IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

Plug-and-Play Shape Matching Module for Zero-Shot Mesh-Free Grasp Refinement on Unknown Objects

Ju Yong Hong, *Graduate Student Member, IEEE*, Yeong Gwang Son, *Graduate Student Member, IEEE*, Seung Hwan Um, *Graduate Student Member, IEEE*, and Hyouk Ryeol Choi, *Fellow, IEEE*

***Abstract*—Reliably grasping unknown objects in logistics automation remains a major challenge. While most approaches rely on 3D CAD models or large-scale training, their applicability to novel items is limited. This letter proposes a plug-and-play geometric refinement module that can be appended to any existing grasp planner. The module operates in a training-free and mesh-free manner, estimating an object's approximate centroid from a single RGB-D image to enhance grasp stability. Its core mechanism involves using an initial grasp candidate as an automatic prompt for segmentation, followed by geometric primitive fitting to the isolated object's point cloud. By rescoring grasp candidates based on proximity to the estimated centroid, our module improves physical stability. Experimental results demonstrate that our module improves the success rate of baseline grasp planners by up to 25%p enhancing real-world pick-and-place performance without requiring any offline training or prior object models.**

***Index Terms*—Perception for grasping and manipulation, RGB-D perception, object detection, segmentation and categorization.**

I. Introduction

ROBOTIC manipulation has made significant strides in recent years, but object picking and placing remains one of the most time-consuming and failure-prone processes in logistics automation, particularly in cluttered and unstructured environments [1], [2]. To address this, numerous grasp planning [3], [4], [4], [5], [6], [7] and 6D pose estimation [8], [9], [10], [11], [12], [13] algorithms have been proposed. However, a critical limitation persists: most state-of-the-art approaches rely heavily on pre-defined 3D CAD models or large-scale training datasets, making them impractical for novel objects in real-world logistics.

The challenge of handling previously unseen objects has led to a focus on "zero-shot grasping." However, the term is applied inconsistently across the literature. Many recent data-driven methods, while claiming zero-shot capabilities, still require extensive pre-training on massive datasets of objects and grasps [14], or depend on semantic priors from large vision-language models [15]. In contrast, the goal of this work is to pursue a stricter definition of zero-shot: a genuinely training-free, model-free, and mesh-free system that requires no prior exposure to object models, categories, or offline learning. This distinction is the core differentiator between our approach and the dominant data-driven paradigm.

As a solution, we propose a lightweight, plug-and-play refinement module. The novelty of this module lies not in its individual components—such as the use of foundation segmentation models or geometric fitting algorithms—but in their unique integration into a novel system architecture. Our method begins by taking a candidate grasp point from any standard planner and using it as an automatic prompt for a segmentation model [16], [17], [18]. This grasp-driven segmentation isolates the target object, whose partial point cloud is then fitted to geometric primitives to estimate its pose and centroid. Finally, a centroid-aware rescoring strategy refines the initial grasp to minimize gravitational torque and improve stability. This system architecture is designed to achieve both real-time performance and broad generality, offering a practical way to enhance existing grasp planners without modification or retraining.

In summary, this letter makes the following key contributions:

- A lightweight shape matching module that estimates an object's approximate 6D pose and geometric shape from a single RGB-D image without a 3D model.
- A grasp-driven segmentation technique that utilizes the output of an existing grasp planner as an auto point prompt to effectively isolate the target object in cluttered scenes without a separate object detector.
- A rescoring strategy that enhances the stability of candidate grasps by minimizing the gravitational torque using the estimated centroid.
- A plug-and-play module that improves the performance of existing grasp planners without requiring retraining.

Fig. 1 illustrates our proposed shape-matching-based grasp refinement pipeline. The pipeline adapts based on object complexity. For primitive-shaped objects (top row), a suction grasp candidate serves as a point prompt for segmentation, which is followed by primitive fitting and centroid-based grasp refinement. For non-primitive objects (bottom row), an initial finger-based grasp is refined by first segmenting the object into parts, fitting each part to a primitive, and computing a composite centroid.

Received 1 June 2025; accepted 26 September 2025. Date of publication 17 October 2025; date of current version 4 November 2025. This article was recommended for publication by Associate Editor D. Seita and Editor A. Valada upon evaluation of the reviewers' comments. This work was supported by the National Research Foundation of Korea (NRF) funded by the Korea Government (MSIT) under Grant RS-2023-00207772. *(Corresponding author: Hyouk Ryeol Choi.)*

The authors are with the School of Mechanical Engineering, Sungkyunkwan University, Suwon 16417, South Korea (e-mail: juyong0000@skku.edu; syoungk20@g.skku.edu; seunghwanum@gmail.com; choihyoukryeol@gmail.com).

This article has supplementary downloadable material available at https://doi.org/10.1109/LRA.2025.3623004, provided by the authors.

