Multimodal Grasp Planner for Hybrid Grippers in Cluttered Scenes

Salvatore D'Avella, *Student Member, IEEE*, Ashok M. Sundaram, Werner Friedl, Paolo Tripicchio, and Máximo A. Roa, *Senior Member, IEEE*

***Abstract*—Grasping a variety of objects is still an open problem in robotics, especially for cluttered scenarios. Multimodal grasping has been recognized as a promising strategy to improve the manipulation capabilities of a robotic system. This work presents a novel grasp planning algorithm for hybrid grippers that allows for multiple grasping modalities. In particular, the planner manages two-finger grasps, single or double suction grasps, and magnetic grasps. Grasps for different modalities are geometrically computed based on the cuboid and the material properties of the objects in the clutter. The presented framework is modular and can leverage any 6D pose estimation or material segmentation network as far as they satisfy the required interface. Furthermore, the planner can be applied to any (hybrid) gripper, provided the gripper clearance, finger width, and suction diameter. The approach is fast and has a low computational burden, as it uses geometric computations for grasp synthesis and selection. The performance of the system has been assessed with an experimental campaign in three manipulation scenarios of increasing difficulty using the objects of the YCB dataset and the DLR hybrid-compliant gripper.**

***Index Terms*—Grasping, grippers and other end-effectors, perception for grasping and manipulation.**

I. Introduction

AUTONOMOUS and reliable robotic grasping is a desirable functionality in robotic manipulation, as it is required for a broad range of applications including pick and place in service robots [1], manufacturing [2], and logistics [3]. Autonomous grasping of a variety of objects is a complex problem due to its multidisciplinary nature, involving end-effector design, perception, planning, and control to guarantee a reliable and robust grasp of the target object. Manipulating objects with different shapes, sizes, surface properties, and weights using a single end-effector is not a trivial problem. The dominant end-effectors used in industry are suction and parallel-jaw grippers, due to their simplicity, precision and low control complexity [4]; however, each modality has its own drawbacks. A promising venue is the design of an end effector combining multiple grasping modalities, which would allow handling different kinds of objects.

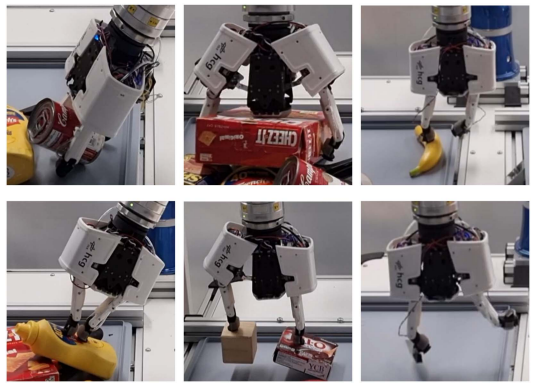

Fig. 1. Examples of different grasping modalities offered by DLR HCG. First row: two-finger grasp, two-finger grasp on a wide object, single suction grasp. Second row: double suction grasp on a single heavy object, double suction grasp on two different objects, magnetic grasp.

After the success obtained during the Amazon Picking Challenges (APC), where the top placing teams used a combination of suction and parallel jaw approaches [5], the idea of using multiple grasping modalities gained popularity. However, even if multimodal grasping has been recognized as an interesting strategy to improve manipulation capabilities, robust integration of such modalities with multimodal grasp planning is scarce. Moreover, most such grasp pipelines exploit a single-stage network that embeds some parameters of the gripper, thus making it difficult to apply them to a different gripper. Other useful capabilities of hybrid grippers, for instance, grasping two objects simultaneously with the use of two suction cups, are also not addressed within planning frameworks so far.

This work presents a novel multimodal grasp planning framework for hybrid grippers that exploits only geometrical (3D bounding box) and material information of the objects. The grasping pipeline takes as input the RGB and depth images, and outputs the best grasp among the possible grasping modalities, depending on the arrangement of the objects in the scene. The main novelties and properties of the proposed approach compared to other multimodal grasp planners are:

- modularity: the planner performs the geometrical computation starting from the 3D bounding box and the material information of the objects. Thus, the 6D pose estimation and the material segmentation networks can be changed

Manuscript received 28 October 2022; accepted 4 February 2023. Date of publication 22 February 2023; date of current version 28 February 2023. This letter was recommended for publication by Associate Editor H. Dong and Editor H. Liu upon evaluation of the reviewers' comments. This work was supported by Project INTELLIMAN through European Union's Horizon Europe research and innovation program under Grant 101070136. *(Corresponding author: Salvatore D'Avella.)*

Salvatore D'Avella and Paolo Tripicchio are with the Department of Excellence in Robotics & AI, Mechanical Intelligence Institute, Scuola Superiore Sant'Anna, 56124 Pisa, Italy (e-mail: s.davella@santannapisa.it; p.tripicchio@santannapisa.it).

Ashok M. Sundaram, Werner Friedl, and Máximo A. Roa are with the German Aerospace Center (DLR), Institute of Robotics and Mechatronics, 82234 Wessling, Germany (e-mail: ashok.meenakshisundaram@dlr.de; werner.friedl@dlr.de; maximo.roa@dlr.de).

This letter has supplementary downloadable material available at https://doi.org/10.1109/LRA.2023.3247221, provided by the authors.

Digital Object Identifier 10.1109/LRA.2023.3247221

to use the latest advances in computer vision, adapted to satisfy the required module interface, i.e., the cuboid (8 points) for the detection module and a mask for the material segmentation one that should be able to identify metallic objects;
- generality: it is not restricted to a specific (hybrid) gripper but can be easily adapted by providing the required gripper information such as gripper clearance, finger width, suction cup diameter, or the Unified Robot Description Format (URDF) for multi-DoF (Degrees of Freedom) grippers, in order to solve the IK for the joint positioning;
- new grasping modalities: besides the two-finger and single suction grasps present in existing literature, double suction modality for big and/or heavy objects and magnetic grasps for metallic objects (Fig. 1) are considered;
- effectiveness: as it is based on geometrical computation, it does not require huge training datasets that are critical for learning-based approaches.

