Noname manuscript No.
(will be inserted by the editor)

Semantic-Aware Active Perception for Next-Best-View Grasp Planning

Tae Hyeon Kweon · Soo Jeon

Received: date / Accepted: date

Abstract Robotic grasping is a cornerstone of manufacturing automation, and recent advances in deep learning have brought data-driven adaptability to vision-based grasping. However, achieving human-like performance in cluttered environments requires additional capabilities, such as correctly perceiving the object to be retrieved and efficiently planning viewpoints to reconstruct the target object for better grasping under heavy occlusion. To address these challenges, we propose a semantic-aware Next-Best-View (NBV) planning framework that integrates geometric and semantic information gains for targeted exploration. The proposed method maintains a semantic–geometric voxel representation that incrementally accumulates semantic detections across views, guiding viewpoint selection toward regions most likely to reveal graspable target surfaces. We evaluate the framework in simulation and real-world experiments using a Franka Emika Panda arm under heavy occlusion. The proposed approach achieves an 84% success rate in simulation and 10/10 successful grasps in real-world experiments, outperforming baselines in simulation while matching the real-world performance of a geometric NBV method despite requiring no prior knowledge of object locations.

Keywords robotic grasping · next-best-view planning · semantic segmentation · active perception · volumetric mapping

T. H. Kweon · S. Jeon
Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON N2L 3G1, Canada
E-mail: {thkweon, soojeon}@uwaterloo.ca

1 Introduction

Vision-guided robotic grasping in unstructured environments remains a fundamental challenge in robotics, particularly when target objects are partially occluded or have significant hidden surfaces [1,2]. In real-world settings such as manufacturing, logistics, domestic assistance, and agricultural harvesting, robots must frequently manipulate objects in cluttered scenes where target objects are often partially occluded by other items. A typical configuration of vision-guided robotic grasping uses an RGB-D camera mounted near the end-effector (i.e., eye-in-hand camera). With this configuration, single-view grasp planning methods cannot observe hidden surfaces that may contain the most graspable regions, leading to frequent failures when the visible surfaces offer poor grasp affordances [3].

Grasp detection methods predict feasible grasp poses from visual observations. Early analytical methods relied on geometric heuristics [4,5], while modern learning-based approaches leverage deep neural networks trained on large-scale datasets [6,7]. Volumetric methods such as the VGN (Volumetric Grasp Network) [8] and GIGA (Grasp detection via Implicit Geometry & Affordance) [9] employ 3D convolutional or implicit neural networks on Truncated Signed Distance Function (TSDF) representations to predict 6-DoF (Degrees-of-Freedom) grasps from partial observations. Multi-view approaches have been explored to improve grasp detection under occlusion. Morrison et al. [3] propose a sequential observation strategy that aggregates grasp predictions from multiple viewpoints. However, these methods typically rely on a single fixed viewpoint or predefined multi-view trajectories rather than actively optimizing view selection through quantitative measures (e.g., entropy or information gain).

Active perception approaches, particularly Next-Best-View (NBV) planning, provide a principled framework to address partial observability by iteratively selecting informative viewpoints that reveal hidden regions [10]. By moving the camera to observe the scene from different viewpoints, NBV methods improve visibility of occluded regions and obtain more informative observations for grasp planning. Early NBV methods for grasping focused on geometric reconstruction. Breyer et al. [11] propose a closed-loop NBV framework that maximizes volumetric information gain via ray casting, while Schaub et al. [12] extend this with probabilistic TSDF formulations, improving robustness to noise and occlusion. More recent approaches incorporate task-oriented objectives beyond geometry. ACE-NBV [13] proposes an affordance-driven policy based on implicit grasp detection, while Ma et al. [14] propose a neural graspness field that estimates grasp quality over the scene for NBV planning. However, most NBV methods optimize either geometric completeness or grasp-oriented metrics without explicitly modeling semantic information about object identity. In cluttered environments with multiple objects, this can lead to exploration that unnecessarily observes occluders and irrelevant surfaces rather than focusing on the target object.

Recent advances in semantic segmentation and object detection have enabled real-time identification of objects of interest [15, 16]. Several works incorporate semantic information into active perception for manipulation tasks. Dengler et al. [17] maintain semantic volumetric maps with evidential uncertainty estimates, combining uncertainty-aware viewpoint selection and push actions for semantic mapping in cluttered shelf environments. Koc and Sariel [18] enrich volumetric representations with object-aware semantic information to evaluate active and interactive actions for tabletop scene exploration. However, these approaches focus on complete scene understanding rather than guiding exploration toward a specific target object for grasping.

Other works have integrated semantic information into NBV planning frameworks to focus exploration on task-relevant objects. Kay et al. [19] extend volumetric information gain with semantic weighted entropy for aerial reconstruction, multiplying geometric entropy by class-specific utility weights to distinguish target from nuisance objects. Burusa et al. [20] develop a semantics-aware NBV planner that maintains semantic voxel grids with class labels and confidence scores, using entropy-based semantic information gain to prioritize viewpoints that reduce uncertainty about task-relevant plant parts. For manipulation-oriented tasks, Song et al. [21] propose a geometry-based, semantics-aware NBV planner for avocado harvesting that evaluates candidate viewpoints using semantic entropy under task-specific geometric constraints. While these approaches demonstrate the value of semantics for NBV, they typically formulate viewpoint selection using per-voxel semantic uncertainty or entropy-based objectives. In contrast, our method uses semantic detections to progressively localize the target region and guide geometric information gain toward target-focused exploration in cluttered environments.

Fig. 1 The proposed semantic-aware NBV framework actively selects informative viewpoints to refine the geometric and semantic representation of a target object in a cluttered scene prior to grasp execution.

