A System for Autonomous Grasping and Manipulation in Unstructured Environments

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

von der KIT-Fakultät für Informatik
des Karlsruher Instituts für Technologie (KIT)

genehmigte

DISSERTATION

von

M.Sc. Christoph Pohl

aus Klausdorf, Brandenburg

Tag der mündlichen Prüfung:	11. November 2024
Erster Gutachter:	Prof. Dr.-Ing. Tamim Asfour
Zweiter Gutachter:	Ao. Univ.-Prof. Dipl.-Ing. Dr. techn. Markus Vincze

Abstract

The rapid advance of autonomous robots in the personal sector is driven by their potential to transform the service and hospitality, healthcare and nursing, and domestic sectors, offering improvements in efficiency and quality of life. For example, in the service and hospitality industry, robots can automate repetitive tasks, manage inventories, and enhance customer service, reducing labor costs and increasing operational efficiency. In the healthcare and nursing domains, robotic assistants can improve patient care, medication management, and routine check-ups, alleviating the burden on healthcare workers. Their ability to assist with daily living activities and support rehabilitation can become crucial for maintaining the *autonomy* of individuals with impairments. In domestic settings, personal robots can manage everyday chores, providing practical assistance and emotional support, particularly for those living alone.

Despite their potential, the deployment of robotic assistants is hindered by challenges related to their versatility, reliability, and adaptability in *unstructured environments*. Overcoming these challenges is essential to enhance *task generality* and reduce the dependency on *task-specific knowledge*. Versatility is crucial for robotic assistants in order to deal with the large variety of tasks in the personal sector, reducing the need for pre-specified task information. Reliability is fundamental for ensuring the safe and efficient operation of robots in the presence of *uncertainties* inherent to *unstructured environments*, especially in human-centric applications. Enhancing adaptability will allow robots to adjust to the changing circumstances and requirements for their operation, improving their ability to handle the *variability* of real-world applications.

Addressing these core capabilities will enable robots to realize their full potential in applications across the personal sector. Therefore, this thesis will investigate ways to increase the versatility of action *discovery*, improve the reliability of *selected* actions despite *uncertainty*, and enhance the adaptability of task *execution*.

Versatile *Discovery* of Interaction Possibilities

Flexible grasp synthesis approaches are proposed to enable interaction in *unstructured environments* with minimal *task-specific* information. A novel model

encoding multiple spatial features in an implicit neural field is introduced for objects with shared properties, termed similar objects, which improves geometric consistency and reconstruction from partial observations. By creating a detailed descriptor space for precise correspondences and accurate grasp transfer, it enhances pick-and-place tasks in unstructured settings. A grasp evaluation network further refines grasp poses, with evaluations in both simulated and real-world environments demonstrating the approach's efficacy. The model and the associated framework effectively handle variability and partial observations, increasing versatility in diverse and dynamic scenarios. Additionally, a novel approach for manipulating unknown objects based on local surface geometry is introduced, using point cloud analysis and heuristics to identify potential actions without accurate scene information. This method, designed for *unstructured environments*, operates efficiently by approximating local surface geometry and defining affordances like *graspability* using surface metrics. Actions are generated within a local coordinate system, providing a reliable basis for execution, which is demonstrated in multiple real-world grasping experiments.

Selection of Robust Action Candidates

Several probabilistic methods to enhance the reliability of mobile manipulation tasks are proposed, particularly for *grasping* unknown objects in *unstructured environments*, addressing the high *uncertainty* inherent in these tasks. First, a method using Bayesian state estimation is presented, combining an Unscented Kalman Filter with a Hidden Markov Model to estimate and update both pose and existence certainty of action hypotheses over multiple observations, thereby improving the *grasping* of unknown objects despite noisy sensor data. Experimental results show that this approach can enhance the grasp success rate in real-world *grasping* experiments. Secondly, a probabilistic method to optimize the grasp selection is proposed that evaluates grasp candidates based on Gaussian-distributed metrics and derives a ranking score to maximize the grasp success rate. Experiments involving over 1100 grasps on unknown objects demonstrate a significant increase in grasp success rates using this optimized selection method, highlighting the effectiveness of probabilistic approaches in improving the reliability in mobile manipulation tasks.

Adaptable Task *Execution*

The adaptability of mobile manipulation tasks to handle missing information or changing conditions is enhanced by proposing methods and representations that facilitate knowledge transfer between low-level sensory and motor information and

high-level symbolic representations. A task description and execution framework is integrated with a high-level planning framework. The task execution framework allows the flexible and autonomous generation and execution of manipulation actions, thus facilitating the transfer of skills across different tasks, environments, and robots. The framework's affordance-based representation supports collaborative learning and allows for accumulating and sharing mobile manipulation experiences, thereby improving adaptability. Real-world experiments, including uni- and bimanual *grasping*, placing, and memory-enabled skill transfer, validate the effectiveness of these methods. On the other hand, the planning system employs Large Language Models to interpret natural language commands and generate adaptable plans based on affordance-based scene representations, demonstrating the system's ability to handle diverse tasks in service and assistance scenarios. The real-world experiments demonstrate the framework's ability to generate successful plans despite missing information and execute complex tasks in *unstructured environments*.

Deutsche Zusammenfassung

Die rasche Verbreitung autonomer Roboter im privaten Sektor wird durch ihr Potenzial vorangetrieben, das Dienstleistungs- und Gaststättengewerbe, das Gesundheitswesen und die häusliche Pflege zu reformieren, indem sie die Effizienz steigern und die Servicequalität verbessern. Im Dienstleistungs- und Gaststättengewerbe beispielsweise können Roboter repetitive Tätigkeiten automatisieren, Lagerbestände verwalten und den Kundenservice verbessern, wodurch die Arbeitskosten gesenkt und die Betriebseffizienz gesteigert werden. Im Gesundheits- und Pflegesektor können Roboterassistenten die Patientenpflege, die Verabreichung von Medikamenten und Routineuntersuchungen verbessern und damit das Pflegepersonal entlasten. Ihre Fähigkeit, bei Aktivitäten des täglichen Lebens zu helfen und die Rehabilitation zu unterstützen, kann für die Erhaltung der Unabhängigkeit von Menschen mit Behinderungen entscheidend sein. Im häuslichen Umfeld können persönliche Roboter alltägliche Aufgaben übernehmen und insbesondere allein lebenden Menschen praktische Hilfe und emotionale Unterstützung bieten.

Trotz ihres Potenzials wird der Einsatz von Assistenzrobotern durch Herausforderungen im Zusammenhang mit ihrer Vielseitigkeit, Zuverlässigkeit und Anpassungsfähigkeit in unstrukturierten Umgebungen behindert. Die Bewältigung dieser Herausforderungen ist von entscheidender Bedeutung, um die Generalisierbarkeit der Anwendungen zu verbessern und die Abhängigkeit von aufgabenspezifischem Wissen zu verringern. Vielseitigkeit ist für Roboterassistenten von entscheidender Bedeutung, um die große Vielfalt von Aufgaben im persönlichen Bereich zu bewältigen und den Bedarf an vordefinierten Aufgabeninformationen zu verringern. Zuverlässigkeit ist angesichts der Unvorhersagbarkeit unstrukturierter Umgebungen von grundlegender Bedeutung für den sicheren und effizienten Betrieb von Robotern, insbesondere in Anwendungen, bei denen der Mensch im Mittelpunkt steht. Die Verbesserung der Anpassungsfähigkeit wird Roboter in die Lage versetzen, sich an die wechselnden Umstände und Anforderungen ihres Einsatzes anzupassen, und damit ihre Fähigkeit verbessern, mit der Variabilität realer Anwendungen umzugehen.

Die Berücksichtigung dieser Kernfähigkeiten wird es Robotern ermöglichen, ihr volles Potenzial in Anwendungen im gesamten persönlichen Bereich auszuschöp-

fen. In dieser Arbeit werden daher Möglichkeiten untersucht, die Vielseitigkeit der Handlungsfindung zu erhöhen, die Zuverlässigkeit durch Auswahl der besten Handlung trotz Unsicherheit zu verbessern und die Anpassungsfähigkeit der Aufgabenausführung zu steigern.

Vielseitige Erkennung von Interaktionsmöglichkeiten

Es werden flexible Ansätze zur Griffsynthese vorgeschlagen, um Interaktionen in unstrukturierten Umgebungen mit minimaler aufgabenspezifischer Information zu ermöglichen. Für Objekte mit gemeinsamen Eigenschaften, die als ähnliche Objekte bezeichnet werden, wird ein neuartiges Modell eingeführt, das mehrere räumliche Merkmale in einem impliziten neuronalen Feld kodiert, was die Konsistenz und die Rekonstruktion aus Teilbeobachtungen verbessert. Durch die Schaffung eines detaillierten Deskriptorraums für präzise Korrespondenzen und genaue Greiftransfers werden Pick-and-Place-Aufgaben in unstrukturierten Umgebungen verbessert. Ein Griffbewertungsnetzwerk verfeinert die Griffkandidaten weiter, wobei Evaluationen in simulierten und realen Umgebungen die Effektivität des Ansatzes belegen. Das Modell und das dazugehörige Framework gehen effektiv mit Variabilität und teilweisen Beobachtungen um, was die Vielseitigkeit in unterschiedlichen und dynamischen Szenarien erhöht. Darüber hinaus wird ein neuartiger Ansatz zur Manipulation unbekannter Objekte vorgestellt, der auf der lokalen Oberflächengeometrie, der Punktwolkenanalyse und Heuristiken zur Identifizierung potenzieller Aktionen ohne genaue Szeneninformationen basiert. Diese Methode, die für unstrukturierte Umgebungen entwickelt wurde, arbeitet effizient, indem sie die lokale Oberflächengeometrie approximiert und Affordanzen wie *Greifbarkeit* mit Hilfe von Oberflächenmetriken definiert. Die Aktionen werden in einem lokalen, abstrakten Koordinatensystem generiert, das eine zuverlässige Basis für die Ausführung bietet, was in mehreren realen Greifexperimenten demonstriert wird.

Auswahl robuster Handlungskandidaten

Zur Verbesserung der Zuverlässigkeit mobiler Manipulationen, insbesondere beim Greifen unbekannter Objekte in unstrukturierten Umgebungen, werden mehrere probabilistische Methoden vorgeschlagen, um der hohen Unsicherheit dieser Aufgaben zu begegnen. Zunächst wird eine Bayes'sche Zustandsschätzungsmethode vorgestellt, die einen Unscented Kalman Filter mit einem Hidden Markov Modell kombiniert, um sowohl die Pose als auch die Existenzwahrscheinlichkeit von Aktionshypothesen über mehrere Beobachtungen zu schätzen und zu aktualisieren, und so das Greifen unbekannter Objekte trotz verrauschter Sensordaten zu verbessern. Experimentelle Ergebnisse zeigen, dass dieser Ansatz die Erfolgsrate beim Greifen

in realen Experimenten erhöhen kann. Basierend darauf wird eine probabilistische Methode zur Optimierung der Greifauswahl vorgeschlagen, die Greifkandidaten mithilfe von Gauß-verteilten Metriken bewertet und daraus eine Wertung zur Maximierung der Greiferfolgsrate ableitet. In Experimenten mit mehr als 1100 Griffen an unbekannten Objekten konnte mit dieser optimierten Auswahlmethode eine signifikante Steigerung der Greiferfolgsrate nachgewiesen werden, was die Wirksamkeit probabilistischer Ansätze zur Verbesserung der Zuverlässigkeit mobiler Manipulationen unterstreicht.

Adaptive Ausführung von Aufgaben

Die Anpassungsfähigkeit mobiler Manipulationsfähigkeiten an fehlende Informationen oder sich ändernde Umstände wird verbessert, indem Methoden und Darstellungen vorgeschlagen werden, die den Wissenstransfer zwischen sensorischen und motorischen Informationen auf niedriger Ebene und symbolischen Darstellungen auf hoher Ebene verbessern. Ein Verfahren für die Aufgabenbeschreibung und -ausführung wird mit einen Planungssystem kombiniert. Das Framework für die Aufgabenausführung ermöglicht die flexible und autonome Generierung und Ausführung von Manipulationsaktionen und erleichtert die Übertragung von Fähigkeiten auf verschiedene Aufgaben, Umgebungen und Roboter. Die affordanzbasierte Repräsentation des Frameworks unterstützt kollaboratives Lernen und ermöglicht das Sammeln und Teilen von Manipulationserfahrungen, wodurch die Anpassungsfähigkeit verbessert wird. Reale Experimente, einschließlich ein- und zweihändigem Greifen, Platzieren und dem Transfer von erlernten Fähigkeiten über ein kognitives Roboter-Gedächtnis, bestätigen die Effektivität dieser Methoden. Auf der Planungsseite verwendet das entwickelte Verfahren Large Language Models, um natürlichsprachliche Befehle zu interpretieren und adaptive Pläne zu generieren, die auf affordanzbasierten Szenenrepräsentationen aufbauen. Die realen Experimente zeigen die Fähigkeit des Systems, trotz fehlender Informationen und komplexer Aufgaben in unstrukturierten Umgebungen erfolgreiche Pläne zu generieren.

Contents

1. Introduction

The rise of robotic assistants is driven by their potential to revolutionize the tertiary sector, including applications in the service and hospitality, healthcare and nursing, and domestic domains. These robots promise to enhance efficiency, improve quality of life, and provide essential support across various domains. The development of more autonomous and intelligent systems, which are a prerequisite for the effective deployment of robotic assistants in *unstructured environments*, has advanced rapidly in the last decades. Service robots, like robot wheelchairs, surveillance drones, therapy robots, and entertainment robots, are increasingly being integrated into diverse sectors, offering significant advancements in automation and user interaction (Lee, 2021). The global market for service robots was predicted to grow 7 times as fast as that of industrial robots by 2022 due to the rapid adoption of robotics in the service and hospitality industry (Xiao and Kumar, 2021). In healthcare and nursing, robotic assistants are becoming more important for personal care, as they assist with activities of daily living and support individuals with impairments (Bilyea et al., 2017). Furthermore, personal service robots can help balance the impact of the demographic change and support the aging population in their own homes (Fischinger et al., 2016).

In the service and hospitality industries, robotic assistants have the potential to automate repetitive tasks, manage inventories, and significantly improve customer service (Murphy et al., 2017). Tuomi et al. (2021) emphasize the transformative impact of service robots in the hospitality industry, which can lead to reduced labor costs, increased operational efficiency, and improved brand differentiation. Similarly, robots in education can enhance student engagement, foster interdisciplinary learning, and develop essential problem-solving and teamwork skills, ultimately preparing students more effectively for future technological challenges (Miller and Nourbakhsh, 2016). However, the customer's perception of the quality of service and acceptance of robots depends to a large degree on the assurance (i. e., ability to perform tasks with expertise, politeness, and trust) and reliability of the service robot (Chiang and Trimi, 2020).

Robotic assistants hold significant potential for enhancing various aspects of the healthcare and nursing industry, including patient care, medication management, and routine check-ups. Furthermore, they can alleviate the burdens on healthcare workers by automating repetitive tasks. Robotic dispensing systems, for example, have improved patient safety, inventory management, and staff satisfaction in outpatient hospital pharmacies by minimizing medication dispensing errors (Rodriguez-Gonzalez et al., 2019). In nursing care, potential tasks for robotic assistants include delivering medication, processing patient data, aiding with daily living activities (Ohneberg et al., 2023), and supporting rehabilitation and cognitive therapy (Feil-Seifer and Matarić, 2009). Robots have proven crucial in helping sustain *autonomy* in users by addressing age-related challenges such as cognitive, motor, and perceptual declines (Bilyea et al., 2017; Smarr et al., 2014). However, adaptability is essential to comply with personal preferences (Martinez-Martin and del Pobil, 2018) and the individual needs of care recipients, such as mobility and communication capabilities (Pineau et al., 2003).

In domestic environments, personal robots have the potential to handle everyday chores such as cleaning, cooking, and organizing, thereby freeing up time for individuals and families (Young et al., 2009). These robots can also provide companionship and support for those living alone, addressing both practical and emotional needs (Feil-Seifer and Matarić, 2009; Fischinger et al., 2016). The acceptance of social robots in domestic settings is influenced by factors such as versatility, usability, and perceived usefulness. Studies have shown that older adults prefer a robot's help over that of a human for chores, manipulating objects, and information management (De Graaf et al., 2019). However, current applications in domestic environments include floor, pool, and window cleaning robots, lawnmowers, ironing robots, intelligent refrigerators, and digital wardrobes (Prassler et al., 2016), disregarding the social and manipulation aspects almost entirely.

Even though robots are already able to manipulate and grasp objects under certain conditions and in simple situations, contact-rich or even bimanual manipulation in cluttered environments is still a major challenge in robotics (Billard and Kragic, 2019). Existing systems often struggle with adapting to new and varied environments, requiring significant human intervention, or lack the reliability needed for safe operation in unstructured settings. Brock et al. (2016) identify *high dimensionality*, *uncertainty*, and task *variability* as the main obstacles in mobile manipulation research. These must be thoroughly addressed to increase *task generality* and minimize the dependency on *task-specific knowledge*. On an architectural level, robotic systems need to manage *uncertainty*, handle various temporal scopes, and integrate high-level planning with low-level *uncertainty* to enhance the *autonomy* of robotic assistants (Kortenkamp et al., 2016). In order to cope with the inherent *variability* robots encounter in real-world applications, the transfer of knowledge and skills across tasks, environments, and robots is a major opportunity for adapting to novel situations (Jaquier et al., 2024).

To address the challenges in real-world applications and truly benefit humans, robotic assistants must achieve a higher degree of *autonomy*. This *autonomy* enables them to (inter-)act independently without constant human supervision, which is essential for their effective operation. Robots must be capable of analyzing diverse scenarios and perceiving their environment to make optimal decisions for each situation. Increasing *autonomy* ensures that robots can safely and effectively coexist with humans, providing assistance even when users are unable to communicate efficiently (e. g., in the case of disabilities) or assess potential risks accurately (e. g., due to lack of expertise, Van der Loos et al., 2016). In addition, the degree of *autonomy* significantly impacts Human-Robot Interaction (HRI), influencing both the quantity of interaction required to complete tasks and the nature of interactions between humans and robots. Higher *autonomy* allows robots to work unsupervised for extended periods and facilitates more sophisticated HRI (Beer et al., 2014). Robots must infer appropriate actions independently to handle complex, dynamic, and partially known environments. This ability to reason about actions, intended effects, and unintended side effects is crucial for autonomous decision-making in real-world applications (Beetz et al., 2016).

Despite their potential benefits, the widespread deployment of robotic assistants is hindered by several obstacles, as outlined in Brock et al. (2016). This thesis addresses three primary challenges to enhance the operation of robots in *unstructured environments*, increase *task generality*, and minimize dependency on *task-specific knowledge*. These challenges primarily revolve around promoting the *autonomy* of robots in *unstructured environments* by improving their versatility, reliability, and adaptability. Versatility is essential for robots to perform effectively in diverse and dynamic environments and a variety of tasks, therefore being able to handle the *high dimensionality* of information in the real world (e. g., Sawyer et al., 2021; Young et al., 2009). Reliability ensures consistent and safe performance under *uncertainty* in perception and scene understanding, crucial for building trust and fostering acceptance of robotic assistants (e. g., Zhang et al., 2022). Meanwhile, adaptability enables robots to leverage their knowledge, experience, and acquired skills to adjust to various situations and circumstances, helping them cope with the broad *variability* of objects, tasks and environments (e. g., Wirtz et al., 2018).

Versatility is a cornerstone for the effectiveness of robotic assistants in diverse and dynamic environments, enabling them to handle various tasks and changing surroundings while reducing the need for pre-specified task information. Assistive systems must be capable of managing a wide range of tasks for individuals with diverse conditions, irrespective of specific needs or environments (Chen et al., 2013). Example applications with such requirements include supporting caregivers and individuals in need of care in nursing scenarios (Ohneberg et al., 2023), assisting individuals with upper limb impairments in performing daily living activities (Bilyea et al., 2017) and handling a range of tasks, from providing emotional support to assisting with daily activities, in elderly care (Martinez-Martin and del Pobil, 2018). Versatility not only enhances performance but could provide a strategic advantage in the future marketing of domestic robots, as highlighted by Young et al. (2009). The ability to cope with all context variations, such as different handling requirements for objects based on their state and environment, is essential for maintaining effectiveness across various applications (Beetz et al., 2016). Moreover, increased versatility enhances the *autonomy* of robots by allowing them to work with imperfect knowledge about the task and their environment and maintain performance without constant human intervention. This facilitates the generalization of robotic systems to different tasks and supports human operators or supervisors in maintaining control over these systems in dynamic and unforeseen situations (Sawyer et al., 2021).

Reliability is fundamental to ensuring the safety and efficiency of robotic assistants in *unstructured environments*. In human-centric applications, robustness to unknown disturbances is of great importance, as unmodeled disruptions can lead to catastrophic failures. Resilient systems that can adapt and reorganize in response to unknown disturbances are crucial for maintaining consistent and safe performance under *uncertainty* (Prorok et al., 2021). Trustworthy robots capable of dependable and safe interaction with humans are essential for reducing errors in hospital pharmacies (Rodriguez-Gonzalez et al., 2019), assisting individuals with disabilities (Van der Loos et al., 2016), the elderly in nursing homes (Pineau et al., 2003) or at their own homes (Fischinger et al., 2016), and maintaining customer trust and satisfaction in the hospitality industry (Chiang and Trimi, 2020). To handle real-world variations such as diverse materials, lighting, and clutter, shared *autonomy* can serve as an intermediate step, directly impacting the reliability of robots (Chen et al., 2013). Integrating modalities like segmentation, object recognition, and collision-free motion planning into *grasping* and manipulation can increase robustness in *unstructured environments* (de Jong et al., 2018). However, error correction during execution might become necessary to address the *uncertainties* introduced through visual perception and proprioception (Ciocarlie et al., 2014). Ultimately, by enhancing reliability, the *autonomy* of robots is promoted, enabling them to operate more independently and effectively in unstructured real-world applications.

Adaptability is a critical challenge for robotic assistants, as they must learn and operate effectively in diverse environments and tasks. Adjusting to new tasks or unexpected situations while maintaining predictable behavior is of great importance for multi-purpose robotic assistants in scenarios with changing conditions and incomplete knowledge (Hawes et al., 2017). For instance, a robot used in domestic settings might need to adapt its skills after moving to a different house, where the environment and tasks differ but still relate closely to its previous experience. In an evolving world, transferring knowledge across different tasks and environments is crucial for coping with the *variability* of objects and scenarios encountered in realistic applications. Unlike special-purpose robots (e. g., vacuuming and lawn-mowing robots), general-purpose robotic assistants must handle multiple tasks in any environment. While current robots can adapt to small variations in object properties during manipulation, these adaptations are often limited to a number of expected changes (Billard and Kragic, 2019). Therefore, leveraging experience from one task to improve performance in another and generalizing tasks across different environments will be necessary for truly versatile robotic assistants (Jaquier et al., 2024). Moreover, adaptability is essential for robots to infer appropriate actions in partially unknown environments or tasks that are too complex for defining the correct behavior in all cases, as they need to reason about constraints and context-specific variations based on their previous experience (Beetz et al., 2016). In socially assistive robots, particularly in the context of elderly care, where robots must dynamically adjust their roles to align with the evolving needs and preferences of users (Huber et al., 2014), adaptability ensures that these robots can provide personalized and contextually appropriate support, thereby enhancing user satisfaction and promoting long-term acceptance. Therefore, robotic assistants should be able to transfer and adapt their knowledge and skills to meet the particular requirements and constraints of each user and their environment (De Graaf et al., 2019; Sawyer et al., 2021). By developing comprehensive representations of tasks and leveraging symbolic reasoning, robots can effectively encode both low-level motions and high-level action sequences, enhancing their ability to adapt skills across different environments (Billard et al., 2016). Thus, increasing adaptability is not only tightly coupled with the versatility of robots but also promotes their *autonomy* by enabling them to independently handle a broader spectrum of tasks and scenarios in real-world applications.

Improving the versatility, adaptability, and reliability of mobile manipulation skills is crucial for the advancement of robotic assistants. Addressing these core capabilities will enable robots' safe and effective integration into real-world applications, offering significant benefits across various sectors. The transformative potential of assistive robots in these areas motivates overcoming ethical and societal challenges to ensure their beneficial integration (Wirtz et al., 2018). Therefore, research in mobile manipulation has tried to improve the robustness, *task generality*, and *autonomy* in *unstructured environments* while decreasing the amount of *task-specific or hardcoded knowledge* necessary for these scenarios (Brock et al., 2016).

1.1. Problem Statement

This thesis aims to improve general-purpose robots' mobile manipulation capabilities in the context of *service* and *assistance* tasks, as commonly found in applications in the personal sector (e. g., domestic, healthcare and nursing, and service and hospitality domains). Therefore, it follows along the lines of research in this topic and has the objective of maximizing the *task generality* of autonomous systems while minimizing the amount of *task-specific knowledge* required, as pointed out by Brock et al. (2016). Specifically, this thesis has the goal of increasing the *autonomy* of robots in *unstructured environments* so that they can work in human-centric environments without constant supervision. Therefore, the main objective of the thesis is formulated as follows:

> **Main Objective**
>
> Increase the *autonomy* of robotic assistants by improving the *task generality* and decreasing the amount of *task-specific knowledge* required in order to deploy them to real-world, human-centric applications in the personal sector.

As the objects involved in realistic applications in these domains are often not precisely known, robots should not rely too much on the availability of explicit object models. Gibson (1966, 1979) introduced the concept of affordances[1], which represent interaction possibilities that the environment offers to an agent. Applying this concept from cognitive psychology to mobile manipulation offers the opportunity to think of the environment and objects in terms of what can be *done* with them instead of what specific *kind* of object it is. Şahin et al. (2007) introduce the *representationalist* formalization of affordances as the tuple $(\mathit{effect}, (\mathit{entity}, \mathit{behavior}))$, which states that an affordance is an *effect* generated by applying a *behavior* to an *entity* (e. g., an object). This representation has the advantage of separating the agent-specific (i. e., *behavior*) and agent-agnostic (*entity*) parts of mobile manipulation tasks. In doing so, it makes the concept of affordances very well mappable to the three stages of a *discriminative* approach to mobile manipulation: (i) *discovery*

[1] For a more in-depth account of affordances, see Appendix A.

of potential actions, where the environment is scanned for interaction possibilities (i. e., *entities*) (ii) *selection* of the best action, where the robot decides on the most promising *behavior* to execute, and (iii) *execution* of the action in order to produce the desired *effect*.

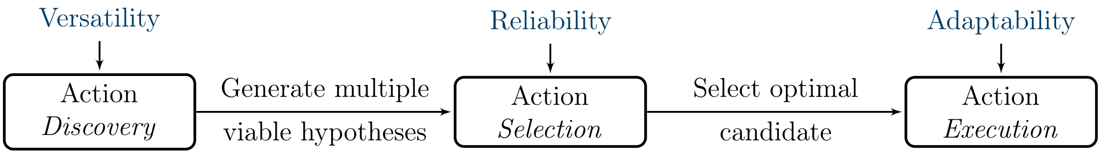


Figure 1.1.: Overview of the core capabilities that this thesis addresses with respect to the *discriminative* approach to mobile manipulation.

Therefore, to contribute to the main objective, this thesis focuses on improving the three stages of *discriminative* manipulation by targeting a specific *core capability* relevant to each stage in order to enhance the *autonomy* of robotic assistants. These core capabilities, visualized in Figure 1.1, aim at overcoming the task *variability* and *uncertainty* that robots have to deal with in *unstructured environments*.

1.1.1. Core Capabilities

The core capabilities of robotic assistants that are in the center of interest for this thesis, already shortly introduced in the beginning of Chapter 1, are versatility, reliability, and adaptability. By improving aspects related to these core capabilities, this thesis aims at enhancing the *autonomy* of robotic assistants in *unstructured environments*.

Core Capability 1: Versatility

Versatility refers to the ability of an individual, system, or object to perform a wide range of tasks or functions. It implies having multiple skills, capabilities, or uses, making one competent in various situations. Versatility is about breadth – being able to handle many different scenarios effectively. Versatility is essential for robotic assistants as they must operate effectively in diverse and dynamic environments and solve a multitude of tasks. These robots need to be able to handle new assignments, various objects, missing prior information, and changing surroundings without extensive reprogramming to remain useful and efficient. The goal is to have systems that can seamlessly transition between different tasks and environments and maintain high performance levels.

Core Capability 2: Reliability

Reliability refers to the consistency and dependability of a person, system, or process to perform its intended function over time without failure. A reliable entity consistently produces the expected outcomes under various conditions and is trusted to work correctly whenever needed. Reliability is critical for ensuring the safe and robust operation of robotic assistants. They must cope with *uncertainties* in their environment, visual perception, and available information to avoid accidents and maintain trustworthiness. Achieving robust and dependable performance is vital for gaining user confidence and ensuring that robots can consistently support various tasks without failure and without the need for constant supervision from human operators.

Core Capability 3: Adaptability

Adaptability is the ability to adjust to new conditions, environments, or situations. It refers to how well someone or something can change or be changed in response to new challenges or changing circumstances. Adaptability is about change – how effectively one can modify behavior or function in response to evolving external factors. Adaptability involves the ability of robotic assistants to apply learned skills across different tasks and environments. Often, this requires robots to leverage and transfer experience from one context to enhance their performance in another, ensuring versatility and reliability in their service delivery. The goal is to have systems that can generalize learned behaviors and adapt to a wide range of applications, thus maximizing their utility and effectiveness.

1.1.2. Research Questions

From the core capabilities, defined in Section 1.1.1, this thesis derives three concrete research questions that are tackled in order to contribute to the main objective.

Research Question 1: How Can Robotic Assistants Extract Versatile Manipulation Actions from the Visual Perception of Unstructured Environments?

The ability of robotic assistants to handle diverse and dynamic environments is crucial for their effective deployment in real-world applications. This research question addresses the task of enabling robots to perform flexible grasping and manipulation tasks in unstructured settings, which is essential for their *autonomy*. More specifically, by facilitating the *discovery* of hypotheses for mobile manipulation actions in *unstructured environments* with minimal pre-specified information, robots can become more flexible and perform a wide range of tasks and objects. Therefore, this thesis leverages machine learning and computer vision approaches to extract information from the visual perception of the robot in order to find high-quality interaction possibilities in the environment. These advancements ensure that robotic assistants can operate efficiently in *unstructured environments* and handle the large *variability* encountered in these scenarios, ultimately improving their practical applicability and *autonomy*.

Research Question 2: How Can Probabilistic Approaches Enhance the Reliability of Mobile Manipulation?

The reliability of mobile manipulation skills is crucial for the effective deployment of robotic assistants in real-world applications. While robots have become very adept at grasping and manipulation in static and known environments, it remains a difficult problem if the visual inputs deviate from the anticipated or *uncertainty* is present (Vincze et al., 2020). This research question addresses the challenge of ensuring that manipulation actions are robust to noise and *uncertainties* in their perception and available information. By employing probabilistic methods, the system can make informed decisions in *unstructured environments*, increasing the robustness of mobile manipulation actions. Taking the various *uncertainties* that arise into account, this thesis investigates methods for the optimization of the *selection* process with regard to the success rate of the selected actions, thereby making them more robust against external influences. This enhances the reliability of robotic manipulation, making robotic assistants more dependable and effective in diverse and dynamic settings and increases their *autonomy*.

Research Question 3: How Can the Execution of Mobile Manipulation Skills be Adapted to Different Conditions and Situations?

Developing adaptable mobile manipulation skills is crucial for increasing the *autonomy* of general-purpose robots in real-world applications. This research question addresses the challenge of enabling robots to apply their manipulation skills across different situations and adjust their *execution* to the current conditions. By focusing on an adaptable *execution* of mobile manipulation tasks, the dissertation aims to enhance the performance, operational capability, and practical applicability of robotic assistants. This involves creating systems that can generalize learned skills to new scenarios and adapt to missing information. Addressing this question is essential for ensuring that robotic assistants can operate effectively in diverse and dynamic environments, adapt to new tasks and surroundings, and learn from previous experience or even other robots. Through increasing the adaptability of task *execution*, robots can retain their *autonomy* despite changing circumstances and evolving requirements for their operation in *service* and *assistance* scenarios.

1.2. Contributions

By addressing the research questions from Section 1.1.2, this thesis aims to make significant contributions to the *autonomy* of robotic assistants in the personal sector. In the following, the concrete contributions will be introduced and associated with several scientific publications in robotic journals and conferences that originated in the course of the thesis. An overview of the relevant publications of this thesis in the Contributor Role Taxonomy (CRediT)[2] can be seen in Table 1.1.

Table 1.1.: CRediT roles for publications of this thesis.

