2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Seg2Grasp: A Robust Modular Suction Grasping in Bin Picking

Hye-Jung Yoon[1*] Juno Kim[1*] Yesol Park[1*] Jun-Ki Lee[2] Byoung-Tak Zhang[1,2,3]

***Abstract*— Current bin picking methods that rely heavily on end-to-end learning often falter when confronted with unfamiliar or complex objects in unstructured environments. To overcome these limitations, we introduce Seg2Grasp, a modular pipeline designed for robust suction grasping in dynamic and cluttered bin scenarios. Seg2Grasp is built on a three-step process: *Segmentation*, *Grasping*, and *Classification*. The *Segmentation* module employs a Transformer-based model to generate class-agnostic object masks from RGB-D images, ensuring accurate detection across various conditions. The *Grasping* module uses surface normals and mask proposals to determine the optimal suction points, enhancing grasp success. Finally, the *Classification* module leverages fine-tuned open-vocabulary Mask-CLIP for precise object identification, enabling versatile handling of diverse objects. Real-world robotic experiments demonstrate that Seg2Grasp outperforms existing methods in success rates and adaptability, establishing it as a powerful tool for automated bin picking in industrial settings.**

I. INTRODUCTION

In the field of industrial automation, bin picking is a critical yet challenging task, especially when applied to environments that are dynamic and contain unknown objects. Such settings are characterized by a variety of unpredictable factors, including fluctuating lighting, different camera viewpoints, and the presence of objects that the system has not previously encountered. These challenges demand a solution that is both precise and adaptable, capable of handling the complexities of real-world industrial scenarios.

Existing bin picking systems predominantly utilize end-to-end learning-based methods [1]–[5], which aim to directly map sensory inputs to outputs. Although these approaches have shown success in controlled environments, they frequently fall short in terms of adaptability—the ability to function effectively across diverse conditions—and robustness, which refers to maintaining performance despite variations in the environment. The relatively simplistic nature of these end-to-end methods limits their generalization capabilities, particularly when faced with novel objects or unfamiliar configurations.

To address these limitations, we propose Seg2Grasp, a novel pipeline designed to enhance the robustness and adaptability of suction-based bin picking systems in dynamic environments. Unlike traditional end-to-end models, Seg2Grasp leverages a modular architecture comprising three core components: *Segmentation, Grasping,* and *Classification*. This modular design, depicted in Fig. 1, allows each component to specialize in a specific task, providing greater flexibility and improved performance across a wider range of scenarios.

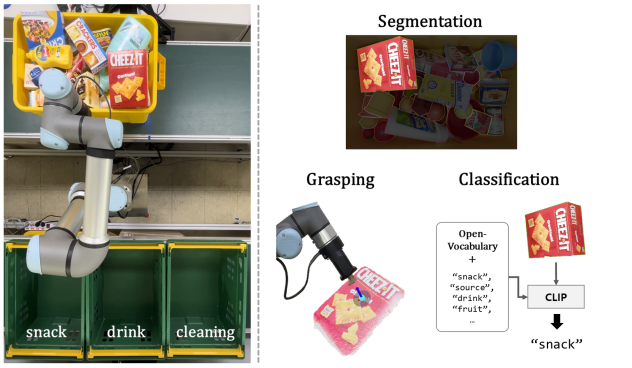


Fig. 1. **Illustration of proposed system.** Our bin picking system can segment variously shaped, class-agnostic objects in a dynamic environment and proceed with grasping and classifying them.

The Segmentation module utilizes a Transformer-based model to generate class-agnostic object masks from depth-weighted RGB images. This method enables accurate object segmentation across diverse conditions, overcoming one of the primary challenges in bin picking—reliable detection and segmentation of objects regardless of environmental variability.

These object masks are then passed to the Grasping module, which, combined with surface normal data, identifies optimal grasping points. This process significantly improves the precision of object manipulation, enabling the system to successfully grasp a wide range of objects with varying geometries and orientations.

Finally, the Classification module employs an open-vocabulary classification system to categorize a wide array of objects. This capability allows the system to adapt to and handle objects without prior explicit knowledge of each object, greatly expanding the operational versatility of the robotic system.

To validate the effectiveness of Seg2Grasp, we conducted a series of robotic experiments in dynamic environments, utilizing various bins and camera configurations to simulate realistic bin picking scenarios. Our approach consistently outperformed existing end-to-end methods [3], [4] in terms of success rates, particularly in unstructured and dynamic environments.