More recently, vision-language models (VLMs) have been applied to semantic-aware viewpoint planning. Wang et al. [22] and Liu et al. [23] leverage VLMs for viewpoint scoring based on semantic relevance. APeG [24] incorporates CLIP [25] text-image similarity for language-guided grasping and selects semantic-aware viewpoints that reveal occluded regions around the target. VISO-Grasp [26] reasons about occlusions by identifying likely occluding objects when the target object is not detected and uses NBV fields to guide sequential grasping. However, object-centric approaches such as VISO-Grasp and APeG handle occlusion locally, without performing global scene exploration or maintaining a persistent semantic representation over unobserved regions.

To address this limitation, we propose a semantic-aware Next-Best-View (NBV) planner for grasping that leverages semantic information to focus sensing on objects of interest (see Fig. 1 for the overall procedure). We maintain a unified volumetric representation that

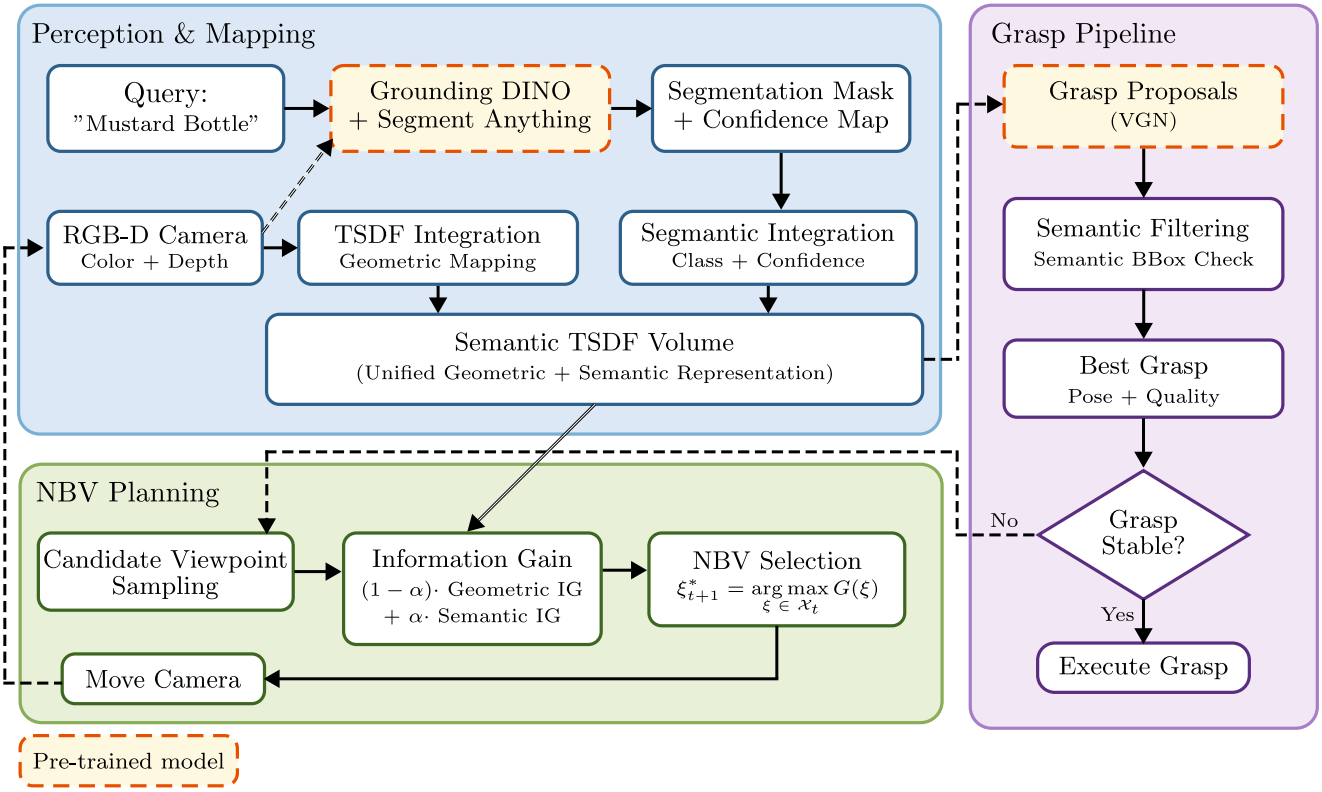


Fig. 2 Overview of the proposed semantic-aware Next-Best-View (NBV) pipeline for active grasping. The system integrates semantic and geometric information for voxel-based mapping, viewpoint sampling, and grasp execution.

fuses geometric TSDF and semantic voxel grids, accumulating semantic labels and confidence scores across multiple viewpoints. Semantic detections are used to construct an adaptive spatial bounding box, within which geometric information gain is evaluated to guide target-focused exploration. Finally, we introduce semantic grasp filtering that leverages the accumulated semantic representation to ensure that grasps target the intended object rather than occluders.

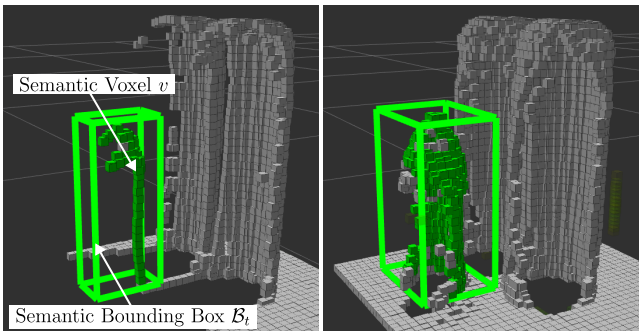
The main contributions of this paper are as follows:

- A semantic-aware NBV framework that combines workspace-level and target-centric geometric information gain through a semantic bounding box that expands as target detections accumulate.
- A unified volumetric representation with aligned geometric and semantic grids that preserves persistent semantic labels and confidence scores across multiple viewpoints.

We validate these contributions in both simulation and real-world cluttered scenarios, demonstrating effective grasp performance in heavily occluded environments. The proposed semantic-aware NBV algorithm is built on top of Volumetric Grasping Network (VGN) [8] and integrated with open-vocabulary semantic segmentation, with full details described in Sect. 2.