CRediT Role	1	2	3	4	5	6	7	8	9
Conceptualization	●	●	●	●	●	●	●	◐	○
Data Curation	●	◐	●	○	●	○	●	○	○
Formal Analysis	●	○	●	○	○	○	●	○	○
Funding Acquisition	○	○	○	○	○	◐	◐	◐	○
Investigation	●	●	●	●	●	●	●	●	●
Methodology	●	●	●	●	●	●	●	◐	○
Project Administration	●	●	●	●	●	●	●	●	○
Resources	○	○	○	○	○	○	○	○	○
Software	●	○	●	◐	●	○	●	○	●
Supervision	○	●	○	●	○	●	○	●	○
Validation	●	●	●	◐	●	◐	●	◐	○
Visualization	●	◐	●	●	●	○	●	○	●
Writing - Original Draft	●	●	●	●	●	●	●	●	●
Writing - Review & Editing	●	●	●	●	●	◐	●	●	●

1: Pohl et al., 2020 2: Birr et al., 2022 3: Pohl and Asfour, 2022
4: Baek et al., 2022 5: Pohl et al., 2022 6: Birr et al., 2024
7: Pohl et al., 2024 8: Cai et al., 2024 9: Suarez et al., 2024

[2] `https://credit.niso.org/`

Contribution 1: Versatile Action Discovery in Unstructured Environments using Visual Perception

In order to address Research Question 1, approaches for the versatile *discovery* of interaction possibilities in *unstructured environments* using visual perception are proposed. This is crucial to be able to operate within real-world scenarios with a minimal amount of *task-specific knowledge* required. Versatile action hypotheses, which do not require full knowledge about the objects under consideration, are necessary to be able to work in dynamic and cluttered environments with incomplete information about the scene available. For example, in domestic applications robotic assistants should be able to handle daily tasks in the kitchen independent of which concrete bottle of milk was bought or which specific cup is available for *pouring* the coffee.

For objects sharing some properties with other objects of the same class, so-called similar objects, the Multi-feature Implicit Model (MIMO, Cai et al., 2024) is introduced, which is an object representation model that encodes multiple spatial features between a point and an object in an implicit neural field. MIMO's capability to predict dense correspondences across slight variations in geometry enables robots to learn from visual demonstrations and apply these learned behaviors in dynamic, unpredictable settings. The improved consistency of the geometric features of partially observed objects enhances MIMO's shape reconstruction accuracy and enables more accurate transfer of poses across instances of the same class. Additionally, a grasp evaluation network is introduced that predicts the probability of success of grasp poses, refining them if necessary. Extensive real-world experiments demonstrate the efficacy of MIMO and the proposed grasping framework in one-shot and few-shot visual imitation learning of manipulation tasks.

As the amount of information about objects is very limited in real-world applications, the main focus of this contribution is the manipulation of objects without any prior knowledge available, so-called unknown objects. To this end, an approach based on the local surface geometry of the environment is presented. By analyzing the local surface structure of a point cloud and using heuristics on the averaged local surface information, the Geometry-based Action Extraction (GAE, Pohl and Asfour, 2022) identifies potential manipulation actions without relying on precise information about the scene, which is often noisy and inaccurate in real-world applications. The method is specifically designed to operate efficiently in cluttered and *unstructured environments* by not depending on any object- or *task-specific knowledge*. Instead, it approximates the local surface of a raw point cloud using quadrics, which allows the calculation of properties of the local surface geometry. By using surface metrics such as curvature and normal direction, different affordances like *graspability*, *pushability*, and *placability* are heuristically defined. Actions are then generated in a uniquely defined local coordinate system called the Abstract Affordance Frame. This frame uses the averaged normal and minimal curvature direction of supervoxels to create a coherent and reliable basis for action execution. The approach is evaluated through extensive experiments with over 900 grasp executions using the humanoid robot ARMAR-6 (Asfour et al., 2019) in various unstructured scenes. The experimental results demonstrate a significant improvement in grasp success rates (almost 10% compared to a baseline approach using Object-Oriented Bounding Boxes) and robustness in handling scenes with varying degrees of clutter. This shows that the use of local surface geometry for defining affordances significantly improves the robot's ability to perform versatile manipulation tasks in *unstructured environments*.

Contribution 2: Probabilistic Methods for Reliable Grasping of Unknown Objects

Employing probabilistic methods to enhance the reliability of mobile manipulation, particularly in the context of the *selection* of the most promising action to execute, can significantly increase robots' ability to cope with the high *uncertainty* present in *unstructured environments*. In order to address Research Question 2, two contributions to improve the robustness of *grasping* unknown objects using probabilistic methods are proposed by this thesis: (i) the spatiotemporal fusion or filtering of grasp candidates to reduce *uncertainty* in the perception of the robot, and (ii) the probabilistic treatment of the grasp *selection* process to account for *uncertainties* in the scene representation and understanding.

In order to address the first aspect, the Probabilistic Action Extraction and Fusion (PAEF, Pohl and Asfour, 2022), a robust method for improving the reliability of *grasping* unknown objects by employing Bayesian state estimation to handle the *uncertainties* in action hypotheses, is presented. Specifically, the system combines an Unscented Kalman Filter with a Hidden Markov Model to estimate both the pose and the existence certainty of action hypotheses over multiple observations. This probabilistic approach allows the system to refine its understanding and improve the reliability of the actions despite noisy and uncertain sensor data. Each action hypothesis is represented as a pose and its associated *uncertainty*, which are updated recursively as new observations are made. To associate new observations with existing hypotheses, the method uses Gaussian models for the position and orientation, facilitating the computation of correspondence likelihoods. Once a correspondence is found, the system updates the hypothesis using the

Unscented Kalman Filter, which adjusts both the pose and the *uncertainty*. The Hidden Markov Model further refines the hypothesis by determining the likelihood of the action's existence, enhancing robustness against spurious detections and disappearing objects. In the experiments with ARMAR-6, PAEF could improve the grasp success rate by another 5% over GAE, showing that accounting for *uncertainties* in perception can indeed improve the robustness and quality of *grasping* and manipulation.

To address the second aspect, the Uncertainty-Aware Sensitivity Optimization (UASO, Baek et al., 2022), a probabilistic approach to optimize the *selection* of grasp candidates, is introduced. Each grasp candidate is evaluated based on specific metrics that are treated probabilistically. Different Gaussian-distributed metrics are used to characterize each grasp and account for *uncertainties*. A scalar ranking score is derived to rate each grasp candidate. This score is calculated based on the sensitivities of the metrics towards grasp success and is aimed at maximizing the grasp success rate. The ranking score incorporates both global weighting (influence of each metric on success) and local weighting (likelihood of success for each candidate). The experiments involve over 1100 grasp attempts on unknown objects using the humanoid robot ARMAR-6. These grasps were executed under real-world conditions to provide a robust dataset for optimization and validation. The evaluation shows a significant improvement in grasp success rates using the proposed UASO method. Out of 932 randomly performed grasps, only 32.6% were successful. However, using the optimized grasp selection method based on the ranking score, the success rate increases to 73.8% for 187 additional grasps. This demonstrates the effectiveness of the probabilistic approach in enhancing the reliability of mobile manipulation.

Contribution 3: Methods and Representations for Adapting the Execution of Mobile Manipulation Skills

In an effort to increase the adaptability of manipulation skills, this thesis proposes methods and representations for adjusting the *execution* of tasks to varying conditions and evolving requirements. Vernon and Vincze (2017) present a list of 11 fundamental capabilities that a cognitive robot should have. "*Adaptive planning*" and "*High-level instruction and context-aware task execution*" are two of these capabilities, which need to be improved for robotic assistants to be useful in real-world applications. Focusing on these two capabilities, and thereby addressing Research Question 3, a task description and execution framework is combined with a high-level planning framework to increase the adaptability of mobile manipulation skills.

The Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills (MAkEable, Pohl et al., 2024) integrates an affordance-based task description into the memory-centric cognitive architecture[3] of the ARMAR humanoid robot family. This supports the *transfer* of knowledge and experience across different tasks, environments, and robots. By representing mobile manipulation actions through affordances, the framework provides a unifying structure for the autonomous uni- and multi-manual manipulation of known and unknown objects in various environments. Incorporating this representation into a universal task description fosters collaborative learning among robots. This description is flexible enough to adapt to various tasks and scenarios, facilitating the autonomous and semi-autonomous generation and execution of manipulation actions. The integration into a memory-centric cognitive architecture allows for the accumulation and sharing of rich repositories of mobile manipulation knowledge. This supports learning from previous experiences and enables the practical *transfer* of skills among different robots and tasks. The improved adaptability is demonstrated through real-world experiments involving *grasping* known and unknown objects, object placing, bimanual object *grasping*, and memory-enabled skill transfer, such as a drawer opening scenario across different robots (ARMAR-6 and ARMAR-DE).

Furthermore, the planning system AutoGPT+P (Birr et al., 2024) is introduced, which utilizes Large Language Models (LLMs) to select tools that support generating a plan to accomplish tasks based on the affordance-based scene representation. The integration of affordances additionally facilitates the generalization of the planning domain and even allows handling missing objects, thereby relaxing the closed-world assumption of conventional planners. In doing so, AutoGPT+P complements the task execution capabilities of MAkEable on a semantically high abstraction level by improving and adapting the action sequences that MAkEable can execute. Therefore, the combination of these two systems presents the unique opportunity to increase the adaptability of the *execution* of mobile manipulation tasks from the formulation of plans to their execution. This was demonstrated in multiple experiments conducted on the humanoid robots ARMAR-6 and ARMAR-DE, which validate the system's feasibility in real-world scenarios. AutoGPT+P's ability to perform tasks such as *picking*, *placing*, *handover*, *pouring*, and *wiping*, even with missing objects, showcases its potential for practical applications in various sectors like healthcare and nursing, domestic settings, and the service and hospitality industries.

[3] Peller-Konrad et al., 2023.

1.3. Structure of Thesis

In the pursuit of the main objective, this thesis is structured as follows. In Chapter 2, the state of the art with a focus on *discriminative grasping* and software frameworks for mobile manipulation is introduced. In the *grasping* specific part, relevant works for the *discovery* of grasp candidates for unknown objects and similar objects in Section 2.1, as well as grasp *selection* based on different grasp quality measures in Section 2.2, are listed. Subsequently, robotic software frameworks for the *execution* of manipulation tasks are investigated in Section 2.3.

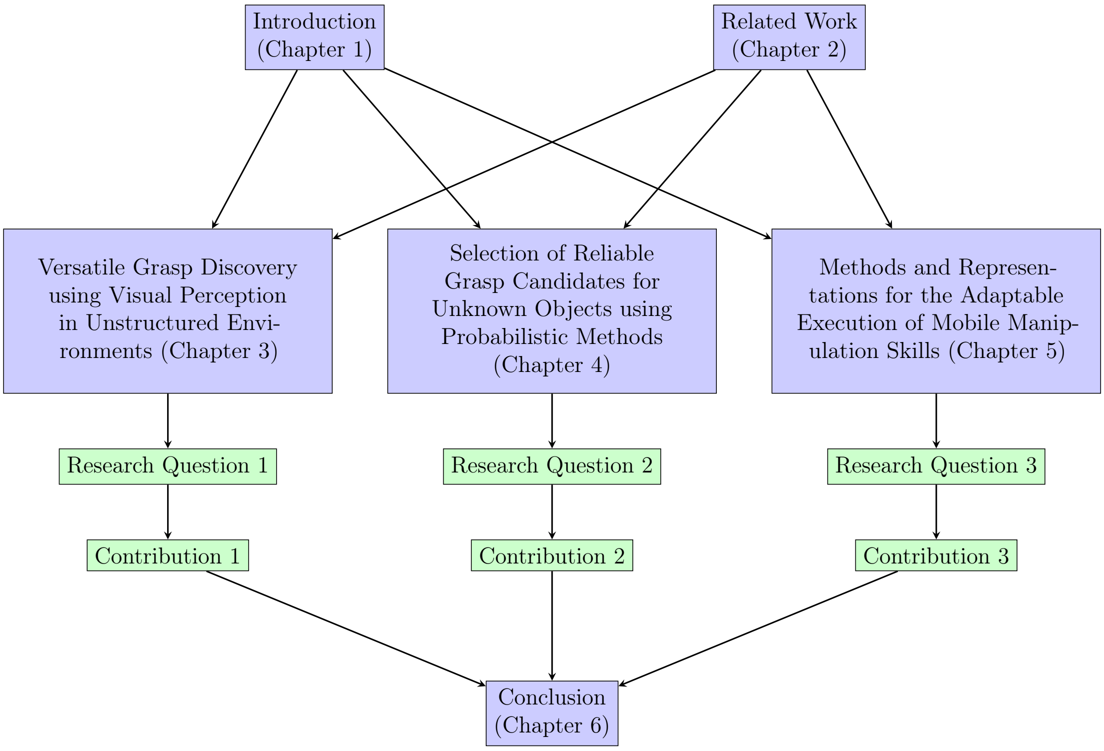

Introduction (Chapter 1)
Related Work (Chapter 2)
Versatile Grasp Discovery using Visual Perception in Unstructured Environments (Chapter 3)
Selection of Reliable Grasp Candidates for Unknown Objects using Probabilistic Methods (Chapter 4)
Methods and Representations for the Adaptable Execution of Mobile Manipulation Skills (Chapter 5)
Research Question 1
Research Question 2
Research Question 3
Contribution 1
Contribution 2
Contribution 3
Conclusion (Chapter 6)

Figure 1.2.: Structure of the thesis.

Chapter 3 focuses on answering Research Question 1 and introducing the works relevant for Contribution 1. To this end, it investigates methods for the *discovery* of potential interaction possibilities in *unstructured environments* using visual perception methods. To minimize the amount of *task-specific knowledge* required for autonomous manipulation, the chapter proposes methods that can extract grasping and manipulation hypotheses with little to no prior knowledge about the objects involved. Therefore, Section 3.1 introduces MIMO and its framework for the task-oriented grasping and rearrangement of similar objects. To further decrease the amount of object information required, Section 3.2 introduces GAE that extracts action hypotheses for multiple different affordances using only the local surface structure, therefore completely decoupling action generation from the concept of objects.

Chapter 4 introduces the research relevant for Contribution 2 that tries to answer Research Question 2. Using the surface-based action hypotheses that were extracted using GAE, Section 4.1 introduces PAEF in an attempt to increase the reliability of mobile manipulation in *unstructured environments*. Through the spatiotemporal fusion of the poses of the related Abstract Affordance Frame, a covariance as well as an existence certainty of this geometrically-inspired, coherent frame are calculated. This is used in Section 4.2 to calculate an optimized grasp score using UASO. This grasp score can be used to account for the various *uncertainties* involved in the grasping process in order to find the most robust and likely-to-succeed action in the current scene.

Chapter 5 is concerned with answering Research Question 3 using Contribution 3. The main focus of this chapter is the increase of adaptability in mobile manipulation skills that stems from the combination of a task description and execution framework with a high-level planning system. To this end, Section 5.1 introduces MAkEable and its general task description. This facilitates the transfer of knowledge and experience for the *execution* of skills across varying requirements and situations that can arise in real-world applications of general-purpose robots. Subsequently, Section 5.2 demonstrates how the flexible planning system AutoGPT+P can be combined with MAkEable to facilitate and adapt the generation and execution of plans under missing information in realistic scenarios.

At the end, in Chapter 6, a short summary and conclusion of the thesis will be given with an outlook to future research directions of interest.

2. Related Work

This thesis aims to make progress towards the main objective of increasing the autonomy of robotic assistants in *unstructured environments* – as frequently encountered in the domestic, healthcare and nursing, and service and hospitality industries – by answering Research Questions 1 to 3. To this end, the Contributions 1 to 3 are proposed, which focus on improving the *task generality* of mobile manipulation tasks while decreasing the *task-specific knowledge* necessary (see Brock et al., 2016). In the context of this thesis, these contributions are based on a *discriminative* approach to manipulation under *variability* and *uncertainty* in *unstructured environments*. Furthermore, the evaluation of the contributions was performed considering the *graspability* affordance in most cases. Therefore, the first part of this chapter will discuss only approaches relevant to *discriminative grasping*. The chapter has a similar structure as Figure 1.1 indicates: The first part of related work is concerned with the *discovery* of potential grasp opportunities (Section 2.1), while the second part discusses the *selection* of suitable grasp candidates (Section 2.2). Finally, software architectures for robotic systems with a special focus on the *execution* of mobile manipulation tasks are reviewed in Section 2.3.

2.1. Grasp Synthesis

This section delves into robotic grasping, focusing specifically on the *discovery* of grasp hypotheses in *unstructured environments*. It highlights the different approaches commonly found in the literature regarding the amount of information available for the object under consideration. Generally, objects can be categorized into three distinct classes, depending on how much about the object's properties and features is known (Bohg et al., 2014):

(a) For known objects, full knowledge about the geometry and the object's properties is available, including object meshes, textures, or visual features.

(b) For similar objects, full knowledge about the object's class is available, and the object's geometry does not vary much compared to other instances of the same class.

(c) For unknown objects, no prior information about the object is available.

To analyze and subsequently categorize the works for grasp synthesis, this thesis uses an extended version of the scheme from Newbury et al. (2023). There, four categories of approaches – *sampling*, *direct regression*, *reinforcement learning*, and *exemplar* methods – for deep learning-based grasp synthesis, as well as three types of scenes – *singulated*, *structured* and *piled* clutter – were introduced. To extend this to classical approaches (i. e., not employing deep learning), a fifth category of *geometric analysis*[1] is added for this analysis. However, it is noteworthy that the *sampling* and *geometric analysis* categories are very similar and sometimes even identical, as many of the *geometric analysis* approaches generate multiple candidates and select one for execution. Therefore, in this thesis, methods that focus on the shape and geometry of an object are categorized as *geometric analysis*, while generative models and other learning-based sampling methods are categorized as *sampling* approaches. An overview of the features of interest for this section's analysis can be seen in Figure 2.1.

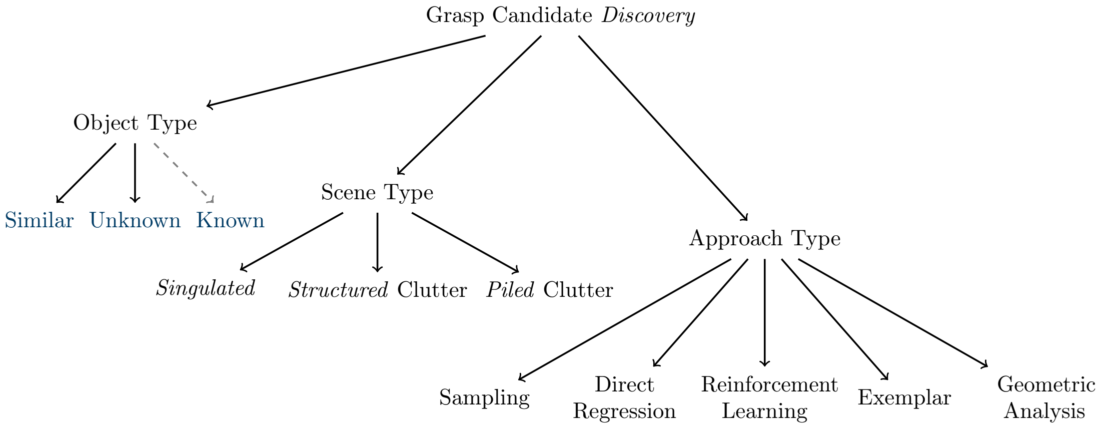


Figure 2.1.: Overview of the different features of grasp *discovery* approaches of interest for this section.

Grasp synthesis for known objects in cluttered environments has seen significant advancements, with methodologies focusing on precise object detection, pose estimation, and collision-free manipulation. For example, Ge et al. (2023) presented a novel network for grasp detection in cluttered trays specifically designed for medical test tubes. Focusing on deformable objects, de Farias et al. (2022) transferred grasps based on shape similarities through functional map correspondence. Logothetis et al. (2018) employ a model-predictive control approach for vision-based object grasping, calculating optimal grasping areas using the tracked object's point cloud data.

[1] see e. g., Kragic and Vincze, 2009, Section 4.1.

These approaches rely on the availability of detailed object knowledge, including geometry and class, for practical grasp synthesis in *unstructured environments*.

However, in scenarios related to commonplace activities as encountered in *service* and *assistance* applications, complete object knowledge cannot be guaranteed. Therefore, robotic assistants must handle tasks autonomously despite partial or missing object information. To this end, Section 2.1.1 will introduce works for grasping similar objects and Section 2.1.2 works for grasping unknown objects.

2.1.1. Grasping Similar Objects

The ability to grasp similar objects represents a significant challenge and opportunity for advancing *autonomy* of robotic assistants. This subsection delves into various methodologies developed to generate grasp hypotheses for objects classified as similar based on their class or primitive shape. The aim is to provide a detailed exploration of recent works to gain insights into how contemporary approaches facilitate object grasping without requiring instance-specific knowledge, instead relying on general characteristics shared within object categories (e. g., handles of cups, the neck of bottles, etc.).

Table 2.1.: Overview of grasp *discovery* approaches for similar objects.

	sampling	direct regression	reinforcement learning	exemplar	geometric analysis
Bohg et al. (2012)	○	○	○	●	○
Chen et al. (2022)	◐	○	○	○	●
Chen et al. (2023a)	○	●	○	○	○
Detry et al. (2017)	○	◐	○	○	●
Ficuciello et al. (2019)	◐	◐	●	○	○
Hidalgo-Carvajal et al. (2023)	○	●	○	○	●
Huang et al. (2023)	○	●	○	●	○
Kurenkov et al. (2017)	○	○	○	●	○
Li et al. (2024)	○	○	○	○	●
Madry et al. (2012)	○	○	○	●	○
Rodriguez et al. (2018)	○	○	○	●	○
Simeonov et al. (2023, 2022)	○	●	○	●	○
Tang et al. (2024a)	○	●	○	○	○
Tekden et al. (2023)	○	○	○	◐	●
Tsagkas et al. (2024)	○	○	○	○	●
Vahrenkamp et al. (2016)	○	○	○	●	●
Wen et al. (2022)	●	●	○	●	○
Wu et al. (2023c)	○	○	○	●	○
Cai et al. (2024)	●	●	○	●	○

Exemplar Methods

Grasping similar objects in robotics often involves leveraging knowledge from known examples to facilitate grasp synthesis on novel items. For instance, Madry et al. (2012) introduce a probabilistic system that integrates task-oriented reasoning with object categorization for grasp transfer from a 2D-3D object database, capitalizing on visual properties. Bohg et al. (2012) leverage a database of object models, categorized by type, which is annotated with task-specific grasp hypotheses. When a new object is encountered, the system identifies its category, retrieves the most similar exemplar, and selects the appropriate grasp based on the given task. Furthermore, Kurenkov et al. (2017) explore grasp transfer through 3D shape deformation, enabling the application of known grasps to novel objects without requiring instance-specific knowledge. Similarly, a method for transferring grasping skills to novel instances within a category, using latent space non-rigid registration to handle partially occluded shapes effectively, is introduced in Rodriguez et al. (2018). While these approaches, rooted in the *exemplar* category, underscore the potential of using knowledge from a canonical object model for a category, other approaches focus more on objects' shapes than their class.

Geometric Analysis

Grasp synthesis for similar objects through *geometric analysis* enables robots to handle objects with slight variations in shape by focusing on their geometric similarities. Chen et al. (2022) propose a transformer-based shape completion module to enhance grasping interaction by restoring sparse point clouds, underlining the importance of geometric information in grasp synthesis. Through optimization of latent shape codes and aligning object poses, Tekden et al. (2023) present an approach that transfers grasp knowledge across objects with geometrically similar surfaces and shows the applicability of their approach even across classes. Moreover, Hidalgo-Carvajal et al. (2023) leverage Neural Networks (NNs) for first completing the object shape and a subsequent grasp posture prediction, targeting specific object categories. By using visual diffusion descriptors, *geometric analysis*, and user-defined interaction points, Tsagkas et al. (2024) create an approach that allows for zero-shot precise manipulation.

Direct Regression & Reinforcement Learning

While most approaches for *grasping* similar objects focus on *exemplar* or *geometric analysis*, there are also approaches in the *direct regression* and *reinforcement* *learning* categories. For example, Ficuciello et al. (2019) demonstrate the use of *reinforcement learning* to adaptively grasp objects with primitive shapes, employing a synergy-based control framework for an anthropomorphic hand-arm system. Chen et al. (2023a) introduce Keypoint-GraspNet, leveraging *direct regression* via Convolutional Neural Networks (CNNs) to predict grasp poses from RGB-D input, focusing on objects of known classes.

Task-Oriented Synthesis

Using category-level knowledge has proven valuable for generating task-oriented grasps, enabling robotic systems to adapt to diverse environments by reducing the *task-specific knowledge* necessary for *grasping*. Leveraging part segmentation and semantic labeling, Vahrenkamp et al. (2016) focus on shape and functionality similarities for grasp planning on familiar objects. Similarly, Detry et al. (2017) combine semantic and geometric scene understanding to plan task-oriented grasps, relying on geometric models to align the gripper with object surfaces. Wen et al. (2022) introduces CaTGrasp, a framework that learns task-relevant grasping in clutter from simulation, using the NUNOCS representation for dense correspondences across object instances, showing promise for industrial applications in cluttered scenarios. Wu et al. (2023c) present a method for transferring functional grasp information across objects within the same category using touch codes. Li et al. (2024) further this concept with ShapeGrasp, which uses geometric decomposition and LLMs to assign semantic meanings to the decomposed parts. In a subsequent step, the LLM is used to decide which part to grasp for task-oriented grasping based on *geometric analysis* of the parts. Similarly, Tang et al. (2024a) leverage foundation models to encode semantic and geometric knowledge, which is then used by a Transformer-based evaluator to directly predict the task relevancy of a set of generated grasp candidates to satisfy both stability and task-compatibility constraints.

Neural Descriptor Fields

Simeonov et al. introduce Neural Descriptor Fields (NDFs, Simeonov et al., 2022), which are continuous functions mapping 3D spatial coordinates to category-level descriptors that are SE(3)-equivariant, meaning they remain consistent under arbitrary translations and rotations of the object. NDFs are computed using a neural network trained via a 3D reconstruction task, allowing the network to encode spatial relationships and key features of objects without requiring manual keypoint annotation. This approach is extended to relational tasks by introducing the Relational-Neural Descriptor Field (R-NDF, Simeonov et al., 2023), which uses NDFs to model interactions between multiple objects, solving for relative transformations necessary for tasks like stacking or arranging objects. A similar approach, Neural Interaction Field and Template (NIFT, Huang et al., 2023), encodes object interactions by using a Neural Interaction Field to capture spatial features around objects and a Neural Interaction Template derived from the Interaction Bisector Surface to optimize object poses for imitation learning. Unlike NDF, which focuses on individual object representation, NIFT emphasizes interaction between objects, improving generalization in manipulation tasks.

The exploration of grasp synthesis and candidate extraction for similar objects, as discussed in this subsection, underscores the diversity of current methodologies. By focusing on the shared attributes of object classes or shapes, these approaches demonstrate a robust capacity for enhancing robotic manipulation in environments populated with a multitude of similar yet distinct items.

2.1.2. Grasping Unknown Objects

In scenarios and applications where there is no prior knowledge about the objects available, different approaches to those presented in Section 2.1.1 are needed. Grasping unknown objects presents a significant challenge due to the absence of information about the objects' shapes, sizes, or materials. Therefore, this subsection focuses on the methodologies dealing with generating grasp hypotheses without prior knowledge, exploring various approaches from *geometric analysis* and *direct regression* methods to adaptive grasping strategies and the integration of multi-modal data.

Geometric Analysis

Early approaches focus on geometric properties of 3D image data for grasping *singulated*[2] unknown objects. The work by Dune et al. (2008) introduces a method for estimating the rough shape of unknown objects through contour analysis and 3D reconstruction using quadrics. Bohg et al. (2011) and Kraft et al. (2009) both address the incompleteness of object observation, with the former predicting full shapes from partial observations and the latter learning objects and grasping affordances through autonomous exploration. Similarly, Schiebener et al. (2016) leverage symmetry of objects and scene context for 3D object shape completion,

[2] The concepts of *singulated* objects in contrast to *structured* and *piled* clutter are described in Newbury et al., 2023.

Table 2.2.: Overview of grasp *discovery* approaches for unknown objects.

	sampling	direct regression	reinforcement learning	exemplar	geometric analysis
Ala et al. (2015)	○	○	○	○	●
Barad et al. (2023)	●	○	○	○	○
Bohg et al. (2011)	○	○	○	○	●
Chen et al. (2016)	○	○	○	○	●
Chen et al. (2023c)	○	●	○	○	◐
Cheng et al. (2020)	○	●	○	○	○
Cheng et al. (2022)	○	●	○	○	○
Danielczuk et al. (2020)	○	○	●	○	○
Deng et al. (2019)	○	●	○	○	●
Dune et al. (2008)	○	○	○	○	●
Eppner and Brock (2013)	○	○	○	○	●
Fischinger and Vincze (2012); Fischinger et al. (2013, 2015)	●	○	○	○	●
Gabellieri et al. (2020)	○	●	○	○	●
Grimm et al. (2021)	○	○	○	○	●
Guo et al. (2024)	○	●	○	○	◐
Hoang et al. (2022)	◐	●	○	○	○
Jiang et al. (2021)	●	◐	○	○	◐
Kopicki et al. (2019, 2016)	●	○	○	●	○
Kraft et al. (2009)	○	○	○	○	●
Li et al. (2022)	○	●	○	○	○
Liu et al. (2022)	◐	◐	○	○	●
Mahler et al. (2017)	●	◐	○	○	○
Marton et al. (2010)	○	○	○	○	●
Mosbach and Behnke (2024)	○	○	●	○	○
Ni et al. (2021)	○	●	○	○	○
ten Pas et al. (2017)	●	○	○	○	●
Patten et al. (2020)	○	○	○	●	○
Player et al. (2023)	○	●	○	○	○
Popović et al. (2010)	○	○	○	○	●
Qin et al. (2023)	○	●	○	○	○
Rao et al. (2010)	●	○	○	○	●
Sabzejou et al. (2023)	○	○	○	○	●
Saxena et al. (2008)	○	●	○	○	●
Schiebener et al. (2012)	○	○	○	○	●
Schiebener et al. (2016)	○	○	○	○	●
Schmidt et al. (2018)	○	●	○	○	○
Song et al. (2018a)	○	●	○	○	○
Su et al. (2024)	○	○	●	○	○
Sundermeyer et al. (2021)	○	●	○	○	○
Suzuki et al. (2022)	○	○	○	○	●
Tang et al. (2024b)	●	◐	○	○	●
Wei et al. (2022)	●	○	○	○	○
Wu et al. (2023a)	●	○	○	○	○
Wu et al. (2019)	○	○	●	○	○
Xu et al. (2022)	○	●	○	○	○
Xu et al. (2023)	○	●	○	○	○
Zhang et al. (2021)	○	●	○	○	○
Pohl and Asfour (2022)	●	○	○	●	●

enhancing the grasp synthesis process for *singulated* objects. Meanwhile, Suzuki et al. (2022) propose a real-time grasp-stability evaluation using proximity sensing, dynamically adjusting the hand pose for stable grasps on unknown objects.

Grasping unknown objects in *structured* clutter environments, on the other hand, requires approaches that can handle occlusions and incomplete shapes. Rao et al. (2010) use depth segmentation to identify and classify graspable segments, constructing a triangular mesh for shape completion and employing a supervised learning method to select the most stable antipodal grasp points based on the partial 3D information obtained. Similarly, Popović et al. (2010) leverage co-planarity and color information from visual cues to formulate grasping strategies. Marton et al. (2010) focus on reconstructing 3D models from single views using geometric model fitting, facilitating classical grasp planning for novel objects. Furthermore, Schiebener et al. (2012) present an integrated approach for discovery, segmentation, and reactive grasping of unknown objects, employing *geometric analysis* to iteratively segment and refine object hypothesis by pushing. Eppner and Brock (2013) simplify perception by exploiting the shape adaptability between the hand and objects, as well as the environmental constraints, allowing for robust grasping without detailed object models. Ala et al. (2015) propose a 3D grasp synthesis algorithm based on *geometric analysis* of cloud points to identify stable contact points, facilitating grasping in complex environments. Chen et al. (2016) introduce a probabilistic approach to grasp planning, utilizing a probabilistic Signed Distance Function to address sensor uncertainty.

For the extraction of robust grasp hypotheses for *piled* unknown objects, the Height Accumulated Features (HAF, Fischinger and Vincze, 2012) that analyze the vertical structure of point cloud data have proven effective. HAF enable the abstraction of shape information for grasp synthesis by calculating the sum of height values within defined regions of a discretized point cloud grid. The approach is extended with Symmetry Height Accumulated Features (SHAF, Fischinger et al., 2013), which enhance the grasp synthesis process by incorporating symmetry information, improving grasp accuracy in complex environments. In Fischinger et al. (2015) the HAF and SHAF methods are applied to a broader range of scenarios, emphasizing topographic analysis to refine grasp quality and adaptability across different robotic platforms. Grimm et al. (2021) follow a different approach, where grasp candidates are generated by segmenting the scene based on depth data, approximating object shapes with 3D bounding boxes derived from 2D projections, filtering out noise and outliers through spatiotemporal clustering over multiple frames, and aligning the grasp poses to maximize execution robustness, all while ensuring computational efficiency suitable for resource-constrained robots. Finally, Liu et al. (2022) focus on robotic picking in *piled* clutter through domain-invariant learning from synthetic data. They use *geometric analysis* to reason about suitable suction regions, from which possible candidates are sampled and then rated according to their quality based on a NN. Similarly, Sabzejou et al. (2023) employ object skeletons generated from the 2D contour of unknown objects for keypoint generation without prior object knowledge, demonstrating adaptability across diverse objects.

