Anthropomorphic Grasping with Neural Object Shape Completion

Diego Hidalgo-Carvajal*,1,2, Hanzhi Chen*,3, Gemma C. Bettelani[1], Jaesug Jung[1], Melissa Zavaglia[1], Laura Busse[4], Abdeldjallil Naceri[1], Stefan Leutenegger‡,3, Sami Haddadin‡,1,2

***Abstract*— The progressive prevalence of robots in human suited environments has given rise to a myriad of object manipulation techniques, where dexterity plays a paramount role. It is well established that humans exhibit extraordinary dexterity when handling objects. Such dexterity seems to derive from a robust understanding of object properties (such as weight, size and shape), as well as a remarkable capacity to interact with them. Hand postures commonly demonstrate the influence of specific regions on objects that need to be grasped, especially when objects are partially visible. In this work, we leverage human-like object understanding by reconstructing and completing their full geometry from partial observations, and manipulating them using a 7-DoF anthropomorphic robot hand. Our approach has significantly improved the grasping success rates of baselines with only partial reconstruction by nearly 30% and achieved over 150 successful grasps with three different object categories. This demonstrates our approach's consistent ability to predict and execute grasping postures based on the completed object shapes from various directions and positions in real-world scenarios. Our work opens up new possibilities for enhancing robotic applications that require precise grasping and manipulation skills of real-world reconstructed objects.**

I. INTRODUCTION

Achieving human-like dexterous manipulation is a sought-after goal in robotics. Although significant progress has been made to attain this aim, current solutions are limited by both methodologies and hardware constraints. In order to dexterously manipulate an object, two main aspects need to be considered: i) the understanding of the visual object scene, and ii) the grasping strategy. Despite extensive research on these two aspects separately, limited focus has been placed on integrating them into a complete human-like grasping approach.

In humans, visual object recognition is a complex cognitive process that involves the ability to identify and categorize objects based on their visual features despite substantial ambiguity. It is thought that the brain tackles this challenge with a cascade of *feedforward* processing stages that extract and represent increasingly complex information [1], [2]. These computations are supported by *top-down* mechanisms [3], that can adjust the neural representations based on an internal model of the world derived from prior experience. These mechanisms are often thought to underlie the remarkable ability of humans to robustly recognize objects, even when faced with cluttered environments, ambiguous stimuli, or incomplete information due to partial occlusion [3], [4].

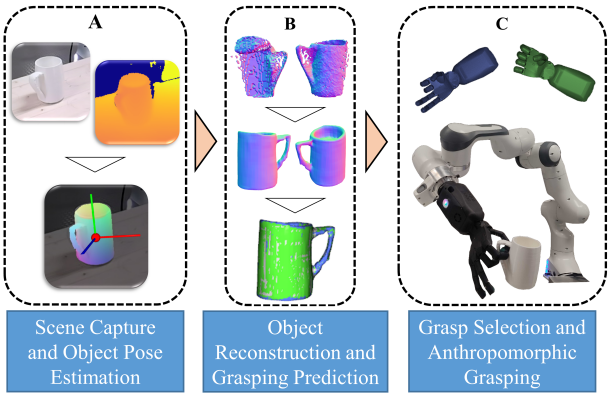

Fig. 1. Overview of our approach. (A) shows sample RGBD images of the captured scene and the object pose estimation, (B) shows the object 3D shape reconstruction, completion, and grasping prediction, respectively. (C) shows the possible grasps for the robot hand and the robotic arm-hand system grasping an object in an anthropomorphic manner.

Contrary to this, the approaches of the classical artificial vision systems are fundamentally different from human mechanisms. They tend to lead to partial or incomplete object models because they only fuse observable information [5]–[7]. Recent methodologies have improved those systems by incorporating more human-like approaches by utilizing the complete geometric and semantic information of objects from partial visual information (e.g. [8]). While such approaches have performed well in a variety of applications, including robotic grasping, it remains unclear how to leverage them into more end-to-end deep learning grasping approaches, as such concepts exhibit the potential to leverage learned geometries for more dexterous tasks at hand, where the deployment of anthropomorphic robot hands is needed.

Robotic grasping and manipulation have been extensively studied. However, the majority of studies have focused on non-anthropomorphic simplified hands, such as parallel grippers [9]. The reason for this is twofold: i) it is difficult to mimic the human hand motion in an artificial robot hand, and ii) controlling such a convoluted system is an arduous task. In an attempt to find a compromise for these issues, underactuated tendon-driven robotic hands have been proposed [10]. These allow the execution of anthropomorphic manipulation strategies, which can be used to manipulate objects with higher dexterity. Recently, studies have focused on developing generalizable methods to robustly grasp objects

* and ‡ Equal first and last authorships respectively.
[1] {diego.hidalgo-carvajal, djallil.naceri, gemma.bettelani, melissa.zavaglia, jaesug.jung, haddadin}@tum.de, Technical University of Munich, Germany; TUM School of Computation, Information and Technology (CIT); Chair of Robotics and Systems Intelligence (RSI); Munich Institute of Robotics and Machine Intelligence (MIRMI)
[2] also with the Centre for Tactile Internet with Human-in-the-Loop (CeTI)
[3] {hanzhi.chen, stefan.leutenegger}@tum.de Technical University of Munich, Germany; TUM School of Computation, Information and Technology (CIT) and MIRMI; Smart Robotics Lab.
[4] busse@bio.lmu.de LMU Munich, Germany; Division of Neuroscience, Faculty of Biology

from different regions, while understanding the relationship between their geometry and grasping contexts [11]–[13].

