3D Mixed Reality Object Drag-and-Drop

Petar Lukovic[1,⋆], Federica Bruni[1,⋆], Pablo Soler[1], Martin Brun[1], Lucie Reynaud[1], Andrea Sgobbi[1], Marc Zünd[1], Riccardo Bollati[1], Marc Pollefeys[1,2], Hermann Blum[1,3,†] and Zuria Bauer[1,†]
[1]*ETH Zurich* [2]*Microsoft* [3]*Lamarr Institute / Uni Bonn*
{plukovic, fbruni, psoler, mabrun, lreynaud, asgobbi, mazuend, rbollati
jiaqchen, pomarc, blumh, zbauer} @ethz.ch

***Abstract*— Human-robot interaction through mixed reality (MR) technologies enables novel, intuitive interfaces to control robots in remote operations. Such interfaces facilitate operations in hazardous environments, where human presence is risky, yet human oversight remains crucial. Potential environments include disaster response scenarios and areas with high radiation or toxic chemicals. In this paper we present an interface system projecting a 3D representation of a scanned room as a scaled-down 'dollhouse' hologram, allowing users to select and manipulate objects using a straightforward drag-and-drop interface. We then translate these drag-and-drop user commands into real-time robot actions based on the recent spot-compose framework. The Unity-based application provides an interactive tutorial and a user-friendly experience, ensuring ease of use. Through comprehensive end-to-end testing, we validate the system's capability in executing pick-and-place tasks. Our findings highlight the advantages of this interface in improving user experience and operational efficiency. This work lays the groundwork for a robust framework that advances the potential for seamless human-robot collaboration in diverse applications.**

I. INTRODUCTION

Human-robot interaction (HRI) has become increasingly important due to the complementary nature of human and machine capabilities. Robots are now commonly deployed in scenarios where human presence is risky or impractical, such as hazardous environments and remote locations. Examples include industrial settings with toxic materials, areas affected by natural disasters, and space exploration missions. These deployments minimize the need for on-site human intervention, enhancing operational efficiency and safety [13], [4].

A significant challenge in HRI lies in the interfaces used to control robots. Traditionally, three-dimensional (3D) environments are represented on two-dimensional (2D) screens. This representation can complicate the operator's ability to maintain situational awareness and precision, often leading to difficulties in effectively controlling the robot and performing complex tasks. Traditional 2D interfaces can limit the operator's depth perception and spatial understanding, making it challenging to interact with and manipulate objects accurately in a 3D space [2].

Mixed reality (MR) technology offers a promising solution by providing a more immersive and intuitive interface for human-robot interaction. By integrating MR, control systems for robots can be significantly enhanced, making them more user-friendly and efficient. This approach addresses the limitations of traditional 2D interfaces by providing a 3D visual representation that aligns with the real-world environment, thereby improving situational awareness and operational precision.

Fig. 1: **Visualization of our system.** The picture shows actually scene where robot is operating along with a holographic image user sees in the HoloLens 2 [12] headset. Robot is in the start location and ready for user with a headset to issue a task command by moving object in its virtual scene.

Recent research has explored the use of MR headsets to control robots. For example, Chen et al. demonstrate real-time visualization and intuitive navigation of robots in 3D, highlighting the benefits of MR interfaces over traditional methods [5]. Similarly, "OpenMask3D: Open-Vocabulary 3D Instance Segmentation" by Takmaz et al. showcases the use of instance segmentation in 3D, which is essential for understanding and interacting with complex 3D environments [14]. These studies underscore the advantages of MR technology in enhancing the interaction between humans and robots, making it possible to perform more complex and precise tasks.

Building on this, our project presents an innovative strategy for developing a user-friendly application on a Microsoft HoloLens 2, enabling users to command Spot robot from Boston Dynamics. The primary goal of our project is to create an interface that allows users to control a mobile manipulator through natural, intuitive actions in a MR environment.

⋆Shared first authorship.
†Equal supervision.

Our system achieves this by implementing several key features. First, we scan and reconstruct the deployment environment in 3D. This digital twin provides a comprehensive spatial understanding that is crucial for accurate robot control. Second, we segment the reconstruction into individual object instances through 3D object instance segmentation [**?**], [14]. Third, we make the digital twin interactive based on the object segmentation and link any interaction in the digital twin to a robot that reproduces the same environment interaction in the real environment. Therefore, by simply dragging and dropping virtual objects, users can instruct the robot to pick up and move these objects to specified locations in the real world. To ensure ease of use, our application includes an interactive tutorial that guides users through the interface and functionality. This tutorial helps users quickly get familiar with the system, making it accessible even to those with minimal technical expertise.