Direct Regression

Direct regression methods have significantly advanced the grasping of unknown objects in cluttered environments by utilizing depth and point cloud data. For example, Gabellieri et al. (2020) leverage human expertise and *geometric analysis*, using a reduced database of human-performed grasps and a Oriented Bounding Box decomposition algorithm to regress grasps for unknown objects. Sundermeyer et al. (2021) and Player et al. (2023) demonstrate approaches to generate stable 6-Degree of Freedom (DoF) grasp poses in real-time for objects in dynamic, cluttered settings, including underwater and tabletop scenes, by directly regressing from depth video or point cloud data to grasp configurations. Zhang et al. (2021) use a CNN-based architecture to estimate grasp quality for suction grippers, trained on a large synthetic dataset of point clouds and grasp poses, and integrate a closed-loop control algorithm with feedback from a 6-DoF force-torque sensor for optimizing grasp execution. The work by Ni et al. (2021) introduces a method that employs a SPH3D-GCN-based network to predict grasp poses, categories, and qualities directly from point clouds, followed by an iterative refinement process to enhance grasp accuracy without requiring traditional sampling or search processes. Similarly, Li et al. (2022) present HGC-Net, a data-driven method predicting grasp poses from point clouds in cluttered scenes, demonstrating significant improvements in grasp success rates and time efficiency. Furthermore, Hoang et al. (2022) propose VoteGrasp, employing a deep Hough voting mechanism where each point in the cloud votes for potential grasp centers, followed by clustering these votes to form candidate grasp configurations that are refined using a context-learning module to ensure robustness against occlusions and to generate collision-free grasps. The approach of Chen et al. (2023c) involves using a grasp region exploration module to enhance point density around grasp points, followed by a grasp region attention module to dynamically aggregate features and directly regress 7-DoF grasp poses from the point cloud data in cluttered scenes. Finally, Guo et al. (2024)'s PhyGrasp model integrates physical commonsense reasoning with multi-modal data, generalizing robotic grasping for *singulated* objects.

Instead of using 3D depth data, several methods focus on generating grasp poses directly from RGB or RGB-D images, utilizing the advances from computer vision in the field of image understanding. Saxena et al. (2008) present a learning algorithm that uses supervised learning to identify optimal grasp points directly from 2D images and employs geometric triangulation to infer their 3D positions for effective robotic grasping. Song et al. (2018a) introduce a novel approach using multi-level CNNs to imitate human grasping skills for unknown objects, mapping RGB-D images to grasping postures without requiring prior object knowledge. Similarly, Schmidt et al. (2018) employ CNNs to generate grasping actions from depth images, which are trained on synthetic images of objects and grasps generated by classical grasp planning. The work of Deng et al. (2019) proposes an attention-based visual analysis framework that combines a computational visual attention model and a deep CNN to generate grasp-relevant information for guiding grasp synthesis of unknown objects, followed by a *geometric analysis* to optimize grasp configurations. Cheng et al. (2020) propose a dense prediction model that directly generates grasp poses from images, efficiently handling similar objects by analyzing their surface properties and shapes. Furthermore, Cheng et al. (2022) introduce a robot grasping system that employs a single-stage anchor-free deep grasp detector to generate grasp possibility heatmaps and estimate grasp properties directly from RGB-D inputs, showcasing versatility in handling unknown objects in both *singulated* and *piled* clutter scenes. GKNet, introduced by Xu et al. (2022), simplifies grasp candidate detection by treating it as a keypoint detection problem, enabling real-time, accurate grasp predictions in various scenarios, including *piled* clutter. Xu et al. (2023), on the other hand, develop a single-stage grasp synthesis model that integrates grasp representation with instance segmentation, predicting grasp configurations for specific objects using RGB-D images. Similarly, the DG-CAN model (Qin et al., 2023) employs a two-stage approach for grasp candidate generation using RGB-D images, where grasp proposals are first generated by a Grasp Proposal Network based on multi-modal feature maps (from RGB and depth images), and then refined by a Grasp Region of Interest Network to predict precise 6-dimensional grasp configurations including depth, with the depth information being refined through a Local Cross-modal Attention module to enhance the fusion of RGB and depth features.

Reinforcement Learning

On the other hand, adaptive grasping of unknown objects through *reinforcement learning* has shown promising advancements in recent years. Wu et al. (2019) employ an approach that integrates a pixel-attentive policy gradient method to train a multi-fingered robotic grasping policy using depth images, which enhances grasping success in *piled* clutter through an attention mechanism that allows the robot to zoom into relevant sub-regions of the image. Furthermore, Danielczuk et al. (2020) model the task as a Markov Decision Process (MDP), where the BORGES algorithm explores different stable poses of an unknown polyhedral object, iteratively refining grasp actions based on the success of previous attempts and the probabilistic transitions between poses. Mosbach and Behnke (2024) introduced a Teacher-Augmented Policy Gradient method, integrating *reinforcement learning* with instance segmentation for grasping arbitrary objects in cluttered environments, demonstrating strong zero-shot transfer capabilities. Similarly, Su et al. (2024) explored tactile-based *reinforcement learning* for manipulating unknown objects, emphasizing the importance of tactile feedback and zero-shot Sim2Real transfer capabilities.

Sampling

Sampling-based methods have shown promising results in handling the complexities of grasping unknown objects in cluttered environments. *Dex-Net 2.0* (Mahler et al., 2017) uses a deep CNN trained on a large synthetic dataset to predict the success of grasps from depth images, where grasps are specified by their planar position, angle, and depth relative to an RGB-D sensor, and then planning grasps by sampling and ranking candidate grasps based on their predicted robustness. Furthermore, the method presented by ten Pas et al. (2017) directly sample grasp poses from point clouds by using the Darboux frame as the grasp, utilizing a CNN to classify thousands of 6-DoF grasp candidates, showing effectiveness in densely cluttered environments. The FlexLoG framework introduced by Tang et al. (2024b) employs a novel approach that identifies potential grasp points through *sampling*, then extracting local geometric features to predict high-quality grasps within specific regions of a scene that are suitable for specific downstream tasks.

Generative models have emerged as a powerful tool for synthesizing grasps on unknown objects that employ the *sampling* paradigm. Jiang et al. (2021) introduce the GIGA model, which leverages synergies between grasp affordance prediction and 3D reconstruction through the use of deep implicit neural representations, enabling the model to learn geometrically-aware features from self-supervised grasp trials. Wei et al. (2022) employ a Conditional Variational Auto-Encoder (CVAE) to generate high-degree-of-freedom grasps, incorporating a point completion module and iterative refinement to enhance grasp synthesis. Similarly, Wu et al. (2023a) introduce a method combining a CVAE with bilevel optimization to predict and refine contact points on unseen objects, demonstrating high success rates in experiments. Barad et al. (2023) leverage latent diffusion models for generating 6-DoF grasps, highlighting the method's scalability and quality of grasp samples.

Exemplar

Being a popular approach for grasping similar objects, the *exemplar* approach is hard to apply to unknown objects due to its reliance on similarity. However, Kopicki et al. (2019, 2016) utilize Learning from Demonstration (LfD) to transfer kinesthetically taught grasps to generate new grasp candidates for novel objects by *sampling* from a probabilistic model, emphasizing the importance of generative models and a product of experts formulation for adapting to various object orientations. Similarly, Patten et al. (2020) use the Dense Geometrical Correspondence Matching Network to encode the geometry of objects into a feature space through metric learning, allowing the system to retrieve and transfer grasps from a database of past successful grasps to new, geometrically similar objects. This method incrementally builds a database of grasp experiences and uses dense 3D-3D correspondence reconstruction to adapt these grasps to unseen objects, improving grasp success over time as more experiences are accumulated.

The exploration of robotic grasping of unknown objects showcases a broad spectrum of methodologies, each contributing to the field's advancement in distinct ways. From leveraging *geometric analysis* in cluttered environments to employing *direct regression* for real-time grasp prediction and integrating depth and RGB data for enhanced grasp detection, these studies advance the *autonomy* of robotic manipulation in *unstructured environments*. The detailed examination of various approaches underscores the ability of robots to handle a wide array of objects without prior knowledge, marking significant progress towards the main objective of this thesis.

2.1.3. Discussion

To deal with the diverse set of tasks, a general purpose robotic assistant has to cope with everyday situations. Thus, a versatile *discovery* of interaction possibilities based on visual perception is a fundamental requirement. Since applications in real-world scenarios involve a large variety of objects, it is reasonable to assume that a robot will never have access to all the information required for conventional grasp planning. More likely, a robot will have to deal with objects that are similar to ones they have already dealt with or are completely unknown. Therefore, this section investigated approaches for the synthesis of grasp candidates for similar and unknown objects. The field of grasp synthesis has rapidly evolved over the last decade (see Bohg et al., 2014; Newbury et al., 2023). Especially the advance of NNs and deep learning has influenced these developments (e. g., Mosbach and Behnke, 2024; Schmidt et al., 2018; Sundermeyer et al., 2021; Wu et al., 2023a). However, still a large section of the approaches rely on *geometric analysis* for the grasp candidate extraction (e. g., Liu et al., 2022; Sabzejou et al., 2023; Suzuki et al., 2022). Utilizing task-specific knowledge for grasping of objects with the same semantic class is a visible trend for improving the efficacy of functional grasping (e. g., Tang et al., 2024a; Wen et al., 2022; Wu et al., 2023c).

Previous methods for generating task-oriented grasps have primarily relied on large, manually annotated datasets to train NNs, but these approaches fail to generalize to new objects with significant shape variations, and manual annotation is both costly and challenging. While visual imitation learning approaches offer more efficient means to generalize manipulation skills across categorical objects using human demonstrations, they often require multiple views of the object, which may not be available in real-world scenarios, leading to less precise grasp candidates and potential instability. Similarly, affordance-based methods often depend on primitive shapes or large datasets and face challenges such as inaccurate simulation-based representations of real-world physics, labor-intensive manual definitions, and limitations in generalizing to novel scenarios. Furthermore, learning from human demonstrations can be slow and task-specific, with real-world exploration being risky and dependent on noisy sensor data.

Therefore, this thesis will present two novel approaches for grasping similar (Cai et al., 2024) and unknown objects (Pohl and Asfour, 2022), respectively.

2.2. Grasp Selection and Quality Prediction

The second step in *discriminative grasping* – after the generation of a large number of grasp hypotheses – is the *selection* of a candidate for execution that maximizes the chances of success. To this end, different measures of the quality of a grasp have been introduced in the literature. This section revolves around introducing some of these measures and approaches for grasp *selection* and categorizing them with regard to how these measures are obtained. In this context, three different categories will be used: (a) *analytical quality* measures rely on mathematical models and physical principles to assess grasp stability and performance, (b) *heuristic quality* measures use experience-based techniques, simplified rules, and practical guidelines rather than strict mathematical formulations to evaluate grasp quality, and (c) *learned quality* measures leverage machine learning algorithms to predict grasp success based on training data.

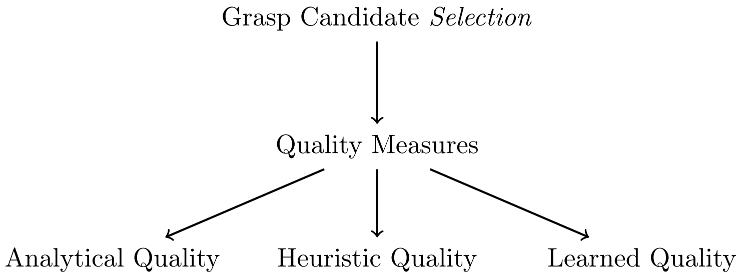


Figure 2.2.: Overview of the different features of grasp *selection* approaches of interest for this section.

Some works use grasp quality prediction in an online fashion to improve finger positioning during grasp execution. For example, the approach of Song et al. (2018b) presents a method for predicting and measuring grasp quality by building a contact score map on a 3D object's voxelization and iteratively adjusting the hand's pose and joint angles to optimize the fit between the hand's geometric shape and the object's surface. Similarly, the work by Arapi et al. (2020) proposes an end-to-end deep learning approach that uses Inertial Measurement Units (IMUs) on soft robotic hands to predict grasp failures by leveraging a NN for real-time prediction, enabling proactive adjustments to prevent failures. Kumar and Mukherjee (2022) use an algorithm to search for an optimal grasp pose with rolling contacts by reducing the geodesic distance between the current and desired manipulability matrices through defining a manipulability measure that characterizes the grasp for multi-fingered robotic handling. The work by Si et al. (2022) presents a simulation framework that predicts grasp stability during execution by using tactile images generated from simulated contact forces and deformation, integrating a vision-based tactile sensor and contact dynamics model, and achieving high prediction accuracy when transferred from simulation to real-world tasks without additional real-world training data. While this line of research uses quality measures to increase the grasp success rate, the focus of the following related work is the use of quality measures to select the best grasp out of several grasp candidates.

To this end, Section 2.2.1 introduces approaches using *analytical quality* measures. Subsequently, Section 2.2.2 focuses on methods using *heuristic quality* and Section 2.2.3 on approaches using *learned quality* measures.

2.2.1. Analytical Grasp Quality Prediction

In grasp selection, the *analytical quality* metrics serve as a fundamental basis for evaluating and selecting potential grasps. *Analytical quality* measures are a class of techniques based on precise mathematical models and physical principles.

Table 2.3.: Overview of grasp *selection* approaches.

	analytical quality	heuristic quality	learned quality
Almeida and Moreno (2021)	●	●	○
Arapi et al. (2020)	○	○	●
Asif et al. (2014)	○	●	○
Baressi Šegota et al. (2022)	○	○	●
Cavalli et al. (2019)	●	○	●
Chen et al. (2018)	○	●	○
DeGol et al. (2016)	○	○	●
Erkan et al. (2010)	○	○	●
Ghalamzan E. et al. (2016)	◐	○	●
Goins et al. (2014)	●	○	●
Gori et al. (2013)	●	●	●
Gravdahl et al. (2019)	○	●	○
Gualtieri et al. (2016)	○	○	●
Herzog et al. (2012, 2014)	○	●	●
Kappler et al. (2015, 2016)	○	○	●
Kent and Toris (2018)	○	●	○
Konrad et al. (2022)	○	○	●
Krug et al. (2016)	●	○	○
Kumar and Mukherjee (2022)	●	○	○
Lin and Sun (2015)	●	○	○
Lu et al. (2020)	○	○	●
Mavrakis et al. (2017)	●	○	○
Mnyussiwalla et al. (2022)	●	○	○
Morales et al. (2003)	○	●	●
Nadon and Payeur (2020)	○	●	○
Pardi et al. (2021)	●	○	○
Qian et al. (2020)	○	●	●
Quispe et al. (2016)	●	●	○
Rohanimanesh et al. (2023)	○	○	●
Rubert et al. (2018, 2017)	○	●	●
Sharif et al. (2019)	○	●	○
Si et al. (2022)	○	○	●
Song et al. (2011, 2015)	○	○	●
Song et al. (2018b)	●	○	○
Vollhardt et al. (2019)	●	○	○
Wakabayashi et al. (2022)	○	○	●
Ying et al. (2018)	●	●	○
Baek et al. (2022)	●	●	●

They often involve the calculation of various metrics derived from the laws of physics and mechanics, such as force closure, torque equilibrium, and wrench space analysis. These measures often provide a high degree of predictability and reliability in assessing grasp quality and can be distinguished by their reliance on exact calculations and theoretical foundations. This subsection delves into various approaches that revolve around physical and geometric properties, dynamic and energy-based evaluations, optimization strategies, and task-specific planning.

Contact-based Approaches

Some approaches for calculating *analytical quality* measures revolve around the physical properties of the hand-object-contact, similar to metrics used in classical grasp synthesis. Krug et al. (2016) evaluate the containment of a set of task wrenches within the Grasp Wrench Space (GWS), using tactile feedback to alleviate contact placement uncertainties and employing a quality criterion that measures the maximum scaling factor of the task wrenches that can be resisted by the grasp, thereby predicting grasp success without the need for extensive modeling or training data. Similarly, the work by Mnyussiwalla et al. (2022) provides a comprehensive analysis of various analytical grasp quality criteria of two categories – either depending exclusively on the contact points, such as force closure and the volume of the GWS, or the kinematics of the hand, e. g., the hand-object Jacobian, which are essential for optimizing dexterous manipulation.

Energy-based Approaches

In addition to these contact-based approaches, dynamic and energy-based analytical methods offer precise evaluations of grasp stability and safety. Mavrakis et al. (2017) propose a method to minimize impact forces in post-grasp manipulations by calculating the effective mass and kinetic energy matrix, aiming to enhance safety in dynamic environments. Similarly, Vollhardt et al. (2019) introduce an energy-based stability analysis for multi-fingered, compliant robotic hands, focusing on stability characterization through metrics like minimum destabilizing energy and grasp stiffness. This approach allows for the *selection* of grasps that are robust against external disturbances.

Task-specific Approaches

Contrary to general analytical strategies in grasp *selection*, other approaches are tailored to task-specific grasp *selection*. Lin and Sun (2015) propose a task-oriented grasp quality metric based on the distribution of task disturbances captured during task demonstrations, ensuring the chosen grasp covers the most frequent disturbances while maintaining computational efficiency by reducing the configuration space using specific thumb placements and directions. Furthermore, Ying et al. (2018) integrate an EEG-based Brain-Computer Interface that detects user interest in visual stimuli, allowing the user to iteratively filter and refine grasp options generated by an online planner until the most suitable grasp is chosen, leveraging both analytical (e. g., reachability and maximum wrench perturbation force) and heuristic (e. g., closeness between the hand and the object's surface) grasp quality metrics to ensure effectiveness in cluttered environments. Cavalli et al. (2019) introduce a framework that evaluates task-oriented grasps through *analytical quality* metrics like grasp robustness and rotational inertia and train NNs to predict these metrics from vision. Additionally, the work of Pardi et al. (2021) presents an optimization-based algorithm that selects the best grasp location by minimizing the effort needed to keep an object stable against external forces. It specifically considers the dynamic interactions and force profiles involved in the manipulation tasks, ensuring that the grasp quality is optimized for the specific requirements of the task being performed.

In summary, *analytical quality* metrics provide a robust framework for evaluating the efficacy of grasps. By leveraging physical properties, dynamic analyses, optimization techniques, and task-specific considerations to provide precise measures of grasp quality, these methods enhance the versatility, reliability, and adaptability of robotic manipulation.

2.2.2. Heuristic Grasp Quality Prediction

Heuristic quality metrics in grasp selection utilize various pre-defined heuristics to evaluate and prioritize grasp candidates. In contrast to *analytical quality* measures, they are usually derived from the experience or intuition of experts and are not always based on precise theoretical modeling of hand-object relations. The advantages of *heuristic quality* measures include their speed and simplicity. They can provide rapid assessments without the need for complex calculations, making them suitable for real-time applications where quick decision-making is essential. For example, the work of Sharif et al. (2019) proposes a framework utilizing particle filters to systematically combine various cues, primarily hand trajectory information, for accurately inferring user intent (i. e., selecting a grasp type) and estimating the remaining time until the hand reaches the object, which enables precise grasp planning and execution for prosthetic robot hands. However, their reliance on general rules means they may not always be as accurate or reliable as more rigorous analytical methods.

Efficiency Optimization

Heuristic measures are often used to optimize the grasp selection process, significantly enhancing efficiency and practicality. For example, Quispe et al. (2016) propose an approach combining both arm and hand metrics, which includes evaluating the arm's ease of reach and comfort (based on analytical Inverse Kinematics (IK) and manipulability measures) in combination with the hand's grasp robustness heuristic, to ensure that the chosen grasp is not only stable but also feasible and efficient in execution. Similarly, Chen et al. (2018) propose a probabilistic framework that leverages heuristic measures for force closure and perceptual uncertainty, optimizing grasp selection through a simulated annealing process. In the study by Gravdahl et al. (2019), a heuristic method is employed to prioritize grasp candidates based on reachability, planning, and execution time, focusing on practical feasibility within the robot's workspace. Almeida and Moreno (2021) introduce heuristics to streamline the computation of the Potential Grasp Robustness metric, significantly reducing the complexity involved in evaluating grasp stability for underactuated hands.

Geometry-based Approaches

Other approaches analyze the object's shape or curvature in order to improve grasp selection from visual perception. Gori et al. (2013) present a grasping pipeline for effectively grasping unknown objects by matching object surface curvature with the robot's palm, utilizing binocular vision to segment 3D point clouds into smooth surfaces and rank potential grasp points. The score function for evaluating these grasp points incorporates a learned component using a Least-Square Support Vector Machine to map the local curvature of object surfaces to the curvature of the robot's hand, an analytical measure assessing the manipulability of the hand configuration, and heuristic rules based on object dimensions and user-defined task preferences. Additionally, Asif et al. (2014) introduce a vision-based approach that utilizes shape descriptors and distances to surfel mean positions as heuristics for selecting grasps for unknown stacked objects. Furthermore, Kent and Toris (2018) present an approach for the pairwise ranking of grasp candidates based on a set of grasp metrics and object features to adaptively select the best grasp with reduced data collection and improved generalization to novel objects. Lastly, Nadon and Payeur (2020) integrate RGB-D computer vision with a three-finger robotic gripper to identify optimal grasp regions on the object's contour based on distance to target shape, followed by a validation of grasp stability through curvature analysis and heuristic force closure criteria. This approach efficiently narrows down potential grasps by eliminating those that do not allow desired reshaping or are unstable, ensuring the grasp respects the robotic hand's mechanical constraints.

Heuristic quality metrics play a significant role in improving the efficiency and effectiveness of grasp *selection* for robotic manipulation. Therefore, they complement *analytical quality* measures by reducing the computational complexity and overall runtime of the grasp *selection* process. In the following, the third category for grasp *selection* approaches will be introduced.

2.2.3. Learned Grasp Quality Prediction

Learned quality measures are techniques for evaluating the effectiveness and stability of a grasp hypothesis using machine learning approaches that are trained e. g., on large datasets of successful and unsuccessful grasps. Relevant features that influence grasp quality, such as object shape, surface texture, and contact points, are extracted from the data. Once trained, the model can predict the quality of new grasps based on the learned patterns. The system can continuously improve by incorporating feedback from new grasps, refining its predictions, and adapting to new objects and conditions. In contrast to *analytical quality* and *heuristic quality* metrics, learned approaches utilize the strengths of machine learning to provide a flexible, adaptive, and data-driven method for assessing grasp quality. This allows them to handle complex and varied grasping scenarios that may be difficult to model analytically or capture through heuristics alone.

Classical Approaches

Early approaches mostly used classical machine learning classifiers to learn from data and improve the grasp success rate. As one example, Morales et al. (2003) focus on predicting grasp performance from heuristic visual features, using a k-nearest neighbor rule to assess grasp reliability. Erkan et al. (2010) explored using Kernel Logistic Regression to map hypothetical grasp configurations obtained from visual descriptors into class conditional probability values and using semi-supervised and active learning techniques to improve the model's performance with limited labeled data. Furthermore, Herzog et al. (2012, 2014) contribute by using a template-based algorithm that learns grasp configurations from demonstrated examples, matches new objects to a library of shape templates, and adapts over time using feedback from previous grasp attempts to improve performance. Lastly, Rubert et al. (2017) analyze and evaluate various grasp metrics using a large-scale database of simulated grasps and different classifiers (including neural networks) to understand their predictive capabilities, while Rubert et al. (2018) build on this by combining these metrics using machine learning classifiers and validating their effectiveness through real-world experiments and a novel 3-category classification system to improve grasp success prediction for robotic manipulation.

Deep Learning-based Approaches

The integration of deep learning and neural network-based grasp selection methods has significantly advanced the field of robotic grasping, enabling more nuanced and effective approaches to object manipulation. For instance, DeGol et al. (2016) explored the use of CNN for automatic grasp selection in prostheses, showcasing the potential of incorporating deep learning on visual data into the daily lives of prosthesis users. Similarly, Gualtieri et al. (2016) classify each candidate using a CNN trained on informative representations of grasp geometry and appearance, leveraging prior object knowledge and pre-training on simulated data to enhance detection accuracy and achieve a 93% grasp success rate in *piled* clutter scenarios. Kappler et al. (2015) introduce a large-scale dataset and validate stability metrics for grasp planning using deep learning, demonstrating that such data can significantly enhance performance. Building on this foundation, Kappler et al. (2016) propose a novel ranking loss method specifically adapted for binary-labeled grasp hypotheses, optimizing the selection of the top grasp hypothesis from noisy sensor data and showing significant performance improvements over the previous approach. Qian et al. (2020) proposed using a neural network to segment key regions of cloth (edges, inner edges, and corners) from depth images, estimating grasp direction based on the correspondence between outer and inner edge points, and computing the directional uncertainty to select the grasp point with the lowest uncertainty. Konrad et al. (2022) present VGQ-CNN, a 6-DoF grasp quality prediction network that evaluates grasp candidates based on depth images and versatile grasp datasets from a wide range of camera poses, allowing for flexible and efficient grasping without the need for network retraining for each new camera setup. Finally, Wakabayashi et al. (2022) introduced a self-supervised learning system, VGP-Net, to optimize grasp poses for tableware objects, demonstrating the flexibility of learned models in adapting to environmental constraints (e. g., avoiding collisions or dirty spots).

Probabilistic Models

On the other hand, the use of probabilistic models and Bayesian inference has greatly improved the robustness and reliability of grasp selection methods. For example, Song et al. (2011) introduces a novel approach using a Bayesian Network, learned from discretized high-dimensional sensory and motor data through Gaussian Process Latent Variable Models and Gaussian Mixture Models (GMMs), to probabilistically evaluate and select the most suitable grasp configuration for a given manipulation task based on the dependencies and relationships among the observed features. Building on this, Song et al. (2015) use probabilistic inference within a Bayesian Network to select the most appropriate grasp configuration that meets the specific constraints and requirements of a given task based on learned probabilistic relationships among task-relevant variables. Similarly, Goins et al. (2014) use a Gaussian Process-based classifier that combines multiple analytical grasp metrics, significantly improving prediction accuracy compared to individual metric thresholding. Ghalamzan E. et al. (2016) build on recent grasp-learning methods and combine a probabilistic model of grasp likelihood with a manipulation capability index, which is computed analytically but relies on learned models for grasp generation, aiming to optimize both stable grasp likelihood and task-relevant manipulability. The paper by Lu et al. (2020) uses a probabilistic inference approach in a learned Deep Neural Network (DNN) which integrates a voxel-based 3D CNN to predict grasp success probabilities based on the object voxel grid and grasp configuration, enhanced by an object-conditional prior modeled as a Mixture Density Network that captures the distribution of grasp configurations relative to the observed object geometry. Finally, Rohanimanesh et al. (2023) present an approach for online tool selection with learned grasp prediction models in robotic bin-picking systems. The primary challenge is selecting the most efficient sequence of grasps and corresponding tool changes to maximize system throughput despite occlusions and the dynamic nature of visible objects, modeling this as a MDP optimized through model predictive control and integer linear programming to handle real-time decision-making efficiently.

Learning from previous grasp executions through machine learning, NN, and probabilistic methods has greatly improved the reliability and versatility of robotic grasping in *unstructured environments*. However, in contrast to *analytical quality* and *heuristic quality* measures, *learned quality* measures often need a large amount of training data, which is hard to obtain, especially in real-world scenarios.

2.2.4. Discussion

Selecting the best action for the current task in a scene can have a large influence on the reliability of a manipulation task. In the literature, this is usually done by calculating some kind of quality measure and selecting the action with the highest quality. To give an overview of related methods, this section categorized approaches according to the way these measures are computed. The exact calculations that are the foundation of *analytical quality* measures (e. g., Krug et al., 2016; Kumar and Mukherjee, 2022; Lin and Sun, 2015) make them very precise if the required information is available. *Heuristic quality* metrics (e. g., Gravdahl et al., 2019; Kent and Toris, 2018; Nadon and Payeur, 2020), in contrast, are easier and faster to calculate as they involve less complicated computations. Finally, *learned quality* measures (e. g., Erkan et al., 2010; Rohanimanesh et al., 2023) heavily rely on the quality of the available datasets learn what a good action is from previous experiences.

Despite the advances in grasp candidate selection methods, significant challenges remain, particularly in dealing with perceptual and systematic *uncertainties*. Many existing approaches lack robust mechanisms for handling these disruptive factors, which can result in unreliable and inefficient grasping processes. For instance, *analytical quality* metrics, while precise when all required information (such as friction coefficients) is available, can be computationally expensive and often assume perfect knowledge of the environment. In contrast, *heuristic quality* metrics offer faster computation but tend to be less accurate due to their inability to account for noise and incomplete information. *Learned quality* methods may address some of these issues by leveraging large datasets, yet they come with their own set of challenges, including the need for extensive training data and the potential lack of interpretability for human operators.

To address these limitations, Chapter 4 will introduce a framework for the *uncertainty-aware* grasp candidate selection using the approach from Baek et al. (2022), based on the spatiotemporal fusion of grasp candidates from Pohl and Asfour (2022).

2.3. Robotic System Architectures

Three-tiered robot architectures (Bonasso, 1991; Firby, 1989) have long been a standard for operating mobile robots in unstructured and uncertain environments and executing mobile manipulation tasks (see e. g., Kortenkamp et al., 2016). The three layers – *Planning*, *Executive*, *Behavioral Control* – of the architecture correspond to different abstraction levels and robot capabilities (e. g., from Jaquier et al., 2024) and are visualized in Figure 2.3.

The *Behavioral Control* layer focuses on real-time, low-level control of the robot's actions and interactions with its environment. This includes sensor processing, motor control, and immediate responses to environmental stimuli to ensure the robot operates safely and effectively. It executes the specific actions and tasks assigned by the *Executive* layer and responds with real-time feedback on the robot's status and environmental conditions, enabling adjustments and ensuring the robot can adapt to changes or unexpected situations. The medium-level *Executive* layer acts as an intermediary between the *Planning* and *Behavioral Control* layers. Its main role is to decompose high-level plans into detailed tasks and manage the

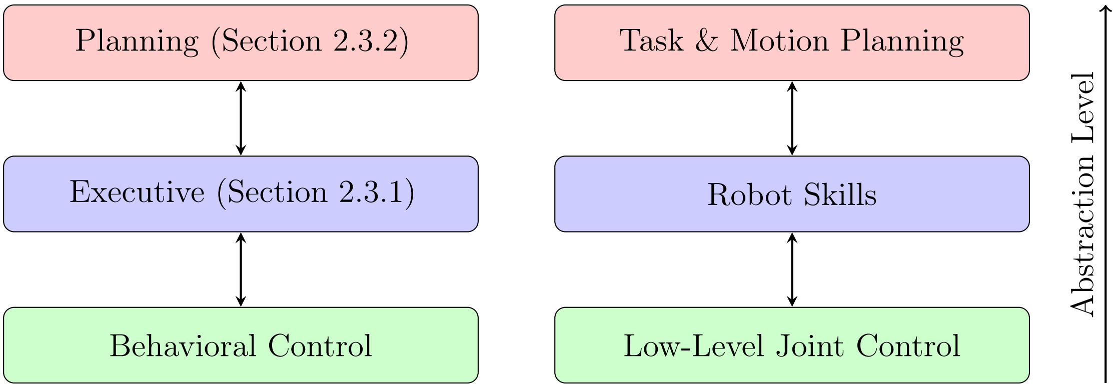


Figure 2.3.: Components of a generic three-tiered robot architecture (e. g., Kortenkamp et al., 2016) combined with the capabilities stack from Jaquier et al. (2024). The different layers are color-coded corresponding to their respective abstraction level as follows: high abstraction (🟥), medium abstraction (🟦), and low abstraction (🟩).

sequence of these tasks. It oversees the execution of plans, handles contingencies, and ensures that the robot's actions align with the strategic goals set by the *Planning* layer. The abstract *Planning* layer is responsible for high-level decision-making and generating long-term goals. It typically involves strategic thinking, such as route planning, task sequencing, and problem-solving based on the robot's objectives and constraints.

The humanoid robots of the ARMAR (Asfour et al., 1999; Asfour et al., 2017) family are an example of the successful implementation of the three-tiered robot architecture for applications in the personal sector. ARMAR-III (Asfour et al., 2006) employs a hierarchically organized control architecture with three levels: task planning (*Planning*), task coordination (*Executive*), and task execution (*Behavioral Control*). This architecture enables dynamic task execution and interaction in domestic environment. ARMAR-6 (Asfour et al., 2019), developed for the assistance of technicians working on maintenance tasks, builds on these concepts and extends the capabilities to advanced cognitive functions like human action recognition, facilitated by the ArmarX (Vahrenkamp et al., 2015) software framework.

A recent trend in mobile manipulation frameworks based on a combination of multi-modal foundational models, so-called Robotic Transformers (Brohan et al., 2023a,b; O'Neill et al., 2024), capitalizes on the generalization capabilities of transformer-based models when trained on large datasets. In contrast to the three-tiered robot architecture, these approaches combine the three layers to directly predict actions based on visual perception and natural language instructions. Even though these models show remarkable emerging capabilities (e. g., generalizing concepts from web-scale data to perform tasks that were not in the training set), they require a tremendous amount of data and result in decreased performance in the control layer (Jaquier et al., 2024). Therefore, this section will focus on the more "traditional" approaches resembling the three-tiered robot architecture.