In summary, Seg2Grasp addresses the significant gaps in

*Authors have equal contributions
[1]Interdisciplinary Program in AI, Seoul National University
[2]Artificial Intelligence Institute, Seoul National University
[3]Department of Computer Science, Seoul National University

This work was partly supported by Institute of Information communications Technology Planning Evaluation (IITP) grant funded by the Korea government(MSIT) [RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] and (RS-2021-II212068-AIHub/10%, RS-2021-II211343-GSAI/15%, 2022-0-00951-LBA/15%, 2022-0-00953-PICA/20%), NRF (RS-2024-00353991/20%, RS-2023-00274280/10%), and KEIT (RS-2024-00423940/10%) grant funded by the Korean government.

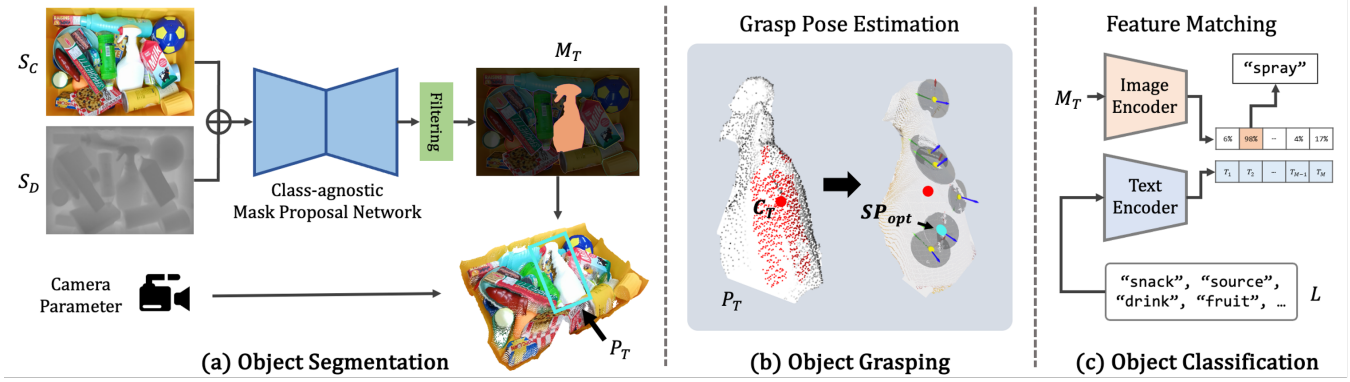


Fig. 2. **Overview of our modular bin picking system.** The system comprises three modules: class-agnostic object segmentation, object grasp pose estimation, and open-vocabulary object classification. The black arrow indicates the flow of operations within the entire system.

adaptability and robustness found in current bin picking technologies. By integrating specialized modules for segmentation, grasping, and classification, Seg2Grasp not only enhances performance in complex environments but also sets a new benchmark for the potential of modular approaches in industrial automation.

II. Related Work

In robotics, bin picking in unstructured environments is a complex task that requires robust object recognition, accurate grasp pose estimation, and flexible object classification. This section reviews the recent advances in these areas, which are essential to the development of our Seg2Grasp framework.

A. Class-agnostic Object Instance Segmentation

Class-agnostic Object Instance Segmentation is critical for enabling robots to identify and segment objects that have not been previously encountered. Traditional segmentation methods often struggle in unstructured and cluttered environments [6], [7]. Unlike category-based instance segmentation, which focuses on known categories [8]–[10], Class-agnostic Object Instance Segmentation aims to generalize segmentation to arbitrary objects [11]–[13].

Recent advancements include UOAIS-Net [14], which addresses occlusion and amodal perception in robotic manipulation but faces challenges with occluded areas, leading to potential misinterpretations in cluttered scenarios. Another approach, MSMFormer [15], improves segmentation accuracy by integrating a clustering method with Mask2Former [9]. However, this comes at the cost of increased computational requirements, highlighting a precision-efficiency trade-off.

Our approach adopts Mask2Former for its robust segmentation capabilities, particularly its ability to maintain high accuracy without the need for clustering. This selection strikes a balance between efficiency and performance, making it suitable for the dynamic environments typical of bin picking tasks.

B. Grasp Pose Estimation

Grasp pose estimation is fundamental in robotic manipulation, where accurate identification of grasp points is necessary for successful object handling. This process typically relies on RGB-D and point cloud data [16]. There are two main approaches: learning-free methods [17] and learning-based methods [2], [4], [5], [18]–[20].

Learning-free approaches often depend on CAD models to determine grasp candidates for recognized objects [17], [21]. On the other hand, learning-based methods, particularly those utilizing deep learning, have gained prominence due to their improved accuracy and reliability in controlled environments [22]. These models are generally trained on depth or RGB-D datasets [3], [4], which allows for precise and consistent grasping. However, their performance tends to degrade in unstructured environments where the conditions differ from the training data.

To address the limitations of traditional methods, our grasping algorithm is specifically designed to maintain robustness in diverse and unpredictable environments. This approach ensures that grasp performance remains effective even under the variable conditions typical of real-world industrial settings, thus filling a significant gap in current robotic manipulation strategies.