Digital Object Identifier 10.1109/LRA.2025.3623004

©2026 IEEE

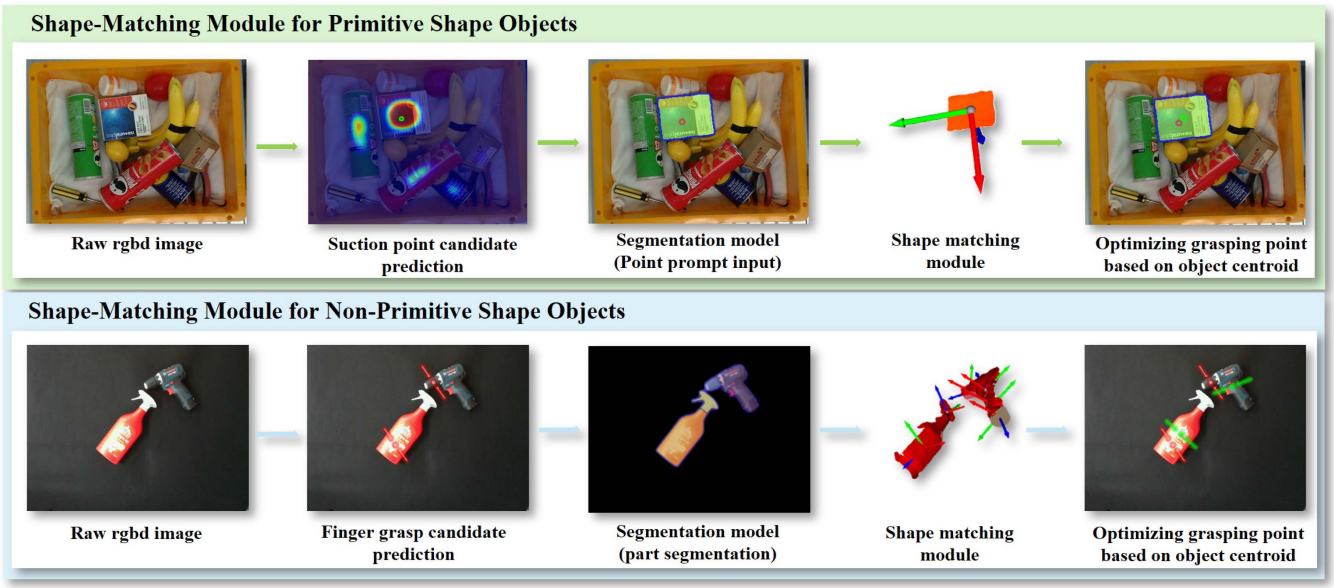


Fig. 1. Overview of the proposed shape-matching-based grasping pipeline. **(Top)** For primitive-shaped objects, the pipeline performs suction-based grasp prediction, point-prompt segmentation, primitive shape fitting, and centroid-based grasp optimization. **(Bottom)** For non-primitive objects, the pipeline performs finger-based grasp prediction, part-level segmentation, part-wise primitive fitting, and centroid-based grasp optimization.

This modular structure allows the module to be plugged into any existing grasping pipeline to improve stability, without requiring mesh models or category supervision.

II. Related Work

Our work integrates concepts from three primary domains: object 6D pose estimation, segmentation, and zero-shot grasping.

A. Objects 6D Pose Estimation

Estimating the 6D pose—translation and rotation—of objects is fundamental to robotic manipulation. Prior work can be divided into methods for known and unknown objects.

For known objects, model-based methods like ICP [19] align observed data to object meshes. Learning-based models such as [11], [12], [13] further improve accuracy using RGB-D inputs and synthetic renderings. These are commonly benchmarked via the BOP Challenge [BOP-2023] [20]. For unknown objects aims for large-scale generalization but still relies on 3D models at inference time. For instance, methods like MegaPose [21], GigaPose [22], Freeze [23], SAM-6D [24] can generalize to novel object instances but fundamentally require a target mesh. Approaches like 6-PACK [25] operate without prior models but are often limited to category-level estimation. Any6D [26] and MFOS [27] achieve instance-level pose estimation for unseen objects, but they require a reference anchor image or employ complex attention mechanisms for correspondence matching. While these methods aim for high-fidelity pose recovery, they often come with the need for specific auxiliary inputs.

Our module's goal is not precise pose reconstruction but rather a "good enough" pose approximation for stability-oriented grasp refinement. By using simple, geometric primitive fitting, our approach sidesteps the need for CAD models, anchor images, or complex network architectures, making it highly efficient and practical for real-world logistics.

B. Segmentation Models

Accurate object segmentation is a critical prerequisite for grasping. While classical semantic [28] and instance segmentation models [29], [30], [31] are powerful, their performance is tied to the classes seen during training. The advent of foundation models like the Segment Anything Model (SAM) [16] has enabled class-agnostic segmentation. Our work utilizes Fast-SAM [17], an efficient variant suitable for real-time robotics. The novelty of our approach, however, is not the segmentation model itself but its integration into our pipeline: we use the output of an upstream grasp planner [3], [5], [6], [7], [32] as an automatic point prompt. This grasp-driven mechanism allows us to isolate the target object for manipulation in cluttered scenes without needing a separate, class-aware object detector, making the entire process highly adaptable and robust.

C. Zero-Shot Grasping Strategies

Grasping unknown objects is a key robotics challenge, with research covering analytical, data-driven, and hybrid strategies. Traditional analytical and geometry-based methods are often impractical as they require detailed object geometry, which is unavailable for unknown or partially visible objects.

To overcome these limitations, data-driven approaches have gained significant traction. Some studies [6] introduced large-scale synthetic grasp datasets and trained deep networks to predict grasp quality directly from depth images.