An experimental campaign has been conducted to assess the performance of the proposed planning framework on the objects of the YCB dataset [6]. For the experiments, the DLR Hybrid Compliant Gripper (HCG) has been used as a gripper, the Deep Objects Pose (DOPE) network as a pose estimation network, and GoogLeNet opportunely modified as a material segmentation network that can recognize metallic objects. In particular, the grasp pipeline has been tested in several scenarios: structured tabletop with sparse clutter, dense clutter, and bin-cluttered scenes. Each of the three scenarios introduces a different level of difficulty to test the ability of all components in the grasp pipeline. Other scenarios for the simultaneous suction grasp of two objects were considered to show additional capabilities of the grasping pipeline.

The remainder of the letter is organized as follows: Section II introduces the related work in the field of multimodal grasp planning and hybrid grippers; Section III describes in detail the proposed approach, explaining the building blocks of the grasp planner; Section IV presents the experimental evaluation and results; and Section V concludes the letter.

II. Related Work

Most current manipulation systems, especially in the industrial sector, use a manipulator equipped with a single gripper capable of providing a unique grasping modality. Such systems have difficulties for grasping a wide range of objects with different shapes, sizes, surface properties, weights, and poses. In recent years, the use of multiple grasping modalities has been proposed to improve the manipulation capabilities of robotic systems. Dex-Net [7] is a popular solution that employes a dual-arm manipulator equipped with a parallel-jaw gripper and a suction gripper, thus achieving a system capable of multimodal grasping. For each gripper, a GQ-CNN is trained on synthetic images generated from 3D object models, and uses the expected wrench resistance as a quality metric to select the best grasp from a depth map of the current scene. Such an approach can be expensive, as it requires using two manipulators, and does not exploit the synergy between the two types of grasping modalities. Our proposed approach, instead, can exploit the capabilities of hybrid grippers that allow multiple grasping modalities within the same end-effector.

During the APC, several teams successfully exploited two-finger parallel grasp and suction grasp modalities simultaneously, using hybrid grippers. The work proposed in [8] uses a hybrid gripper with two fingers, each with a scooping spatula as a distal segment and a passively rotating suction cup on the side. The gripper allows for parallel grasp, suction grasp, and scooping using the spatula. The vision system uses an in-hand camera to estimate the 6D pose of the objects by collecting multiple RGB-D views of the scene, passed through a Fully Convolutional Network that returns a probability distribution of the object classes for each pixel in the image. A high-level motion planner chooses the best grasp based on human-trained heuristics and analytical quality metrics. This design evolved into a gripper with a single retractable suction cup on the side of the gripper and replaced the in-hand camera with multiple separated cameras [9]. The grasping pipeline uses a learned affordance score to find the best contact points for each grasping primitive (suction down, suction side, grasp down, flush grasp) and chooses the best grasp among all options. However, both approaches are not generic, as they were designed taking into account the particular gripper structure.

Other hybrid gripper designs are proposed as proof of concepts and do not include any sensors or have no advanced planning capability for real-world grasping [10], [11].

Recently, soft hybrid grippers have been investigated, since passive compliance allows an intuitive adaptation to the shape of the target object. For example, the gripper in [12] has four flexible fingers actuated through tendons, each with a suction cup on the fingertip, and exploits a deep reinforcement learning algorithm to select between three grasping modes (enveloping, sucking, and enveloping and sucking). The policy function is designed to use the enveloping and sucking grasping modality as often as possible to grasp more than one object at the same time by first caging one target with the fingers and then looking for a flat planar surface to apply the suction on the second target. The inherent bias towards one grasp modality (enveloping and sucking grasp modality) might not be suitable for all scenarios. Rather, a grasp modality decision based on the current object scenario (low or high clutter, neighborhood) is much more versatile.

Differently from other approaches that are designed for a specific (hybrid) gripper or that train an end-to-end network combining different stages like object pose estimation and grasp planning into one stage, including gripper-specific parameters, the presented framework is general and modular, having the advantage of allowing replacement of one module without affecting the others, and is applicable to different hybrid grippers. The planner has been developed for dense clutter and bin-picking scenarios and exploits a neural network for obtaining the 6D pose (translation and orientation) of the objects in the scene and another network to infer the surface properties of the objects using an RGB-D image. Any pose estimation networks that give as output a 3D bounding box or material segmentation methods that provide the 2D mask of the metallic objects can be used. The advantage of the proposed solution is that the method is general and can exploit the latest improvements in computer vision by just changing the backbone modules. Furthermore, it does not need specific training to be adapted to other grippers. In addition, as the grasping pipeline uses only geometrical information provided by the pose estimation module, it has a low computational cost, and is fast compared to other grasp pipeline approaches. In contrast to other works, our approach considers finger joint configuration in the planning framework,