The main contributions of this work can be summarized as follows:

- We present a novel interface for remote robot operations that enables object manipulation using a drag-and-drop feature within a MR environment.
- We implement the proposed interface as an app for the HoloLens2 and Spot robot.
- We perform extensive real-world tests with our interface and show that a short, fully automated training is sufficient to train users to perform complex pick-and-place interactions with a mobile manipulator.

II. RELATED WORKS

The intersection of Mixed Reality and Human-Robot Interaction has become a dynamic research area, with numerous studies exploring the potential of MR to enhance robotic control and human collaboration in complex environments. We highlight a selection of recent and particularly pertinent studies.

A. *Control via Mixed Reality Interfaces*

Recent advancements in MR technologies have significantly impacted HRI, particularly in scenarios requiring intuitive control and enhanced situational awareness. Previous studies have demonstrated the effectiveness of MR interfaces in improving operator control over robotic systems, particularly in environments where traditional 2D interfaces fall short.

For example, [5] developed a VR-based interface that enables real-time visualization and manipulation of robotic systems. Their work demonstrated the advantages of using 3D interfaces to overcome the limitations inherent in 2D displays, allowing operators to control robots with greater precision and efficiency.

Additionally, the study by [9] explores the fusion of MR with robotic systems, particularly focusing on the application of MR in scenarios requiring intricate spatial understanding and collaboration. Their research highlights the use of MR to facilitate HRI in environments that are traditionally challenging for robot navigation and manipulation, underscoring the potential of MR technologies in HRI.

B. *3D Instance Segmentation and Grasp Pose Estimation*

Understanding and interacting with complex environments is a fundamental challenge in robotics, which is often addressed through 3D instance segmentation and grasp pose estimation. The Spot-Compose framework, introduced by [10], is an advanced approach to these tasks. This framework utilizes a combination of sophisticated algorithms to segment 3D scenes and localize objects, making it particularly effective for tasks like object retrieval and manipulation.

Building on the concepts of 3D instance segmentation, the OpenMask3D framework [14] offers a powerful tool for segmenting and identifying objects within complex scenes. OpenMask3D employs visual-language models (VLMs) to perform open-vocabulary segmentation, allowing segmentation and identification of arbitrary objects within a scene, enhancing the robot's ability to interact with diverse and previously unseen items based on natural language descriptions. This capability is critical in environments where the robot must adapt to new and unseen objects, enhancing its operational flexibility and effectiveness.

In parallel, grasp pose estimation has seen substantial progress through the development of AnyGrasp, a framework introduced by [8]. AnyGrasp predicts two-finger grasps directly from 3D point clouds, using a dense supervision strategy that incorporates real perception data and analytic labels. This method accounts for factors such as the object's center-of-mass and environmental constraints, ensuring stable and collision-free grasps.

By combining these advanced segmentation and grasping techniques, robots are equipped with the necessary tools to perform dynamic and precise manipulation tasks in varied and unpredictable settings.

C. *Human-Robot Collaboration in Hazardous Environments*

The deployment of robotic systems in hazardous environments is critical for minimizing human exposure to dangerous conditions while maintaining essential oversight. In such settings, human-robot collaboration becomes indispensable. Research has increasingly focused on developing interfaces that enhance this collaboration, particularly in environments where traditional methods of operation are challenged by factors such as restricted visibility, heightened risks, or the necessity for accurate, real-time decision-making.

A notable example is the work by CERN ATS on developing an MR interface for remote operations in accelerator facilities [4]. This system was designed to allow operators to control robots within hazardous accelerator environments, where human presence is restricted due to safety concerns. By integrating MR, the interface enables users to interact with a virtual representation of the environment, enhancing their ability to perform complex tasks with greater precision and situational awareness.