C. Open-Vocabulary Classification

Open-vocabulary classification represents a significant advancement in image classification, allowing systems to recognize a broader range of visual concepts beyond fixed label sets, enhanced by natural language processing. Language-Image Pre-training models, such as CLIP [23], have been instrumental in bridging the gap between visual data and textual descriptions.

Despite its strengths, CLIP and similar models often experience reduced performance when applied to images that deviate from their training distributions. OVSeg [24] addresses this issue by extracting CLIP features directly from segmentation masks, excluding background elements to enhance accuracy. This refined approach improves the model's performance, particularly when dealing with images significantly different from those in the training set.

In our Seg2Grasp framework, we utilize an open-vocabulary classification system enhanced by Mask-CLIP, which enables effective object recognition across a wide variety of scenarios. This flexibility further expands the operational capabilities of the system in dynamic bin picking environments.

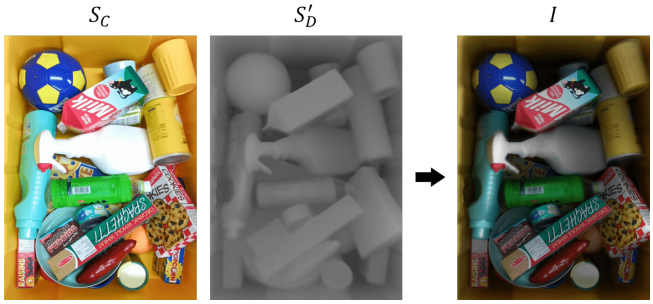


Fig. 3. **Mask proposal network input.** The input I to the mask proposal network consists of a fusion of the color image S_C and the inverted depth image S'_D.

III. Problem Statement

In the context of bin picking, our primary objective is to develop a system capable of accurately predicting a set of feasible grasp poses and their corresponding object labels. This system must effectively transfer arbitrary objects from a source bin to a target bin, even in unstructured and cluttered environments. Achieving this goal requires a comprehensive understanding of the scene, including precise segmentation, grasp pose estimation, and object classification.

Given a scene image S, the task is to segment the image into distinct objects, represented by a set of masks $\{M_i \mid 1 \leq i \leq N\}$, where N is the number of objects identified in the scene. Each mask M_i corresponds to an individual object, facilitating further analysis. The target object in the scene is denoted by T.

The grasp configuration for the target object T is defined as $g = (SP_{opt}, NV_{opt})$, where SP_{opt} represents the optimal suction point on the object, and NV_{opt} is the associated normal vector at that point. Additionally, each object is assigned a label l, which indicates its categorical identity and is essential for its correct placement.

The core challenge we address is to develop a predictive model f that, for each input mask M_i, accurately maps to both a grasp configuration $g_i = (SP_{opt,i}, NV_{opt,i})$ and an object label l_i. This mapping is expressed as:

$$f : M_i \longmapsto (g_i, l_i), \quad \text{for } i = 1, 2, \ldots, N.$$

This formulation is critical for determining not only the optimal grasping approach but also for understanding the object's identity, ensuring its correct transfer and placement.

To solve this problem, we propose a three-step method:

1) **Object Segmentation:** A Transformer-based model processes the scene image S to generate class-agnostic object masks, producing the set $\{M_i \mid 1 \leq i \leq N\}$. The target object T is selected from this set.
2) **Object Grasping:** For the selected target object T, surface normals are used to calculate the optimal grasp pose $g = (SP_{opt}, NV_{opt})$, ensuring the most suitable suction point is identified.
3) **Object Classification:** An open-vocabulary classification module assigns a label l to each object, facilitating accurate identification and enabling precise placement.

Algorithm 1 Mask Filtering

Require: Set of masks M
Ensure: Target object T, Largest planar area A_T

```
for each mask M_i in M do
    Initialize: LP_i ← null, Best inlier count In_best,i ← 0
    for iteration k do
        Select a random set X from M_i
        Estimate plane Pl_k from X
        Determine inliers In_k: Points in M_i fitting Pl_k within tolerance ε
        if |In_k| > In_best,i then
            LP_i ← Pl_k, In_best,i ← |In_k|
        end if
    end for
    Calculate centroid C_i = (x_i, y_i, z_i) of In_best,i
end for
T = arg max_i(z_i)
A_T = Area of LP for T
return T, A_T
```

Our modular approach, which separates segmentation, grasping, and classification into distinct yet interconnected components, offers enhanced adaptability and robustness. This flexibility is particularly advantageous in dynamic and variable environments, where conditions may differ significantly from the controlled settings typically assumed by other methods.

IV. Method

Seg2Grasp is structured into three main modules: (1) *Object Segmentation*, (2) *Object Grasping*, and (3) *Object Classification*. The overall framework is illustrated in Fig. 2

A. Object Segmentation

The object segmentation module processes the scene image S, and distinguishes individual objects within the scene. The target object selected for grasping is denoted as T.