In this paper we proposed a new end-to-end approach (see Fig. 1) which leverages neural object shape completion to conduct objects grasping in an anthropomorphic manner with a low-cost under-actuated robotic hand. Our contributions are as follows:

- We show an application in robotic grasping for partial-view conditioned shape completion together with shape confidence.
- We leverage transferred grasping knowledge from only one to multiple instances as proposed in our previous work [12].
- We further design a grasping solver for a 7-DoF robot arm in combination with under-actuated tendon-driven anthropomorphic robot hands.
- In a series of real-world experiments, we obtained grasping success rates well above 80% that are close to those achieved with entirely known models and substantially beyond what is achievable solely using (partial) reconstruction.

To the best of our knowledge, we are the first to design a system capable of conducting anthropomorphic grasping with object shape completion from single-view visual information in a human-like manner.

II. RELATED WORKS

A. Unseen Object's Regions Grasping

Recent methodologies have provided an alternative solution for visual object scene understanding during robotic manipulation. Such methodologies are based on leveraging categorical semantic consistencies for grasping of novel instances. On this note, the authors of [14] collected a grasping code book in a canonical space from simulation environments and inferred grasp poses through correspondence established by its proposed Non-Uniform Normalized Object Coordinate Space (NUNOCS). The authors of [13] proposed to learn feature-metric descriptors in 3D through shape completion pre-training. Then, they directly transferred grasps to novel instances which shared deep correspondence. Work of [15] leveraged a learnt deformation field to transfer grasping poses to novel instances, and further conducted refinement leveraging the antipodal principle. The authors of [16] focused on a system for fluid human-robot object handovers, where approaching directions were calculated depending on user grasping postures. It is worth noting that although some of these works explored transferable knowledge among categories to conduct grasping on novel instances in a label-efficient manner, they demonstrated such knowledge transfer with parallel-jaw grippers, which have limited dexterity compared to robot anthropomorphic hands. Additionally, they did not consider the effects of shape and pose variance of unseen objects on transferred knowledge uncertainty. These factors were considered in our approach.

B. Human-like Dexterous Manipulation

Grasping strategies are dependent on the dexterity of robot hands and their aptness for specific tasks [9], [11], [17]–[26]. In an attempt to find a compromise for the complexity of mimicking human hand motion, and the difficulties of controlling its multiple degrees of freedom (DOF), under-actuated tendon-driven robot hands have been proposed [10]. They allow the exertion of anthropomorphic manipulation strategies, which have the potential to expand the current grasping capabilities in robotics.

A well established tendency in robotic manipulation is the deployment of grasping strategies for parallel simplified grippers. Works such as [9], [17]–[19] have predicted feasible and stable finger posture placements on objects for such grippers. Unfortunately, their approaches lack generalization for more dexterous end effectors (EE), such as anthropomorphic robot hands. On the contrary, on a more recent trend, research efforts have focused on the development of dexterous manipulation using anthropomorphic robot hands [20]–[22] emulating human capabilities. Endowing robotic grasping systems with human-like dexterity entails the understanding of human-object interactions. In an attempt to do so, the authors of [27] captured a data set of human-hand-object contacts and developed an anthropomorphic grasping predictor on novel objects, similarly to [28]. The authors of [23], [24] exploited such concepts of human grasping contacts to develop deep learning approaches capable of generating grasps in simulation in a human-like manner using an anthropomorphic robot hand. The authors of [25] followed a similar contact-inspired methodology using a real world robot-arm-hand system. On a related note, authors of [26] proposed a deep learning method, where human-informed policies were used to grasp different objects with different anthropomorphic grasps. The authors of [11] went further and used depth images and deep convolutional neural networks to grasp objects using their entire geometry, without the limitation of using parallel grippers. The aforementioned approaches were able to grasp objects stably in simulation and real world experiments. Nevertheless, in most cases, they used simplified grippers, and could not choose grasping regions or approaching directions to the objects. Additionally, they did not analyze relations between different grasping types and the entirety of the objects geometry. This may be a limiting factor when manipulating objects in a dexterous manner, as different object regions may require different grips depending on the task at hand.

III. METHODS

As shown in Fig. 1, our proposed approach started with the capture of a static RGB-D image of the grasping scene. This image was used as an input to estimate present object shapes. The parts of the objects that were not visible to the camera were completed using our object shape reconstruction pipeline inspired by [8]. Specifically, we focused on drinking cups, bottles, and bowls. Our grasp posture estimator, proposed in [12] was used to predict grasping postures in the entire geometry of reconstructed and completed objects.

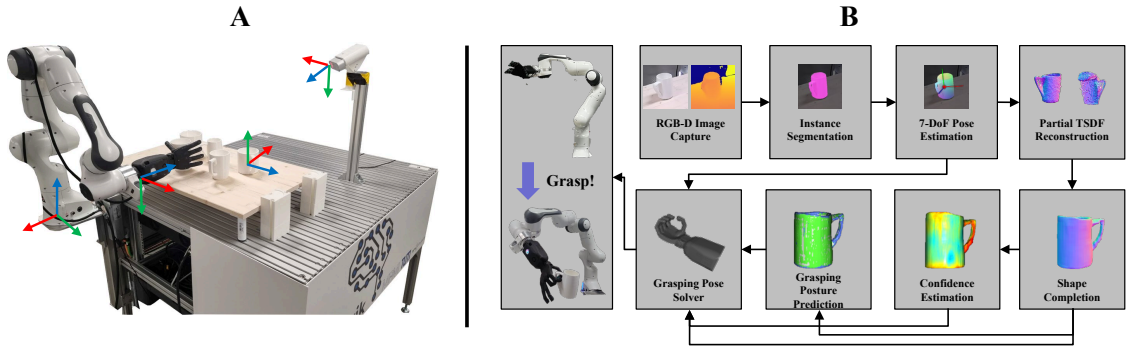

Fig. 2. (A) Physical setup of our anthropomorphic hand grasping system. XYZ-coordinate systems of the robot base, the robot hand, the object, and the camera are marked. All frames were with respect to the robot base. (B) Overview of our proposed vision-based grasping workflow.