Complementing this, [11] investigated the use of MR for robotic operations in environments contaminated with toxic chemicals. The study demonstrated that MR could significantly improve the operator's ability to navigate and

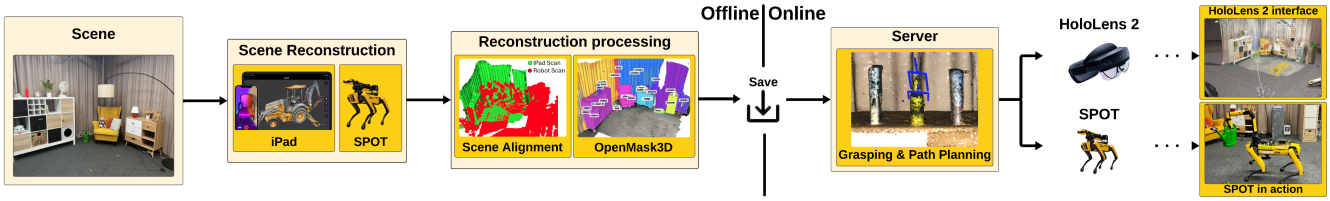

Fig. 2: **System overview.** Our method relies on both offline and online segments. Offline segment (left), is used to construct 3D scene that will be utilised in the subsequent online phase. The online segment (right) is used to control the Boston Dynamics Spot [7] using the Microsoft HoloLens 2 [12] with the help of an online server.

manipulate objects in these environments, thus reducing the risk to human life and enhancing the efficiency of robotic interventions.

III. METHOD

In this section, we present the main components and techniques employed by the system. An overview is illustrated in Figure 2. Starting with the implementation of the offline part, followed with the implementation of the online part of the pipeline and ending with the integration between the two, in the following we go through each part of our proposed method.

A. *Obtaining 3D Scene Reconstruction*

We assume the availability of a pre-scanned 3D representation of the scene. This representation can be optained through SLAM with a suitably equipped robot or handheld scanners. In our experiments, we use a LiDAR-equipped iPad and the *3D Scanner App* [1] (high detail scan). Like in [10], we register the robot to this scan by performing a depth scan (low detail scan) with the onboard cameras of the spot [7]. While the onboard cameras face downward and therefore cannot be used alone to reconstruct the scene, the two registered pointclouds then form a comprehensive and accurate 3D reconstruction of the scene, which is used in all subsequent steps.

B. *3D Semantic Instance Segmentation*

We segment the 3D reconstruction into semantic instances using the OpenMask3D [14] framework. This is illustrated in subsequent steps in Figures 2 and 3. This state-of-the-art method supports open vocabulary queries on 3D scenes, allowing us to predict and delineate distinct object instances within a 3D point cloud while simultaneously identifying their semantic categories.

The result of this step is presented at the end of the Figure 3. The point cloud is segmented into each movable object instance and the static environment. It is worth noting that we provide a predefined list of language prompts to select movable objects, or segmented point cloud for that matter. All other objects are considered static.

C. *Unity App Interface on HoloLens 2*

To enable intuitive interaction with the reconstructed 3D environment, we develop a Unity application (see Figure 4) for the HoloLens 2 [12]. The interface shows the 3D reconstruction as a hologram in front of the user, where they can then with their hands drag-and-drop the segmented movable objects through the HoloLens hand tracking.

The application imports the point cloud of the scene, with each movable object instance as an independent grabbable hologram. To facilitate user manipulation, we furhter add a virtual floor to the hologram. When an object hologram is grabbed and subsequently dropped, custom scripts send the object's new coordinates to the main workstation. To ensure operational feasibility, a control mechanism verifies whether the drop location is within the robot's operational area.

During testing, we found it beneficial to include a visual representation of the robot with a status display to keep users informed of the robot's activities. For users unfamiliar with the HoloLens, we introduce a tutorial to explain the app's features and practice drag-and-drop of holograms. We also add confirmation buttons to avoid sending wrong object placements to the robot and to alert users if the chosen location is not feasible. Additionally, we integrate voice control features: the *"show items"* command highlights movable objects to the user with colored bounding boxes, while the *"reset"* command allows users to return all moved objects to their original positions. All features are illustrated in Figure 4.

D. *Online operation of the Robot*

After an object is selected by the user and dropped at a new location in the interface, the interaction gets translated into a pick-and-place command on and send to the Spot robot. In Figure 5, this information is depicted as *"Navigation & Grasping Instruction"*. This includes the drop coordinates in the 3D scene and the index of the object in the scene representation that is available to both the robot and the interface.