Input Preparation. The input to the network, denoted as I, is constructed by combining the RGB image S_C with a modified depth image S'_D. To enhance the depth information, the depth image S_D is normalized and inverted:

$$S'_D = 1 - \left(\frac{S_D - S_{D_{\min}}}{S_{D_{\max}} - S_{D_{\min}}} \right), \tag{1}$$

where $S_{D_{\min}}$ and $S_{D_{\max}}$ are the minimum and maximum values in S_D. This inversion highlights closer objects, improving the contrast between objects at different depths. The final input I is the product of the color image S_C and the inverted depth image S'_D, which enhances object differentiation. This process is shown in Fig. 3

Mask Proposing. We developed a model for proposing class-agnostic instance masks using the standard Mask Transformer model [9]. The model outputs a set of masks $M = \{M_i \mid 1 \leq i \leq 100\}$ for each input image, where each mask captures an individual object. By using the fused input I

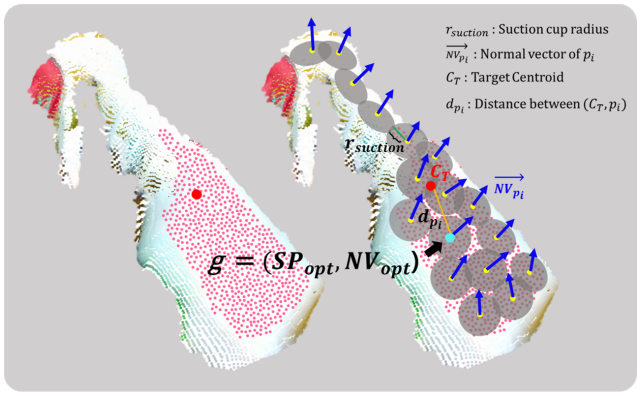


Fig. 4. **Grasp pose estimation.** The cyan dot represents the optimal grasp point on the target object, identified by the highest final score (detailed in Alg. 2).

and focusing on class-agnostic outputs, the model generalizes well across diverse and class-agnostic objects.

Mask Filtering. To identify the target object T, we refine the set of generated masks M using a filtering process. This process leverages the RANSAC algorithm to estimate the largest planar area (LP_i) within each mask M_i. For each mask, RANSAC iteratively selects a subset of points to estimate a plane and identifies inliers—points that fit the plane within a specified tolerance. The plane with the highest inlier count is selected, and the centroid $C_i = (x_i, y_i, z_i)$ of these inliers is computed. The mask with the highest z-coordinate centroid is then chosen as the target object T, ensuring that the most elevated object in the scene, and thus the most accessible for grasping, is selected. The complete mask filtering process is detailed in Alg. 1.

B. Object Grasping

The object grasping module processes the selected target object T to determine the optimal grasp point $g = (SP_{opt}, NV_{opt})$, as depicted in Fig. 4. This process, detailed in Alg. 2, utilizes surface normals derived from the object to identify the most suitable suction pose for grasping.

Our approach differs from traditional methods [3], [4], which often depend on extensive training data and specific environmental setups. Instead, we rely solely on RGB-D imagery to compute the optimal suction point, providing greater flexibility across various objects and scenarios. This is particularly advantageous when using suction grippers, which can attach to suitable surface points on diverse objects.

Suction Point Evaluation. The optimal suction point is determined by analyzing the preprocessed point cloud P_T. The process involves navigating through P_T to identify potential suction areas, with the region of interest adjusted to match the suction cup's dimensions. Each potential grasp point is evaluated based on its alignment with the normal vector NV_T of the central point, considering a tolerance angle ϕ.

To determine the most suitable suction point $g = (SP_{opt}, NV_{opt})$, each candidate site is assessed using three metrics: surface angle θ, distance d from the centroid C_T, and the count of graspable points $\mathscr{G}$. These metrics are normalized and combined into a composite score that prioritizes angular alignment, proximity, and point count:

Algorithm 2 Pose Estimation for Optimal Suction Point

Require: Point cloud P_T, 2D mask M_T, centroid C_T, suction radius $r_{suction}$, angle tolerance ϕ, min grasped points ψ, angle threshold δ
Ensure: Optimal suction point SP_{opt}, normal vector NV_{opt} at SP_{opt}