2 Methodology

Given a text prompt specifying the target object class, a cluttered workspace, and an eye-in-hand RGB-D camera, the system must plan a sequence of viewpoints and execute a grasp of the target object under heavy occlusion. As shown in Fig. 2, the system operates in a closed-loop exploration-grasp cycle targeting a specified object in a cluttered scene, leveraging VGN [8] for grasp candidate generation and GroundingDINO [27] with the Segment Anything Model (SAM) [16] for open-vocabulary semantic segmentation. At each iteration, the robot: (1) acquires RGB-D observations with object detections and segmentation, and updates the geometric TSDF and semantic voxel representations (Perception & Mapping), (2) evaluates grasp candidates on the target object (Grasp Pipeline), and (3) either executes the highest-quality valid grasp if the grasp condition is met, or selects the next best viewpoint via information gain maximization to continue exploration (NBV Planning). The cycle continues until a successful grasp is executed. VGN [8] predicts 6-DoF grasp candidates from the accumulated TSDF volume and outputs grasp candidates with associated quality scores and 6-DoF poses.



(a) (b)

Fig. 3 Semantic region expansion through multi-view observation. (a) Partial observation. (b) Expanded observation. The semantic bounding box (green) expands as more target voxels are observed from different viewpoints.

2.1 Volumetric Mapping with Semantic Integration

We represent the geometry of the scene using a TSDF volume [28], where each voxel stores the signed distance to the nearest surface. The TSDF is incrementally constructed through volumetric integration of depth observations from multiple viewpoints, providing a continuous volumetric representation of the scene geometry. This representation enables the identification of occluded voxels behind observed surfaces for the computation of information gain and serves as input to the VGN [8] for the prediction of 6-DoF grasp poses.

In this work, we take an approach to maintain a unified voxel-based representation that stores both geometric and semantic information in aligned voxel grids. Let $\mathcal{V}$ denote the set of all voxels within a bounding box for the entire workspace $\mathcal{W}$. At each iteration step t, a set of semantic voxels $\mathcal{S}_t \subset \mathcal{V}$ is identified (see Fig. 3). When the semantic voxel set $\mathcal{S}_t$ is constructed (as explained below in 2.1.1), each voxel element $v \in \mathcal{S}_t$ is associated with a class label $c_t(v) \in \{0, 1, ..., N_c\}$ and detection confidence $p_t(v) \in [0, 1]$, where N_c is the number of identifiable classes and $c_t(v) = 0$ indicates unlabeled voxels.

While the semantic voxel representation can accommodate multiple object classes, in this work semantic integration is restricted to the specified target object class, as the active grasping task requires identifying only the grasp target.

2.1.1 Construction of Semantic Voxels

At each iteration time t, the camera pose $\xi_t \in SE(3)$ generates the RGB-D observation. Given a text prompt for a single target object class, GroundingDINO [27] produces a 2D detection box, a class label for the target c_t, and a detection confidence score p_t. The detected regions are then segmented using a segmentation algorithm (e.g., SAM [16]), producing pixel-wise masks in 2D. These masks are projected onto 3D space through ray-casting from the camera pose $\xi_t \in SE(3)$ taking into account the camera intrinsics. This process generates the semantic voxel set $\mathcal{S}_t$, where each voxel $v \in \mathcal{S}_t$ is assigned a class label $c_t(v)$ and a confidence score $p_t(v)$.

Semantic labels are integrated only when the detector confidence exceeds a threshold τ_p. By discarding low-confidence detections, this reduces incorrect semantic labels from false-positive target detections (e.g., occluders misidentified as the target object) and ensures that semantic information gain is driven primarily by high-confidence target observations.

Unlike approaches that tightly couple geometric and semantic reconstruction such as SemanticFusion [29] and Voxblox++ [30], we maintain separate but aligned representations. This modular design preserves standard TSDF reconstruction while restricting semantic integration on detected objects, rather than performing dense semantic labeling of all surfaces.

2.1.2 Update of Semantic Voxels

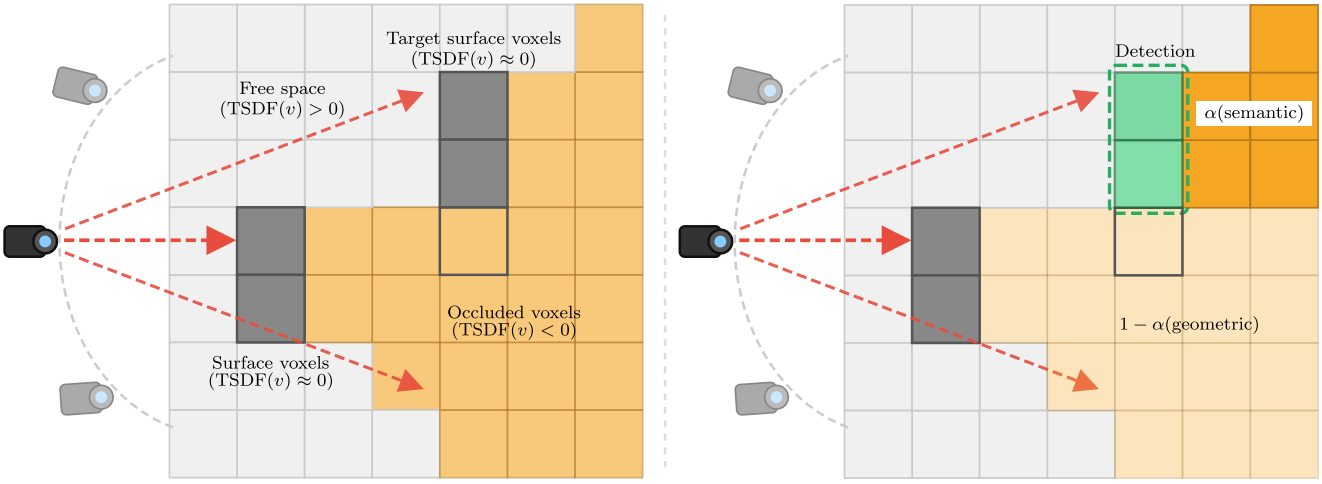
As the robot observes the scene from multiple viewpoints, the same voxel may receive different semantic labels due to occlusion, viewpoint variation, or detector noise. To maintain consistent semantic assignments, the semantic voxel set $\mathcal{S}_t$ is only extended at each iteration step, i.e., $\mathcal{S}_{t-1} \subseteq \mathcal{S}_t$, retaining all previously labeled voxels. When a voxel is re-observed, we update both its class label and confidence only if the new detection has higher confidence than the currently stored value. Specifically, for voxel $v \in \mathcal{S}_{t-1}$ with a new detection $c_t^{\text{new}}(v)$ and $p_t^{\text{new}}(v)$, the stored semantic information is updated as:

$$\begin{aligned} c_t(v) &= \begin{cases} c_t^{\text{new}}(v), & \text{if } p_t^{\text{new}}(v) \geq p_{t-1}(v) \\ c_{t-1}(v), & \text{otherwise} \end{cases}, \quad \forall v \in \mathcal{S}_{t-1}. \\ p_t(v) &= \max\{p_t^{\text{new}}(v), p_{t-1}(v)\} \end{aligned} \tag{1}$$

This maximum-confidence strategy incrementally reinforces semantic assignments as more views are integrated, converging toward the most reliable detection.

2.2 Semantic-Aware NBV Formulation

Our key contribution is extending geometric Next-Best-View (NBV) by focusing exploration on semantically detected target regions through an adaptive bounding



(a) Geometric IG (b) Semantic-aware IG

Fig. 4 Geometric and semantic information gain (IG) via ray casting. (a) Geometric IG counts all occluded voxels (TSDF<0) equally across the workspace. (b) Semantic IG counts occluded voxels within the detected target region, weighted by α, while voxels outside the target region are weighted by $(1-\alpha)$. In 3D, all rear-side voxels (TSDF<0) within the semantic bounding box $\mathcal{B}_t$ contribute to the semantic information gain G_s.

box that guides geometric information gain computation.

2.2.1 Geometric Information Gain

Following Breyer et al. [11], the geometric information gain (IG) quantifies the number of occluded voxels that would become visible from a candidate viewpoint, computed via rear-side voxel counting through ray casting. In our formulation, this gain is computed over the entire workspace region $\mathcal{W}$ to promote broad scene exploration and initial detection of the target object.

Let $\mathcal{R}(\xi) \subset \mathcal{W}$ denote the set of voxels visible from viewpoint ξ within the workspace $\mathcal{W}$, determined via ray casting through the camera. For each candidate viewpoint, rays are cast from the camera center, traversing voxels until the first observed surface is encountered (TSDF ≈ 0). Voxels behind this surface, where TSDF < 0, are considered occluded (rear-side voxels).

The geometric IG is computed as:

$$G_{\mathrm{g}}(\xi) = \frac{1}{|\mathcal{W}|} \sum_{v \in \mathcal{R}(\xi)} \mathbb{I}[\mathrm{TSDF}(v) < 0], \tag{2}$$

where $\mathbb{I}[\bullet]$ is the indicator function whose value is 1 if the statement $\bullet$ is true and 0 otherwise, and $|\mathcal{W}|$ denotes the total number of voxels in the workspace $\mathcal{W}$. This quantifies the proportion of currently occluded voxels within the workspace that would become observable from viewpoint ξ.

2.2.2 Semantic-aware Geometric Information Gain

To focus semantic exploration on the target object, we define the semantic bounding box $\mathcal{B}_t \supset \mathcal{S}_t$ as the smallest axis-aligned bounding box containing the semantic voxel set $\mathcal{S}_t$ (see the green box in Fig. 3). As shown in Fig. 3, the semantic bounding box $\mathcal{B}_t$ dynamically grows as the robot observes the target from multiple viewpoints.

The semantic-aware IG applies the same geometric rear-side voxel counting, but restricts the computation to voxels within the semantic bounding box (see Fig. 4):

$$G_{\mathrm{s}}(\xi) = \frac{1}{|\mathcal{B}_t|} \sum_{v \in \mathcal{R}(\xi) \cap \mathcal{B}_t} \mathbb{I}[\mathrm{TSDF}(v) < 0], \tag{3}$$

where $|\mathcal{B}_t|$ denotes the total number of voxels in the semantic bounding box $\mathcal{B}_t$. This quantifies the proportion of currently occluded voxels within the semantic region that would become observable from viewpoint ξ.

2.2.3 Combined Information Gain and View Selection

This dual-scale formulation enables workspace-level exploration via G_{g} for initial scene understanding, while G_{s} guides exploration toward the detected target region. The total IG denoted by $G(\xi)$ is then computed as:

$$G(\xi) = (1-\alpha) \cdot G_{\mathrm{g}}(\xi) + \alpha \cdot G_{\mathrm{s}}(\xi) \tag{4}$$

where $\alpha \in [0,1]$ balances workspace-level exploration and semantic refinement toward the target region. A

higher α focuses the robot on reducing occlusions within the target region, while a lower α favors broader workspace exploration. Since both G_{g} and G_{s} are normalized as fractions within their respective regions, they lie on the same $[0, 1]$ scale, ensuring α provides meaningful control over the balance between the two terms. When no semantic detections are available (i.e., $\mathcal{B}_t = \emptyset$), the semantic gain G_s is set to zero. In this case, the combined information gain reduces to the geometric information gain G_g.

Candidate viewpoints are uniformly sampled on a hemisphere of radius R_s centered at the workspace center. The hemisphere radius R_s is chosen to ensure the workspace remains visible while satisfying kinematic constraints.