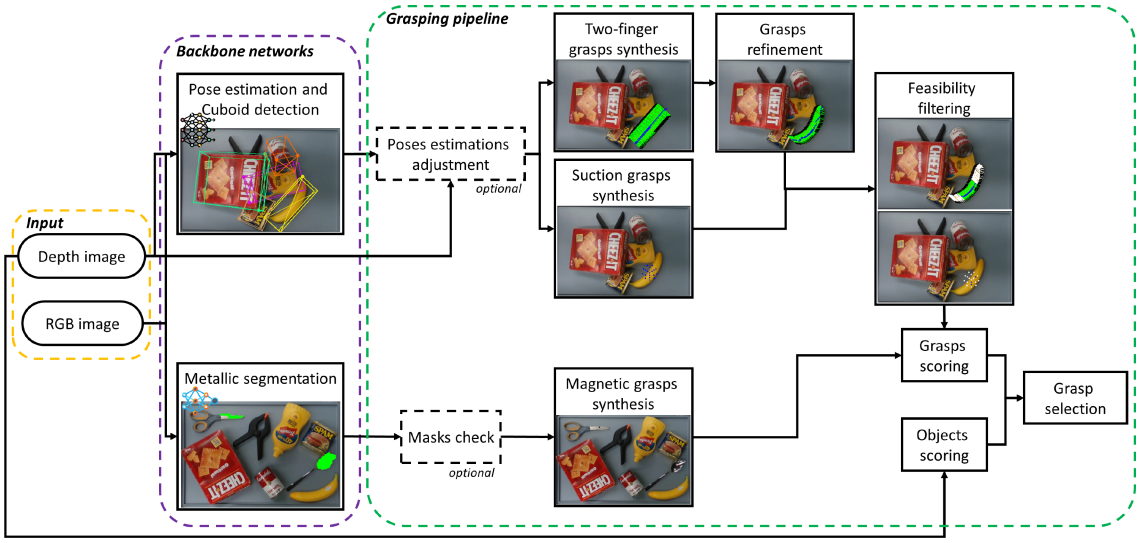

Fig. 2. Block diagram of the grasping pipeline for two-finger, suction, and magnetic grasp modes. The diagram shows the modularity of the design and flow of information. The input images (RGB and depth) are fed to the two backbone networks for pose estimation and metallic segmentation. Optionally, depending on the employed networks, a pose estimation adjustment and the mask check can be applied, respectively. Then, the grasps are computed. The two-finger grasps are refined using the available CAD models of the object or the raw depth information. The two-finger and suction grasps are filtered based on some feasibility criteria. Finally, after a scoring stage, the best grasp candidate is selected. It is worth noticing that for better visualization, two different image scenes are shown for the two-finger and suction grasps and for the magnetic grasp.

in case that the gripper offers multi-DoF fingers, allowing it to adapt for grasping objects of different sizes even in cluttered scenarios.

III. Grasping Pipeline

The proposed grasp pipeline (Fig. 2) takes as input the RGB and depth images, the 3D bounding boxes of the detected objects as given by a pose estimation network, the mask of the metallic objects as given by a material segmentation network, the URDF of the gripper, and optionally, the CAD models of the objects for precise filtering of grasps. Using the CAD models as prior for the detection is useful for complex scenarios such as dense clutter and bin-cluttered scenes, where space for grasping might be very restricted.

For the explanation of the pipeline, we assume that the CAD models are available unless otherwise specified. It is also assumed that the employed pose estimation network directly returns for each detected object a 3D bounding box, or that it is possible to reconstruct for each object a cuboid combining the outputs of the network. Therefore, each detected object is modeled as a set of 9 points: 8 corners and the centroid. Each point is represented as a 3-dimensional vector $\vec{p}^{\,c} = (x, y, z)$ in the camera frame (O_c). Hereafter, if the superscript is not specified, the point is considered to be in the camera frame to reduce notation complexity. Fig. 4 depicts the cuboid, together with the camera and world frames. The extrinsic camera parameter T_o^c (obtained by a calibration procedure) is used to transform a point expressed in the camera frame to the world frame and vice-versa, and the intrinsic camera parameter K is used to project the 3D point in the image frame and vice-versa.

For the cuboid, each face can be defined as a set of 4 points, where the face center for object i is computed as $\vec{c}_k^{\,i} = \frac{1}{4}\sum_{j=1}^{4} \vec{p}_{kj}^{\,i}$, with k representing a face of the cuboid. The faces relevant for grasping are the ones that are visible from the camera's viewpoint. Therefore, for a generic object i, it is possible to define a set of upward faces that satisfy the following condition: $\mathbb{F}_u^i = \{f_k^i : \theta_k^i < \eta\}$ where $k = 1, 2, \ldots, 8$, and θ_k^i is the angle between the vertical vector $-\vec{z}$ and the face normal (Fig. 4). The condition is satisfied if the angle is below a given threshold η (which has been set to 30° for the experiments). The normal $\vec{n}_k^{\,i}$ is computed as $\vec{v}_{k1}^{\,i} \times \vec{v}_{k2}^{\,i}$, where $\vec{v}_{k1}^{\,i} = \vec{p}_{k4}^{\,i} - \vec{p}_{k1}^{\,i}$ and $\vec{v}_{k2}^{\,i} = \vec{p}_{k2}^{\,i} - \vec{p}_{k1}^{\,i}$ (Fig. 4). Thus, after normalizing the normal $\vec{n}_k^{\,i} = \frac{\vec{n}_k^{\,i}}{||\vec{n}_k^{\,i}||}$, θ_k^i is obtained as $\arccos(\vec{n}_k^{\,i} \cdot -\vec{z})$.

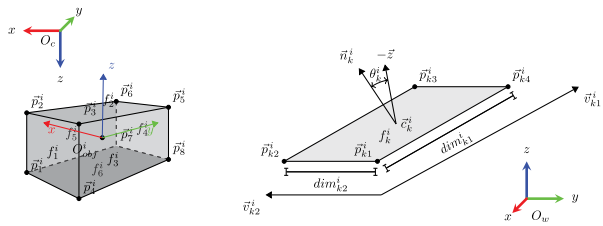

Fig. 3. DLR Hybrid Compliant Gripper, with overlayed representation of the rotation matrix expressed as the three orthonormal vectors: *axis*, *binormal*, *approach*.