Preprocessing:
Filter and downsample P_T using M_T.
Calculate surface normals NV_T and angles Θ_T.
Point Evaluation:
for each point p in P_T **do**
  if angle $\Theta_T(p) > \delta$ **then**
    Define region R_{near} around p within $r_{suction}$.
    Filter points in R_{near} with normals nearly parallel to C_T within ϕ.
    if count of filtered points $> \psi$ **then**
      Determine plane normal N_{plane} for these points.
      Aggregate N_{plane}, its angle, distance to C_T, and inlier count.
    end if
  end if
end for
Aggregate and weight selection criteria S_θ, S_d, S_g with weights w_θ, w_d, w_g.
Identify SP_{opt} and NV_{opt} using weighted criteria.
return SP_{opt}, NV_{opt}

$$\text{FinalScore}(g) = w_\theta S_\theta(\theta) + w_d S_d(d) + w_g S_g(|\mathscr{G}|) \quad (2)$$

Here, S_θ, S_d, and S_g are the normalized score functions for each metric, with w_θ, w_d, and w_g as their respective weights, satisfying the condition $w_\theta + w_d + w_g = 1$.

C. Object Classification

The classification module operates in parallel with the grasping mechanism, identifying the target object T and assigning it a corresponding label l, thereby enabling precise placement.

Preparation for CLIP. We employ a fine-tuned MaskCLIP model for open-vocabulary object classification, following the methodology outlined in previous research [24]. The model processes the mask proposals generated earlier, with a prediction branch specifically tailored for masked inputs. A CLIP-based branch then computes the class probabilities $\hat{p}_{i,k}$ for each mask. A key challenge arises from the fact that CLIP, being trained primarily on images with natural backgrounds, exhibits reduced effectiveness when working with masked inputs that contain large areas of blank space.

Fine-Tuning Mask-CLIP. Masked images, when tokenized for CLIP, often lead to zero tokens due to the extensive blank areas, indicating a domain shift. To address this, we apply mask prompt tuning. This involves enhancing the tensorized masked images with learnable prompt tokens

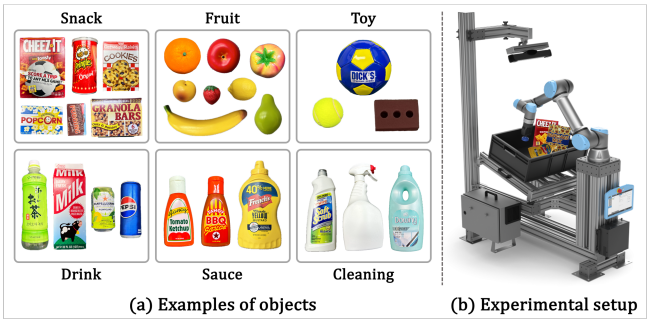


Fig. 5. **Experimental setup.** (a) Examples of objects used in the bin picking experiments. (b) Experimental environment using the UR5e robot.

derived from a binary mask, helping to preserve essential boundary information and improve classification accuracy.

Feature Matching. After fine-tuning the Mask-CLIP model, we measure the cosine similarity between the features of the masked images and the text descriptors of the categories. This approach enables precise object classification by effectively matching the visual features to the corresponding textual descriptions.

V. EXPERIMENTS

This section presents the experimental setup and results, which demonstrate the efficacy of the proposed method in addressing our problem context.

A. Implementation Detail

1) Segmentation Module: Accurate object segmentation within cluttered bin environments, especially when encountering class-agnostic items, requires a model that generalizes well across different scenarios. This capability is crucial for successful bin picking tasks.

Dataset. Given the scarcity of comprehensive real-world datasets for bin picking, we utilized the UOAIS-SIM dataset [14], which includes 50,000 photorealistic RGB-D images of objects in bin settings. This dataset bridges the gap between simulation and reality, offering extensive exposure to various object scenarios critical for training robust models.

Training. We optimized our segmentation module using the Tversky loss function, which addresses class imbalance by focusing on class-agnostic object detection. The final loss combines Tversky loss and a classification loss component to ensure comprehensive learning. Training was conducted on the UOAIS-SIM dataset using the AdamW optimizer with a learning rate of $1e-4$ and a batch size of 4 over five epochs.

2) Classification Module: To enhance the accuracy of open-vocabulary object classification with mask-only inputs, fine-tuning is imperative. This adjustment ensures the system can effectively recognize objects across a wide range of categories.

Dataset. For our experiment, we constructed a product database using the 'Product Image Dataset' [25]. Given the variability in object orientations within bins, we developed a 'mask-category' dataset. This dataset encompasses 720,000 product mask images across 53 main categories (e.g., snack, drink, dairy), generated from photographs of 10,000 objects captured from 72 different angles. This comprehensive approach allows our model to better recognize objects from any orientation, enhancing the robustness of our classification module.

Training. The OpenCLIP framework [26] was fine-tuned using our specialized dataset. We used 90% of the data for training and 10% for evaluation, ensuring balanced exposure across all categories. The model was optimized for accurate classification of masked images, with training conducted over 10 epochs using a ViT-L/14 CLIP variant.