Denoting the set of candidate viewpoints by $\mathcal{X}$, the next best view is selected as:

$$\xi_{t+1}^{*} = \underset{\xi \in \mathcal{X}}{\arg\max}\ G(\xi) \tag{5}$$

The robot moves to ξ_{t+1}^{*} using a Cartesian velocity controller at a fixed linear velocity, continuously updating the TSDF and semantic voxel representations during motion. This allows new observations to be incorporated incrementally rather than only upon arrival at the target viewpoint.

2.3 Semantic Grasp Filtering

For each camera view ξ_t, the VGN generates a set of grasp candidates $\mathcal{G} = \{g_1, g_2, ..., g_m\}$, where m is the number of grasp candidates and $g_i = (\mathbf{t}_i, \mathbf{R}_i) \in SE(3)$. For each grasp g_i, the VGN also computes a quality score denoted by q_i. Then, among grasp candidates in $\mathcal{G}$ predicted by the VGN, we execute the highest-quality grasp g^* whose position lies within the semantic bounding box $\mathcal{B}_t$ of the target object:

$$g^* = g_k \text{ s.t. } k = \underset{i \in \{1,\dots,m\}}{\arg\max}\ q_i, \forall q_i > \tau_q \text{ and } \forall \mathbf{t}_i \in \mathcal{B}_t \tag{6}$$

where τ_q is the quality threshold and $\mathbf{t}_i \in \mathbb{R}^3$ is the position vector associated with g_i. This spatial constraint ensures grasps are placed on the detected target object rather than on occluders or background surfaces. If no valid grasp satisfying these constraints exists, the system selects the next best view to continue exploration. Similar to Breyer et al. [11], we implement the stability window T to avoid premature grasp execution on noisy or transient detections. Specifically, we require a valid grasp candidate to be detected in at least T views before committing to execution, ensuring the selected grasp is supported by sufficient multi-view evidence.

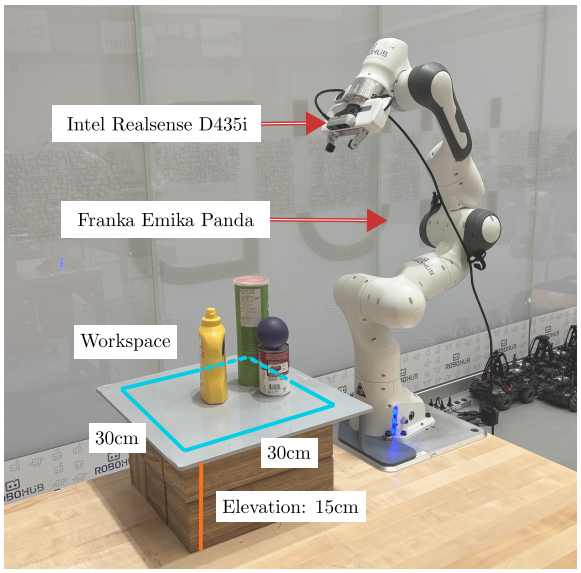


Fig. 5 Experimental setup. The robot observes a 30 cm × 30 cm workspace containing a target object and occluding objects.

3 Validation of Proposed Method

The proposed method has been validated through simulation and experiments. For both cases, vision inference was performed on a remote server equipped with the RTX 4090 GPU from NVIDIA. The simulation was run on a laptop with an Intel i5 CPU, while real-world experiments were conducted on a desktop PC with an Intel i7-8700T.

3.1 Experimental Setup and Performance Metrics

Figure 5 shows our experimental platform, which consists of a 7-DoF Franka Emika Panda robotic arm equipped with a parallel-jaw gripper and an eye-in-hand RGB-D camera (Intel RealSense D435i). We evaluate our semantic active grasping approach in both simulation and real-world settings (Fig. 6), using the PyBullet physics engine [31] for simulation. In simulation, we use ground-truth object instance segmentation masks from the simulator. The semantic segmentation masks are fused with depth images to create a semantic-geometric voxel representation. Test objects are selected from the YCB object set [32]. Objects are placed with random orientations and positions sampled within the workspace, as shown in Fig. 6(b). Table 1 summarizes the implementation parameters used in all experiments.

We evaluate system performance using the following metrics:

(a) **Success Rate (SR)**: The proportion of trials where the target object was successfully grasped.

(a) Real-world setup (b) Simulation environment

Fig. 6 Arrangement of objects for (a) experiment and (b) simulation environments. Top row: scene setup. Bottom row: initial camera view.

Table 1 Implementation Parameters

Parameter	Symbol	Value
TSDF size		$(0.3\,\mathrm{m})^3$
Voxel count per side		40
Hemisphere radius	R_s	0.45 m
Detection confidence threshold	τ_p	0.3
Grasp quality threshold	τ_q	0.7
Semantic IG weight	α	0.5
Stability window	T	5
Max views		15
Linear velocity		5 cm/s

(b) **Abort Rate (AR):** The proportion of trials where no valid grasp was found before the maximum view limit.
(c) **Grasp Failure Rate (GFR):** The proportion of trials where a grasp was attempted but failed during execution.
(d) **Views**: The average number of viewpoints selected per trial.
(e) **Search Time (s):** The average total time spent on viewpoint selection and grasp planning before grasp execution.

The proposed semantic-aware NBV approach has been evaluated against several baseline methods, all of which are reimplemented in our framework with the same perception and grasp evaluation pipeline for fair comparison:

(a) **Initial-view**: Single fixed initial viewpoint without view exploration.
(b) **Top-view**: Single fixed overhead viewpoint without view exploration.
(c) **Random**: Selects viewpoints uniformly at random from the view sphere.