Other works [5], [7] extended this approach to suction-based grasping, using large-scale simulation to generate training data

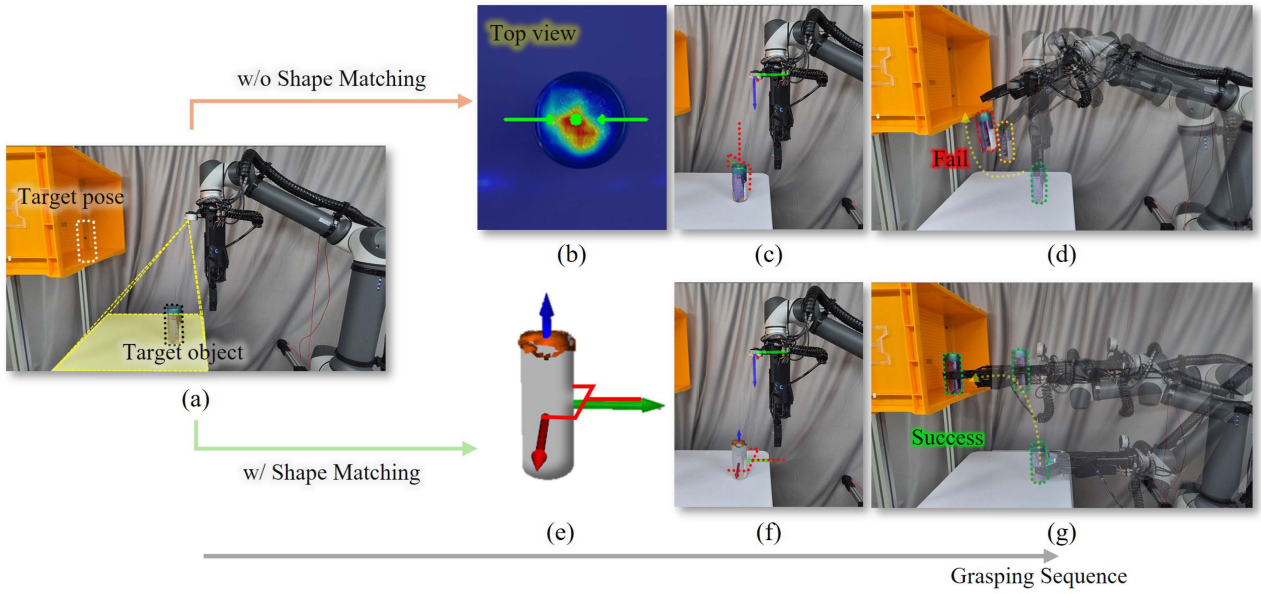


Fig. 2. Comparison of grasping performance with and without the proposed shape matching module. (a) A cylindrical object and its target pose in the shelf. (b) Suction candidate predicted without considering shape. (c)–(d) Grasp attempt and failure due to incorrect approach vector. (e) Shape matching result estimating the object's principal axis and centroid. (f)–(g) Successful grasp and placement with geometry-aware refinement. The shape matching module enables better alignment of the gripper with the object's orientation, resulting in improved grasp stability and precise placement.

across various types of objects. These methods have shown strong generalization, especially for warehouse environments, but remain dependent on synthetic mesh-based data.

Recent efforts broaden the scope: ShapeGrasp [15] and SegGrasp [33] leverage LLMs/VLMs for task-oriented, semantic grasping; LERF-TOGO [34] reconstructs 3D scenes from multi-view images for language-queried manipulation. While powerful, they typically require substantial computation or semantic priors. In parallel, ZeroGrasp [14] and the GraspAnything dataset [35] pursue data-centric generalization, and Click-to-Grasp [36] uses a single user click to guide diffusion-based shape completion before grasping—an interactive contrast to our fully automated pipeline.

Our work offers a distinct, training-free alternative. Unlike semantic or interactive approaches, our module is stability-oriented and uses only geometric cues from a single RGB-D frame, aligning with classical geometry-based reasoning while extending it to the object's overall shape and centroid.

III. Problem Statement

A. Assumption

We consider a robot manipulator performing autonomous pick-and-place of unknown objects in a logistics environment. The system relies solely on a partial point cloud from a single-view RGB-D image, with no prior knowledge of the objects' geometry, mesh models, or class labels.

B. Objective

The objective of this work is to develop a model-free pipeline capable of performing robust grasping and placement of previously unseen objects by estimating their approximate pose and geometric shape using minimal input. In addition to detecting feasible grasp points, the system incorporates a shape-aware refinement step to improve grasp stability. As illustrated in Fig. 2, conventional grasping algorithms often rely on the top-view surface to extract grasp points (Fig. 2(b)), which frequently leads to unstable grasping and placement failures (Fig. 2(c)–(d)). Our method estimates the full geometric shape of the object by fitting primitive models (Fig. 2(e)), allowing the grasp to be refined near the true centroid of the object (Fig. 2(f)–(g)). This refinement process leads to significantly more stable grasping and accurate object placement.

IV. Proposed Method

A. Overview

Our model-free pipeline refines a given grasp point by estimating the approximate 6D pose and shape of an unknown object from a single-view RGB-D image. The process begins with a grasp candidate from an upstream planner, such as CoASNet [5], Dex-Net [6], or SuctionNet [7]. This point serves as an automatic prompt for a foundation segmentation model (e.g., FastSAM [17]) to isolate the target object. This grasp-driven segmentation is a key design choice for our zero-shot goal. Unlike supervised models like YOLO [29] or YOLACT [31], which are limited to pre-trained categories, our approach is category-agnostic and requires no class-specific training. The model returns a mask corresponding to the prompted grasp, which we use to extract a partial 3D point cloud from the depth data. After post-processing to remove noise and outliers, a shape matching algorithm fits the refined point cloud to a primitive (box, sphere, or cylinder) to estimate its approximate 6D pose and parameters.