In order to introduce the state of the art related to Contribution 3 and discuss the adaptability of mobile manipulation skills, this section will consider mainly two fields of research. First, *Executive* mobile manipulation frameworks will be analyzed regarding their capacity to *transfer* knowledge, experience, and skills across varying situations (Section 2.3.1). Subsequently, Section 2.3.2 will review current trends that incorporate and use LLMs in the *Planning* layer.

2.3.1. Task Execution Frameworks

The review of Jaquier et al. (2024) of transfer learning in robotics introduces three main modes of *transfer*: *environment*, *task*, and *robot*. This thesis adopts these categories in an attempt to increase the adaptability of mobile manipulation skills in real-world applications. First, *task transfer* focuses on leveraging the ability of a robot to perform a given task to learn how to execute a different but related task in the same environment. An example would be the reuse of trajectories from grasping an object to place it again. Second, *robot transfer* aims to endow a target robot (e. g., a humanoid robot) with the ability to perform a task known by another source robot (e. g., an industrial manipulator) in the same environment. Finally, *environmental transfer* involves the ability of a robot to perform a task equally well in a target environment compared to a different source environment. That could mean a robot can execute a task in a household scenario as well as an industrial setup. Even though Jaquier et al. (2024) specifically include *Sim-to-Real transfer* in this category, the following analysis of related works does not.

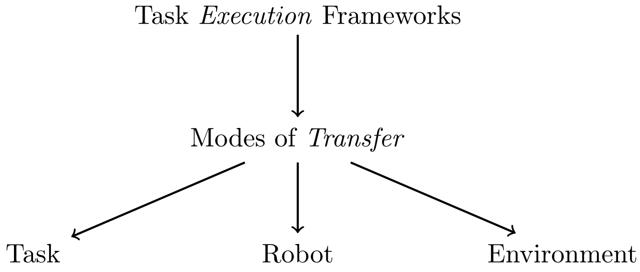


Figure 2.4.: Overview of the different modes of *transfer* for task *execution* frameworks used in this section.

Transfer in the Behavioral Control, Executive, and Planning Layers

The *Behavioral Control* layer is not suited for the *transfer* of knowledge and experience because it operates at a low level of abstraction, dealing with specific sensor processing and motor control tasks that are tailored to individual robots. These tasks are highly dependent on the robot's hardware and environment, making them difficult to generalize across different systems. Additionally, the *Behavioral Control* layer lacks the flexibility to adapt to varied contexts without significant reprogramming. For example, Wächter et al. (2016) discuss the use of hierarchical, distributed statecharts for robot programming. While these enhance programming convenience, they limit skill transfer due to their detailed, robot-specific nature. Similarly, Bohren and Cousins (2010) highlight the challenges in coordinating Robot Operating System (ROS, Quigley et al., 2009) nodes through nested state machines because of the reliance on explicit, robot-specific task scripting. The NUClear framework (Houliston et al., 2016) exemplifies a low-latency, modular message-passing architecture designed to optimize communication and data handling in humanoid robotic systems, enabling efficient real-time processing and component reuse. Despite these advantages, the need for extensive reprogramming to adapt to different tasks or robots remains. Lastly, Iovino et al. (2023a) examine the programming effort required for Finite State Machines (FSMs) and Behavior Trees (BTs) in robotic applications, underscoring the limitations of FSMs in adapting to varied tasks without extensive modifications. Consequently, transferring knowledge from the *Behavioral Control* layer to other robots or settings is impractical and inefficient.

The *Planning* layer is not suited for the transfer of knowledge and experience because it operates at a high level of abstraction, providing broad goals without detailed execution instructions. This generality makes it easy to transfer, but it lacks the specific, actionable information needed to implement tasks directly on different robots. Specifically, Jaquier et al. say that "the difficulty of transfer, as well as the resulting performance, is highly dependent on the capability stack that is made available a priori for each robot." (Jaquier et al., 2024) For instance, Brohan et al. demonstrate the innovative SayCan (Brohan et al., 2023c) framework, which, despite its promise, struggles with skill *transfer* due to its high-level linguistic abstraction. Similarly, the planning framework of Ruiz-Celada et al. (2022) employs perception and ontology-based reasoning to generate Planning Domain Definition Language (PDDL) files, integrates symbolic and geometric planning, and uses BTs for robust task execution while depending on predefined actions and domain-specific configurations without a mechanism for generalizing learned knowledge.

The approach of Mullen Jr. and Manocha (2024) involves leveraging LLMs and affordance scores to align model confidence with task success, reducing human intervention and mitigating LLM hallucinations by evaluating the plausibility and safety of actions within a given scene. However, it relies on specific, context-dependent affordance scores that do not generalize well to different settings and scenarios. Ruiz et al. (2022) use BTs and automated PDDL generation based on ontological reasoning for task and motion planning but fail to enable knowledge transfer due to their dependency on predefined, task-specific predicates and actions. Sun et al. (2023) uses LLMs for interactive planning in uncertain environments by guiding robots to collect observations and update actions but requires task-specific belief states and policies. Generally speaking, the *Planning* layer requires a pre-existing framework of low- to mid-level capabilities for each robot, which limits its practicality for direct knowledge transfer. This gap necessitates additional layers to bridge high-level goals with executable actions.

The *Executive* layer is ideal for transferring knowledge and information due to its intermediate level of abstraction. Unlike the low-level *Behavioral Control* layer, which deals with specific, robot-tailored tasks and instructions that are difficult to transfer, and the high-level *Planning* layer, which often uses abstract natural language descriptions with limited concrete task details, the *Executive* layer has a good balance of task-specific information and abstraction. It bundles actions and executions into skills with specific goals, providing enough detail for execution while maintaining flexibility for adaptation across different *robots*, *environments*, and goals. For instance, high-level plans generated by LLMs describe goals in natural language but lack concrete execution details. The *Executive* layer bridges this gap by grounding these high-level descriptions into robot-specific instructions, facilitating execution across various agents and *environments*. This ability to translate abstract plans into detailed, executable steps makes the *Executive* layer the best suited for the *transfer* of knowledge and experience in robotic systems.

Delimitation of the Analysis' Scope

Some works investigating the transfer of knowledge and experience specifically target task description languages. Even though these are a fundamental requirement for transferring knowledge, a task description language in itself is not able to execute any manipulation task. As an example, the instantaneous Task Specification using Constraints (iTaSC, De Schutter et al., 2007) framework for constraint-based task specification and estimation in robotic systems, focuses on handling geometric uncertainties through feature coordinates and systematic estimation methods. The approach aims to integrate task specification with real-time sensor-based adjustments to manage dynamic environments. Furthermore, the work of Aertbelien and De Schutter (2014) advances this concept with the *eTaSL/eTC* framework, introducing expression graphs for a more modular and composable task specification, along with a clear separation of specification, solver, and execution layers. This newer framework enhances flexibility, enabling knowledge transfer across different *environments*, *tasks*, and *robots* with minimal reprogramming and supports integration with various robotic execution environments. As the aim of this section is to introduce approaches that actually execute mobile manipulation actions and can handle one or more of the modes of transfer, task description languages will not be part of the analysis.

Early *Executive* frameworks for mobile manipulation created integrated solutions that do not facilitate the transfer of knowledge. For example, Bagnell et al. (2012) present an integrated system developed under the DARPA ARM-S program. This system combines open-source packages, BTs, and advanced perception and control algorithms to enable autonomous robotic manipulation. Despite its innovative approach, the study reveals significant limitations in achieving seamless adaptability due to hardware issues, calibration errors, and the absence of dual-hand tracking and bi-manual motion planning. In contrast to this, the following sections will introduce *Executive* frameworks for mobile manipulation that allow the transfer of knowledge and experience across one or more of the modes *task*, *environment*, and *robot*.

Single-Mode Transfer

In contrast to integrated approaches, there exist multiple *Executive* approaches that can handle the *transfer* of knowledge, experience, and skills across a single mode in the field of mobile manipulation. Each modality presents unique challenges and opportunities for enhancing robotic capabilities, from reusing trajectories in different *tasks* to executing skills across varied *environments* and *platforms*. The papers discussed in this section exemplify relevant approaches and methodologies that have contributed towards achieving versatile and autonomous robotic systems capable of adapting to new *tasks*, *environments*, or *robotic platforms* with minimal reconfiguration.

Jaquier et al. (2024) define *single-mode transfer* as knowledge or experience being transferred across a single dimension. In this context, one mode (*robot*, *task*, or *environment*) changes while the other two remain the same. For example, transferring a skill learned by one *robot* to another *robot*, while keeping the *task* and *environment* constant, would be an instance of *robot*-mode transfer. Similarly,

Table 2.4.: Overview of task *execution* frameworks regarding their ability to *transfer* knowledge and capabilities across *tasks*, *environments*, and *robots*.

	tasks	robots	environments
Borghesan et al. (2014)	○	○	●
Burgess-Limerick et al. (2022)	○	◐	●
Chen et al. (2024)	◐	●	●
Dömel et al. (2017)	○	●	●
Garcia et al. (2018)	○	●	◐
Hart et al. (2014, 2015, 2022)	◐	●	●
Hermann et al. (2011)	○	○	●
Iovino et al. (2023b)	●	○	◐
Jiang et al. (2018)	●	◐	◐
Kasaei and Kasaei (2024)	●	○	●
Keleştemur et al. (2019)	○	○	●
Koubaa et al. (2016)	○	●	○
Liang et al. (2022)	●	○	◐
Liu et al. (2024a)	○	○	●
Martins et al. (2023)	●	●	○
Nam et al. (2020)	◐	●	●
Nebot and Cervera (2007)	○	●	◐
Paikan et al. (2015)	○	●	○
Pane et al. (2020)	●	◐	◐
Ren et al. (2024)	○	○	●
Rovida and Kruger (2015);Mayr et al. (2023)	●	◐	◐
Staroverov et al. (2023)	○	●	●
Verma et al. (2021)	○	○	●
Wang et al. (2023a)	●	○	○
Wang et al. (2024)	○	●	○
Wu et al. (2020)	○	●	●
Yang and Zhang (2023)	●	○	○
Yao et al. (2022)	●	●	◐
Yenamandra et al. (2024)	◐	○	●
Yi et al. (2020)	◐	●	●
Yokoyama et al. (2023)	◐	○	●
Pohl et al. (2024)	●	●	●

task-mode or *environment*-mode transfers focus on changing just the *task* or *environment*, respectively.

Task Transfer The *transfer* of manipulation skills across *tasks* represents a critical aspect of advancing robotic versatility and adaptability. Especially in service and hospitality, healthcare and nursing, and domestic robotics, the ability to apply learned skills to new tasks without extensive reprogramming is essential for efficient and flexible operation and acceptance from non-expert users. For example, the SkiROS framework, introduced by Rovida and Kruger (2015), is a skill-based, modular software architecture designed to facilitate intuitive task-level programming for autonomous mobile manipulators in industrial environments, focusing on modularity and scalability within the ROS middleware. Building on this, Mayr et al. (2023) present SkiROS2, which enhances the original framework by incorporating behavior trees, improved knowledge representation using the Web Ontology Language, and support for multiple skill implementations to increase adaptability across different *tasks*, *environments*, and hardware scenarios. SkiROS2 facilitates *transfer* across *tasks* effectively, while both frameworks support partial transferability across *environments* and *robots* due to the need for specific skill implementations and adaptations for different settings and hardware. Similarly, the Layered Architecture for Autonomous Interactive Robots (LAAIR, Jiang et al., 2018) is a three-layer hybrid architecture for autonomous interactive robots, integrating reactive control for dynamic task sequencing, deliberative control for goal planning, and modular skills for interaction. LAAIR supports the reuse of skills across various *tasks*, but the *transfer* across different *environments* and *robots* is only partially supported due to the need for some adjustments in implementation and reprogramming for specific contexts and robot-specific skill implementations. The work of Pane et al. (2020) introduces a constraint-based skill programming framework that separates task and progress constraints, allowing for the composition and reuse of skills in different contexts. This approach enables the creation of complex, reactive robot behaviors adaptable to various *tasks* and potentially different *robotic platforms* and *environments*. Furthermore, Liang et al. (2022) contribute a search-based task planning framework utilizing learned *Skill Effect Models*, which iteratively train on diverse task data to enable flexible, parameterized skill planning and adaptation to new *tasks* and scenarios. This approach allows the planner to efficiently incorporate new skills and *tasks* over time, facilitating lifelong learning and *transfer* of knowledge across varying contexts. Iovino et al. (2023b) propose a framework that combines LfD and Genetic Programming to create and evolve BTs for robotic applications, allowing non-expert users to semi-automatically generate adaptable and efficient robot programs. This approach enables the reuse of learned behaviors across related *tasks* and different settings with minimal reprogramming. Moreover, the work of Wang et al. (2023a) presents a deep reinforcement learning framework that employs Hindsight Experience Replay and a novel knowledge *transfer* technique to enhance dexterous robotic manipulation by leveraging learned strategies from simpler *tasks* to solve more complex *tasks* within the same robotic platform.

Lastly, the Structural-BT framework (Yang and Zhang, 2023) enhances robotic software development efficiency by reusing BT structures for abstracting and implementing task planning paradigms, allowing flexible composition and customization of software components. This framework modularizes the interaction pipelines between sensing, planning, and acting functions, facilitating reuse across different *tasks* while requiring specific adaptations for various *robots* and *environments*.

Environment Transfer The adaptability of robotic systems to diverse *environments* is a critical aspect of their utility, particularly in dynamic, unstructured, or novel settings where no prior knowledge about the scene exists, and the environment has not been prepared particularly for the robot. For instance, Hermann et al. (2011) present a highly integrated hardware and software architecture for a bimanual mobile manipulator, combining multi-sensor perception with fast multi-level planning to enable adaptive and intuitive execution of a wide range of tasks in varying industrial *environments*. The approach illustrates the capability of the manipulator to operate across various *environments*, albeit with some limitations in demonstrating the extent of their environmental diversity. The work of Borghesan et al. (2014) presents a method for specifying and controlling manipulation tasks using constraint- and synergy-based approaches within the iTaSC framework, enabling the adaptation of robotic actions to various *environments* through the definition of objects and robots in different scenes, but requiring explicit programming for each task and robot configuration. The work by Keleştemur et al. (2019) presents a system architecture for autonomous mobile manipulation in domestic *environments*, utilizing NNs for object detection, Natural Language Processing (NLP) for task comprehension and integrated modules for perception, navigation, and motion planning. The framework is implemented on Toyota's Human Support Robot and demonstrates adaptability to various domestic settings. Furthermore, Verma et al. (2021) propose a framework that utilizes automatically generated BTs to enable robust execution and real-time adaptation of robotic manipulation tasks in dynamic environments by integrating symbolic and geometric reasoning for seamless task and motion planning. Burgess-Limerick et al. (2022) present a generalized architecture for reactive mobile manipulation on-the-move, enabling flexible and robust task execution across various *environments* while being adaptable to different robotic platforms with minimal modifications. The work of Yokoyama et al. (2023) presents *Adaptive Skill Coordination*, a framework that utilizes a library of basic visuomotor skills (navigation, picking, placing) alongside a skill coordination policy and a corrective policy to adapt and coordinate these skills for long-horizon tasks in diverse, unstructured real-world environments. The skills are trained entirely in simulation and deployed zero-shot on the Boston Dynamics Spot robot, demonstrating robust performance without needing detailed maps or precise object locations. Yenamandra et al. (2024) present *HomeRobot*, an integrated framework and benchmark for open-vocabulary mobile manipulation, enabling robots to navigate and manipulate a wide range of objects in diverse, multi-room household environments using a combination of simulation and real-world components. This approach leverages reinforcement learning and heuristic baselines to facilitate the generalization of robotic skills across different settings aimed at creating versatile household assistants. Similarly, the *OK-Robot* (Liu et al., 2024a) framework integrates vision-language models with navigation and grasping primitives to enable zero-shot pick-and-drop operations in novel home environments, utilizing pre-trained models and a modular approach for robust performance without additional training. Lastly, Ren et al. (2024) present a dual-arm manipulation framework that integrates a learning-based dexterity-reachability-aware perception module for autonomous bimanual grasping of unknown objects and an optimization-based versatility-oriented control module for real-time cooperative manipulation, ensuring system safety and adaptability. While this approach enables the *transfer* of skills across different *environments* due to its robust perception capabilities, it only parenthetically mentions the possibility to generalize to other *tasks* and *robots*.

Robot Transfer The *transfer* of manipulation skills across different *robotic platforms* is a critical aspect in the personal sector, as it facilitates robots learning from each other's experiences and failures to advance robotic *autonomy* and versatility in these scenarios. As an early example, the Acromovi architecture (Nebot and Cervera, 2007) is a distributed, agent-based software framework designed to seamlessly integrate and coordinate multiple heterogeneous robotic systems, such as mobile bases and manipulator arms, enabling flexible and reusable code across different robotic platforms. By leveraging agent wrappers and middleware, the framework facilitates component interaction and resource sharing, promoting efficient task execution and cooperation between diverse robotic elements. Furthermore, the work of Paikan et al. (2015) presents a framework for transferring object grasping skills between different humanoid robots, utilizing a bridge system to interconnect varied software frameworks and employing reactive correction behaviors to adapt grasp definitions to new robot embodiments. This approach enables the execution of grasping tasks across *robots* with different kinematics and middleware without extensive reprogramming. The paper by Koubaa et al. (2016) presents a service and hospitality-oriented software architecture for robotic assistants using ROS, featuring the COROS framework and ROS Web Services to enable modular, distributed, and easily extensible applications. This architecture abstracts robot control and application logic, facilitating interaction with different

ROS-enabled robots and integrating them with client applications via standard web service protocols. Additionally, Garcia et al. (2018) present SERA, a *Self-adaptive dEcentralized Robotic Architecture*, which enables decentralized collaboration and adaptation among heterogeneous *robots* through a three-layered modular design. This architecture allows for flexible integration and coordination of *robots* in various settings, supporting reusable components and efficient management across diverse robotic systems in varying *environments*. Recently, the MOSAIC framework (Wang et al., 2024) introduced a modular architecture that integrates pre-trained vision-language models and reinforcement learning to coordinate multiple *robots* and a human user for collaborative cooking tasks in a predefined kitchen environment. The system uses a task planner to convert high-level instructions into robot actions, enabling seamless interaction and execution across different *robots* but requiring environment-specific setups and independently trained task modules.

In conclusion, the *transfer* of manipulation skills across a single mode represents a first step in the evolution of robotic systems towards greater *autonomy* and adaptability. The research highlighted in this subsection underscores the diverse strategies and frameworks developed to address the challenges inherent in each *transfer* modality.

Dual-Mode Transfer

According to Jaquier et al. (2024), *dual-mode transfer* refers to changing two of the modes while keeping only a single mode constant. This could, for example, mean transferring knowledge from one robot performing a task in one environment to a different robot performing a similar task in a different environment. Some *Executive* frameworks have a greater focus on the adaptability and facilitate such a *transfer* of knowledge across two modes at the same time. This enables them to perform a wide array of *tasks* across various *environments* and platforms. This subsection delves into the research focusing on the *transfer* of knowledge and capabilities across two out of the three identified modalities.

Task-Robot & Task-Environment Transfer Multiple recent approaches combined *tasks* with another mode of *transfer* to improve the adaptability of robotic systems in *unstructured environments*. For example, the work of Yao et al. (2022) proposes a hierarchical control framework integrating disturbance predictive control with reinforcement learning and a forward model to adapt quadruped robots equipped with various robotic arms to different manipulation tasks. This approach enables the system to predict and mitigate disturbances from different robotic arms, facilitating skill *transfer* across *tasks* and *robots*, with partial adaptability to different *environments*. Additionally, Martins et al. (2023) introduce LOLA, a user-centric framework for robotic manipulation that utilizes a robot-agnostic affordance library and high-level task authoring, enabling seamless task creation and execution across various robotic platforms and related tasks. Contrarily, Kasaei and Kasaei (2024) present a modular approach to robotic manipulation by integrating pushing, grasping, and throwing actions through model-free deep reinforcement learning, enabling robots to autonomously manage cluttered environments. The framework, tested in both simulated and real-world scenarios, demonstrates the effective *transfer* of learned skills across different *environments* and related *tasks* but is tailored to one specific robot.

Environment and Robot Transfer Because *task* transferability is still largely unexplored in *Executive* frameworks, many approaches investigated the *transfer* across *environments* and *robots* to improve versatility in mobile manipulation. Dömel et al. (2017) introduce a modular, hierarchical framework for autonomous mobile manipulation that enhances flexibility and adaptability, enabling the *transfer* of components across different *environments* and robotic platforms with advanced perception and path planning. Furthermore, the *Generative Attention Learning* (GenerAL) (Wu et al., 2020) framework leverages deep reinforcement learning to perform high-DoF multi-fingered grasping by using a single depth image to generate 6-DoF grasp poses and finger joint angles, ensuring robustness across different robotic hands and various cluttered environments with novel objects. This approach demonstrates high adaptability across varied robotic platforms without the need for additional training. The unified software framework for intelligent home service and hospitality robots, introduced by Yi et al. (2020), presents a modular, general-purpose software framework for intelligent mobile manipulation robots, enabling seamless adaptation across different *robots* and *environments* by integrating navigation, perception, manipulation, and HRI modules. This framework supports the robust execution of various service tasks with minimal adjustments, as demonstrated in international robot competitions. Similarly, Nam et al. (2020) present a modular software architecture for autonomous service robots that integrates deep learning-based perception, symbolic reasoning, AI task planning, and geometric motion planning, all implemented in ROS. This architecture enables robots to perform manipulation and navigation tasks in varied settings by autonomously generating and executing task plans based on contextual knowledge without extensive reprogramming. SkillFusion (Staroverov et al., 2023), a hybrid framework combining classical and learning-based modules for visual ObjectGoal Navigation (ObjectNav), dynamically selects the optimal skill for navigation tasks. This approach ensures robustness and adaptability across different *environments* and robotic platforms without the need for extensive retraining or reprogramming. Lastly, the RoboScript framework (Chen et al., 2024) generates deployable robot manipulation code from natural language instructions by integrating perception tools, motion planning, and ROS-based simulation, enabling seamless execution across various *environments* and robotic platforms. It leverages a unified code generation pipeline and hierarchical agent architecture to handle complex tasks with minimal reprogramming, facilitating adaptability and knowledge *transfer* across different *robots* and tasks.

Affordance Templates A special mention deserves the Affordance Templates (ATs, Hart et al., 2014) framework, which places special emphasis on the *transfer* of mobile manipulation skills. Introduced by Hart et al. (2014), it provides a graphical 3D environment for specifying and adjusting robot task goals aimed at improving HRI through shared autonomy. In Hart et al. (2015), the framework is extended by developing a ROS package that standardizes task descriptions in a robot-agnostic format, enabling easy application of the same templates to different *robots* with minimal configuration changes. This package includes integration with motion planning tools like *MoveIt!*, facilitating more efficient task execution across varied robotic platforms. The work of Hart et al. (2022) further advances the framework to support generalized mobile manipulation, introducing autonomous grasp determination and perceptual registration, thus improving the *transfer* of tasks across different *environments* without extensive reprogramming. Throughout these developments, the framework maintains its flexibility and adaptability for different *robots* and related tasks while progressively enhancing environmental *transfer* capabilities. However, an AT has to be created for each task separately, hindering a *transfer* of capabilities and knowledge across *tasks*.

The ability to *transfer* knowledge across two modes constitutes significant progress towards achieving adaptable and versatile robotic systems in service and hospitality, healthcare and nursing, and domestic applications. Through the lens of the discussed research, it is evident that the field is moving towards a future where robots can seamlessly adapt their manipulation skills across different *tasks*, *environments*, and *robotic platforms*. This progress is crucial for the development of robots capable of operating in the dynamic and unpredictable real world. While the achievements in these domains are noteworthy, they also highlight the complexity of achieving full transferability, especially when considering the integration of all three modalities.

2.3.2. Planning Frameworks

The *Planning* layer usually transfers knowledge and adapts to the current circumstances via goal specifications. Conventionally, this is done via the definition of a goal state for a planner (e. g., using PDDL). However, on an even higher abstraction level (Jaquier et al., 2024), this *transfer* happens through verbal instructions. In the domain of NLP, transformer-based LLMs that are trained on a very large corpus of text have been shown to excel at understanding and generalizing natural language instructions (Zhao et al., 2023) – facilitating *transfer* as the by-product of a large amount of training data. As planning is not the focus of this thesis and giving a proper review of the current state of the art for planning frameworks would go beyond the scope of this work, only the recent trends of integrating LLMs in the *Planning* layer will be shortly recapitulated. Therefore, the following section will introduce the current state of the art in task and manipulation planning using LLMs and is taken from the publication about the topic by Birr et al. (2024).

Disclaimer

Parts of the content presented in this section were previously published in:

Timo Birr, **Christoph Pohl**, Abdelrahman Younes, and Tamim Asfour (2024). "AutoGPT+P: Affordance-based Task Planning with Large Language Models". In: *Proceedings of Robotics: Science and Systems*. Robotics: Science and Systems. Vol. 20. Delft, Netherlands

Recently, LLMs have shown significant advancements, even surpassing human performance in numerous areas. Despite these achievements, their capability for coherent reasoning remains limited (Valmeekam et al., 2022). Nevertheless, there are numerous recent instances of LLMs being employed in task planning for robotic mobile manipulation. Sarkisyan et al. (2023) identify three primary operational modes: *subtask evaluation*, *full autoregressive plan generation*, and *step-by-step autoregressive plan generation*. This categorization is only applicable when the LLM itself functions as the planner. Conversely, recent research (such as Liu et al., 2023a) proposes an alternative model where LLMs are used to generate symbolic goal descriptions and are paired with a conventional planner, referred to as *LLM with Planner*. This section aims to delineate the various methodologies and classify them according to Figure 2.5.

Subtask Evaluation

This mode uses the LLM as a scoring model, selecting the optimal subtask from predefined options by scoring all possibilities and choosing the best one based on

Note: Adapted from Birr et al. (2024).

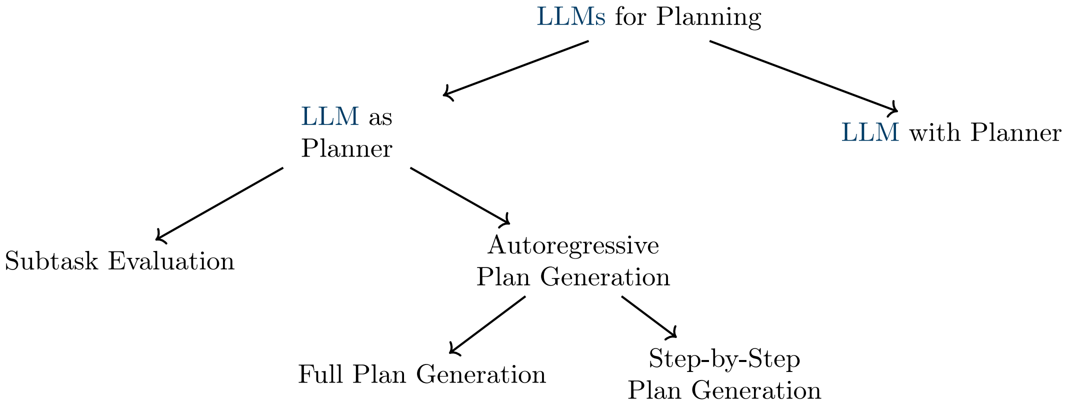


Figure 2.5.: A taxonomy of LLMs in planning tasks with the related work from this section referenced.

the current state. The model output is constrained to specific tokens to ensure executability. While resource-intensive, this method reduces computational complexity by filtering subtasks using environmental constraints and common-sense rules (Sarkisyan et al., 2023). A key example is SayCan (Brohan et al., 2023c), which combines a Reinforcement Learning-based affordance function, which evaluates action viability within the environment, with an LLM to estimate action success. Plans are built incrementally by selecting actions with the highest combined scores. Zhao et al. (2024) extends SayCan with a greedy next-best action evaluation and tree search, enhancing Monte Carlo Tree Search planning through an LLM-based common-sense heuristic.

Full Autoregressive Plan Generation

In this mode, the LLM can generate complete plans from prompts. However, while less resource-intensive than *subtask evaluation*, this often results in unreliable plans due to mismatched actions, free-form outputs, and logical inconsistencies. Wu et al. (2023d) and Wake et al. (2023) address this by grounding the LLM with object lists, restricting actions to listed items. To improve performance, Song et al. (2023) use dynamic in-context retrieval for replanning. Furthermore, Rana et al. (2023) reduce complexity by filtering irrelevant objects via 3D scene graph traversal. Zhou et al. (2023) translate problems into PDDL, enabling plan validation and self-correction through a feedback loop. Proposing a hybrid method combining *subtask evaluation* with *full plan generation*, Lin et al. (2023) use semantic checks and a greedy algorithm to meet goal conditions.

Step-By-Step Autoregressive Plan Generation

Here, the LLM generates and executes subtasks iteratively, balancing feasibility and computational complexity by grounding actions to the environment, which

Table 2.5.: Overview of the taxonomy for planning using LLMs.

	Subtask Evaluation	Full Plan Generation	Step-by-Step Plan Generation	LLM with Planner
Bärmann et al. (2024)	○	○	●	○
Brohan et al. (2023c)	●	○	○	○
Chen et al. (2023b)	○	○	○	●
Ding et al. (2023)	○	○	○	●
Driess et al. (2023)	○	○	●	○
Guan et al. (2023)	○	○	○	●
Huang et al. (2022a)	○	○	●	○
Huang et al. (2022b)	○	○	●	○
Liang et al. (2023)	○	○	●	○
Lin et al. (2023)	●	●	○	○
Liu et al. (2023a)	○	○	○	●
Rana et al. (2023)	○	●	○	○
Singh et al. (2023)	○	○	●	○
Song et al. (2023)	○	●	○	○
Wake et al. (2023)	○	●	○	○
Wang et al. (2023b)	○	○	●	○
Wu et al. (2023b)	○	○	●	○
Wu et al. (2023d)	○	●	○	○
Xie et al. (2023)	○	○	○	●
Zhao et al. (2024)	●	○	○	○
Zhou et al. (2023)	○	●	○	○
Birr et al. (2024)	○	○	●	●

Note: Adapted from Birr et al. (2024).

improves accuracy compared to *full plan generation*, but requires new prompts at each step. As an example for this, Huang et al. (2022a) address the problem of grounding LLMs in real-world scenes via a dual-step approach: a planning-LLM generates an ungrounded plan, which a translation-LLM subsequently adapts to the robot's capabilities. In a follow up work, Huang et al. (2022b) improved upon this by using incremental planning, where feedback after each step enhances performance. Other works, such as ProgPrompt (Singh et al., 2023), use LLMs to generate plans in Python code, with feedback from error messages aiding corrections. Similarly, TidyBot (Wu et al., 2023b) refines code by identifying patterns in prior iterations. Furthermore, Bärmann et al. (2024) present a system that learns from human corrections to improve future plan generation. Wang et al. (2023b) propose a four-step method where the LLM iteratively plans, revises based on failures, and selects actions to optimize execution. Finally, Driess et al. (2023) present an embodied version of the Pathways Language Model (PaLM, Chowdhery et al., 2023), named PaLM-E, enabling multi-modal task planning using visual and robot state inputs.

LLM with Planner

In this mode, the LLM generates a PDDL goal state from a natural language task, which is then used by a conventional planner. Xie et al. (2023) introduced this approach, demonstrating LLMs's effectiveness in translating language into PDDL goals, though performance drops for complex tasks. Expanding on this, LLM+P (Liu et al., 2023a) generates both goals and problems, improving success by using minimal domain examples. Liu et al. (2024b) enhance LLM+P with scene graphs for initial problem states and subgoal decomposition for faster planning. Guan et al. (2023) propose LLM-based problem and domain generation, using syntactic feedback to correct errors and a hybrid approach for faster planning. Furthermore, AutoTAMP (Chen et al., 2023b) replaces PDDL with Signal Temporal Logic, integrating automatic error correction in the form of plans insufficient to achieve the goal, which are identified by the LLM. Ding et al. (2023) extend LLMs to open-world scenarios, allowing for dynamic action generation and affordance-based planning to handle unforeseen situations.

2.3.3. Discussion

The three-tiered robot architecture is a standard paradigm for building robotic software frameworks. Having an adaptable task *execution* is majorly influenced by the flexibility of the software architecture. Therefore, this section focuses on comparing approaches in the *Executive* and *Planning* layer of the three-tiered robot architecture.

In their review about transfer learning in robotics, Jaquier et al. (2024) introduce the three modes of transferability: *robot*, *task*, *environment*. Accordingly, relevant *Executive* frameworks for mobile manipulation were investigated regarding their ability to *transfer* knowledge and experience across these modes. Works like Borghesan et al. (2014); Iovino et al. (2023b); Wang et al. (2023a) facilitate *transfer* across a single mode, while other approaches like Dömel et al. (2017); Kasaei and Kasaei (2024); Staroverov et al. (2023) can handle two modes of *transfer*. However, only very few works try to address all three modes (e. g., Hart et al., 2022; Yao et al., 2022), while none put an explicit focus on the flexibility that this *transfer* can bring to the *execution* process of mobile manipulation. To this end, Section 5.1 introduces an *Executive* framework that explicitly targets the *transfer* of knowledge, experience and skills across the three modes.