B. Real-robot Experiments

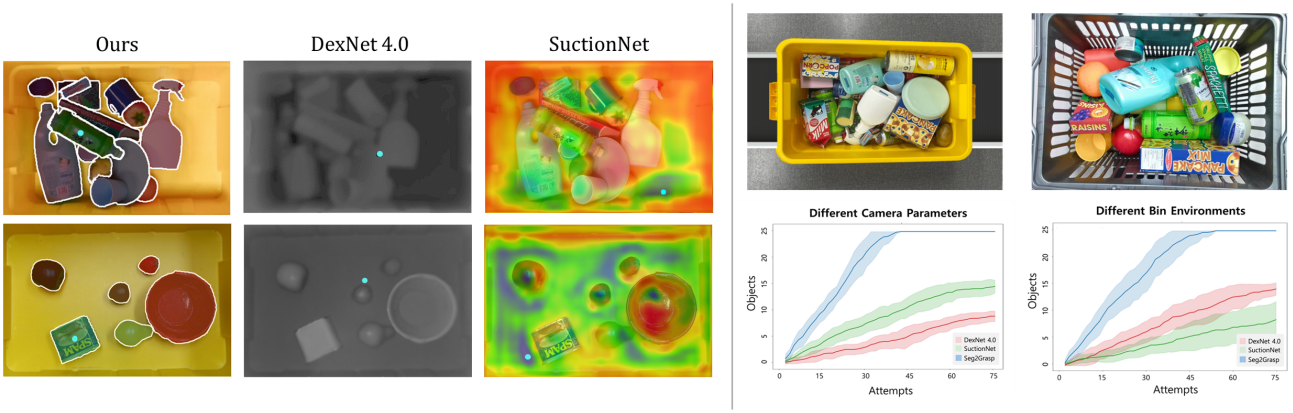
To demonstrate the robustness of our system, we designed three distinct experimental setups: 1) *Optimal Conditions*, to establish a baseline for performance; 2) *Varying Camera Parameters*, to examine how changes in visual input affect system performance; and 3) *Different Bin Environments*, to assess the system's adaptability to changes in the surrounding environment.

Experimental Setup. For our experiments, we selected a diverse range of objects, including boxes, cylinders, spheres, and various irregular shapes, as shown in Fig. 5 (a). This selection ensured a comprehensive evaluation of our system's adaptability. A UR5e robotic arm equipped with a Robotiq AirPick Vacuum Gripper was used for object manipulation, while RGB-D data was captured using an Azure Kinect DK Camera, strategically positioned at an elevated angle, as shown in Fig. 5 (b). The experiment was designed to focus on objects compatible with the vacuum gripper, allowing us to optimize the evaluation of the system's performance in real-world scenarios.

Evaluation Metrics. Our bin picking task focuses on achieving the complete removal of items from a bin, evaluated through three principal metrics: The pick success rate, $pr = \frac{N_g}{N_t}$, quantifies the effectiveness of grasps, where N_g represents the number of successful grasps and N_t the total number of attempts. The object success rate, $or = \frac{N_g}{N_o}$, measures the accuracy of object handling, with N_g indicating the number of objects successfully handled and N_o the total count of objects. The segmentation success rate, $sr = \frac{N_s}{N_o}$, assesses the precision of segmentation, where N_s signifies successfully segmented items, accurate within a tolerance of $\pm 20\%$ of the ground truth area. Together, these critical metrics facilitate a comprehensive evaluation of our system's performance.

1) Optimal Conditions Experiments: We conducted experiments under conditions optimized for each method, ensuring that the environment was adjusted to provide the best possible performance. In this idealized setting, the objects and scenarios used in the experiments closely matched the data on which each model was trained. This approach allowed us to fairly assess the capabilities of each system in an environment that reflects its training conditions.

The evaluation was conducted across three levels of object distribution and complexity, progressively increasing the challenge to each model, as presented in Tab. I. In the simplest conditions, all methods performed comparably,



(a) Limitations of different models – Failure Cases (b) Different Camera Parameters (c) Different Bin Environments

Fig. 6. **Experimental results.** (a) Failure cases highlighting the limitations of different models. (b) Setup and outcomes with different camera parameters. (c) Setup and outcomes with different bin environments.

with Seg2Grasp achieving a *pr* of 89%, slightly higher than DexNet 4.0 at 85% and SuctionNet at 72%. As the complexity increased, Seg2Grasp's performance advantage became more pronounced. Under the most challenging conditions, Seg2Grasp maintained a *pr* of 79%, significantly outperforming DexNet 4.0 at 28% and SuctionNet at 29%.

Similarly, Seg2Grasp excelled in object handling and segmentation accuracy. It achieved an *or* of 86% and a *sr* of 83% in the hardest scenarios, whereas DexNet 4.0 and SuctionNet showed considerable drops in performance, with *or* and *sr* both falling below 31%.