To execute the pick-and-place command, we use the Spot-Compose [10] framework that we briefly outline here: Picking and placing starts with grasp estimation on the object using AnyGrasp [8]. We run inference over multiple rotations of the instance mask, since AnyGrasp identifies poses based on the frontal view. The system then performs joint optimization of poses and grasps based on the AnyGrasp [8] score, the alignment of the robot body with the grasp pose and the vicinity of obstacles. Once the robot has moved to the pose from which it should grasp the object, a local point cloud of the object is captured using the depth camera located in the robot gripper. This local capture is aligned to the initial scan using ICP [3] to obtain a corrective transformation for the grasp pose, compensating any misalignment or drift between robot odometry and scene representation (see *"Optimise grasp"* in Figure 5). Next, the robot grasps the item, picks it

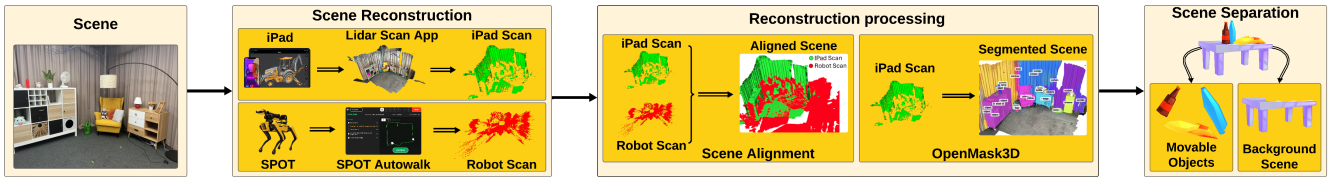

Fig. 3: **Offline pipeline.** Preprocessing done before deploying our system can be separated into three parts: Scene Reconstruction, Reconstruction Processing and Scene Separation, from left to right. Scene reconstruction consists of gathering high and low resolution scans using iPad LiDAR(*3D Scanner App* [1]) and Boston Dynamics Spot robot [7] cameras, respectively. Recorded point clouds are then aligned into same coordinate system and high resolution scan is segmented using OpenMask3D [14]. We will call reconstruction processing these two steps. At the end we manually separate segmented instances into draggable objects and environment.

Fig. 4: **HoloLens 2 Interface.** The above figures show various aspects of the visual interface on HoloLens 2 [12]. Labeling images with number left to right, and top to bottom we have: on the first image we can see bounding boxes surrounding objects following ”show items” voice command, following image shows how robot status is displayed during manipulation, third image show user manipulation of the watering can, on the next image we can see confirmation dialog box. On the fifth and sixth images we can see virtual representation of the robot and menu containing battery percentage and additional status information.

up, and subsequently moves to the drop location. The route to this location is calculated at the same time as the route from start to the object location. We also run joint optimisation for drop location (see *”Optimise grasp”* in Figure 5). This is computed already during the pick operation, since drop location is known a priori. Finally, Spot drops the object and moves back to the start location and localizes itself again. This step is important as to alleviate drift that would otherwise accumulate over multiple interactions.

E. System Integration

Our system is based on a centralized architecture, with all devices connected via wifi (these include the Spot [7], HoloLens 2 [12] and the server). The server is responsible for the entire planning procedure and sends commands to the Spot robot using the Boston Dynamics Python SDK [6]. The AnyGrasp [8] and OpenMask3D [14] models are served by containers also running on this machine. The HoloLens app queries the robot status information and issue commands via a REST API exposed by the workstation.

IV. EXPERIMENTS

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna.