For the *Planning* layer, related works were analyzed with respect to the role that LLMs play in the creation of a plan. Some approaches, like Brohan et al. (2023c); Zhao et al. (2024) use the LLM to select the best subtask to execute next in order to reach a goal. Works like Lin et al. (2023); Rana et al. (2023); Wake et al. (2023) use LLMs directly to generate entire plans, while others like Bärmann et al. (2024); Driess et al. (2023); Huang et al. (2022b) use it to only generate the next step. Finally, some approaches have combined LLMs with conventional planners to capitalize on the advantages of both sides. However, most of these approaches are limited by the closed-world assumption, lack of automated error correction, and deterministic modeling, which restrict their ability to handle dynamic, uncertain environments. Additionally, these approaches struggle with adaptability, feedback integration, and generating long plans, reducing their effectiveness in real-world applications. Therefore, Section 5.2 will introduce a hybrid *step-by-step* approach in combination with a conventional planner that addresses some of these limitations.

3. Versatile Grasp Discovery using Visual Perception in Unstructured Environments

In this chapter, approaches for flexible grasp synthesis in *unstructured environments* using visual perception are revisited to address Research Question 1. This is essential for interacting with *unstructured environments* with minimal *task-specific knowledge*. To this end, methods related to Contribution 1 will be introduced. These methods enable versatile grasp *discovery* that does not require full object knowledge, facilitating work in dynamic and cluttered settings with incomplete scene information. This is especially important for real-world applications in the context of this thesis' main objective. In real-world scenarios, it is often not reasonable to assume complete knowledge of objects for grasping, as the appearance of objects might change over time or depend on the state of the object or its environment. Therefore, reducing the amount of prior knowledge required for grasping and manipulation in these settings can greatly enhance the versatility and subsequently the *autonomy* of robotic assistants. Therefore, this chapter introduces an affordance-based action *discovery* method for unknown objects. Second, for cases where information about the semantic object class is available, an additional method is introduced that capitalizes on available similarities of objects of the same semantic class.

Disclaimer

Parts of the content presented in this chapter were previously published in:

- **Pohl, Christoph** and Tamim Asfour (2022). "Probabilistic Spatio-Temporal Fusion of Affordances for Grasping and Manipulation". In: *IEEE Robotics and Automation Letters* 7.2, pp. 3226–3233
- Cai, Yichen, Jianfeng Gao, **Christoph Pohl**, and Tamim Asfour (2024). "Visual Imitation Learning of Task-Oriented Object Grasping and Rearrangement". In: *Proc. of the 2024 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*. International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, UAE: IEEE/RSJ, accepted for publication

For handling objects that share characteristic traits with other instances of the same class (e. g., "cups" or "bottles"), referred to as similar objects, a framework for task-specific *grasping* based on the Multi-feature Implicit Model is introduced in Section 3.1. The model's capability to encode multiple spatial features and improve object shape reconstruction from partial observations increases the success rate of object grasping and rearrangement tasks in *unstructured environments*. Moreover, an approach for *grasping* and manipulating unknown objects based on the local surface geometry is detailed in Section 3.2. Its efficiency in cluttered environments is established by using surface metrics such as curvature and normal direction to define affordances like *graspability*, *pushability*, and *placability*. In addition to the theoretical foundations, extensive experiments demonstrate the improvements in grasp success rates and versatility and validate the practical applicability and robustness of both methods in diverse and dynamic scenarios.

3.1. Task-Specific Grasp Synthesis for Similar Objects

For applications in the service industry, especially in domestic environments, many objects are not completely unknown. They often have similar shapes and functional parts, which are shared for every instance of that class. For example, cups usually have handles, an opening on top and are flat on the bottom. Along with shared geometrical features, affordances are shared between the instances of a class. In the case of the cup, that would be *graspability*, *fillability*, and *placability*, among others. To allow for versatile grasping and manipulation in scenarios where the class of the object is known, an approach for task-oriented grasping and rearrangement was established.

To this end, a novel neural network based on neural fields – called Multi-feature Implicit Model (MIMO, Cai et al., 2024) – was developed that facilitates the transfer of poses from a canonical object to newly observed (potentially incomplete) instances of the same class. Using this network, it is possible to transfer grasp poses to novel instances that have been extracted from human observation or autonomously generated by a grasp generation network trained in simulation. Therefore, the approach improves the versatility of grasp synthesis by capitalizing on knowledge from a canonical instance.

The following sections will describe the development and application of MIMO for task-oriented grasp generation, which has been previously published in Cai et al. (2024). The aim is to showcase how MIMO addresses Research Question 1 and therefore contributes to the main objective of this thesis.

3.1.1. Motivation

In order to handle everyday tasks in real-world applications, robotic assistants should be able to use knowledge about common objects in these scenarios and generalize it to unseen situations. Truly versatile general-purpose robots need to be able to execute tasks, even if the objects involved have not been encountered before. However, especially in the personal sector, many objects share common features with other instances from the same class. General-purpose robots should be able to handle a task independent of the concrete instance of such a similar object, as the affordances involved do not change. Therefore, robots should be able to transfer the task-specific knowledge from one object instance to another, as long as they are similar. In the case of grasping, the robot must identify the optimal grasps for specific tasks and generate an appropriate motion trajectory to achieve the desired configuration. For example, a side grasp by the mug handle is ideal for pouring water from a mug, while a top grasp by the rim is more suitable when placing the mug into a container to avoid collision between the hand and the container. To engage Research Question 1, a novel neural network is used to transfer task-specific grasping knowledge across different instances of the same semantic class, thereby improving the versatility of the grasp synthesis for similar objects.

Previous methods for generating task-oriented grasps have concentrated on training neural networks using large, manually annotated datasets (see Section 2.1.1). Despite their effectiveness, these methods cannot generalize to new objects with significant shape variations. Furthermore, manual annotation is both costly and challenging to obtain. In contrast, visual imitation learning approaches offer efficient means to teach robots manipulation skills based on human demonstrations, enabling generalization to new scenarios with categorical objects. Neural Descriptor Fields (NDFs), which implicitly encode the spatial properties of objects, have proven very successful in this regard (Simeonov et al., 2023, 2022). They can be trained in a self-supervised manner by leveraging an inherent bias towards object classes, thus eliminating the need for manual annotation (Hidalgo-Carvajal et al., 2023). This bias is crucial for establishing dense 3D correspondences across categorical objects, enabling the adaptation of object manipulation skills to previously unseen object instances (Biza et al., 2023; Huang et al., 2023). However, these approaches often require multiple object views, which are not always available in real-world applications (Kerr et al., 2023; Rashid et al., 2023). When presented with a partial view or categorical objects with significant shape variations, these approaches may produce less precise grasp candidates, potentially leading to collisions or unstable object placements (Hidalgo-Carvajal et al., 2023).

To address these challenges, MIMO is designed to predict multiple spatial properties of a 3D point relative to an object. This allows the model to generate a richer descriptor space and thus more precise dense correspondences than similar approaches, facilitating the accurate transfer of grasps and object target poses to new situations. MIMO can also reconstruct object shapes from partial observations, which is beneficial for coping with task constraints defined on the hidden part of the object. Leveraging MIMO's capabilities, a framework that efficiently learns and generates task-oriented grasps from single or multiple human demonstration videos is proposed. Additionally, an evaluation network is used to predict the success probability of the generated grasps and refine them if necessary.

Therefore, the contributions to improving the versatility of robotic assistants are twofold: (1) A novel neural network called MIMO is introduced that predicts multiple spatial features of a point relative to an object, yielding an informative point and pose descriptor space. It outperforms similar NDF methods in terms of shape reconstruction and pose transfer. The model can be trained in a self-supervised manner without relying on human annotations. (2) MIMO is integrated into a visual imitation learning framework to learn, generate, and refine task-oriented grasps efficiently. It achieves one- and few-shot imitation learning and demonstrates a direct transfer of the learned manipulation tasks to categorical objects.

3.1.2. Multi-feature Implicit Model

Task-oriented grasping is an important skill for robotic assistants to deal with the various scenarios they encounter in everyday applications. Leveraging MIMO's strengths in measuring pose similarities and transferring poses, a framework is introduced to learn task-specific grasping and object rearrangement from human demonstrations. This framework can generate optimal grasp poses for new object instances based on partial observations, addressing the need for robots to handle diverse and dynamic environments. By training a model that can transfer task-specific knowledge between different object instances of the same class in simulation entirely without human labeling, this approach enhances the robot's ability to perform flexible grasping and manipulation tasks in unstructured settings and increases its *autonomy*, thereby answering Research Question 1.

Neural Network Architecture

The network architecture, as visible in Figure 3.1, MIMO employs a shared Vector Neurons-PointNet encoder $\epsilon(\mathbf{P})$ (Deng et al., 2021) to embed the geometric information of the point cloud $\mathbf{P}$ into an equivariant latent code, and a partly shared

Note: Adapted from Cai et al. (2024). © 2024 IEEE.

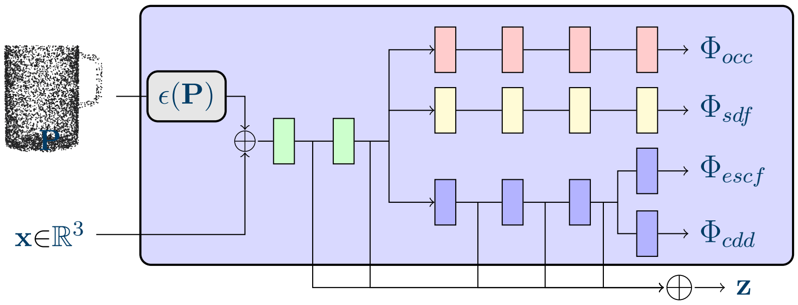


Figure 3.1.: Network architecture of MIMO.

Multi-Layer Perceptron (MLP) decoder with multiple branches to represent SO(3)-invariant spatial relations of a point $\mathbf{x}$ relative to $\mathbf{P}$. The occupancy Φ_{occ} (Mescheder et al., 2019) and signed distance Φ_{sdf} (Park et al., 2019) branches enable precise shape reconstruction, with the occupancy branch facilitating mesh construction using the Multi-resolution IsoSurface Extraction algorithm (Mescheder et al., 2019). To enhance the neural field's capability in capturing geometric details and direction awareness, two novel feature branches, Extended Space Coverage Feature (ESCF) and Closest Distance Direction (CDD), are introduced. The ESCF branch Φ_{escf}, using the coefficients of spherical harmonics expansion across all orders and degrees for supervision, captures finer geometric details compared to the Space Coverage Feature (SCF, Zhao et al., 2016) branch used in similar approaches (e. g., Huang et al. (2023)). The CDD branch Φ_{cdd}, defined as the inner product of unit vectors $\mathbf{v}_d$ and $\mathbf{v}_p$, where $\mathbf{v}_d$ points from a point $\mathbf{x}$ to the closest point on the object and $\mathbf{v}_p$ follows a chosen principal direction, enhances direction-awareness. The combined descriptor space, trained with four branches (Φ_{occ}, Φ_{sdf}, Φ_{escf}, Φ_{cdd}), is more informative and precise in distinguishing fine geometric details and, therefore, better at measuring geometric similarity.

Pose Descriptor and Transfer

The generation of pose descriptors is necessary to transfer task-relevant information from a canonical model of the object class to unseen instances of that same class. For every point $\mathbf{x}$ in the point cloud $\mathbf{P}$, the point descriptor $\mathbf{z}$ is obtained by concatenating the activation layers of the partially-shared decoder for ESCF and CDD. The Basis Point Set (BPS, Prokudin et al., 2019) sampling strategy is then used to create point descriptors for a set of points around an object, which are concatenated to form the pose descriptor $\mathbf{Z}$. Specifically, for a set of points $\mathbf{X} \in \mathbb{R}^{N\times 3}$ sampled from a rigid object $\mathcal{O}_B$ in pose $\mathbf{T}$ around the point cloud $\mathbf{P}_A$ of object $\mathcal{O}_A$, the pose descriptor of $\mathcal{O}_B$ is derived using the trained MIMO of object category A, denoted as ${}^{A}\mathbf{Z}_{B} = \varphi(\mathbf{T}, \mathbf{X}|\mathbf{P}_{A})$. This encodes poses relative to $\mathcal{O}_{A}$ in a way that similar poses exhibit a small L1 distance between their pose descriptors. For instance, in the context of grasping, $\mathcal{O}_{A}$ could be a "mug" and $\mathcal{O}_{B}$ the hand, with ${}^{A}\mathbf{Z}_{B}$ representing a grasp pose. For improved accuracy of the pose descriptor for partially visible objects, the object's mesh is first reconstructed using the occupancy and signed distance branches. Subsequently, a reference point cloud is sampled which is used for the calculation of the pose descriptor ${}^{A}\mathbf{Z}_{B} = \varphi(\mathbf{T}, \mathbf{X}|\mathbf{P}^{r}_{A})$.

To enable the transfer of grasp candidates across similar objects, it is necessary to identify corresponding poses on two different instances of the same class. To this end, the pose of the object in question can be optimized to be as similar as possible to a reference pose descriptor. Given a trained MIMO for object category A and a reference pose descriptor ${}^{A}\hat{\mathbf{Z}}_{B}$, the pose of a new instance of category B ($\bar{\mathcal{O}}_{B}$) is optimized relative to a new instance of category A ($\bar{\mathcal{O}}_{A}$) by solving

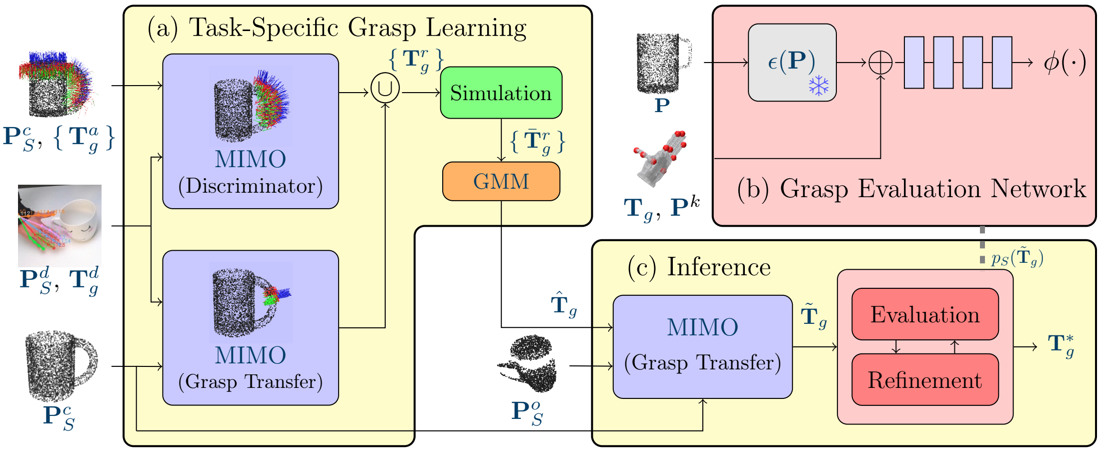
$$\mathbf{T}^{*} = \arg\min_{\mathbf{T}} \|\varphi(\mathbf{T}, \mathbf{X}|\bar{\mathbf{P}}^{r}_{A}) - {}^{A}\hat{\mathbf{Z}}_{B}\|_{1} \,, \tag{3.1}$$

where $\bar{\mathbf{P}}^{r}_{A}$ is the reconstructed point cloud of $\bar{\mathcal{O}}_{A}$. The optimization procedure follows the approach in Simeonov et al. (2022).

In a visual imitation learning setup, the reference pose descriptor can be obtained from human demonstration videos by tracking the hand pose with respect to the object point cloud. In terms of grasping similar objects, the new optimal grasping pose (i. e., the most similar grasp pose to the demonstration), can be found using this reference pose descriptor and Eq. (3.1) to find the pose $\mathbf{T}^{*}$ of $\bar{\mathcal{O}}_{B}$ with respect to the new instance of category A that corresponds to the reference pose. Using the trained MIMO for object category A, a reference pose descriptor ${}^{A}\hat{\mathbf{Z}}_{B}$ (e. g., from human observation), and a new object instance ($\bar{\mathcal{O}}_{A}$), the pose of $\bar{\mathcal{O}}_{B}$ (e. g., the robot's hand) is optimized relative to $\bar{\mathcal{O}}_{A}$ (the new object) by minimizing the L1 distance between the inferred pose descriptor and the reference pose descriptor.

3.1.3. Pick and Place Framework

The framework for performing pick-and-place actions leverages the strengths of MIMO in measuring pose similarities and transferring poses to learn task-specific grasps and object rearrangements from one or multiple human demonstrations. The framework can generate optimal grasp poses for new object instances based on partial observations, as illustrated in Figure 3.2. This capability is crucial for enabling robotic assistants to adapt to diverse and *unstructured environments*, thereby enhancing their *autonomy* and effectiveness in real-world applications.



Note: Adapted from Cai et al. (2024). © 2024 IEEE.

Figure 3.2.: Proposed MIMO-based grasp framework.

To learn grasps from human observation, demonstration videos consisting of sequences of RGB and depth images of a manipulation task are used. Hand poses in all frames are estimated using the approach from Lin et al. (2021). Subsequently, a Via-point Movement Primitive (VMP, Zhou et al., 2019) is trained to represent the hand motion. The grasping time step t_g and the demonstrated grasp pose $\mathbf{T}_g^d \in \mathrm{SE}(3)$ is determined following Gao et al. (2024). The source object $\mathcal{O}_S$ is the one being grasped, while the target object $\mathcal{O}_T$ sets a reference frame for placing $\mathcal{O}_S$ at the last time step t_T. Segmented point clouds of both objects at t_g and t_T are obtained using Grounded SAM (Kirillov et al., 2023; Liu et al., 2023b).

In order to facilitate a task-specific synthesis of grasp candidates, the information obtained from human observation needs to be generalized so that it can be used for all other instances of the same class. To this end, a Riemannian Gaussian Mixture Model (GMM) that represents the task-specific grasps is trained in simulation. For training the GMM, a set of task-agnostic grasp poses $\{\mathbf{T}_g^a\}$ is generated using the method from Sundermeyer et al. (2021) on a canonical point cloud $\mathbf{P}_S^c$ for the class of the source object $\mathcal{O}_S$. From this, two strategies are employed to obtain task-relevant grasp candidates for simulation: (i) using MIMO as a discriminator for pose similarity to find the most similar grasps in $\{\mathbf{T}_g^a\}$ to $\mathbf{T}_g^d$, or (ii) using MIMO to directly transfer the demonstrated grasp $\mathbf{T}_g^d$ to $\mathbf{P}_A$ using Eq. (3.1). The combined set of task-relevant grasp poses $\{\mathbf{T}_g^r\}$ is then simulated with a humanoid hand in Isaac Gym (Makoviychuk et al., 2021). Specifically, a grasp is considered successful if the object is picked up and does not drop after being subjected to random shaking. The rearrangement of the object is subsequently simulated based on the successful grasps. If the task is completed successfully, the grasp is added to the set of successful task-relevant grasp poses $\{\bar{\mathbf{T}}_g^r\}$. These grasps are used to train a GMM on a Riemannian manifold (i. e., $\mathbb{R}^3 \times \mathcal{S}^3$). This GMM can then be used to generate task-oriented grasps during inference.

Even though the GMM would suffice to generate and execute task-oriented grasp candidates, a task-agnostic grasp evaluation network is proposed to ensure the quality of the sampled grasps. This network computes the success probability of a grasp pose $\mathbf{T}_g$ relative to an arbitrary point cloud $\mathbf{P}$. First, $\mathbf{P}$ is encoded using the frozen encoder $\epsilon(\mathbf{P})$ of MIMO. Subsequently, a MLP decoder conditioned on this encoding predicts the success probability given a set of keypoints on the humanoid hand, representing its pose. The model is trained using a binary cross-entropy loss on a dataset that includes all task-agnostic grasp candidates from all tasks and their binary success labels.

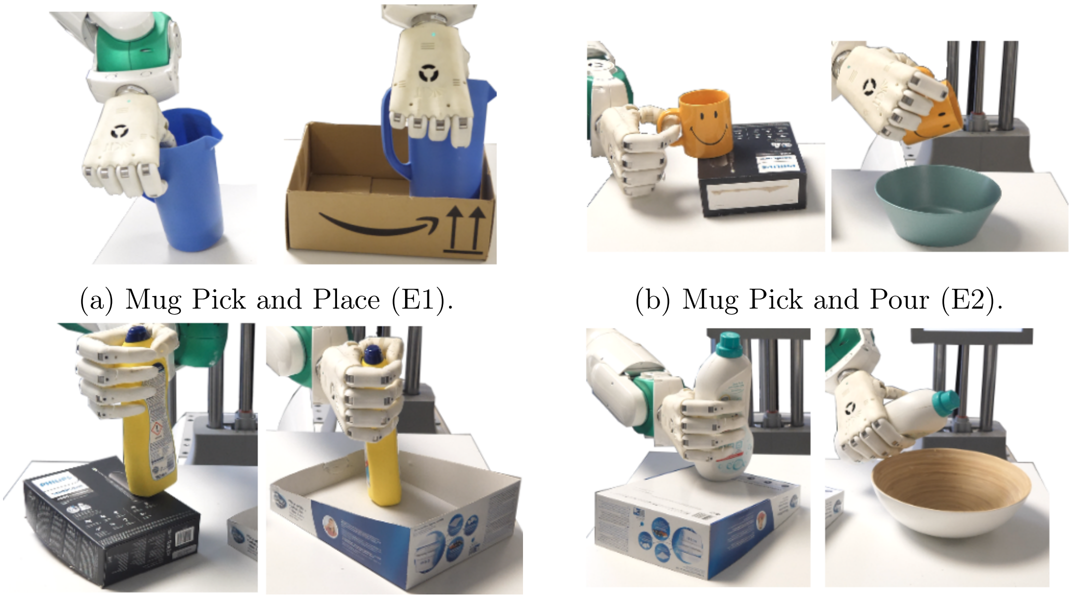
During inference, grasp poses $\hat{\mathbf{T}}_g$ are sampled from the trained GMM for the task at hand relative to the canonical point cloud $\mathbf{P}^c_{\mathcal{S}}$ and transferred to a partially-observed point cloud $\mathbf{P}^o_{\mathcal{S}}$ of a novel categorical instance. The success probability $p_S(\tilde{\mathbf{T}}_g)$ of a transferred grasp pose $\tilde{\mathbf{T}}_g$ is computed using the trained task-agnostic grasp evaluation network. If the success probability is below a certain threshold, the grasp pose is refined by maximizing the grasping success likelihood using the evaluation network, resulting in the optimal grasp pose $\mathbf{T}^*_g$. This inference process ensures that the generated grasps are not only task-relevant but also optimized for success in diverse and *unstructured environments*.

3.1.4. Experiments

To assess the proposed task-oriented grasp generation framework, multiple experiments across various manipulation tasks were conducted using the humanoid robots ARMAR-6 and ARMAR-DE. More details, evaluation videos, and source code are available via the project page[1].

In addition to the real-world *pick-and-place* experiments, multiple ablation studies comparing the efficacy of MIMO to similar state-of-the-art approaches were performed by Cai et al. (2024). For the sake of completeness, they are listed in Appendix B. The ablation studies demonstrate MIMO's superior performance across a range of manipulation tasks when compared to similar approaches such as Neural Descriptor Field (NDF, Simeonov et al., 2022), Relational-Neural Descriptor Field (R-NDF, Simeonov et al., 2023), and Neural Interaction Field and Template (NIFT, Huang et al., 2023). In simulation, MIMO consistently achieves higher grasp and placement success rates, particularly in scenarios involving arbitrary

[1] https://sites.google.com/view/mimo4



Note: Adapted from Cai et al. (2024). © 2024 IEEE.

(a) Mug Pick and Place (E1).

(b) Mug Pick and Pour (E2).

(c) Bottle Pick and Place (E3).

(d) Bottle Pick and Pour (E4).

Figure 3.3.: Example visualizations of the real-world experiments on ARMAR-DE using MAkEable (see Section 5.1).

object starting poses. A key advantage of MIMO lies in its ability to generate accurate grasps and placements even for objects with challenging geometries, such as bottles and mugs, which often pose difficulties for other methods due to failures in distinguishing between top and bottom orientations. Additionally, Cai et al. show that the integration of shape reconstruction in the framework is shown to significantly improve performance, particularly in tasks that require precise object placement. By leveraging enhanced shape descriptors and novel neural features, MIMO excels in SE(3)-equivariant manipulation tasks, providing a more robust and versatile solution for object grasping and rearrangement in complex environments.

The real-world experiments were conducted using the humanoid robots ARMAR-6 and ARMAR-DE to evaluate the proposed MIMO-based task-oriented grasp generation framework. The setup involved four specific tasks with only a single, partial view on the object: (E1) grasping a mug at its rim and placing it upright, (E2) grasping a mug at its handle and pouring into a bowl, (E3) grasping a bottle at its neck and placing it upright, and (E4) grasping a bottle at its body and pouring it into a bowl. For all experiments, MIMO was used to reconstruct object shapes from the partially-observed point cloud. The grasp poses are sampled from the GMM, transferred to the observed objects, and evaluated by the grasp evaluator (see Section 3.1.3). If the estimated success probability dropped below 0.9, the grasp pose was optimized with a learning rate of 10^{-3}.

An Azure Kinect camera mounted on the robot head provided RGB and depth images to extract object point clouds. For ARMAR-DE, the manipulation tasks were validated and executed using the mobile manipulation framework MAkEable (Section 5.1), while ARMAR-6 utilized a task-space impedance controller to execute motions generated by learned movement primitives. The target poses corresponded to the grasp pose during the grasp phase and the object rearrangement pose during the placement or pouring phase. Qualitative results are shown in Figure 3.3 and showcase the applicability of the MIMO-based task-oriented grasping framework to real-world scenarios.

3.2. Affordance-based Action Extraction using the Local Surface Geometry

The MIMO-based approach described in Section 3.1 can successfully transfer manipulation knowledge across instances of the same class. However, it does not facilitate grasping objects without any prior knowledge. To allow for versatile grasping and manipulation in scenes with unknown objects, a flexible method for affordance extraction, called the Geometry-based Action Extraction (GAE), based on the local surface geometry of point clouds, is developed. This approach applies the concept of discriminative grasp synthesis to affordance extraction by analyzing every point in the point cloud, calculating the local surface curvature, and assigning affordances based on heuristics of the principal curvatures and normals at the point. Therefore, this approach completely decouples affordance extraction from the notion of "objects" and is ideal for grasping and manipulation of unknown objects.

Furthermore, by extracting affordances using heuristics on the local surface information of supervoxels, a uniform and coherent state can be defined for different affordances based on a local, geometry-aware coordinate frame (see Section 4.1). The approach is evaluated through grasping experiments with the humanoid robot ARMAR-6, demonstrating an improved grasping success rate. The following sections will delve into the details of the Geometry-based Action Extraction method and its implications for robotic manipulation in *unstructured environments*.

The content of this section has already been published in the paper Pohl and Asfour (2022) and will now be put into the context of this thesis.

3.2.1. Motivation

The interaction of robotic assistants with unstructured and unknown environments based on visual information remains a difficult task. It requires a detailed understanding of the scene and the objects therein to allow the generation of appropriate actions that can be executed in a given situation. Especially for real-world applications in the personal sector, the ability to interact with cluttered, unstructured, and dynamic scenes is a key requirement for robots to operate *autonomously*. In these contexts, robots must handle a wide range of different tasks without relying on *task-specific knowledge*. Therefore, increasing the versatility of grasping and manipulation in *unstructured environments* is a requirement to advance the *autonomy* of robotic assistants (i. e., the main objective of this thesis). To this end, the local surface geometry of point clouds is used to extract action hypotheses without any form of prior knowledge about the scene or its objects.

While Section 3.1 introduced an approach for the grasping and rearrangement of similar objects, it is not always feasible to assume that knowledge about an object's class exists. As a possible solution for this, the *representationalist* view on affordances (see Appendix A) decouples possible actions in a scene from the concept of "objects". By interpreting the scene as a set of *entities* that are connected to possible *behaviors* of the robot, it facilitates an object-agnostic scene understanding and, therefore, the *discovery* of interaction possibilities for unknown objects. Previous affordance-based approaches for grasping and manipulation rely either on overly simplistic scene representations (e. g., primitive shapes; Kaiser et al. (2017) and Kaiser and Asfour (2018)) or require large manual efforts (e. g., for dataset generation; Song et al. (2016)). Accordingly, Yamanobe et al. (2017) find in their review that simulation-based approaches often fail to accurately represent real-world physics, leading to performance gaps, while manually defining affordances (i. e., in an ontology) is labor-intensive and may not generalize well to novel scenarios. Additionally, learning from human demonstration is slow and limited to specific tasks, requiring significant human effort, while real-world interaction and exploration can be risky, slow, and heavily dependent on noisy sensor data. There exists a large body of research that investigates *grasping* of unknown objects (see Section 2.1.2 and Table 2.2). However, these approaches are either tailored to specific scenarios or require large amounts of training data with limited potential for generalization to unseen situations. Contrarily, using the local surface geometry of depth data (as done by e. g., ten Pas et al. (2017)) to extract coordinate frames along the Principal Curvature Directions, so-called Darboux frames, poses a unique opportunity to extract affordances and a corresponding 6D end-effector pose for

Note: Adapted from Pohl and Asfour (2022). © 2022 IEEE.

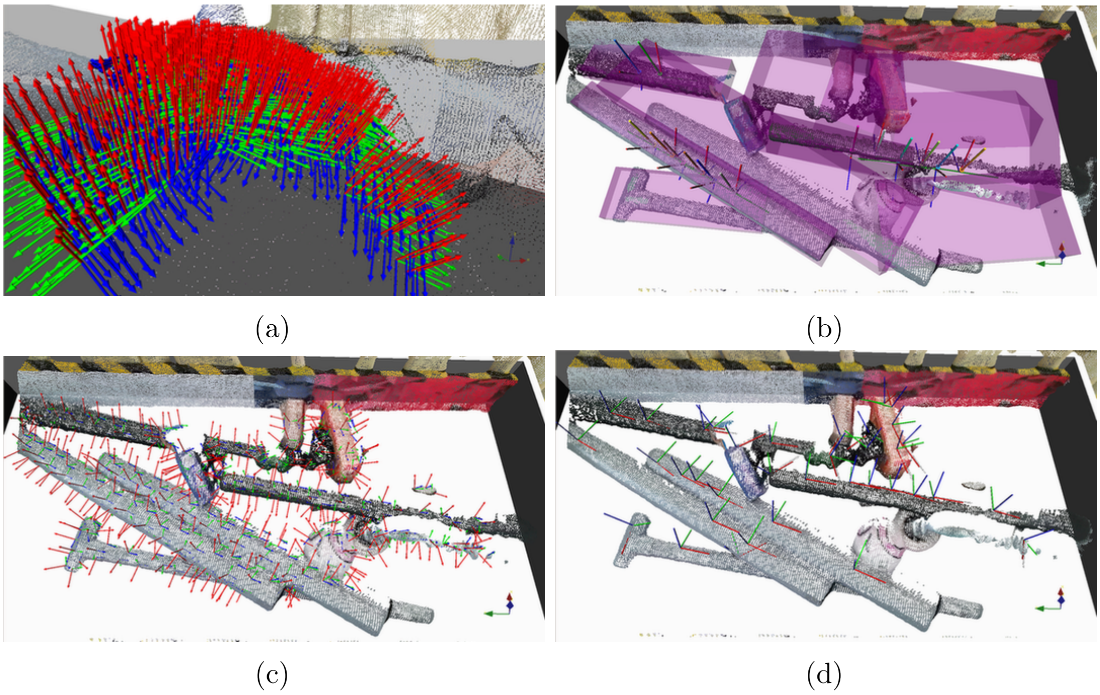


Figure 3.4.: Overview of the Geometry-based Action Extraction. In (a) the surface normals and principal directions for each point in a surface patch are displayed. (b) shows the clustered supervoxels and (c) the averaged surface information. (d) shows the extracted action observations.

the *discovery* of versatile interaction possibilities without the need for *task-specific knowledge*.

To improve the versatility of robotic assistants, the Geometry-based Action Extraction (GAE) – an affordance-based approach to grasping and manipulation of unknown objects in *unstructured environments* using the local surface geometry – is developed. By employing quadric approximations of the local surface of a point cloud, the Principal Curvature Directions are extracted and used to heuristically derive affordances for surface patches. Quadric surface approximations enhance the accuracy of vision-based manipulation in noisy point clouds by using a universal parameterization of surface patches for all shapes and affordances. These affordances are associated with an Abstract Affordance Frame (AAF), a Darboux frame that is uniquely defined for every point of the point cloud, allowing for the direct generation of end-effector poses. By doing so, GAE decouples action *discovery* from the notion of objects, making it ideal for the grasping and manipulation of unknown objects in diverse and dynamic environments. Thus, the geometry-based extraction of affordances ensures versatile action generation and is particularly beneficial for robots operating under realistic conditions and incomplete knowledge. An overview of the approach can be seen in Figure 3.4.