Finally, our robot hand-arm system grasped the objects in areas of interest with predefined anthropomorphic grasping sequences using our proposed hand-grasping posture selection solver. The following sections detail each component of our proposed approach.

A. System Setup and Scene Capture

We used a robot hand-arm system, consisting of a 7-DoF Panda robot (Panda + FCI Lizenz, Franka Emika, Germany) and a 7-DoF (2 for the thumb, 1 for the index, 1 for the middle, 1 for both ring and little fingers, and 2 for the wrist) anthropomorphic robot hand (RH8D, Seed Robotics, Portugal). Our setup consisted of a fixed metallic table of dimensions $1192\times1100\times860$ mm, with an attached secondary metallic base 73 mm beneath. The secondary metallic base had dimensions 226×190 mm and supported the Franka Panda robot. The main metallic table was located on the second octant of the robot's workspace $(-X, +Y, +Z)$. For our experiments, we placed a wooden table of dimensions 800×500 mm on top of the metallic table (120 mm above), where objects were placed during our experiments. Fig. 2-A shows our robot setup.

To capture the grasping scene, we used a programmable multi-mode RGB-D camera (Azure Kinect DK, Microsoft, USA), which was mounted on a fixed platform using a customized 2-DoF assembly, allowing pitch and yaw adjustment (see Fig. 2-A). The camera was located at a height of 515 mm with respect to the metallic table, and was tilted 30° downwards pointing at the reachable workspace of the robot arm in our setup (second octant of its workspace: $-X, +Y, +Z$).

B. Object Pose Estimation and Shape Reconstruction

In this part, our goal was to retrieve geometric information of a novel instance from a pre-defined category. We used an object detector, a pose estimator, and a shape mapper parameterized by three individual deep neural networks, denoted as $f_{\text{detect}}(\cdot)$, $f_{\text{pose}}(\cdot)$, and $f_{\text{map}}(\cdot)$, respectively. Given one single-view RGB-D frame input $[\mathbf{I}, \mathbf{D}]$, we first used the object detector [29] to acquire the foreground mask of the object of interest with $\mathbf{S} = f_{\text{detect}}(\mathbf{I})$. Note $\mathbf{I}, \mathbf{D}, \mathbf{S}$ shared the same resolution ($\mathbf{I}, \mathbf{D}, \mathbf{S} \in \mathbb{R}^{1536\times2048}$) Then we acquired foreground point clouds using the segmented depth $\mathbf{D}_{\text{obj}} = \mathbf{S} \odot \mathbf{D}$, its corresponding pixel coordinates $\mathbf{u}_{\text{obj}}$ and camera intrinsics matrix $\mathbf{K} \in \mathbb{R}^{3\times3}$ with $\mathbf{x}_{\text{obj}} = \pi^{-1}(\mathbf{D}_{\text{obj}}, \mathbf{u}_{\text{obj}}, \mathbf{K})$, and passed $\mathbf{x}_{\text{obj}}$ to the pose estimator [30] to acquire the objects' 7-DoF pose (rotation, translation, and scale) between its pre-defined canonical frame and camera frame: $[\mathbf{R}, \mathbf{t}, s] = f_{\text{pose}}(\mathbf{x}_{\text{obj}})$, where $\mathbf{R} \in \text{SO}(3)$, $\mathbf{t} \in \mathbb{R}^3$, $s \in \mathbb{R}$.

To acquire the complete shape regardless of occlusions, we based the shape mapper from [8], while added gradients regularization for the occupancy of free space during training and removed heavy shape optimization during inference for the sake of speed. We transformed the observed point clouds to the canonical frame with $\mathbf{x}_{\text{cano}} = s\mathbf{R}\mathbf{x}_{\text{obj}} + \mathbf{t}$ and further acquired a partial truncated signed distance function (TSDF) volume $\mathbf{V}_{\text{cano}} \in \mathbb{R}^{64\times64\times64}$ by voxelizing $\mathbf{x}_{\text{cano}}$. Conditioning on $\mathbf{V}_{\text{cano}}$, the shape mapper predicted the complete shape represented by voxelized occupancy probability with $\mathbf{P}_{\text{cano}} = f_{\text{map}}(\mathbf{V}_{\text{cano}})$. Finally, we extracted complete mesh from the voxel grids using multi-resolution iso-surface extraction strategy from [31]. We further inferred shape confidence for each mesh vertex $\mathbf{v} \in \mathbb{R}^3$ represented by the norm of its gradient w.r.t. the occupancy probability ($||\partial\mathbf{P}_{\text{cano}}[\mathbf{v}]/\partial\mathbf{v}||_2$), which were further used as one of the criteria to select reliable grasps introduced in Section III-D.

C. Grasping Posture Prediction

Following the insights for grasping transferability in [12], we adopted our previously proposed grasping posture predictor to select plausible grasping postures for our robot arm-hand system to approach and grasp objects. The grasping posture predictor was also parameterized by a deep neural network denoted as $f_{\text{posture}}(\cdot)$. For each queried vertex $\mathbf{v}$ from the complete mesh provided by Section III-B, it predicted a grasping posture label g corresponding to either a medium wrap grasp (MW), tripod grasp (T), or non-graspable regions (NG) defined in [32] with $g = f_{\text{posture}}(\mathbf{v})$.