A. Grasp Synthesis

Each upward face is then sampled to find potential two-fingers grasps and suction grasps. A two-finger grasp can be defined as a set of a translation vector, rotation matrix, and the required gripper opening needed for grasping the target object. The rotation is represented as a 3x3 matrix where the column vectors are the axis, binormal, and approach orthonormal vectors (see Fig. 3). Therefore, with Fig. 4 as reference, for a two-finger grasp, $\vec{approach}_{ks}^{\,i} = -\vec{n}_k^{\,i}$ and $\vec{axis}_{ks}^{\,i} = \frac{\vec{v}_{ks}^{\,i}}{||\vec{v}_{ks}^{\,i}||}$. The binormal is the cross product between the other two vectors: $\vec{binormal}_{ks}^{\,i} = \vec{axis}_{ks}^{\,i} \times \vec{approach}_{ks}^{\,i}$. The subscript $s = 1, 2$ represents the two

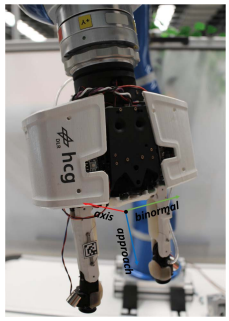

Fig. 4. Left: 6D bounding box representation of an object i. The points $\vec{p}^{\,i}_j$, $j = 1, \ldots, 8$ are the vertexes expressed in the camera frame O_c and f^i_k, $k = 1, \ldots, 6$ are the faces that constitute the cuboid i, composed of 4 points. The two dimensions of the face are dim^i_{k1} and dim^i_{k2}. The vectors $\vec{v}^{\,i}_{k1}$ and $\vec{v}^{\,i}_{k2}$ are used to compute the vector normal to the face $\vec{n}^{\,i}_k$.

sides of the face, since the two-finger grasp can potentially occur along both axis directions. Thus, both sides are sampled starting from the center of the face along the axis direction with a given stride $d = g\lambda$, where $g = \{0, 1, -1, 2, -2, \ldots, \frac{dim^i_k}{2}, -\frac{dim^i_k}{2} : dim^i_k < \xi\}$. λ has been set to 10 mm for the experiments, while ξ is the maximum gripper clearance. Hence, the translation vector of a possible two-finger grasp is derived as $\vec{t}^{\,i}_k(d) = \vec{c}^{\,i}_k + d \cdot \vec{axis}^{\,i}_{ks}$. It is worth noticing that a side of an upward face is sampled only if the dimension dim^i_k of that side is smaller than the gripper opening, and the $opening^i_k$ is set to dim^i_k. Being $R^i_{ks} = \left[\vec{axis}^{\,i\,T}_{ks} \quad \vec{binormal}^{\,i\,T}_{ks} \quad \vec{approach}^{\,i\,T}_{k}\right]$, the set of all the possible two-fingers grasp is defined as $\mathbb{F}^i_a = \{\vec{t}^{\,i}_k(d), R^i_{ks}, opening^i_k\}$.

A suction grasp can be defined with a translation vector and an approach direction. The set of possible suction grasps is computed by sampling the upward face in a circular manner starting from the center ($\vec{c}^{\,i}_k$). The translation vector of a possible suction grasp is derived as: $\vec{t}^{\,i}_k(r, \gamma) = \vec{c}^{\,i}_k + r\cos(\gamma)\vec{v}^{\,i}_{k1} + r\sin(\gamma)\vec{v}^{\,i}_{k2}$, with $r = m\delta$ and $\gamma = n\rho(r)$, where $m = 0, 1, 2, \ldots, \frac{min(dim_{k1}, dim_{k2})}{2}$, δ is a fixed quantity set to 10 mm in the experiments, $n = 0, 1, 2, \ldots, 360$, and $\rho(r) = \frac{60}{r}$, defined as function of r to have an homogeneous sampling independent of the distance to the center. Also for suction grasps, the approach direction is opposite to the normal of the face: $\vec{approach}^{\,i}_k = -\vec{n}^{\,i}_k$. The set of all possible suction grasps is formed as: $\mathbb{S}^i_a = \{\vec{t}^{\,i}_{kg}, \vec{approach}^{\,i\,T}_k\}$.

B. Refinement and Filtering

Depending on the arrangement of the objects in the scene, not all the predicted grasps are feasible, and up to now, the grasps are generated considering the dimension of the cuboids and not the actual size of the objects. Therefore, a refinement process to match the actual size of the objects in the grasp positions and a filtering stage to keep only the feasible grasps are needed. By using the depth information and silhouette of the object, more accurate gripper contact points touching the surface of the objects can be estimated, and the z position of the objects provided by the pose estimation network can be adjusted. The feasibility stage removes grasps that belong to parts of the

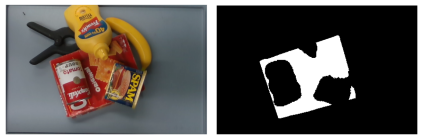

Fig. 5. Result of the visibility computation process. In the left image, the Cheez-it box is occluded by the tomato soup can, the Spam meat, and the Mustard bottle. This occlusion is reflected in the right image, obtained by projecting the CAD model in the depth image for the computed 6D pose. White pixels are considered visible for the next steps in the pipeline.

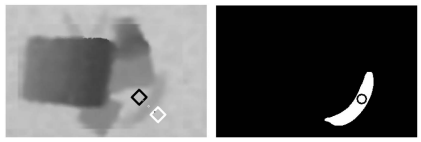

Fig. 6. Grasp feasibility check. Left: example of the process that checks if there is enough space for the two fingertips (represented as two black and white rectangles) to afford in depth. Right: an example of the process that checks if there is enough space for the suction cup (represented as a circle) to completely adhere to the object surface.

object that are invisible to the camera and are occluded by other objects in the scene. In order to compute the visible region of an object, the CAD model is projected in the depth image through a render pass knowing the object's 6D pose and the camera's intrinsic parameters. The pixels of the projected mask that are near the corresponding values of the depth image within a given threshold, which can be tuned depending on the accuracy of the depth sensor, are considered visible. Fig. 5 gives an example result of the procedure.