3.2.2. Geometry-based Action Extraction

The Geometry-based Action Extraction for flexible grasping and manipulation in *unstructured environments* is divided into four primary steps: (i) First, the local surface geometry of a point cloud is examined by fitting quadric patches to the neighborhood of every point to compute the surface normal and principal curvatures. (ii) This enables the extraction of locally consistent surface patches through a modified supervoxel clustering technique. (iii) Next, the averaged geometrical features of these supervoxels are used to heuristically define affordances and (iv) establish a spatiotemporally consistent coordinate system (referred to as the AAF in this thesis) for each patch. By focusing on the local surface geometry and extracting a shared state for all affordances, GAE improves the versatility of grasping and manipulation actions and builds the foundation for applying recursive Bayesian estimation in Section 4.1.

Local Surface Approximation

In order to analyze the local surface structure of the raw point clouds and obtain a geometrical representation of the scene, quadric patches are fit to each point in the point cloud. Quadrics (see e. g., Kobayashi and Nomizu (1996) or Hartshorne (2013)) are D-dimensional hypersurfaces embedded in a space of dimension $(D+1)$, where $D = 2$ in this case. Fitting quadrics to the local neighborhood of each point allows the noisy surfaces to be treated as functions. Consequently, methods of differential geometry allow for the closed-form calculation of essential metrics for the local surface geometry, like the surface normal and the principal curvatures (see e. g., Patrikalakis and Maekawa, 2010, Chapter 3). An overview and description of the used metrics can be found in Table 3.1.

Table 3.1.: Overview of the most important surface metrics.

Metric	Symbol	Description
Surface Normal	$\mathbf{n}$	Vector of the direction that is perpendicular to the surface at the origin
Principal Directions	$\lambda_\pm$	Direction of the extremal values of curvature of a surface at the origin. Representing the local extrema of the curvature values.
Principal Curvatures	$\kappa_\pm$	Amount of curvature at the origin of the surface in the directions $\lambda_\pm$, respectively.

Note: Reprinted from Pohl and Asfour (2022). © 2022 IEEE.

To efficiently approximate the local surface of a raw point cloud as quadrics, the GPU implementation of Spek et al. (2017) is employed. This method yields the surface surface normal $\mathbf{n}$, the principal curvatures $\kappa_\pm$, and the second fundamental form coefficients L, M, N.

In order to calculate the AAF in later steps, the principal directions $\lambda_\pm$ are of great importance, as they are used in combination with $\mathbf{n}$ to calculate the orientation of the frame. To be able to calculate $\lambda_\pm$, the *first* and *second fundamental form* of a parametric surface $\mathbf{r} = \mathbf{r}(u, v)$ are needed:

$$\begin{aligned} \mathbb{I} &= E\,du^2 + 2F\,du\,dv + G\,du^2 \\ \mathbb{II} &= L\,du^2 + 2M\,du\,dv + N\,du^2 \end{aligned}$$

In their work, Spek et al. (2017) use a local neighborhood around each point in the cloud to fit a parametric paraboloid of the form

$$\begin{aligned} \mathbf{r}(u,v) &= \begin{pmatrix} u \\ v \\ \frac{L}{2}u^2 + Muv + \frac{N}{2}v^2 \end{pmatrix}, \text{or} \\ z &= \frac{L}{2}u^2 + Muv + \frac{N}{2}v^2 \,. \end{aligned}$$

From this, it can be seen that at the origin:

$$E = \mathbf{r}_u \cdot \mathbf{r}_u = 1; \qquad F = \mathbf{r}_u \cdot \mathbf{r}_v = 0; \qquad G = \mathbf{r}_v \cdot \mathbf{r}_v = 1$$

The vector $\mathbf{k}_{\lambda_\pm} \in \mathbb{R}^3$ of the principal directions $\lambda_\pm = \frac{du}{dv}$ represents the three-dimensional direction of the greatest and smallest curvatures in Euclidean space. $\mathbf{k}_{\lambda_\pm}$ can be obtained from the definition of the principal curvatures through the surface parameters:

$$\begin{aligned} \kappa_\pm &= \frac{M + N\lambda_\pm}{F + G\lambda_\pm} = \frac{L + M\lambda_\pm}{E + F\lambda_\pm} \\ \kappa_\pm &= \frac{M + N\lambda_\pm}{\lambda_\pm} = L + M\lambda_\pm \\ \lambda_\pm &= -\frac{M}{N - \kappa_\pm} = -\frac{L - \kappa_\pm}{M} \end{aligned} \tag{3.2}$$

Now, $\mathbf{k}_{\lambda_\pm}$ can be calculated as

$$\mathbf{k}_{\lambda_\pm} = \begin{pmatrix} 1 \\ \lambda_\pm \\ \frac{\mathbf{n}_x + \lambda_\pm\,\mathbf{n}_y}{\mathbf{n}_z} \end{pmatrix}$$

due to the orthogonality of the principal directions to the surface normal $\mathbf{n}$.

Supervoxel Clustering

Calculating the principal directions and principal curvatures would be theoretically enough to extract affordances and the AAF. However, as only a small neighborhood is taken into account for the calculation of the surface metrics, they are still strongly affected by noise in the point cloud. Therefore, the next step consists of extracting small surface patches with uniform appearance and geometry to calculate averaged surface metrics. This is achieved by using the clustering algorithm described in Papon et al. (2013), which adheres to object boundaries and provides an over-segmentation of the scene into so-called supervoxels. As it is reasonable to assume that points with similar local surface properties, such as the principal directions, belong to one semantic segment and, therefore, share affordances, the implementation in the Point Cloud Library (PCL, Rusu and Cousins, 2011) was modified to incorporate the previously extracted local surface metrics.

A supervoxel $V = (\mathbf{t}, \mathbf{c}, \mathbf{n}, \mathbf{k}_{\lambda_-}, K)$ represents a cluster of similar points and is defined using the averaged features of all points within it. Here, $\mathbf{t} \in \mathbb{R}^3$ denotes the position, $\mathbf{c} \in [0 \ldots 255]^3$ the color, $\mathbf{n} \in \mathbb{R}^3$ the surface normal, $\mathbf{k}_{\lambda_-} \in \mathbb{R}^3$ the direction of minimal curvature, and $K = \kappa_+ \cdot \kappa_-$ the Gaussian curvature. Starting with an initial seeding, supervoxels are iteratively grown based on the distance d_{vccs} between two adjacent voxels V_1 and V_2 in the feature space[2], given by

$$d_{vccs} = \alpha||\mathbf{t}_2 - \mathbf{t}_1|| + \beta||\mathbf{c}_2 - \mathbf{c}_1|| + \gamma(1 - |\mathbf{n}_1 \cdot \mathbf{n}_2|),$$

where α, β, γ are scaling constants.

To better account for the local surface structure, the feature space was extended by incorporating the principal directions and Gaussian curvatures of the point cloud. Consequently, the new distance metric in the feature space is defined as

$$d_{aug} = d_{vccs} + \delta(1 - |\mathbf{k}_{\lambda_-,1} \cdot \mathbf{k}_{\lambda_-,2}|) \cdot |K_2 - K_1|,$$

where δ is an additional scaling constant. This augmented distance metric ensures that the clustering process is more sensitive to the local surface geometry.

Heuristic Affordance Extraction

Threshold-based decision functions are used for defining affordances based on the local surface geometry, similar to Varadarajan and Vincze (2013). In the prior work of Kaiser et al. (2016), such functions were utilized to determine affordances on geometric primitives extracted from the environment. In this work, heuristics

[2] Note that this is the distance used in the PCL implementation, in Papon et al. (2013) a 39-dimensional feature space is used.

derived from the local surface geometry are employed to extract affordances. For instance, flat surfaces with surface normal anti-parallel to the direction of gravity are assumed to afford *placability* or *supportability*. Additionally, convex objects, as defined by the curvature direction convention in Patrikalakis and Maekawa (2010) with $\kappa_{-} \leq 0$, are considered to afford *graspability*. An overview of the heuristics used for affordance extraction can be found in Table 3.2.

Table 3.2.: Surface metrics for the definition of affordances.

Surface Metric	**Affordances**		
	Graspability	***Pushability***	***Placability***
Mean Surface Curvature	Convex	–	Flat
Mean Normal Direction	Upper Hemisphere	Horizontal	Upwards
Volume	< Grasp Volume	–	> Object/ Threshold

Note: Reprinted from Pohl and Asfour (2022). © 2022 IEEE.

Using this approach, every point in the point cloud has the necessary information to extract affordances. However, as this is often not necessary (except for e. g., a teleoperated affordance extraction procedure as the one used in Section 5.1.3), the averaged metrics of the supervoxel can be used to extract more robust affordances. This reduces the noise in the extracted affordances and does not influence the accuracy too much, as the supervoxels adhere to object boundaries.

Action Generation in Local Coordinate System

In addition to the extraction of affordances, the local surface geometry facilitates the definition of a unique, spatiotemporally coherent, local coordinate system, referred to as Local Curvature Frame in Pohl and Asfour (2022), that can be used as an abstract reference frame in which end-effector poses can be constructed based on the assigned affordances of the supervoxel. In the context of this thesis, such a frame that is connected to and represents one or more affordances at a specific point of an object's surface will be called Abstract Affordance Frame (AAF), as the concept extends to other applications, where the frame is not necessarily extracted from the local curvature. This AAF – which is a Darboux frame – can be uniquely defined at any non-umbilical point on a parametric surface, using the surface normal and the direction of curvature, as these are always orthogonal for differential surfaces (Patrikalakis and Maekawa, 2010).

Using the extracted surface metrics, the pose $\mathbf{T}$ of the AAF in the global coordinate frame can be derived as

$$\mathbf{T} = \begin{bmatrix} | & | & | & | \\ \mathbf{k}_{\lambda_+} & \mathbf{k}_{\lambda_-} & \mathbf{n} & \mathbf{t} \\ | & | & | & | \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

To show that the pose $\mathbf{T}$ is uniquely defined for any non-umbilical point on a parametric surface, an alternative formulation of the principal curvatures to Eq. (3.2) in terms of the Gaussian curvature K and the mean curvature $H = \frac{\kappa_+ + \kappa_-}{2}$ is used:

$$\kappa_\pm = H \pm \sqrt{H^2 - K} \tag{3.3}$$

Note that the argument of the square root in Eq. (3.3) is always ≥ 0. Depending on the geometry of the surface, there are three different cases to consider, which influence the definition of $\mathbf{T}$:

1. $\sqrt{H^2 - K} = 0$ and $H = K = 0$
2. $\sqrt{H^2 - K} = 0$ and $H \neq K \neq 0$
3. $\sqrt{H^2 - K} \neq 0$

Case 1 is the trivial flat case, where both principal curvatures are equal to 0. The second case happens when $H^2 = K$, and implies that $\kappa_+ = \kappa_- = \kappa_\lambda$. In this case, the curvature is the same for every direction λ, which happens, for example, on the surface of a sphere or on saddle points. The last case implies $\kappa_+ \neq \kappa_- \neq 0$. Here, the principal directions are uniquely defined, as is $\mathbf{T}$. In the other two cases, which correspond to umbilical points on surfaces, $\mathbf{T}$ is defined up to a rotation around the surface normal $\mathbf{n}$.

The AAF serves as a universal reference frame for manipulation actions regardless of their affordance. In this way, it provides the possibility of constructing end-effector poses for different actions using the same reference pose. Therefore, it facilitates the definition of a universal and coherent state for affordance-based manipulation actions in this framework, which will be fundamental for the spatiotemporal fusion of action observations in Section 4.1. For example, a grasp candidate could be generated in such a way that the fingers of the hand align with the minimal curvature direction (i. e., y-axis of the AAF), while the forward direction of the hand for a push candidate could be aligned with the surface normal.

3.2.3. Experiments

To assess the performance of GAE, a series of real-world grasping experiments were conducted using the humanoid robot ARMAR-6. The experiments were carried out in two distinct cluttered setups: a box-emptying and a table-clearing scenario. In both scenarios, unknown objects were randomly placed, and the robot was tasked with grasping and manipulating these objects to either empty the box or clear the table. These setups were designed to test the versatility of the approach in *unstructured environments*, directly addressing the Research Question 1. A video detailing the approach and experiments is available online[3].

In both experimental setups, top-grasp candidates were generated using GAE, as well as the approach based on Object-Oriented Bounding Box (OOBB) from Grimm et al. (2021) combined with a region-growing segmentation as a comparison. The OOBB-based grasp candidate extraction was chosen as a baseline in order to compare the efficiency and precision of GAE. For the supervoxel clustering, the parameters were set to $\alpha = \beta = \gamma = \delta = 5$. The respective feature distances were normalized, ensuring that each parameter contributed equally. Once grasping action candidates were generated, each one was tested for reachability by solving the IK. Based on the grasp candidate's orientation, the mobile robot base was positioned appropriately, and the optimal hand was chosen for executing the grasp. Among all valid candidates, the highest was selected, executed, and recorded for reference. Future candidates were favored if they fell outside a small region around each previous candidate, enhancing grasp variability and preventing repetitive execution. Reaching motions for execution were generated using VMPs.

Box Emptying Experiments

For the box-emptying experiments, a varying number of unknown objects were randomly placed inside a box. For each candidate generation method, 30 grasp attempts were performed, and the results were recorded. This process was repeated across five different setups, with the number of objects ranging from 6 to 14 per setup. The objects included simple shapes like boxes and cylinders, as well as more complex items such as bent pipes, a hammer, and a spray bottle. To increase variability, the object configuration was changed after every five grasp attempts by rearranging or exchanging objects through a human operator. Successful grasps involved lifting the object and dropping it from a height of 30 cm before the next attempt. This interaction aimed to reduce the bias of the human operator when creating the scenes and increase the randomness of the object configurations.

Grasp attempts were categorized based on their execution outcomes and failure reasons. The categories included "grasped", "stable lifted", "lifted", "collision",

[3] https://youtu.be/lXxWtTIySB0

Table 3.3.: Possible outcomes of grasping attempts.

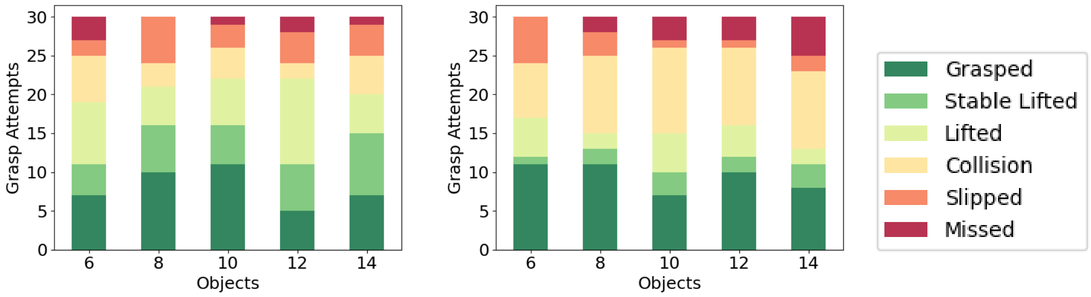
Outcome	Description
Grasped	The object does not touch the ground for 5 seconds
Stable Lifted	The object is lifted for 5 seconds but parts of the object still touch the ground
Lifted	The object is visibly lifted for less than 5 seconds
Collision	The object is not lifted because the hand collides with other objects or the environment (e. g., box)
Slipped	The object is not lifted because the hand slipped off the object / was misaligned
Missed	The grasp is generated incorrectly, no object is close enough to be grasped or no executable grasp is found after 2 minutes

Note: Reprinted from Pohl and Asfour (2022). © 2022 IEEE.

"slipped", and "missed"; each with specific criteria for classification. The descriptions of all categories can be found in Table 3.3. Additionally, the time from candidate generation to selection was measured for each attempt.

The results of these experiments, as depicted in Figure 3.5 show that GAE performs better than the OOBB-based grasp extraction across all degrees of clutter. Counting only the "grasped" and "stable lifted" categories as successful, the GAE method achieved an average success rate of 46.0%, while the OOBB method had a success rate of 38.7%. There was no strong correlation between the number of objects and successful grasps for both methods ($\rho_{GAE} = 0.18$ and $\rho_{OOBB} = -0.42$), while for grasps in the "missed" category the OOBB method showed a significant correlation ($\rho_{GAE} = -0.28$ and $\rho_{OOBB} = 0.96$). This indicates that while both methods can generate good candidates in cluttered scenes, GAE consistently performed well even in difficult scenarios. The use of local surface geometry, independent of point cloud segmentation, positively influenced the accuracy of the approach, confirming the initial hypothesis.

Note: Adapted from Pohl and Asfour (2022). © 2022 IEEE.



(a) GAE-based Grasping (b) OOBB-based Grasping

Figure 3.5.: Results of the box emptying experiments on ARMAR-6

Table Clearing Experiments

The table-clearing experiments' aim was to investigate the ability of GAE to perform well across different *environments* and *tasks*, showcasing the versatility of the approach. The experiment involved clearing an 80 cm × 80 cm table cluttered with 18 diverse objects, including boxes, cups, plates, and various fruits. The process required the robot to stow these objects into a box. If a grasp attempt failed or an object was dropped, the grasping process was manually restarted by a human operator. The grasp candidate generation, selection, and execution followed the same protocol as in the box-emptying scenario. Both methods were tested by clearing the table five times, recording the number of objects successfully stowed, the total time taken, and the number of grasp attempts required. The experiment concluded either when all objects were removed or if no executable grasp was found within five minutes.

Table 3.4.: Results of the table clearing experiments.

	GAE	**OOBB**
Stowed Boxes	4.6 ± 1.7	4.2 ± 0.8
Stowed Plates	2.8 ± 0.4	1.6 ± 1.1
Stowed Cups	2.0 ± 1.0	2.2 ± 0.8
Stowed Fruit	2.4 ± 1.1	1.8 ± 1.3
Total Stowed	11.8 ± 1.5	9.8 ± 1.1
Remaining Objects	0.8 ± 0.8	5.2 ± 2.4
Grasp Attempts	37.0 ± 2.2	29.4 ± 6.1
Total Time [min]	33:11 ± 2:31	25:24 ± 2:50

Note: Adapted from Pohl and Asfour (2022). © 2022 IEEE.

The results of the table-clearing experiments, as shown in Table 3.4, again highlight the performance of the GAE method compared to the OOBB method. The GAE method stowed on average 11.8 ± 1.5 objects and required 37.0 ± 2.2 grasp attempts with 0.8 ± 0.8 remaining on the table. The OOBB method showed slightly lower performance, with an average of 9.8 ± 1.1 objects stowed and a much higher number of remaining objects (5.2 ± 2.4). The total number of grasp attempts, however, was significantly lower as for GAE. This was caused by the OOBB experiments being aborted early because either all objects dropped from the table or no executable grasp candidates could be found. These results underscore the versatility and effectiveness of the GAE method in *unstructured environments*, despite the higher number of grasp attempts required.

3.3. Conclusion

General-purpose robots need to be able to deal with incomplete or missing information regarding the various objects that are commonly found in *unstructured environments*. For some objects, knowledge about the semantic class is enough to infer possible interaction possibilities with other instances. In those cases, the Multi-feature Implicit Model and the associated pick-and-place framework can successfully provide the required versatile task-oriented grasping and rearrangement capabilities. However, for cases where robots have to deal with completely unknown objects, the Geometry-based Action Extraction represents a viable solution for increasing the versatility of the *discovery* of interaction possibilities.

Section 3.1 introduced a novel implicit neural field designed to enhance task-oriented grasp generation for similar objects, called the Multi-feature Implicit Model (MIMO). By providing SE(3)-equivariant point and pose descriptors, MIMO enables a more precise shape similarity measure, which is crucial for effective grasping and manipulation. The network is trained on multiple spatial features, including occupancy and signed distance, as well as the novel Extended Space Coverage Feature and Closest Distance Direction feature, allowing it to detect finer correspondences and achieve more accurate pose transfers compared to existing methods. Additionally, MIMO supports shape reconstruction to handle partial observations, further improving its robustness in *unstructured environments*. MIMO is integrated into a task-oriented grasping and object rearrangement framework, which includes a novel evaluation and refinement network to boost success rates. The versatility of MIMO for manipulating similar objects is demonstrated through its performance in one- and few-shot visual imitation learning experiments for pick-and-rearrangement tasks on the humanoid robots ARMAR-6 and ARMAR-DE.

For the more general grasping and manipulation of unknown objects, Section 3.2 introduced the Geometry-based Action Extraction (GAE) method for extracting scene affordances based on the local surface geometry of point clouds and subsequently generating manipulation actions. The approach involves analyzing the local surface curvature and assigning affordances using heuristics derived from the principal curvatures and surface normal at each point. By defining a unique and spatiotemporally coherent abstract reference frame, termed the Abstract Affordance Frame in this thesis, which is derived entirely from the local surface geometry, the method decouples affordance extraction from the concept of "objects", making it suitable for the grasping and manipulation of unknown objects.

GAE was evaluated in various real-world grasping scenarios using the humanoid robot ARMAR-6, demonstrating its versatilitys in *unstructured environments*. The proposed method consistently outperformed the OOBB-based baseline approach across different levels of scene clutter, resulting in an increase of almost 10% in grasp success rate. Additionally, the success rate of this method was largely independent of the degree of clutter of a scene, highlighting its ability to enable versatile manipulation in complex and cluttered environments.

The ability to handle similar and unknown objects, and, therefore, deal with incomplete object knowledge, directly contributes to the overarching main objective of this dissertation, which is to increase the *autonomy* of robotic assistants for deployment in real-world applications. By enabling more versatile and accurate grasp synthesis using visual perception, MIMO and GAE address the critical challenge of versatility in dynamic and *unstructured environments* by decreasing the amount of *task-specific knowledge* required. Therefore, in response to the Research Question 1, this work demonstrates that robotic assistants can adapt to diverse and *unstructured environments* for effective grasping and manipulation by leveraging local surface geometry for affordance extraction or similarities that are shared across instances of the same class.

4. Selection of Reliable Grasp Candidates for Unknown Objects using Probabilistic Methods

Building on the contributions from Chapter 3 (see also Contribution 1), this chapter deals with improving the reliability (Core capability 2) of grasping and mobile manipulation in *unstructured environments*. To this end, the details of Contribution 2 are revisited and further elaborated. Specifically, the main focus of this chapter is to enhance the robustness of the second step of *discriminative grasping* (see Figure 1.1) – the *selection* of the best grasp or action hypothesis in the current scene. Doing so improves the trustworthiness and safety of robotic assistants in applications that revolve around humans by increasing the degree of *autonomy* such a system has. Integrating probabilistic and statistical methods into the grasp *selection* process enhances the tolerance to *uncertainties* in perception and increases the success rate of executions. Therefore, by answering Research Question 2, it becomes possible to decrease the susceptibility to errors and environmental influences.

Disclaimer

Parts of the content presented in this chapter were previously published in:

- **Pohl, Christoph** and Tamim Asfour (2022). "Probabilistic Spatio-Temporal Fusion of Affordances for Grasping and Manipulation". In: *IEEE Robotics and Automation Letters* 7.2, pp. 3226–3233
- Baek, Woo Jeong, **Christoph Pohl**, Philipp Pelcz, Torsten Kroger, and Tamim Asfour (2022). "Improving Humanoid Grasp Success Rate Based on Uncertainty-Aware Metrics and Sensitivity Optimization". In: *IEEE-RAS International Conference on Humanoid Robots*. Vol. 2022-Novem, pp. 786–793

In Section 4.1, the shared state for affordances introduced in Section 3.2 is used in combination with recursive Bayesian estimation to track an action hypothesis over multiple observations of a scene in order to increase its robustness. Specifically, this facilitates the estimation of the covariance and the existence certainty for the associated pose of an action hypothesis. These probabilistic measures for the certainty and accuracy of an extracted affordance are subsequently used in Section 4.2 to calculate a grasp score that combines multiple *uncertainty-aware* grasp quality metrics that can be used to select the most reliable grasp candidate in a scene.

4.1. Spatiotemporal Action Fusion using Bayesian Recursive State Estimation

To increase the robustness of manipulation actions for unknown objects, a spatiotemporal fusion approach for affordances and their respective actions was developed. To this end, the coherent, uniquely defined reference frame introduced in Section 3.2 is used in combination with techniques from recursive Bayesian estimation to get a probabilistic estimate of the perceptual uncertainty involved in the affordance extraction. Using a Unscented Kalman Filter (UKF, Wan and Van Der Merwe, 2000) on Lie groups and a Hidden Markov Model (HMM), the state of affordances is tracked in the scene, and an updated estimation of the involved *uncertainty* is provided. Therefore, the approach described in this section aims at answering Research Question 2 and contributes towards the main objective of increasing the *autonomy* of robotic assistants.

The content of the following sections has already been published in the paper of Pohl and Asfour (2022). Toward the objective of situating the Probabilistic Action Extraction and Fusion within the context of this thesis, the most important aspects will be reexamined hereafter.

4.1.1. Motivation

The interaction of autonomous robots with unstructured and unknown environments presents significant challenges, particularly when relying on visual perception alone and dealing with incomplete information. This task requires a comprehensive and precise interpretation of the scene to select appropriate actions for execution, which is crucial for enhancing the *autonomy* of robots in human-centric, real-world applications. Affordances represent a valuable framework for enabling robots to identify potential actions based on visual perception. However, existing affordance-based methods often suffer from noise and perceptual *uncertainties*, leading to unreliable actions. To improve the reliability of robotic interactions in dynamic and unstructured environments, the Probabilistic Action Extraction and Fusion (PAEF, Pohl and Asfour, 2022) is proposed, a probabilistic approach for estimating the pose and existence certainty of action hypotheses by tracking a related coherent frame through multiple observations.

Current affordance-based methods for robotic manipulation face multiple limitations, especially in dealing with noisy perception and missing information. PAEF builds on prior work on the formalization of affordances as *Dempster-Shafer* belief over the space of end-effector poses (Kaiser et al., 2018) and their extraction (Kaiser and Asfour, 2018). Using this formulation, information from different sources can be fused to hierarchically define affordances on primitive shapes (e. g., spheres, cylinders, and boxes). Even though this formulation allows for the calculation of a degree of certainty in the existence of an affordance for a discrete end-effector pose, it cannot improve the state estimate over multiple observations, and, therefore, correct for perceptual and proprioceptive *uncertainties*. The use of the local surface geometry of point clouds to improve the versatility of affordance extraction for unknown objects has been described in Section 3.2. There, a uniquely defined and spatiotemporally consistent reference frame referred to as the Abstract Affordance Frame (Section 3.2.2), was extracted and used as a universal and continuous state for all affordances. Although *Recursive State Estimation* is a well-understood problem in robotics (Thrun et al., 2006), estimating the 6D pose remains difficult, as conventional recursive filters in Euclidean space cannot easily handle orientations. However, ready-to-use algorithms for the fusion of poses have recently been developed based on recursive Bayesian estimation (Brossard et al., 2017; Sjøberg and Egeland, 2021).

By employing these methods, PAEF combines multiple observations of a scene to improve the state estimate of affordances, and, thereby, increases the reliability of grasping and manipulation. It uses a coherent, geometrically inspired reference to fuse information in a shared state for affordances over multiple observations to correct for *uncertainties* using a combination of a UKF and HMM. By doing so, PAEF enhances the robustness and fault-tolerance of the extracted action hypotheses, in turn increasing the degree of *autonomy* in grasping and manipulation tasks. Multiple real-world grasping experiments in industrial and domestic scenarios were conducted on the humanoid robot ARMAR-6 to showcase the improved success rates. Section 4.1.2 will introduce the theoretical background for the spatiotemporal fusion of action hypotheses, while Section 4.1.3 showcases the real-world experiments.

4.1.2. Probabilistic Action Extraction and Fusion

In Section 3.2, the Geometry-based Action Extraction (GAE), a method for the extraction of scene affordances and the definition of their unique and coherent state based on the local surface geometry of point clouds, was presented. The basic approach of GAE is split into the four steps (i) estimation of the principal curvatures, (ii) the extraction of locally consistent surface patches, (iii) heuristic definition of affordances, and (iv) establishment of a spatiotemporally consistent coordinate system (referred to as the Abstract Affordance Frame (AAF) (Section 3.2.2)). Additionally, it was shown that the AAF is uniquely (up to a rotation around the surface normal in edge cases) defined for every point of the point cloud, and, therefore, spatially and temporally coherent. Consequently, the AAF is an ideal way of tracking affordance-related action candidates over multiple observations of a scene.

These properties of the AAF can now be used to define the coherent state of an action observation $\mathbf{A}$:

$$\mathbf{A}_t = (\mathbf{T}, t, \{a_1, \ldots, a_n\}).$$

Formally, an action observation $\mathbf{A}$ is linked to the AAF with pose $\mathbf{T}$ at time step $t \in \mathbb{R}^+$ and is associated with n affordances a_i. In terms of the *representationalist* formulation (see Appendix A) of affordances, it represents a potential *behavior* of the robot that can be executed at a certain point on an *entity* in the scene at one specific time (i. e., a single point cloud captured by the camera).

Given that interaction possibilities in a scene can appear and disappear at any time, such as when an object is removed, it is insufficient to model only the action's pose $\mathbf{T}$ and its uncertainty. Therefore, PAEF introduces an additional measure for the existence certainty of an action. Hence, a combination of a UKF (for the spatial filtering of action observations) with an HMM (for the temporal tracking of the existence certainty) is used for the complete probabilistic state estimation of the actions. To this end, an action hypothesis $\bar{\mathbf{A}}$ is formed by combining multiple distinct action observations $\mathbf{A}$:

$$\bar{\mathbf{A}}_t = (\bar{\mathbf{T}}, \mathbf{\Sigma_T}, t, \{p_E^{a_1}, \ldots, p_E^{a_m}\}),$$

where $p_E^{a_i}$ represents the existence certainty of the hypothesis for the i-th affordance a_i, and $\mathbf{\Sigma_T} \in \mathbb{R}^{6\times6}$ is the covariance matrix of the filtered mean pose $\bar{\mathbf{T}}$ with the time of the last observation t.

When estimating hypotheses from multiple observations of a scene, two main challenges arise:

(a) For each new action observation, an appropriate action hypothesis must be identified. Specifically, a hypothesis $\bar{\mathbf{A}}_{t-1}$ that matches the current action $\mathbf{A}_t$ needs to be found:

$$\mathbf{A}_t^i \rightarrow \bar{\mathbf{A}}_{t-1}^j \, .$$

(b) The current observation $\mathbf{A}_t$ must be combined with the previous state estimation $\bar{\mathbf{A}}_{t-1}$ to estimate the new state $\bar{\mathbf{A}}_t$ of the action hypothesis:

$$\bar{\mathbf{A}}_t^i \leftarrow \bar{\mathbf{A}}_{t-1}^i \oplus \mathbf{A}_t^j \, .$$

Correspondence Search

To identify correspondences between different action observations, the fundamental assumption is that observations of the pose $\mathbf{T}$ of an action hypothesis $\bar{\mathbf{A}}$ are Gaussian-distributed around the mean pose $\bar{\mathbf{T}}$. For the positional component $\mathbf{t}$ of the pose, this is expressed by a multivariate Gaussian Probability Density Function (PDF), conditioned on the correspondence C:

$$p(\mathbf{t}|C) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma}_\mathbf{t})}} \exp\left(-\frac{1}{2}(\mathbf{t}-\bar{\mathbf{t}})^T {\boldsymbol{\Sigma}_\mathbf{t}}^{-1}(\mathbf{t}-\bar{\mathbf{t}})\right) \, , \tag{4.1}$$

where $\bar{\mathbf{t}}$ is the mean position of the action observation and $\boldsymbol{\Sigma}_\mathbf{t}$ is its covariance matrix.

Since the orientational part $\mathbf{R}$ of the pose of an action is an element of the special orthogonal group SO(3) and cannot naturally be represented by Eq. (4.1), the standard PDF requires adaptation. Following Solà et al. (2018), local perturbations on the Lie group SO(3) can be used to model orientation uncertainty. A Lie group $\mathcal{G}$ is a smooth manifold $\mathcal{M}$ that locally resembles a linear space, with a unique Euclidean tangent space at each point $\mathbf{Y}$. Probability distributions on $\mathcal{G}$ can be modeled by defining $\mathbf{Y}$ as a perturbation with $\boldsymbol{\tau}$ around the mean point $\bar{\mathbf{Y}}$ in its tangent space $\mathcal{T}_{\bar{\mathbf{Y}}}\mathcal{M}$. Thus, $\mathbf{Y}$ and its covariance matrix $\boldsymbol{\Sigma}_\mathbf{Y}$ can be expressed in terms of $\boldsymbol{\tau}$:

$$\begin{aligned}
\mathbf{Y} = \bar{\mathbf{Y}} \oplus \boldsymbol{\tau} \quad &:= \bar{\mathbf{Y}} \circ \operatorname{Exp}(\boldsymbol{\tau}) && \in \mathcal{M} \\
\boldsymbol{\tau} = \mathbf{Y} \ominus \bar{\mathbf{Y}} \quad &:= \operatorname{Log}(\bar{\mathbf{Y}}^{\text{-1}} \circ \mathbf{Y}) && \in \mathcal{T}_{\bar{\mathbf{Y}}}\mathcal{M} \\
\boldsymbol{\Sigma}_\mathbf{Y} = \boldsymbol{\Sigma}\left[\boldsymbol{\tau}\boldsymbol{\tau}^T\right] \; &\triangleq \mathbb{E}\left[(\mathbf{Y} \ominus \bar{\mathbf{Y}})(\mathbf{Y} \ominus \bar{\mathbf{Y}})^T\right] && \in \mathbb{R}^{m \times m},
\end{aligned}$$

where $\operatorname{Exp}(\boldsymbol{\tau})$ is the retraction of $\boldsymbol{\tau}$ onto the manifold $\mathcal{M}$ and Log is the inverse operation that maps an element of $\mathcal{M}$ to its tangent space $\mathcal{T}_{\bar{\mathbf{Y}}}\mathcal{M}$. This approach allows for a natural expression of Gaussian-distributed variables on Lie groups as

$\mathbf{Y} \sim N(\bar{\mathbf{Y}}, \mathbf{\Sigma_Y})$, which can now be used to adapt Eq. (4.1) to the orientational case. Therefore, the likelihood of the orientation $\mathbf{R} \in \mathrm{SO}(3)$ conditioned on C is calculated as:

$$p(\mathbf{R}|C) = \frac{1}{\sqrt{\det(2\pi\mathbf{\Sigma_R})}} \exp\left(-\frac{1}{2}(\mathbf{R} \ominus \bar{\mathbf{R}})^T \mathbf{\Sigma_R}^{-1} (\mathbf{R} \ominus \bar{\mathbf{R}})\right) , \tag{4.2}$$

where $\mathbf{\Sigma_R} \in \mathbb{R}^{3\times3}$ is the orientational covariance matrix, and $\bar{\mathbf{R}}$ is the mean orientation of an action hypothesis. The mean values are derived from filtered action hypotheses of the UKF.