Table 2 Performance of simulation results

Method	SR (%) ↑	AR (%) ↓	GFR (%) ↓	Views ↓	Time (s) ↓
Initial-view	33	63	4	1	1.14 ± 0.54
Top-view	80	8	12	1	9.44 ± 1.61
Random	70	15	15	9.81 ± 4.46	8.47 ± 3.53
Breyer et al. [11]	74	5	21	6.80 ± 3.23	4.77 ± 3.23
Ours	84	6	10	8.85 ± 3.72	8.74 ± 3.89

(d) **Breyer et al.** [11]: Geometric NBV using rear-side voxel counting as IG on a predefined target bounding box. The target bounding box is provided at the start of each trial.

To evaluate robustness to heavy occlusion, the target object is surrounded by occluding objects such that at least 70% of its surface is occluded from the initial viewpoint. We conduct 100 trials per method with varying object poses and clutter configurations.

3.2 Simulation Results

3.2.1 Performance Evaluation

Table 2 presents simulation results over 100 trials per method in heavily occluded scenes. Our method achieves the highest success rate (84%), outperforming all baselines while maintaining a moderate number of viewpoints and execution time. Single-view baselines struggle under heavy occlusion. The Initial-view baseline achieves only 33% success rate with a 63% abort rate, since the target is often not visible from the fixed viewpoint under heavy occlusion. Top-view achieves 80% success, consistent with VGN's preference for top-down grasps [24], but remains limited by incomplete scene reconstruction from a single observation. These results demonstrate the necessity of active viewpoint exploration in cluttered environments.

Random exploration improves over Initial-view baseline but requires more viewpoints on average (9.81), as it lacks target awareness when selecting views. Breyer et al. achieves 74% success rate with fewer viewpoints. However, their method requires a predefined target bounding box provided at the start of each trial, assuming prior knowledge of object location. In contrast, our method uses semantic segmentation to identify and localize the target during execution, without any prior location information. The semantic TSDF incrementally accumulates detections across views, dynamically constructing a target bounding box that guides exploration toward uncovered regions of the target object. The higher average number of views (8.85 vs. 6.80) likely reflects the additional exploration needed to first detect and localize the target before views can be directed toward it. Furthermore, semantic grasp filtering ensures can-

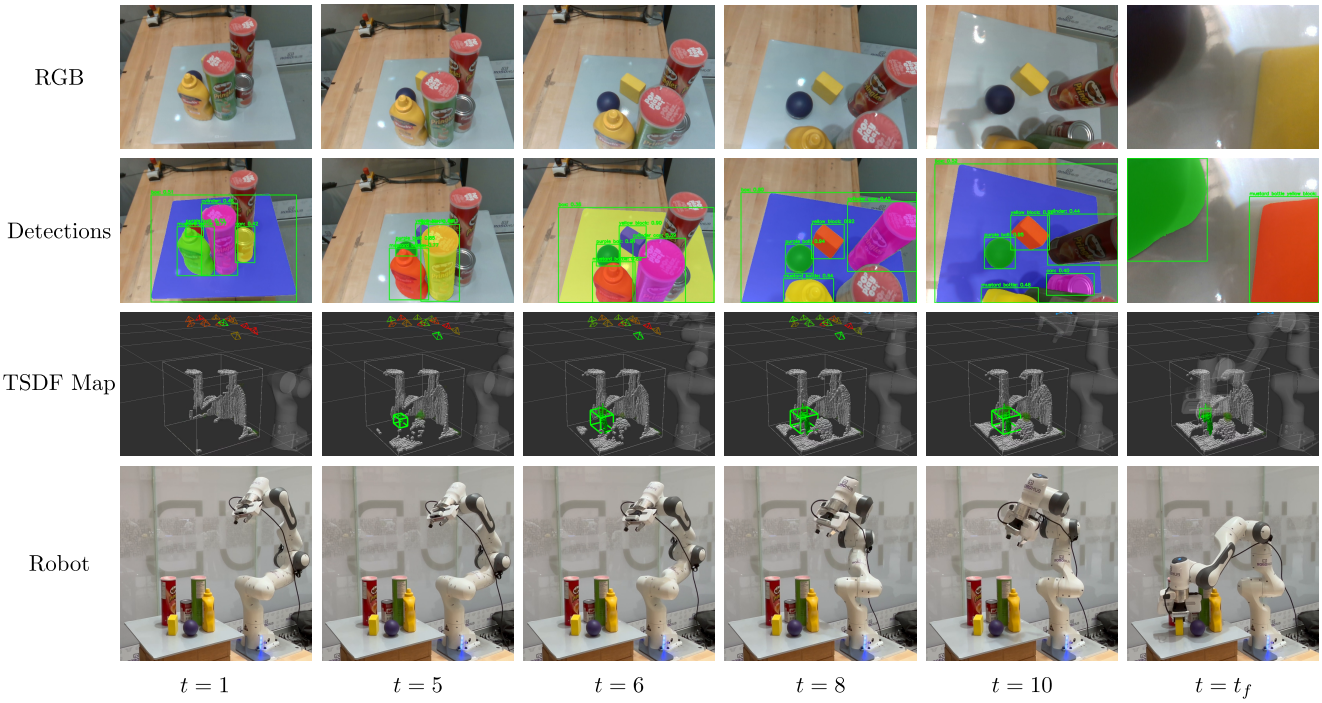


Fig. 7 Sequence of RGB observations, semantic detections, TSDF map updates, and robot motion across multiple viewpoints during semantic-aware NBV planning. The last column ($t = t_f$) shows the final grasp execution.