D. Hand grasping posture selection solver

We used the setup mentioned in Section III-A to exert anthropomorphic grasps with different orientations in different areas of objects of interest. Once the objects had assigned grasping posture predictions for their entire geometry, we selected the vertices of interest for grasping. Due to the constraints of our setup, we chose vertices whose normal vectors pointed on the first octant of the robot base frame $(+X, +Y, +Z)$. We then selected candidate grasping points based on two criteria, namely, according to the highest shape confidence, and in an arbitrary manner, as an alternative

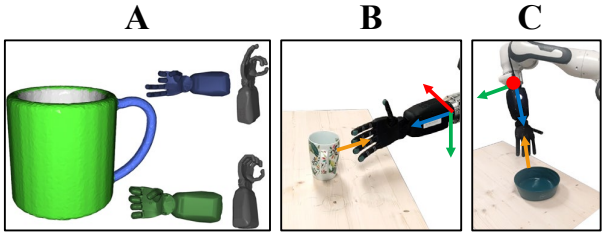

Fig. 3. (A) Object-centric grasping posture labeling. Green indicates medium wrap, and blue indicates tripod grasps. (B) Side-approaching grasp. (C) Top-approaching grasp. We marked the normal of the selected vertex and XYZ indicates the robot hand frame.

when confidence values were not available (see Section III-F). We then used a threshold of $45°$ on the inclination of the vertex's normal with respect to the XY plane of the robot's frame, to select either a side-approaching (Fig. 3-B) or top-approaching grasping (Fig. 3-C). For side-approaching grasps (e.g. for bottles, or drinking cups being grasped from the side): we used vertices whose $Z >= 45$ mm above the table. This safety distance (from the center of the robot hand palm to its ulnar end along its coronal plane [1]) guaranteed that the center of the robot hand's palm could approach the chosen grasping point without colliding with the table. We used top-approaching grasps (e.g. for bowls, or drinking cups being grasped from above) for vertices whose normal formed an angle of $>= 45°$ with the table. Normally normal vectors with such inclination were located at the top of the objects. In such instances a side-grasping approach was unfeasible.

Our grasping strategy consisted of the following steps (see Fig. 4):

1) First, the robot arm-hand system moved to an idle initial position at the top of the manipulation workspace visible by the camera.
2) Then we calculated a suitable robot hand wrist frame, so that its axis perpendicular to the hand transverse plane pointing on the distal direction was opposite to the normal vector of the selected grasping vertex (see Fig. 3). The other two axes on the wrist frame were calculated so the Z-axis of the robot base frame lay on the coronal plane of the robot hand (in flat configuration) for side-approaching grasps, and on the sagittal plane for top-approaching grasps, respectively. All frames and vectors were expressed in the robot base frame.
3) Following this, the robot's EE (tips of the middle and ring fingers of the robot hand in a flat configuration) was moved to an approaching position. This position was calculated as the point where the tips of the middle and ring finger were at a distance of 50 mm to the vertex of interest along its corresponding normal vector.
4) Subsequently, we prepared the robot hand for grasping as follows. We developed an algorithm to keep either the sagittal plane (side-approaching grasp) or the coronal plane (top-approaching grasp) of the robot hand parallel to the XY plane of the robot frame (horizontal leveling). This was achieved by abducting (side-approaching) or extending (top-approaching) the wrist an angle γ (angle between the normal vector of the chosen grasping point and the horizontal XY plane). Since leveling the hand horizontally moved the hand upwards, we added compensation distances, so the robot hand's palm center could reach the selected grasping vertex on the object. These distances were calculated as shown in equations 1, 3. The distance $d_{e_{x'}}$ was computed as:

$$d_{e_{x'}} = l(1 - \cos(\gamma)), \tag{1}$$

where l is the distance from the robot hand's wrist to the tip of the middle finger (EE) when the hand is in a flat configuration. $d_{e_{x'}}$ was compensated along the $\mathbf{e}_{x'}$ axis

$$\mathbf{e}_{x'} = \left(\begin{bmatrix}0\,0\,1\end{bmatrix} \times \mathbf{n}\right) \times \begin{bmatrix}0\,0\,1\end{bmatrix}, \tag{2}$$

where $\mathbf{n}$ is the axis opposing the normal vector to the selected grasping vertex (see Fig. 3). The distance d_z was compensated along the $[0, 0, -1]$ axis:

$$d_z = l\sin(\gamma). \tag{3}$$

Finally, these compensation distances were combined into the compensation displacement vector $\mathbf{c}$ and added to the end effector position. For side-approaching grasps we extended the hand's wrist $45°$ (its maximum extension angle), and fully abducted the thumb in the palmar direction.
5) Afterwards, the EE was moved a distance of 150 mm along the direction opposed to the normal vector at the vertex of interest ($\mathbf{n}$). This distance was set as the sum of the aforementioned approaching distance of 50 mm and the distance from the center of the palm to the tip of the middle finger or EE (100 mm). The aforementioned displacement compensation ($\mathbf{c}$) guaranteed that the palm of the hand was positioned next to the vertex of interest during grasping.
6) Finally, a robot hand pre-programmed finger grasping sequence, corresponding to either a medium wrap or a tripod grasp [32] was executed. These grasping sequences were independent of the aforementioned approaching types. Once the hand was closed, the object was lifted vertically (a distance of 100 mm for side-approaching, and 200 mm for the top-approaching grasps). We used the motor currents on the robot hand to detect a successful grasping. A successful medium wrap and tripod grasp were considered when the motor currents of at least four and three motors were higher than 400 mA for ≥ 4 s, respectively.

E. Network Training Protocol

We used pre-trained off-the-shelf network weights for the object detector and pose estimator, provided by [33] and [30], respectively. For the shape mapper for completion, we used 3D models provided by ShapeNet repository [34] to first render single-view depth train data using BlenderProc2 [35].

[1]Terminology in regard to the robot hand corresponds to the anatomical conventions for human hands.