Assuming conditional independence of the orientation $\mathbf{R}$ and position $\mathbf{t}$, the joint probability of the pose conditioned on C can be written as $p(\mathbf{R}, \mathbf{t}|C) = p(\mathbf{R}|C) \cdot p(\mathbf{t}|C)$. Using Bayes' rule, the correspondence likelihood $p(C|\mathbf{R}, \mathbf{t})$ that the observed action $\mathbf{A}$ at position $\mathbf{t}$ and orientation $\mathbf{R}$ corresponds to the action hypothesis $\bar{\mathbf{A}}$ is:

$$\begin{aligned}
p(\mathbf{R}, \mathbf{t}|C) &= \frac{p(C|\mathbf{R}, \mathbf{t}) \cdot p(\mathbf{R}, \mathbf{t})}{p(C)} \\
p(C|\mathbf{R}, \mathbf{t}) &= \frac{p(C) \cdot p(\mathbf{R}, \mathbf{t}|C)}{p(\mathbf{R}, \mathbf{t})} \\
p(C|\mathbf{R}, \mathbf{t}) &= \frac{p(C) \cdot p(\mathbf{R}|C) \cdot p(\mathbf{t}|C)}{p(\mathbf{R}) \cdot p(\mathbf{t})} \propto p(\mathbf{R}|C) \cdot p(\mathbf{t}|C)
\end{aligned}$$

A k-dimensional tree search is employed to efficiently identify correspondences between hypotheses and observations in a scene. For each hypothesis, a search radius $r = 3 \cdot \sigma_m = \max \operatorname{diag}(\mathbf{\Sigma_t})$ is utilized, justified by the multivariate normal distribution of the hypothesis position with independent components. The confidence region, defined by the three times scaled *Standard Deviational Hyper-Ellipsoid*, is enclosed by this sphere, ensuring a probability greater than $\sim 97\%$ of finding a corresponding observation within this radius (Wang et al., 2015).

Once a corresponding observation $\mathbf{A}_t$ for the filtered action hypothesis $\bar{\mathbf{A}}_{t-1}$ is identified, the estimated state of the filtered action is updated using this observation. This involves updating both the existence certainty p_E^a and the mean pose $\bar{\mathbf{T}}$ of the hypothesis.

Estimation of Existence Certainty

The existence certainty p_E^a is determined using the previously calculated $p(C|\mathbf{R}, \mathbf{t})$. To this end, a Continuous Density Hidden Markov Model (CDHMM) with two states is employed (see e. g., Rabiner, 1990). The hidden states are S_1 (action hypothesis exists) and S_2 (action hypothesis does not exist). Initially, the CDHMM is assumed to be in S_1, and the state can only transition from existing (S_1) to not existing (S_2), meaning a hypothesis can only vanish. The state $\lambda = (\mathbf{A}, \mathbf{B}, \pi)$ of the CDHMM is defined with $\pi = \begin{pmatrix} 1 & 0 \end{pmatrix}$ and

$$\mathbf{A} = \begin{pmatrix} a_{11} & 1 - a_{11} \\ 0 & 1 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} p(C|\mathbf{R}, \mathbf{t}) & 1 - p(C|\mathbf{R}, \mathbf{t}) \\ 1 - b_{22} & b_{22} \end{pmatrix}.$$

A visual representation of the CDHMM can be found in Figure 4.1

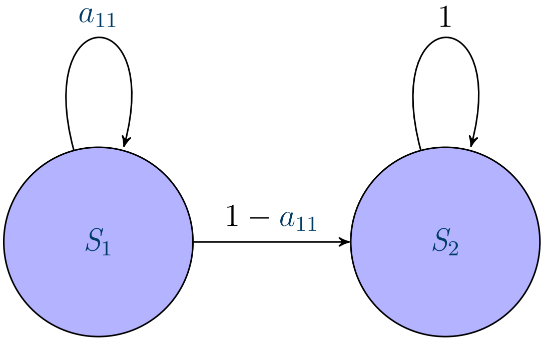


Figure 4.1.: Representation of the existence certainty as a 2-state CDHMM

Now, the *forward-backward* algorithm can be applied to calculate the probability p_E^a of being in state S_1 at time t for the affordance a_i. Let

$$\mathbf{f}_{0:t} = \begin{pmatrix} f(S_1|\lambda, O) \\ f(S_2|\lambda, O) \end{pmatrix}$$

be the probabilities of the HMM being in states S_1 and S_2, respectively, after observations $0 \ldots t$. The state probabilities of the current observation can be calculated via the probabilities of the previous observation and the correspondence likelihood:

$$\mathbf{f}_{0:t} = c_t^{-1} \mathbf{f}_{0:t-1} \mathbf{A} \mathbf{O}_t \,,$$

where the diagonal observation matrix for the event (i. e., an action was observed or not observed) $j \in \{1, 2\}$ is $\mathbf{O}_{j,t} = \mathrm{diag}(\mathbf{B}_{*,j})_t$. If one is only interested in an unnormalized certainty of the existence (or: the HMM being in state S_1), this can be calculated using the probability of an identity transition a_{11} as:

$$p_E^a \propto a_{11} \cdot f(S_1|\lambda, O)_{t-1} \cdot O_{1j,t} \,.$$

Pose Estimation Using UKF on Lie Groups

As mentioned before, the orientational part $\mathbf{R}$ of the pose of the AAF cannot easily be estimated using Gaussian distributions in $\mathbb{R}^3$. Therefore, a simple Kalman filter is inadequate for pose estimation in robotic applications due to the complexity of modeling orientations in Euclidean space. However, recent research has highlighted the advantages of using Lie groups for recursive Bayesian estimation, as they provide a natural and smooth representation of poses (Brossard et al., 2017; Lee, 2018; Sjøberg and Egeland, 2021; Solà et al., 2018).

To address the limitations of conventional Kalman filters, Gaussian distributions on manifolds are employed in Brossard et al. (2017) to implement a UKF for generic Lie groups. This approach uses retractions onto the tangent space, allowing standard UKF algorithms to update and propagate the state. For the spatial fusion of action observations, the open-source implementation of a UKF on manifolds (*UKF-M*, Brossard et al., 2020) was used to combine multiple action observations $\mathbf{A}$ to obtain the mean pose $\bar{\mathbf{T}}$ of an action hypothesis $\bar{\mathbf{A}}$ and its covariance matrix $\mathbf{\Sigma_T}$ over multiple observation of a scene.

Combined with the temporal filtering of the HMM for estimating the existence certainty p_E^a of an action hypothesis, the spatial filtering of the mean pose $\bar{\mathbf{T}}$ using a UKF on manifolds constitutes the foundation of the PAEF approach. This way, PAEF answers Research Question 2 by increasing the reliability of grasping and manipulation as a consequence of accounting for perceptual and proprioceptive *uncertainties*.

4.1.3. Experiments

To showcase the increase in reliability of manipulation actions, PAEF was evaluated alongside GAE in a series of real-world grasping experiments using the humanoid robot ARMAR-6 with more than 900 grasp executions. For a more detailed description of the experimental setup and the results for the GAE method, see Section 3.2.3. The experiments were carried out in two cluttered setups: a box-emptying setup and a table-clearing setup. In both scenarios, unknown objects were randomly placed, and ARMAR-6 was tasked with grasping and manipulating these objects to either empty the box or clear the table. A video detailing the approach and experiments is available online[1].

For the spatiotemporal fusion of action hypotheses, action observations were extracted using GAE for a new point cloud at every time step t. Afterwards, the

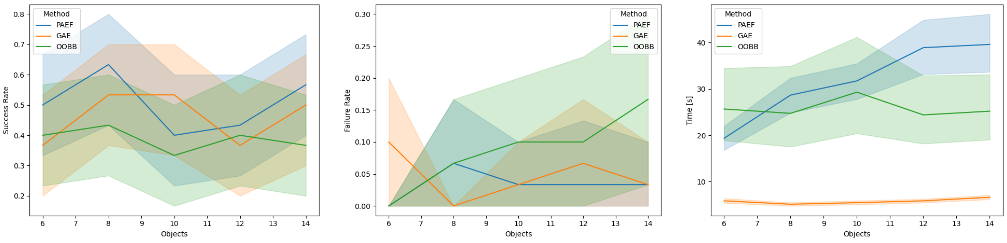
[1] https://youtu.be/lXxWtTIySB0

methods detailed in Section 4.1.2 were used to find correspondences between a list of already observed hypotheses $\bar{\mathbf{A}}_{t-1}$ and the newly extracted action observation $\mathbf{A}_t$. Once a correspondence was identified, the existence certainty p_E^a and the mean pose $\bar{\mathbf{T}}$ of the action hypothesis $\bar{\mathbf{A}}_{t-1}$ were updated using the HMM and UKF, respectively, to obtain $\bar{\mathbf{A}}_t$. The initial position covariance for the UKF was set to 1 cm, and the initial orientation covariance to 0.1 rad. The HMM parameters were chosen as $a_{11} = 0.9$ and $b_{22} = 0.5$. After fusion, the action hypotheses were checked for validity and executed analogously to Section 3.2.3.

Box-Emptying Experiments

The results of the box-emptying experiments comparing the grasping success rates of PAEF and GAE to a baseline approach using OOBBs (Grimm et al., 2021) can be seen in Figure 4.2. As in Section 3.2.3, grasp attempts were categorized according to Table 3.3, and the categories *"grasped"* and *"stable lifted"* were counted as successful grasps, while the categories *"slipped"* and *"missed"* were counted as failed grasp attempts. Additionally, the time from point cloud capture to selection was measured for each attempt. Since the PAEF and OOBB methods rely on previous scene observations, they were reset after each attempt to estimate the worst-case time required for grasp candidate generation and selection in novel scenes.

Note: Adapted from Pohl and Asfour (2022). © 2022 IEEE.



(a) Success Rate (b) Failure Rate (c) Extraction Time

Figure 4.2.: Results of the box-emptying experiments on ARMAR-6 comparing the OOBB-based grasp generation with GAE (Section 3.2) and PAEF

The PAEF method achieved a success rate of 50.7%, outperforming the GAE and OOBB methods, which had success rates of 46.0% and 38.7%, respectively. The PAEF method showed no strong correlation between the number of objects and successful grasps (Person's correlation coefficients: $\rho_{PAEF} = -0.11$, $\rho_{GAE} = 0.18$, $\rho_{OOBB} = -0.42$), indicating its robustness even in very cluttered scenes. While the failure rates of GAE ($\rho_{GAE} = -0.28$) and PAEF ($\rho_{PAEF} = 0.22$) show also no correlation to the number of objects in the box, the OOBB-based grasping ($\rho_{OOBB} = 0.96$) performed worse for more cluttered environments. The PAEF method's improved success rate over the non-filtered GAE and consistent performance in difficult scenes confirmed the positive impact of spatiotemporal fusion on the reliability of mobile manipulation in *unstructured environments*.

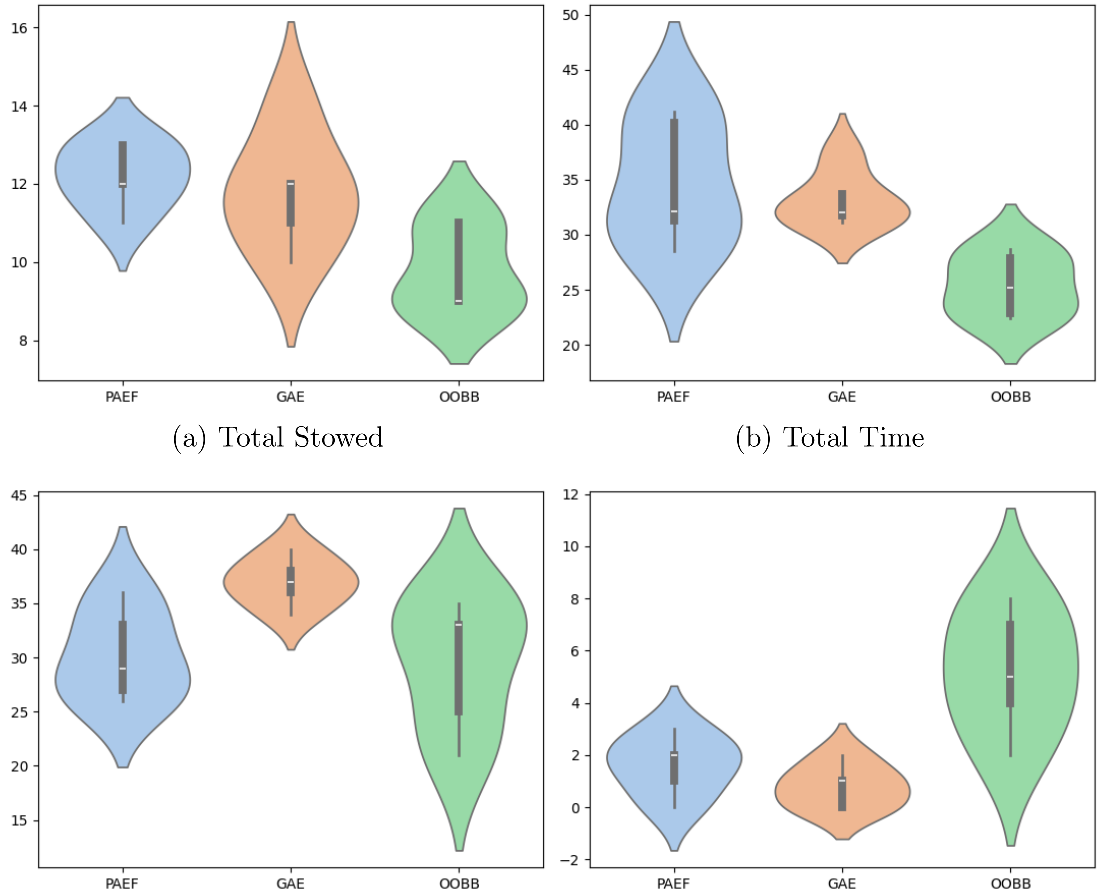
As shown in Figure 4.2c, the time required for grasp selection varied among the methods. The PAEF method generally required more time than GAE and OOBB-based grasp extraction due to the increased computational effort of finding correspondences (i. e., with a tree-search) and the spatiotemporal fusion. Additionally, there is a clear correlation with the number of items in the box, as the more objects there are in the box, the more action observations are extracted by GAE (i. e., more AAFs are connected to the *graspability* affordance), and therefore, the tree-search grows more complex. The OOBB method required almost constant time across different clutter levels due to its efficient grasp pose generation. The GAE method was the fastest, as it processed a fixed number of surface patches regardless of the scene's complexity.

Table Clearing Experiments

The table-clearing experiment was designed to evaluate the robustness of the spatiotemporal affordance fusion approach in a realistic kitchen scenario. The setup remained the same as in Section 3.2.3. The results are visualized in Figure 4.3.

The PAEF method demonstrated a higher success rate in stowing objects compared to the GAE and OOBB methods. Although the total number of objects stowed by the PAEF method was only slightly higher than that of the GAE method (Figure 4.3a), the PAEF method required significantly fewer grasp attempts (Figure 4.3c). This means that the "raw" extracted candidates from GAE failed a lot more often than the filtered candidates from PAEF, validating the claim that the spatiotemporal fusion of grasp candidates indeed improves the reliability of grasping. This, in turn, results in the total time required for clearing the table almost being the same for both methods (Figure 4.2c), even though the box-emptying experiments showed that the grasp synthesis process for PAEF is slower than for GAE. Despite the PAEF method's longer extraction times, its higher accuracy nearly offset this, resulting in a total table clearing time comparable to the GAE method. The OOBB method, while having fewer grasp attempts and a shorter clearing time, left an average of more than 5 objects on the table due to the time constraint (Figure 4.3d), underscoring its lower effectiveness.

Note: Adapted from Pohl and Asfour (2022). © 2022 IEEE.



(a) Total Stowed

(b) Total Time

(c) Grasp Attempts

(d) Remainig Objects

Figure 4.3.: Results of the table-clearing experiments on ARMAR-6 comparing the OOBB-based grasp generation with GAE (Section 3.2) and PAEF

4.2. Uncertainty-aware Grasp Candidate Selection

The previously described Probabilistic Action Extraction and Fusion method (Section 4.1) established a probabilistic approach to affordance-based action extraction for mobile manipulation in *unstructured environments* that improved the reliability of grasping in real-world experiments. The grasp candidates in the evaluation with ARMAR-6 were chosen based on height and improved the success rate by about 5% compared to the GAE method. However, PAEF provides two probabilistic measures, the existence certainty and the pose covariance, which can add valuable insights about the quality of a grasp candidate. Therefore, this section investigates how PAEF can be used in combination with other *uncertainty*-affected metrics to improve the grasp *selection* in *unstructured environments*. To this end, a *sensitivity*-optimized grasp score is calculated, which indicates the quality of a grasp candidate, that is used to select the most reliable grasp in the scene.

To derive this grasp score, multiple uncertainty-affected metrics that could be relevant for the outcome of grasping were selected. Subsequently, a large dataset of real-world grasp executions was collected on the humanoid robot ARMAR-6. Using this dataset, the *sensitivities* of the metrics towards the grasp outcome were analyzed and used to obtain global and local weighting factors for each metric. With these weighting factors, it then becomes possible to calculate a grasp score that can be used to select the grasp candidate with the highest likelihood of success. In a real-world evaluation on ARMAR-6, grasp selection using the optimized grasp score showed large improvements in the grasp success rate compared to randomized grasping.

The content of the following sections will present the details of the Uncertainty-Aware Sensitivity Optimization for grasp selection in *unstructured environments*, which have already been published in the paper by Baek et al. (2022).

4.2.1. Motivation

For the interaction with *unstructured environments*, as they appear in the personal sector, robotic assistants need to be able to handle *uncertainty* and missing or incomplete information. For example, Brock et al. (2016) define *uncertainty* as one of the three main challenges in mobile manipulation research. Likewise, Research Question 2 designates reliability in uncertain situations as one of the core capabilities that this thesis tries to address. In Section 4.1, the Probabilistic Action Extraction and Fusion (PAEF), an approach for the spatiotemporal filtering of affordance-based action hypotheses, was introduced that determines the existence certainty and pose covariance for each hypothesis over multiple observations. In Pohl et al. (2020), it was shown that the selection of autonomously generated grasp candidates by a human operator in a semi-autonomous setup can largely enhance the success rate of grasping. By simply incorporating human intuition and scene understanding into the *discriminative grasping* process via Virtual Reality-based grasp selection, the reliability of grasping in *unstructured environments* improved. Therefore, in an effort to address Research Question 2 and contribute to the main objective of the thesis, this section tries to increase the level of scene understanding and the awareness of *uncertainty* in the *selection* of grasp candidates in order to improve the reliability of mobile manipulation tasks. An overview of the approach can be seen in Figure 4.4.

Perceptual and systematic *uncertainties* that impair the robot's ability to extract and execute reliable grasps are problematic to current approaches in autonomous grasp selection. Many methods lack robust mechanisms for handling these *uncertainties*, resulting in fragile and slow grasping processes that are unsuitable for complex, real-world tasks (see Section 2.2 and Table 2.3). Approaches using *analytical quality* metrics as a basis for selecting grasp candidates can incorporate probabilistic approaches in their models. However, their calculation can become time-consuming and often they require perfect knowledge of the environment (e. g., friction coefficients). *Heuristic quality* metrics, on the other hand, are fast to compute but often do not account for missing information and noise, and might therefore be less precise. *Learning*-based methods for grasp candidate selection might be able to handle these cases, but often require large, time-consuming datasets for training and lack insight for human operators. Taking these limitations and advantages into account, the proposed approach uses a combination of all three categories to select the most reliable grasp candidate using visual perception under *uncertainty* and missing information.

This section presents a probabilistic approach to enhancing autonomous grasping by integrating traditional statistical tools to maximize grasp success rates and, therefore, the reliability of *grasping*. The proposed method introduces an Uncertainty-Aware Sensitivity Optimization (UASO, Baek et al., 2022) framework that derives a scalar ranking score for grasp candidates based on the *sensitivities* of predefined grasp metrics. These metrics, modeled as Gaussian distributions, encapsulate high-level scene understanding and the associated *uncertainties*. By analyzing a dataset of 932 grasps executed by the humanoid robot ARMAR-6 under real-world conditions, the method assigns weights to each metric according to its influence on the grasp success rate. Validation experiments demonstrate a significant improvement in success rates by explicitly considering *uncertainties*. Additionally, this approach allows for detailed correlation studies and statistical analyses, providing deeper insights into the impact of individual metrics on grasp reliability. Therefore, the contribution lies in presenting an explainable, generalizable method that improves the reliability of robotic grasping by accounting for the influence of perceptual *uncertainties*. In doing so, UASO addresses Research Question 2 and contributes to the main objective of this thesis.

4.2.2. Sensitivity Optimization for Grasp Candidate Selection

In order to derive a grasp score for autonomously selecting grasp candidates, probabilistic techniques in the form of a uncertainty-aware sensitivity optimization are employed to improve the reliability of the selected grasp candidates in presence of noise and *uncertainty*. To this end, a number of Gaussian-distributed and

Note: Adapted from Baek et al. (2022). © 2022 IEEE.

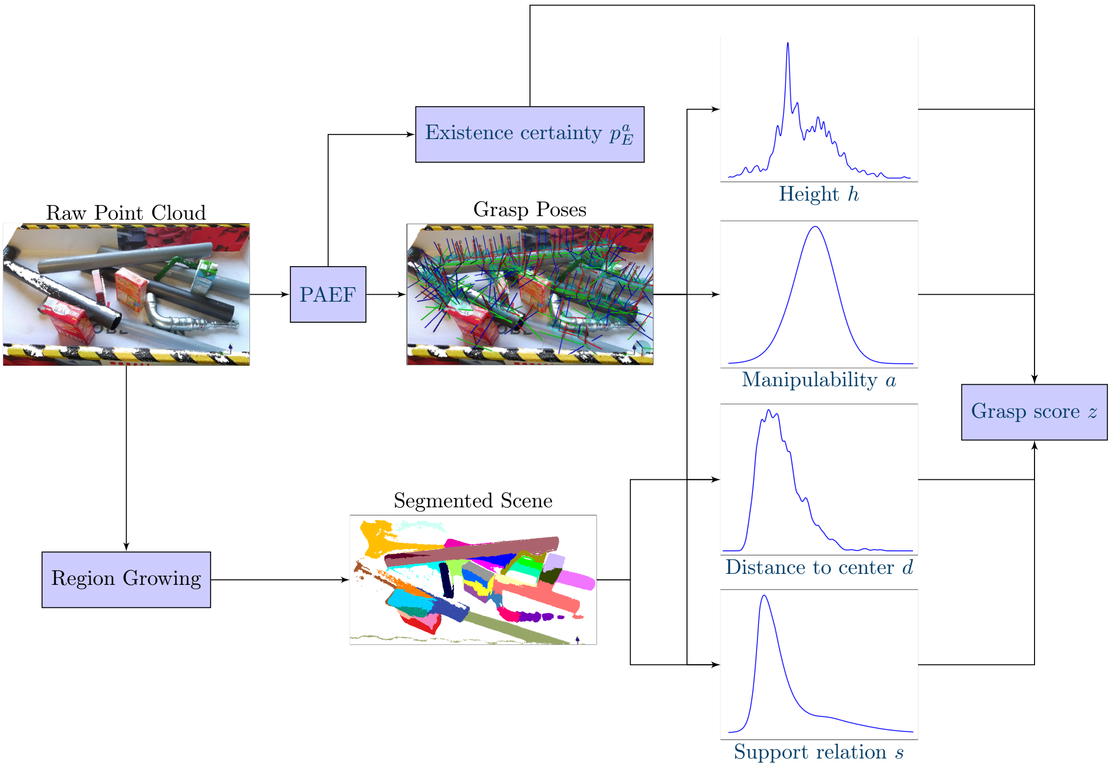


Figure 4.4.: Overview of the Uncertainty-Aware Sensitivity Optimization approach.

uncertainty-affected grasp metrics relevant to the success rate of grasping are chosen and analyzed towards their respective correlation to the outcome of a grasp execution. Based on a dataset detailing the outcome of grasps and their respective grasp metrics, a global weighting factor and a local weighting factor for each grasp metric are calculated, which represent the *sensitivities* of that grasp metric.

Probabilistic Framework

Each grasp g is characterized by n specified grasp metrics m_i, represented as Gaussian distributions $m_i \sim N(\mu_i, \sigma_i)$ with mean value μ_i and standard deviation σ_i. The key idea behind this is to somehow capture the *uncertainties* of these metrics, despite the lack of detailed knowledge about their specific behaviors. For the derivation of a scalar grasp score z, a functional model y: $\mathbb{R}^+ \rightarrow \mathbb{R}_0^+$ of the relations between the different grasp metrics is required:

$$\begin{aligned} z &:= y\left(m_1, ..., m_n\right) \\ &= y\ \left(N(\mu_1, \sigma_1), ..., N(\mu_n, \sigma_n), c\right), \end{aligned} \tag{4.3}$$

In the interest of obtaining a grasp score that maximizes the success rate r^s of grasps, with

$$r^s := \frac{g^s}{g^{\mathrm{tot}}},$$

Eq. (4.3) should capture the *sensitivities* of the grasp metrics m_i towards r^s, i. e., the amount of influence these grasp metrics have on a successful outcome. Technically speaking, the *sensitivity* describes how the output of a system behaves with respect to changes in the input parameters (BIPM et al., 2008). With this in mind, each grasp candidate is classified as either succeeded (g^s) or failed (g^f) after execution. Assuming there exists a dataset of grasp outcomes and the corresponding, *uncertainty-afflicted* grasp metrics m_i, it becomes possible to derive weighting factors for each m_i linked to their *sensitivity*. To get meaningful results, this dataset needs to be of adequate size, cover the respective ranges of the grasp metrics and sufficiently distinguishes the successful from the failed grasps. Given the dataset and the associated information about the grasp metrics and their *uncertainty*, one PDF p_i for the successful grasps (p_i^s) and one for failed grasps (p_i^f) can be calculated by simply summing over the individual, Gaussian-distributed observations in the dataset.

For an optimal grasp score z, i. e., one that takes the *sensitivities* of the grasp metrics into account, a local and global weighting factor are calculated based on p_i^s and p_i^f. These weights represent the local and global importance of the respective grasp metric for the grasp outcome. The global weighting factor f^{glob} can be calculated using using the KL divergence D_{KL}, which measures the difference between two distributions P and Q:

$$D_{KL}(P\|Q) := \int_{-\infty}^{\infty} p(x) \log\left(\frac{p(x)}{q(x)}\right) dx.$$

Here, $p(x)$ and $q(x)$ denote the PDFs of P and Q, respectively. Therefore, the global weighting factor is defined as

$$f_i^{\mathrm{glob}} := D_{KL}(P_i^s || P_i^f), \tag{4.4}$$

where P_i^s and P_i^f represent the PDFs for grasp metric i of successful and failed grasps, respectively. In the second step, the likelihood of a candidate grasp g belonging to the set of successful grasps g^s, denoted as the local weighting factor f^{loc}, is derived. This is calculated as

$$f_i^{\mathrm{loc}}(g) := \frac{p_i^s(m_i|g)}{p_i^s(m_i|g) + p_i^f(m_i|g)}, \tag{4.5}$$

by referring to the specific values of the PDFs p_i^f and p_i^s for the measured metrics of g. Therefore, the total weighting factor f^{tot} for the grasp metric i can be calculated as

$$f_i^{\text{tot}} := f_i^{\text{glob}} \cdot f_i^{\text{loc}}.$$

Having calculated the total weighting factor that can be used to calculate an optimized grasp score z with respect to the *sensitivities* of the grasp metrics m_i, all that is left is to chose a suitable functional model y. For the context of this thesis, the form

$$y_{exp}(\alpha, m_i, \beta) := \alpha \cdot \sum_{i=1}^{n} f_i^{\text{tot}} + \beta \tag{4.6}$$

was chosen, where α, β are scalar constants.

Grasp Selection

The general form Eq. (4.6) facilitates the selection of any *uncertainty-afflicted* grasp metric and calculating an optimized grasp score for the *selection* of grasp candidates. This highlights the strength of UASO, being a general framework for the optimization of data with respect to one category. In theory, it is possible to exchange a grasp g for any other kind of binary variable. In addition, only deterministic methods were applied to derive Eq. (4.6), making the method explainable, i. e., it is possible to trace and analyze the decisions of the model. Therefore, UASO combines the advantages *learned*, *analytical*, and *heuristic* grasp quality measures while remaining lightweight, explainable, and precise.

Relying on *uncertainty* estimations of the grasp metrics for the calculation of the grasp score makes UASO naturally compatible with PAEF: By filtering grasp candidates over time, the *uncertainty* connected to each hypothesis is quantified by the existence certainty p_E^a and pose covariance $\mathbf{\Sigma_T}$. The existence certainty – representing the overall confidence that an action can be executed at the pose of the AAF – influences all other grasp metrics for the grasp, and, therefore, intuitively fits the role of $\alpha = p_E^a$. Additionally, based on the *uncertainty* of the pose of a grasp, the following grasp metrics were selected:

1. **Height (h)**: This metric indicates how high a grasp candidate is above the floor, favoring objects on top of clutter. The mean and variance (μ_h, σ_h^2) are derived from the grasping pose and its covariance, computed using PAEF.
2. **Distance to center (d)**: This metric measures the distance from the grasp position to the center of the object's bounding box, favoring grasps near the object's center of mass. It is calculated by combining PAEF with scene

segmentation. The variance σ_d^2 is estimated as 10% of the bounding box length, while the mean μ_d is the distance to the bounding box center.

3. **Support relation (s)**: This metric counts the number of objects supported by the point cloud segment closest to the grasping pose, favoring objects not covered by others. The mean and variance (μ_s, σ_s^2) are obtained from a probabilistic support graph, as described in Paus and Asfour (2020). This graph is based on Random Sample Consensus shape estimations and details the support relations between the estimated shapes in a segmented scene.
4. **Manipulability (a)**: This metric reflects how freely an end-effector can move at a certain position, favoring easily reachable grasps. It uses the extended manipulability score from Vahrenkamp et al. (2012), calculated solely from the grasping pose using PAEF. The mean and variance (μ_a, σ_a^2) are derived from a manipulability map based on how many hits of randomly sampled joint configurations the corresponding map entry has.

With these grasp metrics, the following assignments are used for Eq. (4.6):

$$\begin{aligned} m_1 &\mapsto h; \quad m_2 \mapsto d; \quad m_3 \mapsto s; \quad m_4 \mapsto a; \\ \alpha &\mapsto p_E^a; \quad \beta = 0; \quad . \end{aligned}$$

Therefore, the final form of the grasp score z is

$$\begin{aligned} z &= y_{exp}(p_E^a, h, d, s, a) \\ &= p_E^a \cdot \left(f_h^{\mathrm{tot}} + f_d^{\mathrm{tot}} + f_s^{\mathrm{tot}} + f_a^{\mathrm{tot}}\right) . \end{aligned} \tag{4.7}$$

4.2.3. Data Collection and Evaluation

To assess the effectiveness of the proposed Uncertainty-Aware Sensitivity Optimization in enhancing the reliability of grasp *selection*, extensive real-world experiments using the humanoid robot ARMAR-6 were conducted. These experiments included over 1100 grasp attempts on various unknown objects, creating a comprehensive dataset for optimizing and evaluating the reliability and success rate of the proposed probabilistic grasp grasp metrics. A video showcasing the approach and experimental procedures is available online[2].

The experimental setup for data collection and the evaluation of the UASO approach is very similar to the one of Section 3.2.3 and Section 4.1.3. For each experiment, 11 objects (five plastic pipes, four boxes, and two metal pipes) were

[2] https://youtu.be/puJmGsK6hSE

Note: Reprinted from Baek et al. (2022). © 2022 IEEE.