Table 3 Ablation study

Method	SR ↑	AR ↓	GFR ↓	Views ↓	Time (s) ↓
Ours ($\alpha = 0.0$)	79	8	13	9.00 ± 3.71	8.66 ± 3.14
Ours ($\alpha = 0.5$)	**84**	**6**	**10**	8.85 ± 3.72	8.74 ± 3.89
Ours ($\alpha = 0.9$)	82	**6**	12	9.12 ± 3.77	8.14 ± 3.77
Ours ($\alpha = 1.0$)	78	7	15	**8.62 ± 3.56**	**8.02 ± 3.01**
Ours ($T = 1$)	76	7	17	**5.14 ± 3.91**	**4.96 ± 3.16**
Ours ($T = 5$)	**84**	**6**	**10**	8.85 ± 3.72	8.74 ± 3.89

Table 4 Performance of experimental results

Method	SR ↑	AR ↓	GFR ↓	Views ↓	Time (s) ↓
Initial-view	0/10	10/10	0/10	1.00	—
Top-view	8/10	0/10	2/10	1.00	—
Random	7/10	0/10	3/10	12.57 ± 1.51	10.66 ± 2.35
Breyer et al. [11]	10/10	0/10	0/10	10.90 ± 1.79	10.43 ± 2.19
Ours	10/10	0/10	0/10	8.00 ± 1.15	10.38 ± 2.00

didates are restricted to the target object, improving grasp precision in cluttered scenes.

3.2.2 Ablation Study

Table 3 presents the ablation study to investigate the effects of the semantic IG weight α and the stability window T. The stability window T has a clear impact on performance. A smaller window ($T = 1$) allows immediate execution after a single detection. This reduces search time and requires fewer views but increases grasp failure rate. Increasing to $T = 5$ improves success rate and reduces grasp failure by ensuring grasp candidates are confirmed across multiple views before execution.

The semantic IG weight α controls the balance between geometric and semantic information gains. Increasing α from 0 to 0.5 improves success rate from 79% to 84%, suggesting that biasing viewpoint selection toward target object regions helps reveal occluded surfaces once the object is detected. However, setting $\alpha = 1.0$ (pure semantic IG) degrades performance, as the geometric term provides a necessary exploration signal when semantic detections are sparse in early views. Based on this ablation, we use $\alpha = 0.5$ and $T = 5$ for all experiments in Tables 2 and 4.

3.3 Experimental Results

Figure 7 shows snapshots of a representative robot arm trajectory from the experimental trial. As shown in the sequence of TSDF maps, the target object was not visible from the initial view when $t = 1$ but progressively detected as the semantic-aware NBV algorithm guides the end-effector for better views. Table 4 presents results from 10 trials with varying occlusion and clutter configurations. The initial-view baseline fails entirely (0/10 SR) due to heavy occlusion, while top-view achieves partial success by capturing a broader scene perspective. Random NBV reaches a moderate success rate, but with higher grasp failure and more views required. Our method matches Breyer et al. [11] (10/10)

in search time (10.38 s vs. 10.43 s), while requiring fewer views (8.00 vs. 10.90).

Although our method requires fewer viewpoints, semantic integration and voxel map updates introduce additional computation per iteration, resulting in similar overall search time. Unlike Breyer et al. [11], which require predefined target bounding boxes, our approach uses semantic segmentation to identify and localize target objects, making it applicable to cluttered scenarios where object locations are unknown. These results validate that our approach generalizes to real-world settings while relaxing the assumption of known object location.

4 Conclusion

We presented a semantic active grasping approach that integrates semantic segmentation with next-best-view (NBV) planning to grasp target objects in heavily occluded environments without prior knowledge of object location. By maintaining a semantic TSDF that accumulates semantic detections across views, our method guides viewpoint selection toward regions most likely to reveal graspable surfaces of the target object. Simulation and experiments demonstrate that our approach outperforms single-view, random, and geometric NBV baselines in simulation, and matches geometric NBV performance in real-world trials, without requiring predefined object location information.

A current limitation is the dependence on accurate semantic segmentation, as failures in the visual grounding pipeline directly impact grasp success. Future work will investigate the approach in dynamic environments with moving obstacles, and extend to mobile manipulator platforms where target objects may be distributed across larger, unstructured spaces.

Acknowledgement

The work in this paper has been supported in part by Natural Sciences and Engineering Research Council (NSERC) of Canada under the grant RGPIN-2020-04746.

References

1. Xia, Y., Ding, R., Qin, Z., Zhan, G., Zhou, K., Yang, L., Dong, H., & Cremers, D. (2024). Targo: Benchmarking target-driven object grasping under occlusions. arXiv preprint arXiv:2407.06168
2. Xiao, Y., Katt, S., Pas, A.t., Chen, S., & Amato, C. (2019). Online planning for target object search in clutter under partial observability. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 8241–8247). https://doi.org/10.1109/ICRA.2019.8793494
3. Morrison, D., Corke, P., & Leitner, J. (2019). Multi-view picking: Next-best-view reaching for improved grasping in clutter. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 8762–8768). https://doi.org/10.1109/ICRA.2019.8793805
4. Bohg, J., Morales, A., Asfour, T., & Kragic, D. (2014). Data-driven grasp synthesis—a survey. IEEE Transactions on Robotics *30*(2), 289–309. https://doi.org/10.1109/TRO.2013.2289018
5. Miller, A., Knoop, S., Christensen, H., & Allen, P. (2003). Automatic grasp planning using shape primitives. In *2003 IEEE International Conference on Robotics and Automation (ICRA)*, vol. 2 (pp. 1824–1829). https://doi.org/10.1109/ROBOT.2003.1241860
6. Fang, H.S., Wang, C., Gou, M., & Lu, C. (2020). Graspnet-1billion: A large-scale benchmark for general object grasping. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 11,441–11,450). https://doi.org/10.1109/CVPR42600.2020.01146
7. Mahler, J., Liang, J., Niyaz, S., Laskey, M., Doan, R., Liu, X., Ojea, J.A., & Goldberg, K. (2017). Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Proceedings of Robotics: Science and Systems (RSS)*. https://doi.org/10.15607/RSS.2017.XIII.058
8. Breyer, M., Chung, J.J., Ott, L., Siegwart, R., & Nieto, J. (2021). Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Proceedings of the 2020 Conference on Robot Learning (CoRL)* (pp. 1602–1611). https://proceedings.mlr.press/v155/breyer21a.html
9. Jiang, Z., Zhu, Y., Svetlik, M., Fang, K., & Zhu, Y. (2021). Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. In *Proceedings of Robotics: Science and Systems (RSS)*. https://doi.org/10.15607/RSS.2021.XVII.024
10. Isler, S., Sabzevari, R., Delmerico, J., & Scaramuzza, D. (2016). An information gain formulation for active volumetric 3d reconstruction. In *2016 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 3477–3484). https://doi.org/10.1109/ICRA.2016.7487527
11. Breyer, M., Ott, L., Siegwart, R., & Chung, J.J. (2022). Closed-loop next-best-view planning for target-driven grasping. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 1411–1416). https://doi.org/10.1109/IROS47612.2022.9981472
12. Schaub, H., Wolff, C., Hoh, M., & Schöttl, A. (2024). Probabilistic closed-loop active grasping. IEEE Robotics and Automation Letters *9*(4), 3964–3971. https://doi.org/10.1109/LRA.2024.3371328
13. Zhang, X., Wang, D., Han, S., Li, W., Zhao, B., Wang, Z., Duan, X., Fang, C., Li, X., & He, J. (2023). Affordance-driven next-best-view planning for robotic grasping. In *Proceedings of The 7th Conference on Robot Learning* (pp. 2849–2862). https://proceedings.mlr.press/v229/zhang23i.html
14. Ma, H., Shi, M., Gao, B., & Huang, D. (2024). Active perception for grasp detection via neural graspness field. Advances in Neural Information Processing Systems *37*, 38,122–38,141. https://doi.org/10.52202/079017-1205
15. He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. In *2017 IEEE International Conference on*