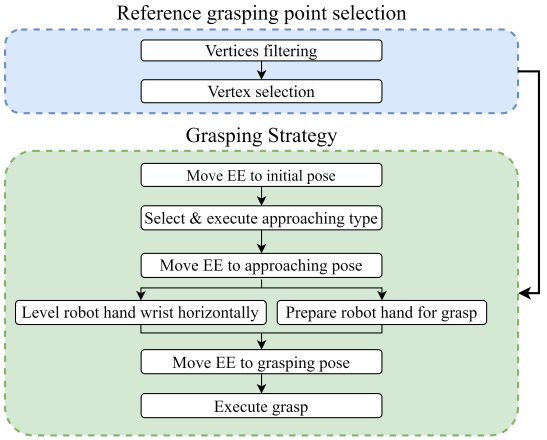

Fig. 4. Hand grasping posture selection solver. See Section III-D for details.

Then we adopted the same loss function as in [8] and trained the network for 20K iterations with a batch size of 48.

For our grasping predictor, we manually labeled one known object instance with the three grasping posture labels introduced in Section III-C. The known objects were represented by a triangular mesh with an average triangle side length of 2.5 mm. We followed the same training protocol as [12].

F. Experiments

Our experiments consisted of three parts. We tested our grasping framework using five 3D-printed drinking cups in the first part. The second part was done with five arbitrarily selected real-world drinking cups. We conducted grasping experiments for five bottles and five bowls in the third part. We calibrated the camera of our setup before the experiments using a modified version of the calibration pipeline proposed in[2]. The camera calibration was done with respect to the robot base frame and it allowed the direct representation of the reconstructed meshes in the robot base frame.

1) Grasping 3D-printed drinking cups: The experiments in this section were threefold. For each part, the 3D-printed drinking cups were placed on top of a wooden plate (simulating a regular table) in the camera's visible area. First, we tested our grasping strategy using the fully known mesh models of the objects (available to us, as we 3D printed them). We then arbitrarily selected points whose coordinates and normal vectors satisfied our setup grasping constraints, and would not result in a singular configuration or a collision. This was done because no object completion was required and, therefore, completion confidence values were not available. We then executed our grasping routine. We categorized a grasping attempt as a success if the hand was able to hold the object for four seconds after it was lifted. In the second part, we took a single static RGB-D image of the scene, and passed it to reconstruct the complete shapes of the desired drinking cup as introduced in Section III-B. We then grasped the drinking cups using the points with the highest shape confidence. Finally, we repeated the second part of the experiments with the object 3D reconstruction but without completion. We selected points randomly from the ones satisfying workspace constraints (see Section III-D).

[2]https://github.com/marcoesposito1988/easy_handeye_demo

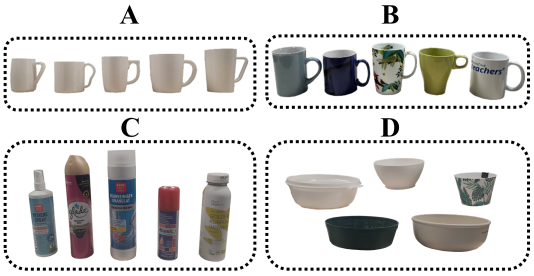

Fig. 5. Tested objects in grasping experiments. (A) 3D-printed drinking cups. (B) Real drinking cups. (C) Real bottles. (D) Real bowls

2) Grasping real-world drinking cups: In this part of the experiments, we recreated the second and third parts of the experiments in section III-F.1 using five real-world drinking cups. We first grasped drinking cups which were reconstructed and completed using depth and RGB information from a static image taken from the camera. We chose grasping points according to the highest shape confidence values and executed our grasping procedure. We then performed our grasping strategy on the same objects and positions, but without 3D shape completion, using points that satisfied the setup constraints.

3) Grasping real-world bottles and bowls: In these experiment trials, we followed the approach of the first part of experiments in section III-F.2 and applied it to two additional object categories. We grasped bottles and bowls, whose shape was reconstructed and completed, choosing points according to the highest shape confidence values and executed our grasping procedure. We tested five different instances for each object category.

IV. RESULTS

A. Shape accuracy of 3D-printed drinking cups

We first used the five 3D-printed drinking cups with ground-truth meshes to showcase the effectiveness of our object shape completion pipeline by comparing the shape accuracy of the extracted mesh after completion and the raw mesh from partial observations. The metrics for shape accuracy evaluation, chamfer distance L1, completeness, and normal consistency were first proposed by [31], and we kindly refer interested readers to it for detailed definitions. For chamfer distance (L1), we significantly improved the accuracy of raw and incomplete geometry from 0.0397 to 0.0157 with our completion pipeline, such gain was also reflected in completeness (from 0.0597 to 0.0124) and normal consistency (from 0.6842 to 0.8969). These results are quantitatively and qualitatively demonstrated in Table I, and in Fig. 6, respectively. 6

B. Grasping 3D-printed drinking cups

, grasping the cups with a medium wrap 16 times (88.89%) and with a tripod grip 13 times (86.67%) . In the second stage, we reconstructed and completed the 3D shape of the objects and were able to grasp them successfully at a rate of 81.82%, with 17 successful grasps using the medium wrap (94.44%) and 10 tripod grips (66.67%). In the third stage, we grasped the objects only with the 3D shape reconstruction but without completion. The success rate for this part of the experiments was 54.6%, being able to grasp the objects 8