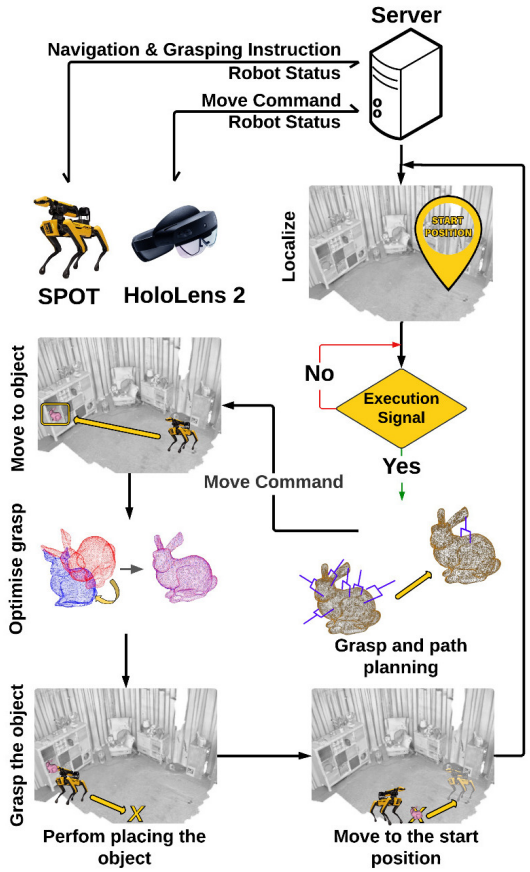

Fig. 5: **Online pipeline.** When system is deployed (online) it follows the given loop. Start of the pick-and-place procedure is triggered by HoloLens 2 [12] user who performs drag-and-drop in the user interface. This triggers the signal to the intermediate server, and gives additional information about the object and its location. After successful information exchange robot is localized. Next, grasp and path are calculated on the server which gives robot move command. After robot arrived to the location, grasp optimisation is performed using ICP algorithm [3]. At the end robot performs the grasp, moves the object and returns to starting position where it localizes itself waiting for another trigger signal.

Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

V. CONCLUSION AND FUTURE WORK

We introduced a novel application of a mixed reality headset to enhance human-robot interaction. We have shown that our system enables intuitive pick and place commands of a Spot robot from Boston Dynamics through a drag-and-drop interface displayed in a 3D mixed reality environment on a Microsoft HoloLens2. Our solution offers an interactive and user-friendly app including a tutorial. Our study shows that *STUDY RESULTS BLABLA*.

Moreover, through experiments, we underlined the practical applicability of our system. Our solution could be integrated in many applications, for instance in disaster management or in industry assistance. Finally, the flexibility and modularity of our project offers numerous exciting opportunities for further improvements. The capabilities of our work could be expanded in various ways: Refining the drag-and-drop mechanics based on the object type, enabling the system to perform sequential pick-and-place tasks - allowing users to move the same object from one location to another repeatedly, and automating the scene generation in the Hololens2 using the semantic information from OpenMask3D, removing the need to hand-pick movable objects.

APPENDIX

REFERENCES

[1] 3D Scanner App. https://3dscannerapp.com/.

[2] Carl Ahlberg and Sara Eriksson. Challenges and developments in human-robot interaction interfaces: Overcoming 2d limitations for effective 3d control. *Current Robotics Reports*, 8(3):215–232, 2022.

[3] P.J. Besl and Neil D. McKay. A method for registration of 3-d shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):239–256, 1992.

[4] CERN ATS. Mixed reality human-robot interface for remote operations in accelerator facilities. https://ats.web.cern.ch/node/145, 2023. Accessed: 2024-07-21.

[5] Jiaqi Chen, Boyang Sun, Marc Pollefeys, and Hermann Blum. A 3d mixed reality interface for human-robot teaming, 2023.

[6] Boston Dynamics. Spot SDK Documentation. https://dev.bostondynamics.com/.

[7] Boston Dynamics. Spot: The agile mobile robot. https://bostondynamics.com/products/spot/.

[8] Hao-Shu Fang, Chenxi Wang, Hongjie Fang, Minghao Gou, Jirong Liu, Hengxu Yan, Wenhai Liu, Yichen Xie, and Cewu Lu. Anygrasp: Robust and efficient grasp perception in spatial and temporal domains, 2023.

[9] Eduardo Iglesius, Masato Kobayashi, Yuki Uranishi, and Haruo Takemura. Mrnab: Mixed reality-based robot navigation interface using optical-see-through mr-beacon. *arXiv preprint arXiv:2403.19310*, 2024.

[10] Oliver Lemke, Zuria Bauer, René Zurbrügg, Marc Pollefeys, Francis Engelmann, and Hermann Blum. Spot-compose: A framework for open-vocabulary object retrieval and drawer manipulation in point clouds, 2024.

[11] Giacomo Lunghi, Raul Marin, Mario Di Castro, Alessandro Masi, and Pedro J Sanz. Multimodal human-robot interface for accessible remote robotic interventions in hazardous environments. *IEEE Access*, 7:127290–127319, 2019.

[12] Microsoft. Hololens. https://www.microsoft.com/en-us/hololens.

[13] Krzysztof Adam Szczurek, Raul Marin, Eloise Matheson, Jose Rodriguez-Nogueira, and Mario Di Castro. Multimodal multi-user mixed reality human–robot interface for remote operations in hazardous environments. *IEEE Access*, 11:17305–17333, 2023.

[14] Ayça Takmaz, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation, 2023.