B. Step 1 – Segmentation With Auto Point Prompt

Our pipeline is designed to be modular and agnostic to the choice of grasp point estimation strategy. Our shape matching module can be seamlessly integrated with any grasp point

Once a grasp point is obtained using any such algorithm, it is used as a point prompt for segmentation. While the original SAM [16] provides excellent segmentation quality, its computational cost is prohibitive for real-time robotic applications. To balance performance and speed, we selected FastSAM [17], which offers a significant speed-up with minimal degradation in accuracy, making it well-suited for the low-latency demands of our pipeline. FastSAM generates a segmentation mask from the prompted region of the RGB image. This mask is then applied to the corresponding depth image to extract the point cloud of the object's visible surface.

C. Step 2 – Post-Processing of Point Cloud

The raw point cloud requires refinement before shape matching. We apply the following steps:

- *Downsampling:* Reduces the total number of points for computational efficiency.
- *Outlier Removal:* Uses KD-tree-based density estimation to remove points with lower local density than the global average.
- *Normal Filtering:* Removes points whose surface normals diverge significantly from the camera's Z-axis, preserving only visible surfaces.

These steps yield a clean, structured point cloud suitable for geometric analysis.

D. Step 3 – Shape Matching

In this work, we fit the refined point cloud to three geometric primitives: a box, a cylinder, and a sphere. These shapes were selected as they are commonly found in logistics environments and provide a sufficient basis for approximating a wide variety of rigid objects. While other primitives like cones are described in geometric fitting literature [37] and could be included, they are less frequent in typical packaging. Furthermore, they can often be partially approximated by a cylinder, so they were excluded from the scope of this work for simplicity. Our framework is extensible, and while initial experiments showed these three primitives provided a robust baseline.

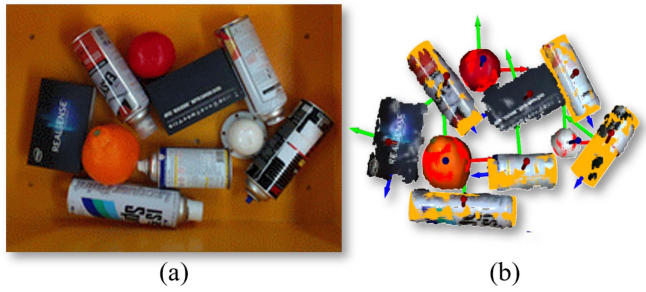
We fit the refined point cloud to these three geometric primitives. Fig. 3 shows that our shape matching module can reliably decompose and estimate object poses under severe occlusion and category variation, without relying on mesh models.

1) Covariance-Based Shape Descriptor: Given N points $\{x_1, x_2, \ldots, x_N\}$, we first compute

$$\frac{1}{N}\sum_{i=1}^{N} x_i \qquad C = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^\top \tag{1}$$

Let the eigenvalues of C be $w_0 \geq w_1 \geq w_2$ We use the conventional *planarity* and *sphericity* scores

$$P = \frac{w_1 - w_2}{w_0} \quad S = \frac{w_2}{w_0 + w_1 + w_2} \tag{2}$$



(a) (b)

Fig. 3. Primitive shape fitting result in cluttered bin environments. (a) RGB img of cluttered scenes with various unknown objects. (b) Estimated primitive shapes with their poses overlaid. (Cylinder : Yellow, Sphere : Red) Despite heavy occlusion and object diversity, our shape matching module robustly decomposes and estimates the pose of each item without requiring mesh models.

If P/S exceeds a threshold τ_{plane}, the patch is classified as planar and its normal is taken as the eigenvector associated with w_2.

2) Huber Loss Formulation: For all primitives we minimize

$$\mathscr{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \rho\big(e_i(\theta)\big) \tag{3}$$

where θ collects the shape parameters and

$$\rho(e) = \begin{cases} \frac{1}{2}e^2 & |e| \leq \delta \\ \delta\big(|e| - \frac{1}{2}\delta\big) & \text{otherwise} \end{cases} \tag{4}$$

with a universal threshold δ.

2) Sphere fitting.: For center C and radius r,

$$e_i = \big|\, \|x_i - C\| - r \,\big| \qquad \min_{C,r} \mathscr{L}(C, r) \tag{5}$$

2) Cylinder fitting.: Let P_0 be a point on the axis, $\mathbf{a}$ a unit axis vector ($\|\mathbf{a}\| = 1$), and r the radius. The shortest distance from M_i to the axis is

$$d_i = \big\|(M_i - P_0) - \big[(M_i - P_0)^\top \mathbf{a}\big]\mathbf{a}\big\| \tag{6}$$

and the residual is $e_i = |d_i - r|$ We solve

$$\min_{P_0,\mathbf{a},r} \mathscr{L}(P_0, \mathbf{a}, r) \tag{7}$$

3) Model Selection: After convergence of each optimization, we compute the mean robust error $\bar{\mathscr{L}} = \mathscr{L}(\theta^\star)$. The primitive with lowest $\bar{\mathscr{L}}$ is selected; if every $\bar{\mathscr{L}}$ exceeds τ_{robust} the object is treated as a *non-primitive* and Step 4 is triggered.