Both DexNet 4.0 and SuctionNet exhibit notable weaknesses under certain conditions. DexNet 4.0 struggles particularly with object overlap, making it difficult to grasp objects when heavier items are stacked on top. On the other hand, SuctionNet is highly sensitive to its environment and the specific objects it encounters, often misidentifying the yellow background as a suction point, which creates significant bottlenecks in the process. These issues are illustrated in Fig. 6 (a).

These results underscore Seg2Grasp's superior robustness and adaptability, particularly under complex and variable conditions. The ability to maintain high accuracy across different levels of difficulty highlights its potential for real-world applications where precision and adaptability are critical.

2) Varying Camera Parameters Experiments: To evaluate the adaptability of the models under varying visual conditions, we conducted experiments focusing on the impact of different camera heights. The camera heights were altered to 50 cm (optimal), 100 cm, and 150 cm to assess how each model handles significant variations in the depth data, which are crucial for reliable object detection and grasping.

Over 10 experimental sets, each model was allowed up to 75 attempts per set to pick objects from the bin. As depicted in Fig. 6. (b), SuctionNet and DexNet 4.0 showed considerable performance degradation at higher camera positions. DexNet 4.0, in particular, struggled due to increased noise in depth perception, successfully picking only 10 to 15 objects. SuctionNet fared slightly better but still

TABLE I
RESULTS OF OPTIMAL CONDITIONS EXPERIMENTS ACROSS THREE LEVELS OF DIFFICULTY.

Level	**Setup** (Layer, Item)	**Method**	*pr*	*or*	*sr*
EASY	Single, Trained	DexNet 4.0 [3]	0.85	0.94	-
		SuctionNet [4]	0.72	0.92	0.93
		Ours	0.89	0.96	0.91
MEDIUM	Double, Mixed	DexNet 4.0 [3]	0.41	0.61	-
		SuctionNet [4]	0.51	0.53	0.43
		Ours	0.87	0.91	0.89
HARD	Complex, Novel	DexNet 4.0 [3]	0.28	0.31	-
		SuctionNet [4]	0.29	0.23	0.26
		Ours	0.79	0.86	0.83

demonstrated a significant drop in performance. In contrast, Seg2Grasp maintained high performance across all tested heights, showcasing remarkable resilience and consistency. These results highlight Seg2Grasp's robustness in adapting to changes in camera parameters, a critical factor in real-world robotic applications where environmental conditions can vary.

3) Different Bin Environments Experiments: We also examined the models' performance across different bin types to further test their adaptability. Starting with a large yellow bin as the baseline, we varied the bin types, including a shopping basket and a small white box, to simulate different real-world conditions. Each experimental set included 25 objects, none of which were encountered during training, with a maximum of 75 attempts allowed per set.

The outcomes, as shown in Fig. 6. (c), reveal that both SuctionNet and DexNet 4.0 faced significant challenges when the bin type was changed. DexNet 4.0's reliance on depth data proved advantageous in scenarios with consistent lighting and minimal RGB variations, slightly outperforming SuctionNet in certain cases. However, neither model succeeded in fully clearing the bin under all conditions. On the other hand, Seg2Grasp consistently achieved high success rates regardless of the bin type, underscoring its superior adaptability and effectiveness in diverse environments.

TABLE II
RESULT OF THE OPEN-VOCABULARY CLASSIFICATION.

METHOD	Top#1 Acc.	Top#3 Acc.
CLIP [23]	66.1%	78.9%
MASK-CLIP [24]	73.8%	84.7%

This consistency reinforces Seg2Grasp's capability to handle varying operational settings, making it a robust solution for industrial bin picking tasks.

C. Open-Vocabulary Classification Experiment

We conducted an experiment to compare the performance of Mask-CLIP against the standard CLIP model using our 'mask-category' evaluation dataset. This involved computing CLIP feature vectors for evaluation mask images with both models and calculating their cosine similarity to 53 predefined categories.

The results, summarized in Tab. II, indicate that Mask-CLIP outperforms the standard CLIP model in open-vocabulary classification tasks for mask-based inputs, achieving a Top#1 Accuracy of 73.8% and a Top#3 Accuracy of 84.7%, compared to CLIP's 66.1% and 78.9%, respectively.

VI. CONCLUSION

This paper presented Seg2Grasp, a modular pipeline for enhancing suction grasping in dynamic, unstructured bin environments. The approach consists of three key modules: (1) Class-Agnostic Object Segmentation, isolating class-agnostic objects; (2) Grasp Pose Estimation using Surface Normals, identifying optimal suction points; and (3) Open-Vocabulary Object Classification, utilizing fine-tuned Mask-CLIP models for accurate identification. Real-robot experiments demonstrated that Seg2Grasp outperforms existing end-to-end bin picking models in both robustness and accuracy, highlighting the effectiveness of a modular strategy for automated bin picking.