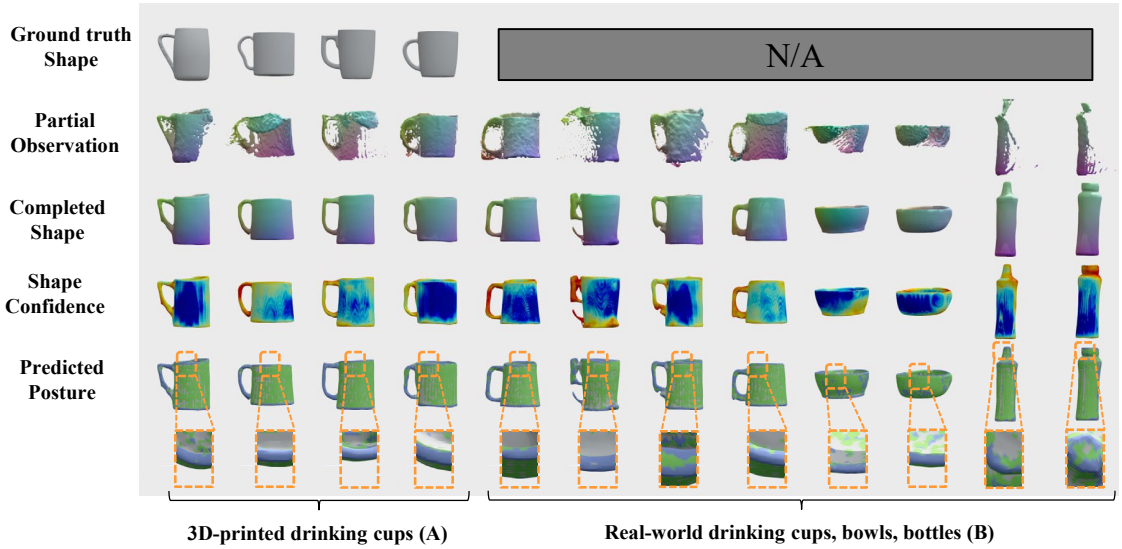

Fig. 6. Ground truth meshes, partial observations fed to the object reconstruction pipeline, and its outputted completed shapes, as well as their shape confidence (blue indicates high confidence, red indicates low confidence). The fifth row shows grasping predictions on the completed shapes (medium wrap in green, and tripod in blue). Last row below (in orange boxes) depicts a zoomed-in view on object rims for top-approaching grasps.

TABLE I

SHAPE ACCURACY (CHAMFER DISTANCE L1, COMPLETENESS AND NORMAL CONSISTENCY) ↓: LOWER IS BETTER, ↑: HIGHER IS BETTER.

Shape Accuracy	3D-printed drinking cups	
	Reconstruction and completion	Reconstruction only
Chamfer Dist. (L1) (↓)	**0.0157**	0.0397
Completeness (↓)	**0.0124**	0.0597
Normal Consis. (↑)	**0.8969**	0.6842

TABLE II

SUCCESS RATES (%) FOR EXPERIMENTS WITH 3D PRINTED AND REAL WORLD DRINKING CUPS. MW = MEDIUM WRAP, T = TRIPOD.

Grasping Type	3D-printed drinking cups			Real-world drinking cups	
	Known models	Unknown models			
		Reconstruction and completion	Reconstruction only	Reconstruction and completion	Reconstruction only
MW	88.89	**94.44**	44.44	**73.33**	46.67
T	86.67	**66.67**	**66.67**	**93.33**	53.33
Total	87.87	**81.82**	54.60	**83.33**	50.00

times with the medium grap (44.44%) and 10 times with the tripod grip (66.67%). Table II shows the success results of the 3D printed objects.

C. Grasping real-world drinking cups

In the second part of our grasping experiments, we grasped five real-world drinking cups in two stages, 30 times each. In each one of the stages, we grasped the objects 15 times with the medium grap and 15 times with the tripod grips. When we used reconstruction and completion, we achieved a success rate of 83.33%, being able to successfully grasp drinking cups with a medium wrap grip with a success rate of 73.33%. When using the tripod grip, the success rate increased to 93.33% (14 times). However, when using non-completed 3D models the overall accuracy was 50.00%. We grasped drinking cups successfully with the medium wrap and tripod grips 46.67% and 53.33% of the time, respectively.

D. Grasping real-world bottles and bowls

Following the experimental procedure of section III-F.3, we conducted grasping experiments for 10 additional objects, namely, 5 commercial bottles, and 5 commercial bowls. We grasped each object 10 times, totalling 100 experiments. For each trial, we executed the grasping posture suggested by our grasping posture predictor (i.e. Medium Wrap or Tripod). Our framework reached successful grasps in 86.00% of the instances for bottles and 82.00% of the instances for bowls.

V. DISCUSSION

In this work we proposed a novel approach which combines an object shape completion algorithm with a grasping strategy. Our approach can be robustly used to grasp partially visible objects in an anthropomorphic manner. The proposed approach benefits from the anthropomorphism properties of a multi-fingered robot hand, resembling more closely human grasping capabilities and strategies. Additionally, our pipeline can be run automatically.

Our object shape reconstruction method demonstrated its strong ability to retrieve smooth, clean, and complete geometric information from raw, noisy, and partial observations, potentially caused by sensed depth noise, imperfect segmentation, or occlusions. To illustrate this, Fig. 6 showed that the partial reconstruction of the handles was significantly deficient compared to other parts of the drinking cups. This is due to their thin structures and high curvatures. Nevertheless, our reconstruction pipeline was still able to retrieve fine details of the geometry even under such imperfect conditions. The estimated shape confidence for the handle regions was low. This is because the conditioned input in these regions was prone to noise. On a similar note, although the drinking cup handles were not visible due to object self-occlusions (see column 6 in Fig. 6), our method was still able to estimate handle shapes reasonably. Nevertheless, such estimations had higher uncertainty, as expected. Regarding the body part of the cups, even though the completed regions had relatively low confidence values compared to those of visible regions (see columns 1 and 2 in Fig. 6), several regions still exhibited high confidence values. This is due to the ”symmetry prior” learnt from our shape mapper network through 3D completion training. We further demonstrated the results of bowls and bottles in columns 9-12, which again

yielded promising full-shape reconstructions.