E. Step 4 – Part Segmentation for Non-Primitive

In cases where an object cannot be accurately modeled using a single geometric primitive, we classify it as a *non-primitive*. Nonetheless, many such objects still exhibit underlying regularities or partial geometric structures that can be exploited for approximate modeling. To address this, we perform **part-level segmentation** of the object's point cloud. Instead of relying on purely geometric clustering, we leverage a foundation model capable of semantic part decomposition, such as Semantic SAM [38]. This model can hierarchically segment an object into meaningful parts without prior training on the specific object class. A key feature of this model is its controllable segmentation granularity (levels 1–6). To maintain a fully autonomous

pipeline without per-object tuning, we employ a fixed, global setting of level 4 for all non-primitive objects. This level was empirically determined to provide the optimal trade-off: it is fine enough to separate functionally distinct components, yet coarse enough to avoid the noisy over-segmentation that occurs at higher levels, ensuring parts are suitable for stable primitive fitting. Each resulting segment is then independently fitted with a geometric primitive, allowing us to construct a composite representation for non-primitive objects. After decomposing the object into primitive segments, we estimate its **centroid** C_{vol} to guide stable grasping. Let C_i and V_i denote the centroid and estimated volume of the i-th part. The centroid is calculated as the volume-weighted mean of all segment centers:

$$C_{\text{vol}} = \frac{1}{\sum_i V_i} \sum_i V_i \cdot C_i \tag{8}$$

This center approximates the object's centroid and provides a key geometric reference for improving grasp stability, particularly in the absence of a full CAD model. Even for globally non-primitive objects, this decomposition strategy enables approximate reasoning about object balance and orientation.

F. Step 5 – Shape-Informed Grasp Refinement

Once C_{vol} is estimated, we select the most stable grasp candidate from a set of predictions provided by a grasp planner. Candidate-based grasp planners typically generate multiple feasible grasps $\{\mathcal{G}_j = (\mathbf{p}_j, \mathbf{R}_j, Q_j)\}_{j=1}^{N}$, where each $\mathbf{p}_j$ and $\mathbf{R}_j$ denote the position and orientation of the j-th candidate, and Q_j is the planner's quality score.

To incorporate geometric stability, we rescore each candidate based on its distance to the estimated centroid:

$$\text{score}_j = \alpha \cdot Q_j - (1 - \alpha) \cdot \|\mathbf{p}_j - C_{\text{vol}}\| \tag{9}$$

where $\alpha \in [0, 1]$ balances the grasp planner's confidence and proximity to the centroid. The hyperparameter α balances the grasp planner's confidence and proximity to the centroid. For all experiments, we use a fixed α of 0.5, empirically chosen to weigh both terms equally. This fixed setting reinforces the module's plug-and-play nature by avoiding per-object tuning, and we found the system to be robust for α values in the range of [0.4, 0.6]. The final grasp point is then selected as:

$$\mathcal{G}_{\text{opt}} = \arg\max_j \ \text{score}_j \tag{10}$$

This approach ensures that the selected grasp is not only feasible, but also close to the object's center of mass, thereby reducing potential gravitational torque:

$$\boldsymbol{\tau}_j = (\mathbf{p}_j - C_{\text{vol}}) \times m\mathbf{g} \tag{11}$$

where $\mathbf{g} = [0, 0, -9.81]^\top$ denotes the gravity vector. As shown in Fig. 4, our shape matching module refines the initial grasp candidates predicted by a finger-based grasp planner by aligning them with the estimated centroids of non-primitive objects. This refinement shifts the grasp points away from unstable edge regions toward more stable and torque-minimizing positions near the object's centroid.

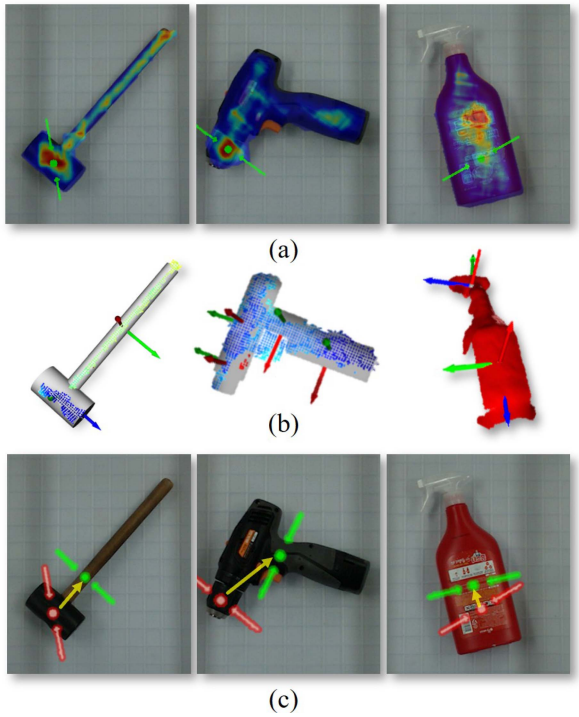


Fig. 4. Shape matching and grasp refinement results on non-primitive objects. (Cylinder : Gray) (a) Initial grasp candidates generated by a finger-based grasp planner. (b) Estimated primitive axes using our shape matching algorithm. (c) Refined grasp positions aligned with the estimated centroids (yellow arrows). Our method effectively guides the grasp away from edge regions toward more stable regions near the object's centroid, improving the likelihood of successful and torque-minimizing grasps.

V. Experiments

We conducted experiments based on a bin-picking and placing task.

A. Experiment Setup

We used a Rainbow Robotics RB10 6-DoF manipulator, an Intel Realsense L515 RGB-D camera, and a Universal gripper [2]. The experiments were conducted using a set of 20 distinct rigid objects commonly found in logistics environments. Deformable objects such as clothing or plastic bags were excluded from this study.