REFERENCES

[1] 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.

[2] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning," in *2018 IEEE International Conference on robotics and automation (ICRA)*, pp. 5620–5627, IEEE, 2018.

[3] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

[4] H. Cao, H.-S. Fang, W. Liu, and C. Lu, "Suctionnet-1billion: A large-scale benchmark for suction grasping," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8718–8725, 2021.

[5] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, *et al.*, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," *The International Journal of Robotics Research*, vol. 41, no. 7, pp. 690–705, 2022.

[6] A. Ückermann, R. Haschke, and H. Ritter, "Real-time 3d segmentation of cluttered scenes for robot grasping," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pp. 198–203, IEEE, 2012.

[7] M. Q. Mohammed, L. C. Kwek, S. C. Chua, A. Al-Dhaqm, S. Nahavandi, T. A. E. Eisa, M. F. Miskon, M. N. Al-Mhiqani, A. Ali, M. Abaker, *et al.*, "Review of learning-based robotic manipulation in cluttered environments," *Sensors*, vol. 22, no. 20, p. 7938, 2022.

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable detr: Deformable transformers for end-to-end object detection," *arXiv preprint arXiv:2010.04159*, 2020.

[9] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

[10] Z. Tian, C. Shen, and H. Chen, "Conditional convolutions for instance segmentation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 282–298, Springer, 2020.

[11] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "The best of both modes: Separately leveraging rgb and depth for unseen object instance segmentation," in *Conference on robot learning*, pp. 1369–1378, PMLR, 2020.

[12] C. Xie, Y. Xiang, A. Mousavian, and D. Fox, "Unseen object instance segmentation for robotic environments," *IEEE Transactions on Robotics*, vol. 37, no. 5, pp. 1343–1359, 2021.

[13] Y. Xiang, C. Xie, A. Mousavian, and D. Fox, "Learning rgb-d feature embeddings for unseen object instance segmentation," in *Conference on Robot Learning*, pp. 461–470, PMLR, 2021.

[14] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee, "Unseen object amodal instance segmentation via hierarchical occlusion modeling," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 5085–5092, IEEE, 2022.

[15] Y. Lu, Y. Chen, N. Ruozzi, and Y. Xiang, "Mean shift mask transformer for unseen object instance segmentation," *arXiv preprint arXiv:2211.11679*, 2022.

[16] A. Cordeiro, L. F. Rocha, C. Costa, P. Costa, and M. F. Silva, "Bin picking approaches based on deep learning techniques: A state-of-the-art survey," in *2022 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*, pp. 110–117, 2022.

[17] F. Spenrath and A. Pott, "Gripping point determination for bin picking using heuristic search," *Procedia CIRP*, vol. 62, pp. 606–611, 2017.

[18] Z. Dong, S. Liu, T. Zhou, H. Cheng, L. Zeng, X. Yu, and H. Liu, "Ppr-net: point-wise pose regression network for instance segmentation and 6d pose estimation in bin-picking scenarios," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1773–1780, IEEE, 2019.

[19] X. Li, R. Cao, Y. Feng, K. Chen, B. Yang, C.-W. Fu, Y. Li, Q. Dou, Y.-H. Liu, and P.-A. Heng, "A sim-to-real object recognition and localization framework for industrial robotic bin picking," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3961–3968, 2022.

[20] H. Zhang, J. Peeters, E. Demeester, and K. Kellens, "A cnn-based grasp planning method for random picking of unknown objects with a vacuum gripper," *Journal of Intelligent & Robotic Systems*, vol. 103, pp. 1–19, 2021.

[21] P. Schillinger, M. Gabriel, A. Kuss, H. Ziesche, and N. A. Vien, "Model-free grasping with multi-suction cup grippers for robotic bin picking," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3107–3113, IEEE, 2023.

[22] J. Chen, L. Zhang, Y. Liu, and C. Xu, "Survey on 6d pose estimation of rigid object," in *2020 39th Chinese Control Conference (CCC)*, pp. 7440–7445, IEEE, 2020.

[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021.

[24] F. Liang, B. Wu, X. Dai, K. Li, Y. Zhao, H. Zhang, P. Zhang, P. Vajda, and D. Marculescu, "Open-vocabulary semantic segmentation with mask-adapted clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 2023.

[25] AI-Hub, S. Korea, The Open AI Dataset Project. All data information can be accessed through 'www.aihub.or.kr'.

[26] G. Ilharco, M. Wortsman, R. Wightman, C. Gordon, N. Carlini, R. Taori, A. Dave, V. Shankar, H. Namkoong, J. Miller, H. Hajishirzi, A. Farhadi, and L. Schmidt, "Openclip," July 2021.