A two-finger grasp can be executed only if the gripper can move deep enough along the approach direction to make contact on the object's side. Therefore, performing a feasibility check to verify if there is enough space for the fingers without colliding with the surrounding objects is crucial. The depth required for a successful grasp, referred as the affordance depth d_{aff}, is chosen as the minimum between a given value that depends on the fingertip size (3 cm for HCG) and 2/3 of the height of one of the two faces perpendicular to the object's upward face along the approach direction. By considering the new location of the two fingers, after the refinement process, a square rectangle, 5 mm bigger than the dimension of the fingertip of the gripper to account for the noise of the depth sensor and the camera calibration, is projected in the depth image using the intrinsic camera matrix K. If 70% of the pixels within the rectangles have a depth difference with respect to the depth of the face central point larger than d_{aff}, that grasp is considered feasible. Fig. 6 graphically explains the feasibility concept. $\mathbb{F}^i_r$ is the resulting set of available grasps removing the non-feasible grasps discarded by the visibility factor and the grasps for which there is not enough space for the fingers.

For suction grasps, unlike finger grasps, there is no need to refine the gripper opening, and the feasibility check to ensure enough space for the fingers is also not applicable. However, a feasibility check is required to verify that the suction cup can adhere to and seal completely on the object's surface. This is achieved by checking if a circle with a radius of the suction cup adequately fits inside the visible silhouette of the object. The feasibility filtering stage concerning occlusion parts of the object is the same as for two-finger grasps. $\mathbb{S}^i_r$ is the resulting

set of available grasps removing the set of non-feasible grasps discarded by the visibility factor and the set of grasps for which the suction cup does not fit completely on the object surface.

C. Grasp Selection

Each object in the scene can be defined as $\mathbb{O}^i = \{\vec{t}^i, R^i, \mathbb{F}^i_u, \vec{c}^i_k, \mathbb{F}^i_r, \mathbb{S}^i_r, CAD^i\}$. After computing the two-finger grasps and suction grasps, it is possible to assign a score to objects and grasps based on desired heuristics to select the grasp for execution. Each score is a numerical value between 0 and 1. The scoring for objects is based on the depth, the visibility percentage, and the number of collisions with other objects. In particular, all N objects are first sorted by depth $\mathbf{O} = \bigcup_{i=1}^{N} \mathbb{O}^i : depth(\mathbb{O}^i) < depth(\mathbb{O}^{i+1})$, and then the depth score is assigned as follows:

$$\sigma^i_{depth} = 1.0 - \frac{depth(\mathbb{O}^1) - depth(\mathbb{O}^i)}{depth(\mathbb{O}^1) - max(depth)}, \tag{1}$$

i = 1, 2,..., N. By doing so, in cluttered scenes, objects at the top are preferred, as they are less occluded by other objects. The visibility score of an object takes into account how much it is visible in the clutter by considering the percentage of visible pixels,

$$\sigma^i_{visibility} = \frac{pixel(mask(CAD^i(\vec{t}^i, R^i)))}{pixel(projection(CAD^i(\vec{t}^i, R^i)))}. \tag{2}$$

Finally, the collision score considers for each object the number of collisions with other objects, privileging the ones that have fewer collisions:

$$\sigma^i_{collisions} = 1.0 - \frac{collision(\mathbb{O}^i) - min(collisions)}{max(collisions) - min(collisions)}, \tag{3}$$

where $collisions = \{\#c^i\}\ i = 1, \ldots, N$ is the set containing the number of collisions for each object. To quickly compute collisions among objects, the separated axis theorem [13] is used by exploiting the cuboid vertices. All scores are combined, weighting each heuristic differently depending on the importance they should receive, and the weighted average is computed to get the final score:

$$\sigma^i_{total} = \frac{w_1\, \sigma^i_{depth} + w_1\, \sigma^i_{visibility} + w_3\, \sigma^i_{collisions}}{w_1 + w_2 + w_3}. \tag{4}$$

In the experiments, the depth score is weighted more to make the planner give the highest importance to the object on the top. Therefore, the target object for grasping is chosen as the object having the highest total score: $target = \mathbb{O}^* : max(\sigma^i_{total})$.

For assigning a grasping score, the relevant metrics for both two-fingers grasp and suction grasp are the distance to the barycenter and the numerosity. The first is meant to privilege grasps that are near the center of the object, to improve the stability of the grasp: $\sigma^*_{baricenter_k} = 1.0 - ||\vec{c}^*_k - \vec{t}^*(d)||$, for two-finger grasp, and, $\sigma^*_{baricenter_k} = 1.0 - ||\vec{c}^*_k - \vec{t}^*_k(r, \gamma)||$ for suction grasp. The numerosity is included considering that if one grasping mode has very few feasible grasps even with a poor barycenter score, then maybe it is possible that also the remaining ones are not so robust, thus, it is better to prefer the other modality. Such a score is simply obtained as the cardinality of the resulting feasible grasps over the total number of grasps, as follows: $\sigma^*_{numerosity_k} = \frac{\#\mathbb{F}^*_r}{\#\mathbb{F}^*_a}$ and $\sigma^*_{numerosity_k} = \frac{\#\mathbb{S}^*_r}{\#\mathbb{S}^*_a}$. This metric is optional and should be considered carefully because a nonuniform sampling between the two grasping modalities can lead to a bias for the one with the highest cardinality. In addition, for the two-finger grasps, it is also possible to add a preference score for selecting the smaller side to grasp, in case it is not too far from the barycenter. Therefore, the total score for two-finger grasps is

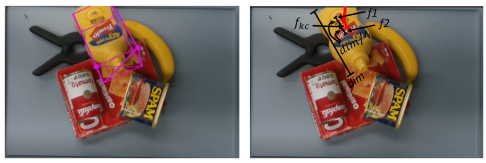

Fig. 7. Fingertip positioning with the double suction single object modality. After the cuboid detection (left figure), being dim the dimension of the principal axis of the upward face, the fingertips are placed equidistant from the center by a factor of $\frac{dim}{4}$ to guarantee a stable grasp aligned with the principal axis of the face.

$$\sigma^*_{total_k} = \frac{w_4\, \sigma^*_{baricenter_k} + w_5\, \sigma^*_{numerosity_k}}{w_4 + w_5}, \tag{5}$$

and for suction grasps is

$$\sigma^*_{total_k} = \frac{w_6\, \sigma^*_{baricenter_k} + w_7\, \sigma^*_{numerosity_k}}{w_6 + w_7}. \tag{6}$$

It is possible to assign different weights w to the metrics, but for the experiments, they are all considered unitary. The selected grasp is the one that has the highest total score among the two modalities and belongs to the target object (in case there is a specific object to be grasped): $grasp\dagger = g\dagger : g\dagger \in \{\mathbb{F}^*_r \cup \mathbb{S}^*_r\}\ s.t.\ \sigma^*_{total_k} = \max(\sigma^*_{total_k})$. If there are no feasible grasps for the target object, the best grasp for the next highest object is considered.