B. Computation Time Analysis

To validate the "plug-and-play" nature of our module, we measured the average latency of each component on an NVIDIA RTX 3090 GPU. Our module utilizes two distinct pipelines based on shape complexity. For primitive shapes, the process involves mask generation via FastSAM [17] (40 ms), followed by point cloud processing (120 ms) and shape recognition (5 ms), resulting in a total latency of approximately 165 ms. For non-primitive shapes, the pipeline is extended for more detailed analysis. After an initial recognition step, Semantic SAM [38] performs level 3 granular segmentation (adding over 400 ms), which is then followed by a secondary shape recognition pass on the newly defined parts (5 ms). This comprehensive process

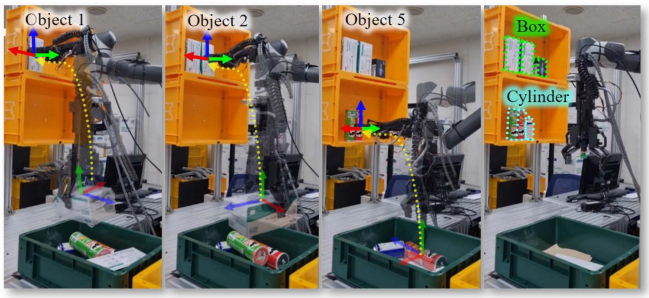


Fig. 5. 6 different boxes and cylinders were grasped and arranged by shape and size on a shelf. The box was placed up, and the cylinder was placed down.

TABLE I
POSE ESTIMATION ACCURACY OF THE SHAPE MATCHING MODULE

Object Type	e_{pose} (cm)	e_T (cm)	e_R (deg)
Box	0.2878	0.32	2.3
Cylinder	0.2986	0.34	2.7
Sphere	0.2840	0.29	–

Note that spherical objects are rotationally symmetric.

TABLE II
COMPARISON OF MEAN POSE ERRORS

Method	Rotation, e_R (deg)	Translation, e_T (cm)
Ours	2.506	0.270
GT+ICP	24.559	0.940

for non-primitive shapes totals approximately 570 ms. Both processing times are well within the 1–2 s cycle time required for industrial robotics, confirming our module's practicality and suitability for real-world applications.

C. Experiment Details

1) Shape Matching Module Accuracy Evaluation: To evaluate the accuracy of our shape matching module, we compare the estimated primitive shape and pose to a manually aligned reference primitive mesh fitted to the observed point cloud acquired using an Intel RealSense L515 depth camera. The evaluation is performed without access to CAD models or markers, relying solely on geometry-based alignment. We also evaluated our method against a standard ICP [19] baseline using single-view RGB-D data with 3 objects for 15 times. The ICP baseline was given an advantage: it received the ground truth shape and dimensions, leaving only the pose (R, T) to be estimated. In contrast, our method estimated all parameters—shape, dimensions, and pose—from the raw input.

We report the average point-to-surface distance between the predicted and reference meshes as a shape-level geometric error, denoted as e_{pose}. Additionally, the translation error e_T and rotation error e_R are separately calculated to provide a more interpretable pose difference. The geometric pose error is computed as the average point-to-surface distance between the transformed reference mesh M_1 and the estimated mesh M_2:

$$e_{\text{pose}} = \frac{1}{N} \sum_{x_i \in M_1} \min_{x_j \in M_2} \left\| \mathbf{R}x_i + \mathbf{T} - (\tilde{\mathbf{R}}x_j + \tilde{\mathbf{T}}) \right\| \quad \text{(cm)} \tag{12}$$

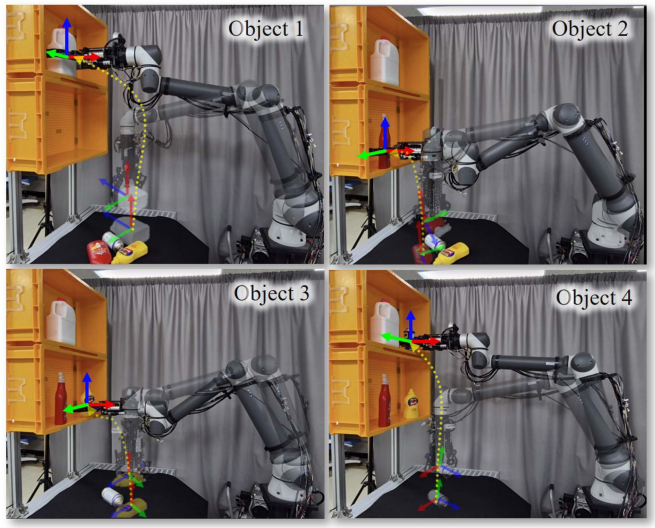


Fig. 6. Placing non-primitive objects from table to shelf using the proposed shape-informed grasping pipeline. Due to the asymmetry and shape complexity, precise control of both grasping orientation and placement position is critical for stable stowing. Our method enables reliable estimation of approximate pose and shape, resulting in consistent success across diverse object geometries.

The translation error is defined as:

$$e_T = \|\mathbf{T} - \tilde{\mathbf{T}}\| \quad \text{(cm)} \tag{13}$$

and the rotation error as:

$$e_R = \cos^{-1}\left(\frac{\text{trace}(\tilde{\mathbf{R}}^\top \mathbf{R}) - 1}{2}\right) \quad \text{(degrees)} \tag{14}$$

2) Grasping Success Rate Comparison With Other Methods: To evaluate the effectiveness of our shape-matching-based grasp refinement module, we conducted three grasping experiments across different object types and settings. The experiments compared our method with widely used grasp planners [3], [5], [6], [7]: All planners were tested both with and without our method, denoted as "+ Ours."