Computer Vision (ICCV) (pp. 2980–2988). https://doi.org/10.1109/ICCV.2017.322
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., & Girshick, R. (2023). Segment anything. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (pp. 3992–4003). https://doi.org/10.1109/ICCV51070.2023.00371
17. Dengler, N., Mücke, J., Menon, R., & Bennewitz, M. (2025). Efficient manipulation-enhanced semantic mapping with uncertainty-informed action selection. In *IEEE-RAS International Conference on Humanoid Robots (Humanoids)* (pp. 768–775). https://doi.org/10.1109/Humanoids65713.2025.11203035
18. Koc, C., & Sariel, S. (2024). Object-aware interactive perception for tabletop scene exploration. Robotics and Autonomous Systems *175*, 104,674. https://doi.org/10.1016/j.robot.2024.104674
19. Kay, S.A., Julier, S., & Pawar, V.M. (2021). Semantically informed next best view planning for autonomous aerial 3d reconstruction. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 3125–3130). https://doi.org/10.1109/IROS51168.2021.9636352
20. Burusa, A.K., Scholten, J., Wang, X., Rapado-Rincón, D., van Henten, E.J., & Kootstra, G. (2024). Semantics-aware next-best-view planning for efficient search and detection of task-relevant plant parts. Biosystems Engineering *248*, 1–14. https://doi.org/10.1016/j.biosystemseng.2024.09.018
21. Song, X., & Karydis, K. (2025). Gs-nbv: a geometry-based, semantics-aware viewpoint planning algorithm for avocado harvesting under occlusions. In *2025 IEEE 21st International Conference on Automation Science and Engineering (CASE)* (pp. 690–695). https://doi.org/10.1109/CASE58245.2025.11164113
22. Wang, S., Dai, M., Su, J., Liu, L., Chen, C., Wu, X., & Lin, L. (2025). Graspview: Active perception scoring and best-view optimization for robotic grasping in cluttered environments. arXiv preprint arXiv:2511.04199. https://doi.org/10.48550/arXiv.2511.04199
23. Liu, S., Li, Z., Wang, W., Sun, H., Zhang, H., Chen, H., Qin, Y., Ajoudani, A., & Wang, Y. (2025). Activepose: Active 6d object pose estimation and tracking for robotic manipulation. arXiv preprint arXiv:2509.11364. https://doi.org/10.48550/arXiv.2509.11364
24. Dai, Y., Chen, S., Yang, K., Hu, D., Xie, P., Li, G., Shen, Y., & Wang, G. (2025). Active-perceptive language-oriented grasp policy for heavily cluttered scenes. IEEE Robotics and Automation Letters *10*(11), 11,094–11,101. https://doi.org/10.1109/LRA.2025.3604750
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)* (pp. 8748–8763). https://proceedings.mlr.press/v139/radford21a.html
26. Shi, Y., Wen, D., Chen, G., Welte, E., Liu, S., Peng, K., Stiefelhagen, R., & Rayyes, R. (2025). Viso-grasp: Vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (pp. 14,931–14,938). https://doi.org/10.1109/IROS60139.2025.11246329
27. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al. (2024). Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision* (pp. 38–55). https://doi.org/10.1007/978-3-031-72970-6_3
28. Curless, B., & Levoy, M. (1996). A volumetric method for building complex models from range images. In *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques* (pp. 303–312). https://doi.org/10.1145/237170.237269
29. McCormac, J., Handa, A., Davison, A., & Leutenegger, S. (2017). Semanticfusion: Dense 3d semantic mapping with convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 4628–4635). https://doi.org/10.1109/ICRA.2017.7989538
30. Grinvald, M., Furrer, F., Novkovic, T., Chung, J.J., Cadena, C., Siegwart, R., & Nieto, J. (2019). Volumetric instance-aware semantic mapping and 3d object discovery. IEEE Robotics and Automation Letters *4*(3), 3037–3044. https://doi.org/10.1109/LRA.2019.2923960
31. Coumans, E., & Bai, Y. (2016). Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org
32. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., & Dollar, A.M. (2015). The ycb object and model set: Towards common benchmarks for manipulation research. In *IEEE International Conference on Advanced Robotics (ICAR)* (pp. 510–517). https://doi.org/10.1109/ICAR.2015.7251504