Besides the normal suction grasp mode, where only one suction cup is used, the planner also considers using both suction cups at the same time for heavy or big objects. In that case, the translation vector of the grasp pose coincides with the center of the upward face, and the rotation matrix is built similarly to the two-finger grasp mode; in such a way the binormal is parallel to the major axis of the face, the approach is opposite to the face normal, and the axis is derived as the cross product between the other two. Then, the suction cups are placed on the surface with an approach perpendicular to the object, having a distance of $dim/4$ from the center on each side, as shown in Fig. 7. If the two suction cups lay on a visible portion of the object, such a grasping mode is preferred over the single suction mode. For grippers with multi-DoF fingers, such as the DLR HCG, an inverse kinematic (IK) is required for positioning the fingertips at the desired pose or, for the two-fingers grasp, for getting the specified opening. Kinematic feasibility is an additional filter used to eliminate non-feasible grasps.

The magnetic mode is meant for thin metallic objects, and is selected when the height and size of the metallic object fulfills a predefined threshold. The metallic mask provided by the material segmentation network is double-checked with the contours of the object retrieved in the depth image to have a better segmentation mask. Indeed, it is not an easy job deriving the material properties of an object from an RGB image, and the accuracy of the existing methods is not so high (around 70%) [14]. The centroid of such a refined mask is re-projected

Fig. 8. Magnetic grasp pipeline. The initial RGB image (left) is fed to the segmentation network that provides a mask of the metallic objects, which is double-checked with the depth information to have more accurate masks (center). The centroids of the masks constitute potential magnetic grasps, as depicted in the right image by the points on the scissor and the spoon.

from the 2D image to the 3D camera coordinate system using the intrinsic matrix, and is used as the translation vector for the magnetic fingertip. Fig. 8 depicts graphically the pipeline used for the magnetic grasp modality.

When a suction cup is available on each finger, such as in the DLR HCG, it is also possible to grasp at the same time two objects using the two suction cups even if the two target objects have different heights, compensating for such a difference with the DoF of the fingers. In this case, the translation vector of the grasp pose is the middle point between the two center points of the two upward faces involved in the task. The rotation matrix has the binormal computed as the normalized difference vector between the two center points, the axis as the cross product between the binormal and the approach vector of the suction grasp in the face center point, and the approach the cross product between the other two. The fingertip placement is handled by the IK solver, which also checks if the configuration is feasible or not. A separate scenario is arranged in the experimental campaign to show such a capability.

The proposed grasping pipeline is general and can be adapted to other hybrid grippers changing the URDF, and the associated parameters such as gripper opening, fingers' width, and allowed grasping modalities. Furthermore, other equivalent backbone networks can be employed as long as they implement the required functionalities.

IV. Experimental Validation and Discussion

Four manipulation scenarios of increasing difficulty are created using the YCB objects. The robot should grasp the items from a table and transport them to a storing area. Demonstration videos can be found at *https://github.com/SalvatoreDAvella/Multimodal-grasp-planning.git*.

A. Hardware Description

The proposed multimodal grasp pipeline has been tested using the HCG gripper (Fig. 3) mounted as end-effector of a DLR LWR robot. The HCG has 8 DoFs and each finger is equipped with a suction cup at the fingertip and an electromagnet at the side of the fingertip, providing four grasping modalities. Its finger design is based on the thumb modules of the DLR CLASH hand [15], which has variable passive stiffness. Each finger has three DoF (1-DoF distal interphalangeal and 2-DOF metacarpophalangeal joint), and the finger stiffness can be controlled independently of the position. The fingers are mounted on a base that provides an additional DoF per finger to tilt them away from the palm, enhancing the grasp span up to 260 mm. The maximum object weight for pinch grasp is about 1.5 kg (for friction coefficients above 0.75), and about 500 g for one suction cup. A Realsense D435 RGB-D camera looking down at the objects has been employed for the vision system.

TABLE I
Experimental Results for the Four Different Scenarios

Scenario	success rate	# two-fingers	# suction	# double-suction	# magnetic
Structured clutter	97%	35	16	10	3
Dense clutter	87.5%	15	5	4	0
Bin clutter	63%	9	4	6	-
Double objects	98%	0	0	6	0

B. Backbone Networks for Object Detection and Segmentation

The grasp pipeline leverages Deep Object Pose (DOPE) [16] for extracting the bounding boxes and 6D poses of the objects involved in the scene for computing two-fingers grasps and suction grasps. DOPE is a model-based approach that only uses an RGB image as input. First, it estimates the belief maps of 2D keypoints of all the objects in the image coordinate system, and then the 6D pose of each object instance with a standard perspective-n-point (PnP) algorithm on the peaks extracted from these belief maps. The final step uses the detected projected vertices of the bounding box, the camera intrinsic parameters, and the object dimensions to recover the final translation and rotation of the object with respect to the camera. All detected projected vertices are used as long as at least four vertices of the cuboid are detected. The network can be trained on synthetic and photo-realistic data without the need for handcrafted labels on real data, and provides satisfactory results on real objects thanks to the domain randomization technique.