- *Bin to Bin (Suction Grasping):* 10 primitive objects were randomly placed in a bin. Each object was grasped and moved to a empty bin using suction-based methods. Each method was tested for 10 trials per object, totaling 100 trials.
- *Bin to Shelf (Primitive Objects):* 6 primitive objects (3 cylinders and 3 boxes) were placed in a bin. The task was to pick and stow each object neatly on a shelf. Fig. 5 shows a bin-to-shelf sorting task involving primitive-shaped objects. The system successfully estimates the pose and shape of each item and places them according to their geometry—boxes on the upper shelf and cylinders on the lower shelf—demonstrating the utility of geometric reasoning for structured placement. Each method was tested for 10 trials per object, totaling 60 trials.
- *Table to Shelf (Non-Primitive Objects):* 4 non-primitive objects were placed on a table. The task was to pick and stow each object on a shelf. Fig. 6 illustrates a table-toshelf placement task using four non-primitive objects. Even with irregular geometries, the proposed method

TABLE III
GRASPING SUCCESS RATE COMPARISON ACROSS TASKS

Method	Gripper Type	Bin to Bin	Bin to Shelf (pm)	Table to Shelf (non-pm)	Total Success	Success Rate (%)
Suction-Net 1B [7]	Suction only	55	–	–	–	–
Suction-Net 1B + Ours	Suction only	**78**	–	–	–	–
Dex-Net 4.0 [6]	Suction + Finger	73	23	15	111	55.5
Dex-Net 4.0 + Ours	Suction + Finger	**82**	**40**	**27**	**149**	**74.5**
FC-GQCNN [3]	Finger only	79	17	13	109	54.5
FC-GQCNN + Ours	Finger only	**83**	**48**	**31**	**162**	**81.0**
CoAS-Net [5]	Suction + Finger	92	31	19	142	71.0
CoAS-Net + Ours	Suction + Finger	**93**	**53**	**33**	**179**	**89.5**

Each column reports the number of successful grasps out of total trials.

TABLE IV
EXECUTION TIME AND PRODUCTIVITY IN BIN-TO-BIN TRANSFER TASK

Test #	Time (s)	Productivity (pieces/hour)
1	50.00	720.00
2	50.84	708.12
3	49.59	726.03
4	48.18	747.16
5	48.61	740.63
6	48.11	748.32
7	49.32	729.97
8	48.68	739.51
9	50.36	714.94
10	49.33	729.75
Average	**49.30**	**730.44**

stable and semantically aware placement in the target shelf region. Each method was tested for 10 trials per object, totaling 40 trials.

For each method, a grasp was considered successful only if the object was firmly picked and placed at the intended location without dropping.

3) Productivity: 10 different objects without prior information were selected and randomly placed in the work space (bin), perform the pick and place operation (until the entire bin is empty) and calculate the productivity from the time it took to move all objects to the bin next to it.

D. Experiment Result

1) Pose Estimation Accuracy: As shown in Table I, our module achieved high accuracy, with an average pose error of less than 0.3 cm and a rotation error of around 2-3 degrees from a single RGB-D view. As summarized in Table II, our approach was substantially more accurate and reliable. Our method achieved lower mean rotation and translation errors compared to the ICP baseline.

2) Grasping Success Rate Comparison: As shown in the Table III, our refinement module consistently and significantly improved success rates across all tested planners and scenarios. The performance gains were statistically notable for all planners ($p < 0.01$ in a two-proportion Z-test). For example, with our module, Dex-Net 4.0's overall success rate increased from 55.5%(95%CI: [48.5%, 62.5%]) to 74.5%(95%CI: [68.1%, 80.9%]).

The module's effectiveness was most notable in precision-placement tasks like the Bin-to-Shelf and Table-to-Shelf scenarios, where grasp and object orientation are critical. In these cases, our module boosted low baseline success rates from a 30-40% range to 65–80%.

3) Computation Time and Productivity: The average processing time for our refinement module was approximately 165ms(primitive)-570ms(non-primitive) per grasp. This latency is well within the cycle time of typical industrial robots, confirming the module's suitability for real-world deployment. The overall system, including robot motion, achieved an average productivity of 730.44 pieces per hour in the bin-to-bin task (Table IV), demonstrating its practical efficiency.

VI. CONCLUSION

We introduced a plug-and-play shape matching module to enhance robotic grasping for unknown objects. Our approach is distinguished by its truly training-free, model-free, and mesh-free nature, enabling immediate deployment with existing grasp planners without retraining. Its novelty lies not in the individual components but in their unique integration into a lightweight and efficient system. By using grasp candidates as automatic prompts for segmentation and refining grasps via a centroid-aware rescoring strategy, our module offers a practical solution for geometry-aware grasp refinement. Real-world experiments demonstrated a statistically notable improvement in grasp success rates, boosting the performance of baseline planners, especially in precision-placement tasks where object orientation is critical. However, we acknowledge our approach's limitations. Its reliance on primitive shape fitting is best suited for rigid objects, while its application to highly irregular or deformable objects remains a challenge. Future work will proceed in two directions: incorporating more complex shape representations to handle a wider variety of geometries, and validating our module on standard large-scale benchmarks for broader comparison with other methods.

