experiments and ordered grasping experiments that we designed to increase the challenge. The computing platform is a machine equipped with an NVIDIA GTX 2080Ti GPU, an Intel i7-7800X CPU, 32GB of RAM, and the Ubuntu 18.04 operating system. We selected twenty-eight objects commonly found in everyday life, as shown in Figure 5. We set up **two settings** to comprehensively evaluate our multi-task object manipulation pipeline. Table IV shows the experiments of unordered grasping, where we execute the grasp with the highest confidence from the pose estimation and re-estimate the best grasp pose after each completion. Table V shows the experiments of ordered grasping, where we use Grounding DINO [56] to sequentially grasp target objects based on pre-defined prompts for different settings.

Scene	Number of objects	Attempt	SR	CR
setting1	10	10	100%	100%
setting2	10	10	100%	100%
setting3	10	11	90.91%	100%
setting4	5	5	100%	100%
setting5	5	5	100%	100%
setting6	5	7	71.43%	100%
Average	7.5	8	93.75%	100%

TABLE V

PHYSICAL ROBOT EXPERIMENTAL RESULTS ON DIFFERENT SETTINGS IN CLUTTERED SCENE. SR IS SUCCESS RATE. CR IS COMPLETION RATE.

VI. CONCLUSION

In this paper, we propose EdgeGrasp, which enhances edge perception by allowing each modality to contribute to the most suitable edge information source for improving grasping performance, achieving state-of-the-art performance on the GraspNet-1Billion benchmark. We use the Internal-External Edge Feature Encoder, which separately encodes the point cloud and RGB image as different edge feature sources. Additionally, to fully integrate the features mentioned above and form edge-enhanced features, we propose an edge spatial encoding attention mechanism. Through sufficient experimental analysis, we show that these key components are indispensable and significantly improve the performance of grasping pose estimation at the edge of the object. The high-precision complex grasping capabilities demonstrated in real-world robotic experiments highlight the potential of our method to improve the ability of intelligent robots to manipulate objects.

ACKNOWLEDGMENT

This work was supported in part by the Key Research and Development Program of Shaanxi Province under Grants 2024GX-YBXM-141 and 2025ZG-JBGS-008, and in part by the National Key Research and Development Program of China under Grant 2022YFB3303800.

REFERENCES

[1] N. Mellado, D. Aiger, and N. J. Mitra, "Super 4pcs fast global pointcloud registration via smart indexing," in *Computer graphics forum*, vol. 33. Wiley Online Library, 2014, pp. 205–215.

[2] J. Yang, H. Li, D. Campbell, and Y. Jia, "Go-icp: A globally optimal solution to 3d icp point-set registration," *TPAMI*, vol. 38, no. 11, pp. 2241–2254, 2015.

[3] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3d object recognition," in *2010 IEEE computer society conference on computer vision and pattern recognition*, 2010, pp. 998–1005.

[4] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.

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

[6] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.

[7] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, "S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes," in *Conference on robot learning*. PMLR, 2020, pp. 53–65.

[8] 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.

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

[10] 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.

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

[12] J. Cai, J. Cen, H. Wang, and M. Y. Wang, "Real-time collision-free grasp pose detection with geometry-aware refinement using high-resolution volume," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1888–1895, 2022.

[13] A. Alliegro, M. Rudorfer, F. Frattin, A. Leonardis, and T. Tommasi, "End-to-end learning to grasp via sampling from object point clouds," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9865–9872, 2022.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] Y. Shi, Z. Tang, X. Cai, H. Zhang, D. Hu, and X. Xu, "Symmetrygrasp: Symmetry-aware antipodal grasp detection from single-view rgb-d images," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 235–12 242, 2022.

[20] Z. Chen, Z. Liu, S. Xie, and W.-S. Zheng, "Grasp region exploration for 7-dof robotic grasping in cluttered scenes," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 3169–3175.

[21] 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*, 2023.

[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.

[23] J. Qiu, F. Wang, and Z. Dang, "Multi-source fusion for voxel-based 7-dof grasping pose estimation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 968–975.

[24] 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*, 2023.

[25] W. Wang, H. Zhu, and M. H. Ang, "Graspcontrast: Self-supervised contrastive learning with false negative elimination for 6-dof grasp detection," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7294–7300.

[26] F. Pomerleau, F. Colas, R. Siegwart *et al.*, "A review of point cloud registration algorithms for mobile robotics," *Foundations and Trends® in Robotics*, vol. 4, no. 1, pp. 1–104, 2015.

[27] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.

[28] 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.

[29] C. Qi, L. Yi, H. Su, and L. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*, Long Beach, California, United States, 2017.

[30] V.-D. Nguyen, "Constructing force-closure grasps," *The International Journal of Robotics Research*, vol. 7, no. 3, pp. 3–16, 1988.

[31] Z. Liu, Z. Chen, and W.-S. Zheng, "Simulating complete points representations for single-view 6-dof grasp detection," *IEEE Robotics and Automation Letters*, vol. 9, no. 3, pp. 2901–2908, 2024.

[32] K. Ma, H. Dong, and Y. Mu, "Local occupancy-enhanced object grasping with multiple triplanar projection," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–18.

[33] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, California, 2019, pp. 3075–3084.

[34] Y. Chen, Y. Di, G. Zhai, F. Manhardt, C. Zhang, R. Zhang, F. Tombari, N. Navab, and B. Busam, "Secondpose: Se (3)-consistent dual-stream feature fusion for category-level pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9959–9969.

[35] R. Zhang, Z. Huang, G. Wang, C. Zhang, Y. Di, X. Zuo, J. Tang, and X. Ji, "Lapose: Laplacian mixture shape modeling for rgb-based category-level object pose estimation," in *European Conference on Computer Vision*. Springer, 2024, pp. 467–484.

[36] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.

[37] 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.

[38] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic

grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.
[39] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 1316–1322.
[40] D. Guo, F. Sun, H. Liu, T. Kong, B. Fang, and N. Xi, "A hybrid deep architecture for robotic grasp detection," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 1609–1614.
[41] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International journal of robotics research*, vol. 37, no. 4-5, pp. 421–436, 2018.
[42] 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.
[43] 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.
[44] U. Asif, J. Tang, and S. Harrer, "Ensemblenet: Improving grasp detection using an ensemble of convolutional neural networks." in *BMVC*, 2018, p. 10.
[45] ——, "Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices." in *IJCAI*, vol. 7, 2018, pp. 4875–4882.
[46] 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.
[47] 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.
[48] S. Kumra and C. Kanan, "Robotic grasp detection using deep convolutional neural networks," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 769–776.
[49] U. Asif, J. Tang, and S. Harrer, "Densely supervised grasp detector (dsgd)," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 8085–8093.
[50] L. Zheng, W. Ma, Y. Cai, T. Lu, and S. Wang, "Gpdan: Grasp pose domain adaptation network for sim-to-real 6-dof object grasping," *IEEE Robotics and Automation Letters*, 2023.
[51] 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.
[52] 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.
[53] Z. Liu, H. Tang, S. Zhao, K. Shao, and S. Han, "Pvnas: 3d neural architecture search with point-voxel convolution," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 11, pp. 8552–8568, 2021.
[54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, Long Beach, California, United States, 2017, pp. 5998–6008.
[55] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, San Diego, CA, USA, 2015.
[56] 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.