For magnetic grasps, metallic objects should be detected and segmented. For this purpose, the Material In Context (MINC) dataset is used for training the network, and the approach presented in [14] is implemented. A GoogLeNet [17] is converted into a sliding CNN without average pooling in order to densely classify a grid across the image. The last fully connected layers are modified into convolutional layers to obtain a fully convolutional network for classifying images of any shape. In post-processing, the dense connected Conditional Random Field (CRF) model [18] is used to predict a label at every pixel, thus achieving material segmentation.

Other equivalent backbone networks can be employed as long as they provide the 6D pose or the bounding box (cuboids) for the objects for the pose estimation stage (e.g. [19] or [20], which is even agnostic to the identity of the objects), and the mask that identifies metallic objects, required for the magnetic grasps.

C. Results

Table I summarizes the results obtained during the experiments for the different scenarios, presented below. The table reports the grasping success rate taking into account all the system's components, starting from the grasp pipeline prediction to the execution of the grasping action by the gripper and the movement of the robotic arm, in order to measure the performance of the pick and place robotic system. For the conducted tests, the planner always provided the correct target and grasp according to the selected metrics when the vision module detected an object. The average object pose estimation time required by the selected vision module DOPE is 1.1 s, while the average time required by the grasp planner for returning the grasping pose and modality for the next target is 1.8 s in scenes containing ten objects.

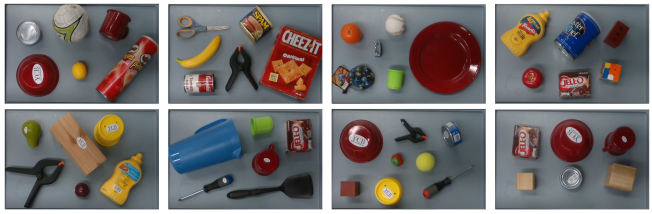

Fig. 9. Some scenes used for the tabletop structured clutter scenarios.

Fig. 10. Some scenes used for the dense clutter scenarios.

Fig. 11. Some bin scenes for the bin picking scenarios.

The first scenario is a tabletop structured clutter scenario, where the objects are disentangled and arranged on the working table to cover at least 70% of the area (Fig. 9). With this scenario, we used a custom vision module that is different from the DOPE network used in the other more complex scenarios. In particular, such a computer vision module considers the following assumptions: a full 6D pose estimation with the detection of the upward faces is not required, as only the top of the object can be included in the grasp synthesis; the CAD model is not necessary since it can be substituted by the raw depth information; no occlusions and collisions have to be handled. Furthermore, the grasping approach can always be top-down, and some rules based on the geometrical properties of the objects are used to drive the grasping selection by considering the characteristics of the gripper. For example, the curvature of the top surface of the object has to be estimated in the depth image to prevent applying the suction grasps on curved surfaces, very big objects should be grasped with the double suction grasp modality, round objects must prefer two-finger grasps, or objects with a flat surface and a low height need to be handled with suction, as the two-finger grasps are more prone to errors in this condition. This shows the modularity of the designed grasp planner in terms of the vision module and is applied to unknown objects.

Ten different scenes with six or seven objects taken randomly from 48 YCB objects are tested. All the objects are selected at least once and then repeated with different poses in the subsequent scenes. From the YCB dataset, only the objects that were too heavy, big, or small and, thus, not feasible for the HCG were discarded. The system achieved a success rate of 97%. The main errors were related to the slippage of thin round objects during the transportation to the placing area. The errors can be explained by the inaccuracies in the final position of the gripper and in the joint control of the HCG fingers, still under development, ending up in a grasping position that is slightly different from the one commanded by the grasping pipeline. Such behavior is critical for thin round objects for the two-fingers grasp mode and narrow objects grasped with the suction modality. For the latter case, a recovery procedure, consisting in trying small positional variations in the area near the one computed by the grasping pipeline, has been implemented to compensate for such placement inaccuracy. It has been shown during the experiments that the recovery strategy helped increase the success rate of task execution. In order to show the generality of the planner for other grippers, the same experiments were conducted using the Franka Emika Panda robot equipped one time with its parallel-jaw gripper achieving a success rate of 47%, and another time with the Schmalz suction gripper achieving a success rate of 36%. The Panda parallel-jaw gripper has a clearance of 8 cm and is not suitable for many of the involved objects. However, the gripper successfully grasped all those objects for which the grasping pipeline found a feasible grasp. The Schmalz suction gripper has a diameter of 3 cm, and it is not suitable for the curved objects involved in the scenes. However, the gripper successfully grasped all those objects for which the grasping pipeline found a feasible grasp. Note that the performance of the single modality grippers is already much lower than the multimodal gripper in the first scenario, which is the simplest one, therefore such a comparison is not repeated for the remaining test scenarios.

The second manipulation scenario presents densely cluttered scenes where the objects can collide, are subject to occlusion, and present challenging poses (Fig. 10). In these cases, all the steps of the grasping pipeline explained in the previous section are required. The complete 6D pose estimation is necessary, and the DOPE network has been trained for seven objects presenting different shapes, sizes, weights, and surface properties. Seven scenes with six objects arranged in different poses have been tested. The proposed system still performs well even for tricky situations where the objects require a particular grasping pose, are on top of each other, or are arranged in such a way that there is limited space to position the fingers. In order to avoid possible robot collisions with the objects during the motion, we move the robot from the starting position to a pre-grasp pose, which is the same as the grasp pose but shifted vertically along the z-axis in the tool center point (TCP) frame so that the affordance for grasping is purely vertical along the approach direction. The system achieved a success rate of 87.5%. The main failures, as in the previous scenario, were related to the slippage of objects during the transportation to the placing area, especially for the suction grasping modality. However, the grasping pipeline provided correct grasping candidates.