REFERENCES

[1] Y. Jaghbeer, R. Hanson, and M. I. Johansson, "Automated order picking systems and the links between design and performance: A systematic literature review," *Int. J. Prod. Res.*, vol. 58, no. 15, pp. 4489–4505, 2020.
[2] S. Um, H. Jeong, C. S. Kim, I. Rhee, and H. R. Choi, "ReC-Gripper: A reconfigurable combined suction and fingered gripper for various logistics picking and stowing tasks," *IEEE Robot. Automat. Lett.*, vol. 9, no. 1, pp. 87–94, Jan. 2024.
[3] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 1357–1364, Apr. 2019.

[4] B. Xu, T. Hassan, and I. Hussain, "Improving reinforcement learning based moving object grasping with trajectory prediction," *Intell. Service Robot.*, vol. 17, no. 2, pp. 265–276, 2024.

[5] Y. G. Son et al., "CoAS-Net: Context-aware suction network with a large-scale domain randomized synthetic dataset," *IEEE Robot. Automat. Lett.*, vol. 9, no. 1, pp. 827–834, Jan. 2024.

[6] J. Mahler et al., "Learning ambidextrous robot grasping policies," *Sci. Robot.*, vol. 4, no. 26, 2019, Art. no. eaau4984.

[7] H. Cao, H. -S. Fang, W. Liu, and C. Lu, "SuctionNet-1billion: A large-scale benchmark for suction grasping," *IEEE Robot. Automat. Lett.*, vol. 6, no. 4, pp. 8718–8725, Oct. 2021.

[8] K. Kleeberger, C. Landgraf, and M. F. Huber, "Large-scale 6D object pose estimation dataset for industrial bin-picking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2019, pp. 2573–2578.

[9] M. Sundermeyer, Z.-C. Marton, M. Durner, M. Brucker, and R. Triebel, "Implicit 3D orientation learning for 6D object detection from RGB images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 699–715.

[10] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1521–1529.

[11] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Proc. Robot.: Sci. Syst. (RSS)*, Pittsburgh, PA, USA, Jul. 2018.

[12] C. Wang et al., "DenseFusion: 6D object pose estimation by iterative dense fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3343–3352.

[13] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "GDR-Net: Geometry-guided direct regression network for monocular 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16611–16621.

[14] S. Iwase et al., "ZeroGrasp: Zero-shot shape reconstruction enabled robotic grasping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 17405–17415.

[15] S. Li et al., "ShapeGrasp: Zero-shot task-oriented grasping with large language models through geometric decomposition," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2024, pp. 10527–10534.

[16] A. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 4015–4026.

[17] X. Zhao et al., "Fast segment anything," 2023, *arXiv:2306.12156*.

[18] C. Zhang et al., "Faster segment anything: Towards lightweight sam for mobile applications," 2023, *arXiv:2306.14289*.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, 1981.

[20] T. Hodan et al., "BOP challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 5610–5619.

[21] Y. Labbé, J. Carpentier, M. Aubry, and J. Sivic, "MegaPose: 6D pose estimation of novel objects via render & compare," in *Proc. Conf. Robot Learn. (CoRL)*, Auckland, New Zealand, 2022.

[22] V. N. Nguyen, T. Groueix, M. Salzmann, and V. Lepetit, "GigaPose: Fast and robust novel object pose estimation via one correspondence," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 9903–9913.

[23] A. Caraffa, D. Boscaini, A. Hamza, and F. Poiesi, "Freeze: Training-free zero-shot 6D pose estimation with geometric and vision foundation models," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 414–431.

[24] J. Lin, L. Liu, D. Lu, and K. Jia, "SAM-6D: Segment anything model meets zero-shot 6D object pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 27906–27916.

[25] C. Wang et al., "6-Pack: Category-level 6D pose tracker with anchor-based keypoints," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 10059–10066.

[26] T. Lee, B. Wen, M. Kang, G. Kang, I. S. Kweon, and K. -J. Yoon, "Any6D: Model-free 6D pose estimation of novel objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 11633–11643.

[27] J. Lee, Y. Cabon, R. Brégier, S. Yoo, and J. Revaud, "MFOS: Model-free & one-shot object pose estimation," in *Proc. AAAI Conf. Artif. Intell.*, 2024, vol. 38, no. 4, pp. 2911–2919.

[28] L. -C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[29] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 779–788.

[30] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2961–2969.

[31] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9157–9166.

[32] T. H. Bui et al., "Deep learning based 6-DoF antipodal grasp planning from point cloud in random bin-picking task using single-view," *IEEE Robot. Automat. Lett.*, vol. 8, no. 8, pp. 5196–5203, Aug. 2023.

[33] H. Li et al., "SegGrasp: Zero-shot task-oriented grasping via semantic and geometric guided segmentation," 2024, *arXiv:2410.08901*.

[34] A. Rashid et al., "Language embedded radiance fields for zero-shot task-oriented grasping," in *Proc. Conf. Robot. Learn.*, 2023 pp. 178–200.

[35] A. D. Vuong et al., "Grasp-anything: Large-scale grasp dataset from foundation models," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 14030–14037.

[36] N. Tsagkas, J. Rome, S. Ramamoorthy, O. M. Aodha, and C. X. Lu, "Click to grasp: Zero-shot precise manipulation via visual diffusion descriptors," in *Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst.*, 2024, pp. 11610–11617.

[37] G. Taylor and L. Kleeman, *Visual Perception and Robotic Manipulation: 3D Object Recognition, Tracking and Hand-Eye Coordination*. Berlin, Germany: Springer, 2006.

[38] F. Li et al., "Semantic-SAM: Segment and recognize anything at any granularity," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Milan, Italy, 2024.