The integration of our shape reconstruction and completion pipeline along with our hand grasping posture selection solver allowed us to robustly grasp different types of objects in spite of significant object shape occlusions. Our approach brings dexterity a step forward by closing the gaps on missing components of previous works [11], [13], where no full object anthropomorphic grasp predictions were analyzed nor exerted. Although we were able to grasp objects using only partially reconstructed meshes from the RGB-D images, the robustness ($\sim 50\%$) of the grasps depended on the quality of the reconstruction, which in turn depended on the position of the objects and on the grasping points selection. Additionally, since the camera was placed on the opposite side of the robot's grasping space, the reconstructed geometries on the robot's grasping space had limited feasible regions for grasping. However, when using 3D shape completion, we were able to select the grasping regions on the robot's grasping space according to two different criteria (highest completion certainty and arbitrarily, as described in Section III-F.1) and grasp the drinking cups with a success rate up to 83.33% using both medium wrap and tripod grips in both 3D printed and regular cups. As expected, the success rate was lower compared to when we used the fully known object models (87.87%), as explained in Section III-F.1. The high success rate for other categories (86% for bottles and 82% for bowls) further verified the scalability potential of our pipeline.

Our physical setup and our right-handed grasping strategy limited our graspable vertices, as explained in section III-D. This vertices selection can be adjusted effortlessly for setups with objects located in other octants of the robot's workspace. Our framework has not been tested in daily real-world scenarios, for instance highly cluttered scenes at home. Nevertheless, the generalization capabilities of our approach shows potential as a real-world pipeline capable of achieving noteworthy results.

The grasping types were selected based on our grasping predictor proposed in [12]. Due to the geometric variance of unseen objects, some regions of the objects could be incorrectly labeled as non-graspable regions, or with an outlier grasping type, which was unfeasible to accomplish (*c.f.* "Predicted Posture" in Fig. 6). Our pipeline mitigated such issue by carefully designing the posture selection solver considering geometric features and shape confidence of the objects, to select reliable approach points for the robot hand. In regard to the postures we used in this work, we selected objects that required the exertion of a medium wrap and tripod grips, as shown in [12]. This selection allowed the demonstration of power and precision grasp, which are crucial for human anthropomorphic grasping. These two grasping postures, additionally, represent motions along the first two synergies proposed by [36], and are among the 5 most common postures in activities of daily living [37]. Since robot hand grasps are treated as an independent module in our framework, additional postures, using more sophisticated robot hands can be implemented. With our approach, we attempt to show that exact contact points or grasping forces are not required for high grasping success rates, emulating human knowledge and manipulation strategies into robotic grasping. Nevertheless, our framework's modularity can be exploited by including robot hands that allow customization of further parameters during grasps such as finger forces.

Since our grasping predictor and shape estimator played a major role in our pipeline, improvements in their accuracy would lead to higher success rates. This may be achieved by adding more category-specific training samples. However, this could bring undesirable inefficiency to the preparation process as annotating grasping postures, for instance, is labor-intensive as discussed in [12]. It is hence worthwhile to explore the trade-off between the training data size and grasping performance required for general grasping frameworks as ours.

VI. CONCLUSION

In this work, we successfully implemented an end-to-end approach (see Fig. 1) that integrates object shape completion, grasping posture prediction, and a robot arm-hand anthropomorphic grasping strategy to grasp unseen objects in a real-world setup. Our approach can automatically reconstruct and infer complete 3D models of novel objects from a single static view under occlusions. We can then infer grasping postures associated with the entire geometry of the objects in a label-efficient manner. Our proposed grasping strategy, shown in Fig. 4, allowed us to robustly grasp and lift a variety of objects over 150 times using a low-cost anthropomorphic robot hand. Given our method's robustness against object occlusions and its aptness to grasp objects in arbitrary regions from arbitrary directions in an anthropomorphic manner, the proposed approach shows high potential for applications where assistive robots play a crucial role, for example, scenarios of activities of daily living for elderly care. Our framework's current state requires the selection of objects that fit within the used robot hand workspace. This might limit the object categories to be grasped, similarly to a human hand. Additionally, the framework does not operate with low-latency constraints. Nevertheless, it has the potential to be deployed with hardware acceleration modules. For future work, we aim to integrate uncertainly estimation for the grasping posture predictor, and further generalize our pipeline to handle additional object categories.

ACKNOWLEDGEMENT

This work was funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden, and by LMUexcellent and TUM AGENDA 2030, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria (ONE MUNICH). This work was also supported by BMBF by funding the project AI.D under the Number 16ME0539K, and by TUM Georg Nemetschek Institute under the project

SPAICR. We acknowledge the funding of the Lighthouse Initiative Geriatronics by LongLeif GaPa gGmbH (Project Y), and as part of the SFB 1233 "Robust vision: Inference Principles and Neural Mechanisms", project number 276693517 (TP13). Please note Sami Haddadin has a potential conflict of interest as a shareholder of Franka Emika GmbH.

REFERENCES

[1] M. H. Herzog and A. M. Clarke, "Why vision is not both hierarchical and feedforward," *Frontiers in computational neuroscience*, vol. 8, p. 135, 2014.

[2] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How Does the Brain Solve Visual Object Recognition?" *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S089662731200092X

[3] C. D. Gilbert and W. Li, "Top-down influences on visual processing," *Nature Reviews Neuroscience*, vol. 14, no. 5, pp. 350–363, 2013.

[4] A. Hollingworth and J. M. Henderson, "Testing a conceptual locus," *Memory & Cognition*, vol. 31, no. 6, pp. 930–940, 2003.

[5] B. Bescos, C. Campos, J. D. Tardós, and J. Neira, "Dynaslam ii: Tightly-coupled multi-object tracking and slam," *IEEE robotics and automation letters*, vol. 6, no. 3, pp. 5191–5198, 2021.