The last scenario is a heavily cluttered bin, where the objects are randomly piled on top of each other (Fig 11). In this setup, the objects suffer from heavy occlusions that pose problems to the vision system. In addition, the very small size of the bin can penalize the movement of bulky grippers, and the objects strongly overlap with each other, making two-finger grasps very difficult. Three trials have been executed. The performance of the system decreased with respect to the previous scenarios, and the system could grasp five, eight, and six objects in the three trials. The main problems are related to the 6D pose estimation network that in a heavy clutter scene with the transparent bin does not detect at all some objects or makes the wrong estimation of the orientation, and to the collisions with the bin walls. The environmental constraints of the bin walls are considered in the space filtering check for the two-finger grasps, which leaves few feasible grasps for final selection. The approach is always to try to grasp the highest object, which is less occluded. To increase the estimation accuracy of the network, the depth image

is cropped with a virtual plane located 5 cm above the bottom of the bin, and an RGB image constituted just from the pixels above that plane is fed to the network. When the cuboids of the objects are computed correctly, the grasping pipeline generates a suitable grasp for the complex scenario. We are using a state-of-the-art algorithm in the pipeline for the 6D pose estimation, and thanks to the modularity of the system, we could upgrade the pipeline by changing the backbone network with future solutions without affecting the overall planning approach.

The fourth scenario is meant to show the capability of the gripper in combination with the grasping pipeline to grasp multiple objects at the same time with the suction cups mounted at the fingertips. In particular, several scenes with six objects having even different heights are created. Using the IK solver, it is possible to check the feasibility of the grasping and compensate for the height difference between the two target objects. The bottom right picture of Fig 9 depicts one of such scenarios. With such a grasping modality, it is possible to half the cycle time of the pick and place task.

V. Conclusion

Autonomous grasping of a variety of objects is still an open problem in robotics, especially for cluttered scenarios. After the success obtained during the Amazon Picking Challenges, multimodal grasping has been recognized as an interesting strategy to improve the manipulation capabilities of robotic systems. This work presented a novel multimodal grasp panning algorithm for hybrid grippers that allow multiple grasping modes. The presented framework is general as it can be used for other hybrid grippers, and it is also independent of the networks used for the 6D pose estimation and material segmentation. The performance of the system has been assessed with an experimental campaign in different manipulation scenarios on the objects of the YCB dataset: structured tabletop sparse cluttered scenes, densely cluttered scenes, and heavily cluttered bin scenes. The results showed high success rates for the first two setups and decay of the performance for the last one, which could be improved by using a future 6D pose estimation network that can better handle cluttered scenes. Future work will be focused on using the properties of the objects from the known CAD models, learning such properties directly from the RGB images, and exploiting environmental constraints in the grasp planner.

References

[1] S. H. Kasaei, J. Melsen, F. van Beers, C. Steenkist, and K. Voncina, "The state of lifelong learning in service robots," *J. Intell. Robotic Syst.*, vol. 103, no. 1, pp. 1–31, 2021.

[2] S. D'Avella, C. A. Avizzano, and P. Tripicchio, "Ros-industrial based robotic cell for industry 4.0: Eye-in-hand stereo camera and visual servoing for flexible, fast, and accurate picking and hooking in the production line," *Robot. Comput.- Integr. Manuf.*, vol. 80, 2023, Art. no. 102453.

[3] S. D'Avella, P. Tripicchio, and C. A. Avizzano, "A study on picking objects in cluttered environments: Exploiting depth features for a custom low-cost universal jamming gripper," *Robot. Comput.- Integr. Manuf.*, vol. 63, 2020, Art. no. 101888.

[4] M. Guo et al., "Design of parallel-jaw gripper tip surfaces for robust grasping," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 2831–2838.

[5] N. Correll et al., "Analysis and observations from the first Amazon picking challenge," *IEEE Trans. Automat. Sci. Eng.*, vol. 15, no. 1, pp. 172–188, Jan. 2018.

[6] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The YCB object and model set: Towards common benchmarks for manipulation research," in *Proc. IEEE Int. Conf. Adv. Robot.*, 2015, pp. 510–517.

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

[8] A. Zeng et al., "Multi-view self-supervised deep learning for 6D pose estimation in the Amazon picking challenge," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2017, pp. 1386–1383.

[9] A. Zeng et al., "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2018, pp. 3750–3757.

[10] C. Tawk, A. Gillett, M. in het Panhuis, G. M. Spinks, and G. Alici, "A 3D-printed omni-purpose soft gripper," *IEEE Trans. Robot.*, vol. 35, no. 5, pp. 1268–1275, Oct. 2019.

[11] L. Chin, F. Barscevicius, J. Lipton, and D. Rus, "Multiplexed manipulation: Versatile multimodal grasping via a hybrid soft gripper," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 8949–8955.

[12] F. Liu, F. Sun, B. Fang, X. Li, S. Sun, and H. Liu, "Hybrid robotic grasping with a soft multimodal gripper and a deep multistage learning scheme," *IEEE Trans. Robot.*, 2023.

[13] S. Gottschalk, M. C. Lin, and D. Manocha, "OBBTree: A hierarchical structure for rapid interference detection," in *Proc. 23rd A. Conf. Comput. Graph. interactive Techn.*, 1996, pp. 171–180.

[14] S. Bell, P. Upchurch, N. Snavely, and K. Bala, "Material recognition in the wild with the materials in context database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3479–3487.

[15] W. Friedl, H. Höppner, F. Schmidt, M. A. Roa, and M. Grebenstein, "CLASH: Compliant low cost antagonistic servo hands," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 6469–6476.

[16] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Proc. Conf. Robot Learn.*, 2018, pp. 306–316.

[17] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, 2015, pp. 1–9.

[18] P. Krähenbühl and V. Koltun, "Parameter learning and convergent inference for dense random fields," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 513–521.

[19] 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.

[20] J. Sun et al., "OnePose: One-shot object pose estimation without CAD models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 6825–6834.

Open Access funding provided by 'Scuola Superiore "S.Anna" di Studi Universitari e di Perfezionamento' within the CRUI CARE Agreement.