[6] B. Xu, W. Li, D. Tzoumanikas, M. Bloesch, A. Davison, and S. Leutenegger, "Mid-fusion: Octree-based object-level multi-instance dynamic slam," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 5231–5237.

[7] M. Runz, M. Buffier, and L. Agapito, "Maskfusion: Real-time recognition, tracking and reconstruction of multiple moving objects," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 2018, pp. 10–20.

[8] B. Xu, A. J. Davison, and S. Leutenegger, "Learning to complete object shapes for object-level mapping in dynamic scenes," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 2257–2264.

[9] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017.

[10] C. Piazza, G. Grioli, M. Catalano, and A. Bicchi, "A century of robotic hands," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 2, pp. 1–32, 2019.

[11] P. Schmidt, N. Vahrenkamp, M. Wächter, and T. Asfour, "Grasping of unknown objects using deep convolutional neural networks based on depth images," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 6831–6838.

[12] D. Hidalgo-Carvajal, C. M. C. O. Valle, A. Naceri, and S. Haddadin, "Object-centric grasping transferability: Linking meshes to postures," in *2022 IEEE-RAS 21st International Conference on Humanoid Robots (Humanoids)*, 2022, pp. 659–666.

[13] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se(3)-equivariant object representations for manipulation," 2022.

[14] B. Wen, W. Lian, K. Bekris, and S. Schaal, "Catgrasp: Learning category-level task-relevant grasping in clutter from simulation," *ICRA 2022*, 2022.

[15] H. Wen, J. Yan, W. Peng, and Y. Sun, "Transgrasp: Grasp pose estimation of a category of objects by transferring grasps from only one labeled instance," in *European Conference on Computer Vision*. Springer, 2022, pp. 445–461.

[16] W. Yang, C. Paxton, M. Cakmak, and D. Fox, "Human grasp classification for reactive human-to-robot handovers," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020, pp. 11 123–11 130.

[17] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[18] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 3406–3413.

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

[20] J. C. V. Tieck, K. Secker, J. Kaiser, A. Roennau, and R. Dillmann, "Soft-grasping with an anthropomorphic robotic hand using spiking neurons," *IEEE Robotics and Automation Letters*, vol. 6, pp. 2894–2901, 2020.

[21] C. Gabellieri, F. Angelini, V. Arapi, A. Palleschi, M. G. Catalano, G. Grioli, L. Pallottino, A. Bicchi, M. Bianchi, and M. Garabini, "Grasp it like a pro grasp of unknown objects with robotic hands based on skilled human expertise," *IEEE Robotics and Automation Letters*, vol. 5, pp. 2808–2815, 2020.

[22] H. Li, J. Tan, and H. He, "Magichand: Context-aware dexterous grasping using an anthropomorphic robotic hand," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9895–9901.

[23] E. Valarezo Añazco, P. Rivera Lopez, N. Park, J. Oh, G. Ryu, M. A. Al-antari, and T.-S. Kim, "Natural object manipulation using anthropomorphic robotic hand through deep reinforcement learning and deep grasping probability network," *Applied Intelligence*, vol. 51, pp. 1041–1055, 2021.

[24] P. Rivera, E. Valarezo Añazco, and T.-S. Kim, "Object manipulation with an anthropomorphic robotic hand via deep reinforcement learning with a synergy space of natural hand poses," *Sensors*, vol. 21, no. 16, p. 5301, 2021.

[25] W. Wei, P. Wang, and S. Wang, "Generalized anthropomorphic functional grasping with minimal demonstrations," *arXiv preprint arXiv:2303.17808*, 2023.

[26] F. Ficuciello, A. Migliozzi, G. Laudante, P. Falco, and B. Siciliano, "Vision-based grasp learning of an anthropomorphic hand-arm system in a synergy-based control framework," *Science robotics*, vol. 4, no. 26, p. eaao4900, 2019.

[27] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas, "GRAB: A dataset of whole-body human grasping of objects," in *European Conference on Computer Vision (ECCV)*, 2020. [Online]. Available: https://grab.is.tue.mpg.de

[28] H. Jiang, S. Liu, and J. U. W. X. W. S. Diego, "Hand-object contact consistency reasoning for human grasps generation." [Online]. Available: https://hwjiang1510.github.io/GraspTTA/.

[29] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9799–9808.

[30] G. Li, Y. Li, Z. Ye, Q. Zhang, T. Kong, Z. Cui, and G. Zhang, "Generative category-level shape and pose estimation with semantic primitives," in *Conference on Robot Learning*. PMLR, 2023, pp. 1390–1400.

[31] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4460–4470.

[32] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[33] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, and R. Girshick, "Detectron2," https://github.com/facebookresearch/detectron2, 2019.

[34] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, "Shapenet: An information-rich 3d model repository," *arXiv preprint arXiv:1512.03012*, 2015.

[35] M. Denninger, D. Winkelbauer, M. Sundermeyer, W. Boerdijk, M. Knauer, K. H. Strobl, M. Humt, and R. Triebel, "Blenderproc2: A procedural pipeline for photorealistic rendering," *Journal of Open Source Software*, vol. 8, no. 82, p. 4901, 2023. [Online]. Available: https://doi.org/10.21105/joss.04901

[36] M. Santello, M. Flanders, and J. F. Soechting, "Postural hand synergies for tool use," *Journal of neuroscience*, vol. 18, no. 23, pp. 10 105–10 115, 1998.

[37] T. Feix, I. M. Bullock, and A. M. Dollar, "Analysis of human grasping behavior: Object characteristics and grasp type," *IEEE transactions on haptics*, vol. 7, no. 3, pp. 311–323, 2014.