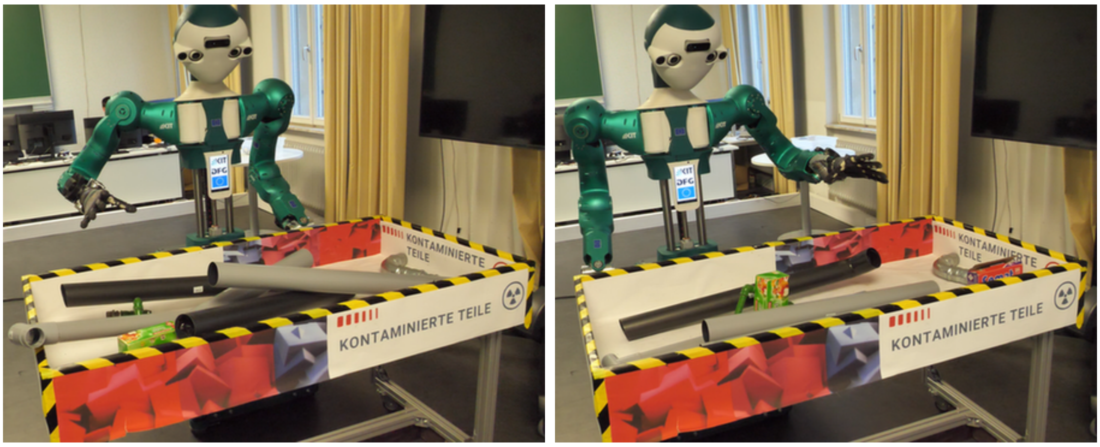


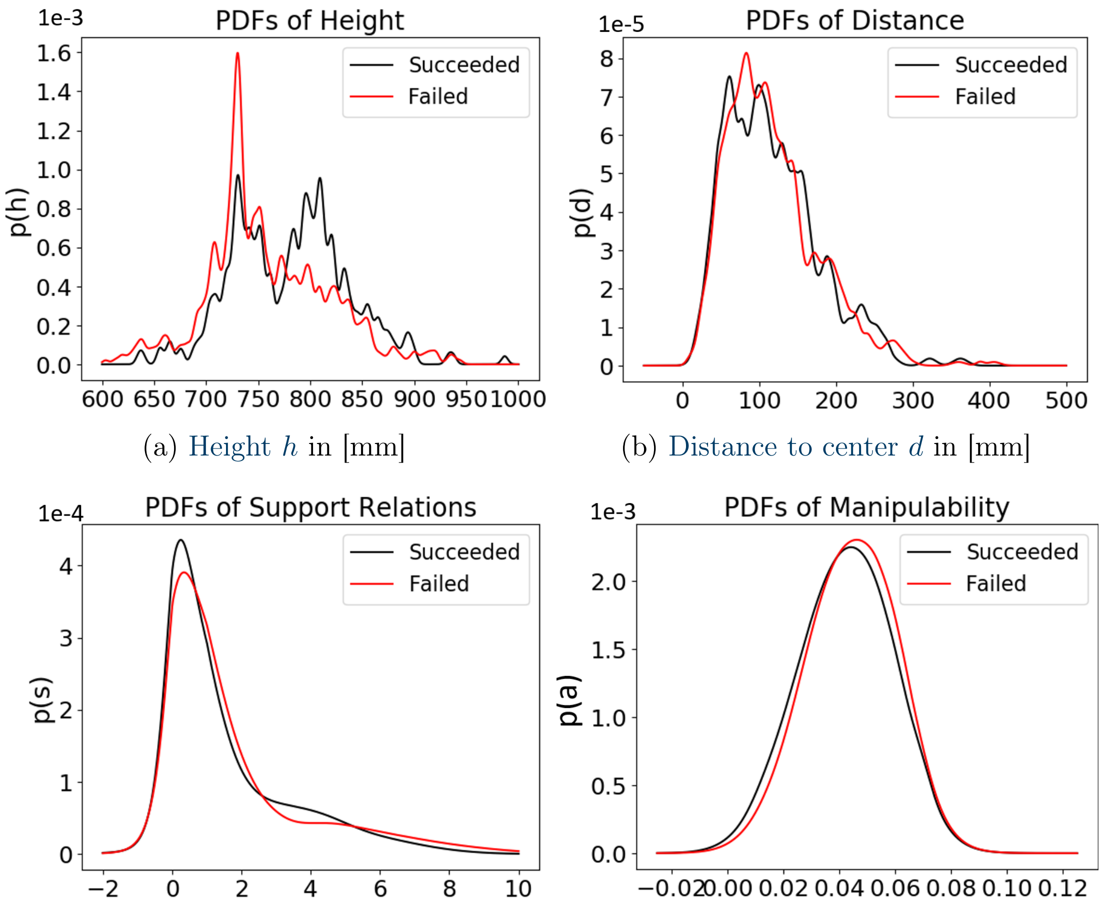
Figure 4.5.: Examples of the experimental setup with ARMAR-6.

placed randomly in a box. Grasp candidates were then extracted using the GAE approach and subsequently filtered using PAEF over multiple consecutive scans of the scene. Additionally, each point cloud was segmented using a region-growing segmentation. Before execution, each candidate's IK and distance to the box borders were checked to ensure reachability and collision avoidance. After each grasp attempt, ARMAR-6 placed the object back into the box to introduce random changes in the scene, ensuring a dynamic and varied dataset. If the scene remained unchanged for multiple attempts, either due to no graspable objects or repeated grasping of the same object, a human operator would randomly rearrange the objects in the box. Two example setups of ARMAR-6 grasping unknown objects can be seen in Figure 4.5.

Data Collection

For the optimization of *sensitivities* of the selected grasp metrics, a dataset of grasp executions and the corresponding mean values μ_i and standard deviations σ_i of the grasp metrics m_i is required. To this end, a set of 932 grasp executions (304 successful and 628 failed, $r^s = 0.326$) was recorded over four consecutive days, during which a grasp was randomly selected from the set of reachable candidates and executed by ARMAR-6. For each executed grasp, the relevant grasp metrics as described in Section 4.2.2 were calculated and recorded along the outcome of the grasp execution (i. e., either failed or successful).

To perform the global and local weighting steps outlined in Equation 4.4 and Equation 4.5, the PDFs of the set of failed and successful grasps were calculated from the dataset. The resulting normalized PDFs can be seen in Figure 4.6.



Note: Reprinted from Baek et al. (2022). © 2022 IEEE.

(a) Height h in [mm]

(b) Distance to center d in [mm]

(c) Number of support relations s

(d) Manipulability a in [arb. units]

Figure 4.6.: PDFs for succeeded (■) and failed (🟥) grasp attempts of the considered metrics from Section 4.2.2 for randomly selected grasps. These distributions provide the basis for calculating the ranking score according to Eq. (4.7).

Sensitivity Analysis

The PDFs of the grasp grasp metrics represent their distribution in the recorded dataset. From these, the local weighting factor f^{loc} and global weighting factor f^{glob} can be calculated using Eq. (4.5) and Eq. (4.4), respectively. The global weighting factor corresponds to the information value that is encoded by the grasp metric m_i with respect to the grasp outcome.

As shown in Table 4.1, the height metric exhibits the highest KL divergence at 0.460, indicating a significant difference between successful and failed grasps. In contrast, the distance to center, support relation, and manipulability grasp metrics have much lower KL divergence values, suggesting less influence on the grasp outcome.

Table 4.1.: KL divergences of the metrics obtained from random grasping.

Metric	KL divergence D_{KL}
Height h	0.460
Distance to center d	0.034
Support relation s	0.014
Manipulability a	0.010

Note: Reprinted from Baek et al. (2022). © 2022 IEEE.

Optimized Grasp Selection

To evaluate the effectiveness of the scoring function in grasp selection, the PDFs from the random dataset were utilized. The same experimental setup and grasp candidate generation process as previously described were employed, with the key difference being the use of the grasp score z to select the grasp candidate. To this end, the relevant metrics and subsequently the grasp score from Eq. (4.7) were computed for each candidate, and the grasp with the highest z was chosen for execution. This approach was tested in a series of experiments, where 187 grasps were performed using ARMAR-6.

The application of the optimized grasp *selection* method led to a significant improvement in performance. Specifically, the success rate of grasp attempts increased from 32.6% (random grasping) to 73.8 %, with 138 successful grasps out of 187 attempts. This improvement underscores the efficacy of UASO in enhancing the reliability of grasp selection. The substantial increase in success rate highlights the potential of probabilistic methods to improve the *autonomy* and robustness of robotic grasping in real-world scenarios.

4.3. Conclusion

This chapter introduced a combined approach for the increase in reliability of the *selection* of manipulation actions. To this end, the Probabilistic Action Extraction and Fusion, a novel method for the probabilistic, spatiotemporal fusion of grasping and manipulation candidates, was integrated into the Uncertainty-Aware Sensitivity Optimization, a comprehensive framework for optimizing the *sensitivity* in scenarios were the outcome is influenced by *uncertainties*. This combination, when applied to the concept of autonomous grasp *selection*, demonstrated a significant increase in the reliability of grasping in multiple real-world experiments using the humanoid robot ARMAR-6, thereby addressing Research Question 2.

Section 4.1 introduced the Probabilistic Action Extraction and Fusion (PAEF). A geometry-aware shared state for all affordances is used in combination with recursive Bayesian estimation techniques to improve the estimate of the pose of an action hypothesis over multiple distinct observations. The approach was rigorously tested in various real-world grasping scenarios using the humanoid robot ARMAR-6. The experimental results clearly demonstrate that the PAEF method significantly enhances the reliability of mobile manipulation actions. By incorporating spatiotemporal affordance fusion, the PAEF method consistently outperformed the GAE and OOBB methods in both the box-emptying and table-clearing scenarios with more than 900 grasp executions of the humanoid robot ARMAR-6. The higher success rates and fewer grasp attempts required by the PAEF method underscore its increased reliability in handling perceptual and proprioceptive *uncertainties* and dynamic changes in *unstructured environments*.

Section 4.2 introduced the Uncertainty-Aware Sensitivity Optimization (UASO) and applied the concept to autonomous grasp selection. Utilizing 932 randomly selected grasps performed by the humanoid robot ARMAR-6 under real-world conditions, a broad dataset was built for the optimization of a grasp score that can be used to select grasp candidates based on *uncertainty-afflicted* grasp metrics. This grasp score integrates four specific metrics, which are combined using a global and local weighting factor for each grasp metric, as well as the existence certainty from PAEF. This scoring function was subsequently employed in a second set of experiments to identify the most reliable grasp candidate for a given scene, resulting in a grasp success rate of 73.8% compared to 32.6% with random selection. This improvement highlights the effectiveness of the selected metrics in predicting successful grasps. By connecting PAEF with UASO, it becomes possible to incorporate global information (i. e., from scene understanding) into the candidates extracted by local surface geometry. Furthermore, the results demonstrate the role of *sensitivity* optimization in enhancing the reliability and *autonomy* of robotic grasping.

In response to Research Question 2, the findings substantiate the hypothesis that probabilistic approaches can significantly enhance the reliability of mobile manipulation. By spatiotemporally tracking and fusing the associated frame of affordance-based action hypotheses, the success rate in *grasping* could be increased by almost 15% compared to a baseline approach using OOBBs. Furthermore, by utilizing these probabilistic measures in the Uncertainty-Aware Sensitivity Optimization that integrates multiple probabilistic grasp grasp metrics, the grasp success rate could be improved by more than 40%, compared to a random grasp selection. This substantial increase underscores the effectiveness of probabilistic methods in handling *uncertainties* in visual perception and proprioception, thereby enhancing the reliability of robotic grasping in *unstructured environments*.

Therefore, robotic assistants can act more *autonomous* in these scenarios, which represents the contribution to the main objective of this section.

5. Methods and Representations for the Adaptable Execution of Mobile Manipulation Skills

After introducing the enhancements to the versatility in grasp *discovery* (Chapter 3) and the reliability of the *selection* of potential grasp candidates (Chapter 4), this chapter is concerned with improving the final step of *discriminative grasping*. One of the core capabilities introduced in Section 1.1.1 is the adaptability of the *execution* of manipulation skills, which is fundamental for the main objective of this thesis, as it enables robotic assistants to adjust their behavior to changing circumstances and the actual, varying conditions at their place of operation. Additionally, having adaptable manipulation skills facilitates the deployment of robots to various domains without the need for extensive reprogramming. Vernon and Vincze (2017) present a list of 11 priorities for cognitive robotics in industrial setups, with "High-level instruction and context-aware task execution" and "Adaptive planning" being two of these. This chapter argues that these priorities also apply to robotics in the personal sector and introduces a context-aware task execution framework, as well as an adaptive planning framework to address these priorities. To this end, this chapter reports on the contributions to the adaptability of the *execution* of mobile manipulation skills in *unstructured environments*. Specifically, the main focus will be on the advantages of using an affordance-based, memory-centric and *execution* framework. Additionally, the opportunities of a combination of this *Executive* framework with an adaptable *Planning* system using LLMs will be detailed. Therefore, this section explains how Contribution 3 addresses the Research Question 3 and, by doing so, increases the adaptability of robotic assistants in real-world applications.

Disclaimer

Parts of the content presented in this chapter were previously published in:

- **Pohl, Christoph**, Fabian Reister, Fabian Peller-Konrad, and Tamim Asfour (2024). "MAkEable: Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills". In: *Proc. of the 2024 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*. International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, UAE: IEEE/RSJ, accepted for publication
- Timo Birr, **Christoph Pohl**, Abdelrahman Younes, and Tamim Asfour (2024). "AutoGPT+P: Affordance-based Task Planning with Large Language Models". In: *Proceedings of Robotics: Science and Systems*. Robotics: Science and Systems. Vol. 20. Delft, Netherlands

Section 2.3 introduced the three-tiered robot architecture that separates the responsibilities of mobile manipulation frameworks into three separate layers: *Behavioral Control*, *Executive*, and *Planning*. As the *Behavioral Control* layer is very hardware-dependent and consists of *Situated Behviors* (Kortenkamp et al., 2016), adaptability in this layer is limited to specific situations and involves some form of *reaction* to sensory stimuli, which is not the focus of this thesis. Contrarily, this chapter investigates the adaptation of the *execution* of mobile manipulation skills to different external influences, which has to originate in the medium- and high-level tiers (see e. g., Jaquier et al., 2024). Therefore, this chapter will investigate how to improve the adaptability of task *execution* by improving the *Executive* and *Planning* layers.

To this end, a mobile manipulation framework that facilitates the *transfer* of skills across different *tasks*, *environments*, and *robots* will be introduced in Section 5.1. Subsequently, in Section 5.2, the increased adaptability that results from integrating affordances and LLMs into the *Planning* layer will be detailed.

5.1. Memory-centered and Affordance-based Mobile Manipulation Framework

In order to adapt to the different situations and changing circumstances that robotic assistants have to face in realistic applications in the personal sector, it will be necessary to take their knowledge and experience and *transfer* them to the conditions of the current task. The ability to generalize and abstract their knowledge, experience, and skills directly influences the *autonomy* of robotic assistants by maximizing the *task generality* of autonomous systems while minimizing the amount of *task-specific knowledge* required (Brock et al., 2016). Additionally, it enables robots to learn from each other and decreases the manual effort required to adapt to changes in their *environment*, *task*, and *embodiment*.

Therefore, this section presents a Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills (MAkEable, Pohl et al., 2024) designed to enable the *transfer* of mobile manipulation skills across different modes (i. e., *robots*, *environments*, and *tasks*; as defined in Jaquier et al., 2024). The framework features a universal, affordance-based *task description* and supports the customization of individual aspects to the user's needs and various scenarios. The framework is centered around the cognitive memory architecture of Peller-Konrad et al. (2023), which fosters a collaborative learning environment among robots and supports explainability through introspection. It demonstrates the *transfer* across the three modes through use cases such as uni- and bimanual *grasping*, *placing* of known and unknown objects, and transferring a drawer-*opening* skill to another robot. Additionally, the framework's versatility is showcased by executing a pouring task in simulation with different robots.

The content of the following sections has already been published in the paper Pohl et al. (2024). Hereafter, the beneficial impact of the *transfer* of knowledge and experience across *robots*, *environments*, and *tasks* on the adaptability of the execution of mobile manipulation skills will be detailed.

5.1.1. Motivation

In the dynamic and various applications that robots encounter in the service industries, the efficient *transfer* of learned skills and experiences between robots or across various environments is crucial. This ability not only speeds up the deployment of robots into new settings but also significantly improves their adaptability and functionality. For example, in domestic scenarios, this could mean that a robot can seamlessly transition from one home to another, adapting to different layouts and task requirements without extensive reprogramming. Additionally, sharing their experiences through a central knowledge base or memory system can facilitate the learning and generalization of new behaviors.

Jaquier et al. (2024) suggest that this *transfer* can happen on different abstraction levels that correspond partially to the layers in the three-tiered robot architecture (see Figure 2.3). Furthermore, they argue that *transfer* comes naturally in the *Planning* layer, i. e., in the form of semantic descriptions, while the *Executive* layer facilitates the translation of these descriptions to the low-level, hardware-dependent instructions required by the *Behavioral Control* layer. For this reason, the *Executive* layer is the central place that needs to support the *transfer*, and therefore, the adaptation of skills across the three modes of *robots*, *environments*, and *tasks*.

However, most current *Executive* frameworks for mobile manipulation are designed with specific scenarios, applications, or contexts in mind, making it challenging to reuse skills in different circumstances (see Section 2.3.1). In fact, most frameworks facilitate only *single*- (e. g., Borghesan et al., 2014; Hermann et al., 2011; Keleştemur et al., 2019) or *dual*-mode (e. g., Dömel et al., 2017; Martins et al., 2023; Staroverov et al., 2023) *transfer* of mobile manipulation skills (see Table 2.4). However, there is no framework yet that explicitly focuses on transferring knowledge, experience, and skills across all three modes. As a special mention, the Affordance Template (AT, Hart et al., 2014) framework has a similar focus on facilitating the *transfer* of mobile manipulation skills, with a key difference being that an AT has to be created for each task separately, hindering the *transfer* of capabilities and knowledge across *tasks*.

To address the three modes of *transfer* – *robot*, *environment*, and *task* – there is a need for a universal framework that facilitates the flexible design and implementation of mobile manipulation skills involving known and unknown objects in *unstructured environments*. To this end, MAkEable, a memory-centered and affordance-based framework for mobile manipulation that unifies the description and execution of actions across the different modes, is introduced in Section 5.1.2. MAkEable is the first *Executive* framework to explicitly tackle all three modes of *transfer*, allowing for the autonomous and semi-autonomous execution of uni- and multi-manual manipulation actions. Its flexibility is demonstrated in several use cases in Section 5.1.3 using the humanoid robots ARMAR-6 and ARMAR-DE, including uni- and bimanual *grasping*, *placing* of known and unknown objects, and transferring a drawer-*opening* skill to another robot. Additionally, a *pouring* task in simulation using an industrial manipulator confirms MAkEable's ability to accommodate different robots.

5.1.2. Memory-centered and Affordance-based Executive Framework

The MAkEable framework is designed to facilitate the autonomous execution of mobile manipulation tasks in *unstructured environments* across various robotic platforms. The framework is embedded within the cognitive memory architecture of ArmarX (Vahrenkamp et al., 2015) introduced by Peller-Konrad et al. (2023). This integration fosters the accumulation of a rich repository of mobile manipulation experiences for collaborative learning among robots. Using an interpretable data format, an affordance-based *task description* is defined to facilitate a universal and generalizable formulation of mobile manipulation skills. Based on this *task description*, a five-*stage* system architecture is realized that is modular and easily adaptable to the current circumstances. A visualization of MAkEable can be seen in Figure 5.1.

Note: Reprinted from Pohl et al. (2024). © 2024 IEEE.

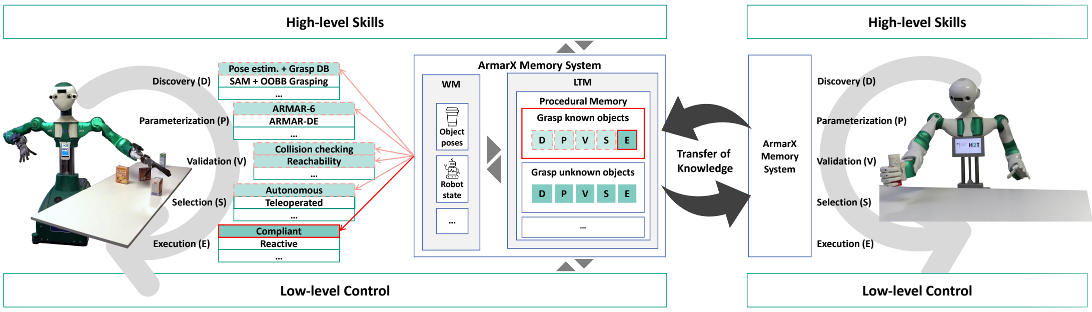
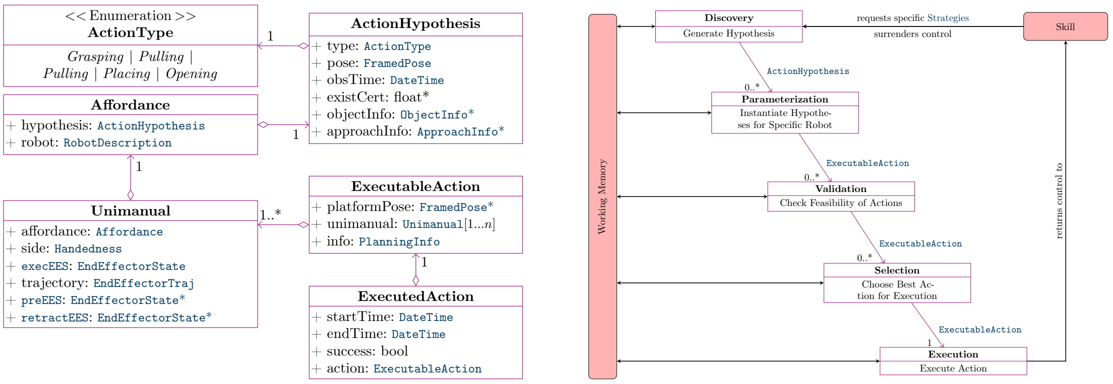


Figure 5.1.: Embedding of MAkEable into the memory-centric cognitive architecture (Peller-Konrad et al., 2023) implemented in ArmarX. Several strategies that implement the five *stages* of the architecture (see Section 5.1.2) are connected to the robot's memory.

Affordance-based

The foundation of MAkEable is the universal, affordance-based *task description* formulated using the Interpretable Data Format (IDF, Peller-Konrad et al., 2023). It consists of multiple data structures that represent the different stages of MAkEable's manipulation cycle. In its center is the `Affordance` class, which – analogously to the definition of an affordance in cognitive psychology (Gibson, 1979) – corresponds to an interaction possibility of an agent with its environment. It, therefore, acts as the liaison between the environment and the robot. Since affordances are per definition agent-specific, the notion of the action hypothesis from Section 4.1.2 is used as a robot-agnostic counterpart for the representation of the environment. An `ActionHypothesis` therefore represents an end-effector pose in an Abstract Affordance Frame (AAF; see Section 3.2.2) associated with an `ActionType`, such as *Grasping*, *Placing*, or *Pulling*. This abstraction allows for the extraction of action candidates from visual perception, independent of the specific robot, thus facilitating the *transfer* of skills. Furthermore, by assigning action possibilities to relevant objects and locations, the affordance-based formulation of the *task description* allows reasoning about what can be done with objects rather than focusing on which specific robot is used. This makes affordance-based representations easier to *transfer* across different agents and environments.

Note: Adapted from Pohl et al. (2024). © 2024 IEEE.



(a) Simplified class diagram of the IDF task description. Types marked with a "*" are optional.

(b) Data flow visualizing the interaction of the IDF task description with the system architecture.

Figure 5.2.: Class and data flow diagrams for MAkEable.

In order to execute mobile manipulation actions based on an `Affordance`, the `ExecutableAction` encapsulates all necessary information tailored to a specific robot. It can include up to n `Unimanuals`, each containing details pertinent to a single end-effector. By generating an `ExecutableAction` with a `Unimanual` for each end-effector, MAkEable supports the execution of complex multi-manual manipulation actions. A `Unimanual` action is defined as an `EndEffectorTraj`, which is a trajectory consisting of `EndEffectorState` as the keypoint's type, executed at the pose of the `execEES`. Each `EndEffectorState` consists of a `FramedPose` and optionally includes finger-joint values or a hand-shape name. Optional `preEES` and `retractEES` can be specified to define safe poses for the end-effector before and after the action has been executed. Post-execution, the result, and all relevant information are stored in an `ExecutedAction` object for memory introspection and continual learning, enhancing the robot's ability to adapt and improve over time. A class diagram of the *task description* can be seen in Figure 5.2a.

Memory-centric System Architecture

The memory-centered system architecture of MAkEable is designed to support the discovery and execution of mobile manipulation actions in *unstructured environments*. It has to be flexible enough to adapt to various *tasks*, *environments*, and *robots*. The architecture divides the overall task of generating and executing actions into five distinct *stages*, ensuring a structured and modular approach to task execution. This design allows for the seamless integration of new tasks and the adaptation of existing ones, promoting the *transfer* of knowledge and experience across different contexts and circumstances. The five *stages* of the MAkEable's manipulation cycle are:

1. ***Discovery*** of potential actions candidates (`ActionHypothesis`) from various, multi-modal inputs (e. g., visual perception or prior-knowledge).

2. ***Parameterization*** of `ExecutableActions` by enriching an `ActionHypothesis` with robot-specific data. This involves deriving all necessary information required for execution (e. g., `EndEffectorTraj`, robot base poses, etc.).

3. ***Validation*** of all created `ExecutableActions` by e. g., checking for collision, calculating the IK, in order to ensure the feasibility of an action before executing it.

4. ***Selection*** of the `ExecutableAction` that is best suited for execution through autonomous ranking or teleoperation

5. ***Execution*** of the `ExecutableAction` on the specific robot and storing the results in an `ExecutedAction`

Each *stage* is implemented using the Strategy behavioral design pattern (Gamma et al., 1993) to ensure enough flexibility of the architecture to be adapted to specific use cases. For example, there might be a Strategy for the *Discovery stage* that takes the `FramedPose` and `ObjectInfo` of an object to be grasped from the *Object Memory* to create a `ActionHypothesis`, and another Strategy might take raw camera images from the *Vision Memory* to directly derive a `ActionHypothesis` using one of the methods introduced in Chapter 3. In doing so, a combination of fitting Strategies for the task can be selected during runtime using a simple Finite State Machine. Each *stage* is embedded within the cognitive memory architecture of Peller-Konrad et al., 2023, meaning that all internal and external communication – i. e., between *stages* and with the robot, respectively – runs through the memory. This ensures that every step of the manipulation cycle is logged and introspectable by default. This modular design facilitates the easy integration of new Strategies (e. g., the methods developed in Chapter 3 for the *Discovery* and Chapter 4 for the *Selection stage*). Additionally, each *stage* in the manipulation cycle is equipped with specific interfaces to accommodate externally implemented Strategies, enhancing the system's flexibility for various use cases.

The data flow within the framework is designed to ensure seamless integration and execution of tasks. Figure 5.2b illustrates MAkEable's data flow, which is managed through a FSM that dynamically selects the appropriate Strategies based on the concrete type of IDF. This FSM orchestrates the workflow by triggering and parametrizing the *stages*, allowing for real-time decision-making and adaptability. Users can leverage high-level skills (such as "grasp an unknown object" or "place the object at a certain location"), which are stored in the robot's procedural memory (see e.g., Peller-Konrad et al., 2023), to request specific Strategy combinations, thereby tailoring the system to meet diverse and complex requirements. The behavior of each *stage* is then dynamically adapted based on the memory's content, such as object poses or common knowledge like typical fetching and placing positions. Since the knowledge within the memory is generalized and all robots share the same memory structure (i. e., distributed working and long-term memory in the form of memory servers and segments), execution knowledge can be seamlessly transferred from one robot to another.

5.1.3. Use Cases and Experiments

To assess the effectiveness of and adaptability to different circumstances of MAkEable, various real-world experiments using the ARMAR humanoid robots were performed. These experiments aimed to demonstrate the framework's ability to *transfer* knowledge across various *tasks*, *robots*, and *environments*. The scenarios included a table-clearing, a box-picking, and a drawer-*opening* task, each selected to validate MAkEable's design principles. Additionally, a simulation experiment was performed to illustrate the transferability across entirely different robot architectures, including industrial manipulators. Videos documenting the execution of all experiments are available on MAkEable's project page[1].

An `ExecutableAction` was executed using the approach described by Pohl et al. (2022), regardless of its `ActionType`. To this end, the execution is structured into four key stages: (a) positioning the robot's Tool Center Point (TCP) to a safe `preEES` close to where the action is executed, (b) moving the TCP into contact with the object (i. e., to the `execEES`), (c) executing the specific `EndEffectorTraj` associated with the `ActionType`, and (d) retracting the end-effector to the `retractEES` after the action is completed. The system uses VMPs in combination with a variable-stiffness impedance controller, where the stiffness is adjusted at each stage to balance precision and compliance—being more rigid during positioning and highly compliant during contact and execution to prevent damage and accommodate the interaction with the environment.

Table-Clearing Task

The table-clearing setup was designed similar to the experiments of Sections 3.2.3 and 4.1.3 to demonstrate the generalization and transfer of manipulation tasks across different robots using the two humanoid robots ARMAR-6 and ARMAR-DE.

[1] `https://sw.pages.h2t.iar.kit.edu/makeable/project_page/`

The primary objective was to evaluate the robots' ability to grasp both known and unknown objects and place them at common locations depending on the object, thereby covering various use cases. Both robots, ARMAR-6 and ARMAR-DE, are humanoid robots with two anthropomorphic 8 DoF arms and underactuated five-finger hands with 2 DoF (ARMAR-6) and 4 DoF (ARMAR-DE). The setup involved seven different rigid and deformable household objects placed arbitrarily on a table in *structured* clutter. Known objects were associated with specific common places such as the *sink*, *kitchen countertop*, or *workbench*, and the robots prioritized manipulating these known objects before addressing unknown ones. This scenario aimed to test MAkEable's ability to handle a mixed setup of known and unknown objects and place them at the correct locations.

The technical execution of the experiments involved utilizing RGBD-based pose estimation on both robots and additional stereo-based pose estimation on ARMAR-6 for 6D object pose estimation. ARMAR-DE was capable of recognizing objects such as *mustard*, *bio-milk*, *apple-tea*, and *spraybottle*, placing them on the *countertop* or *workbench* as appropriate. In contrast, ARMAR-6 could recognize additional objects like the *screwbox* and *sponge*, placing them on the *workbench* and *sink*, respectively. Objects that were unrecognized or unknown were placed on a free table next to the kitchen. The OOBB-based grasp candidate extraction from Grimm et al. (2021) was used to grasp the unknown objects. This allowed the robots to identify and manipulate objects with varying degrees of familiarity.

Note: Reprinted from Pohl et al. (2024). © 2024 IEEE.

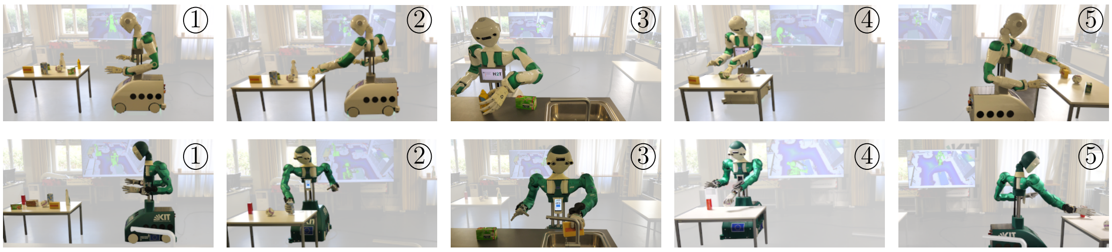


Figure 5.3.: Table-clearing of known and unknown objects with ARMAR-DE and ARMAR-6. ① Initial setup, ② *grasping* of known objects, ③ *placing* of known objects, ④ *grasping* of unknown objects, and ⑤ *placing* of unknown objects.

The results of the experiments, illustrated in Figure 5.3, showed that both ARMAR-6 and ARMAR-DE successfully cleared the table by recognizing and placing known objects in their designated locations and handling unknown objects appropriately. The differences in 6D object pose estimation between the two robots led to variations in object recognition, but both robots were able to complete the task effectively. Addtionally, for all `ActionTypes` (i. e., *Grasping* and *Placing*) and both robots, the same `EndEffectorTraj` was used: a TCP-finger-trajectory that gradually closes the fingers and curls the wrist, which was executed in reverse for *placing* an object. This showcases the *transfer* of knowledge across *robots*, as well as *tasks*.

Bimanual Pick-and-Place Task

The bimanual pick-and-place experiments were designed to demonstrate MAkEable's ability to execute multi-manual actions. Additionally, these experiments were conducted in a semi-autonomous setup, showing the capability to incorporate user feedback and experience through teleoperation (in this case in the *Discovery stage*). In this case, a bimanual grasp consisted of simply synchronizing a unimanual grasp for each end-effector of ARMAR-6. After successfully lifting an object, ARMAR-6 autonomously navigated to a designated location to place the object, employing a strategy similar to unimanual placing but again synchronized for bimanual execution.

Note: Reprinted from Pohl et al. (2024). © 2024 IEEE.

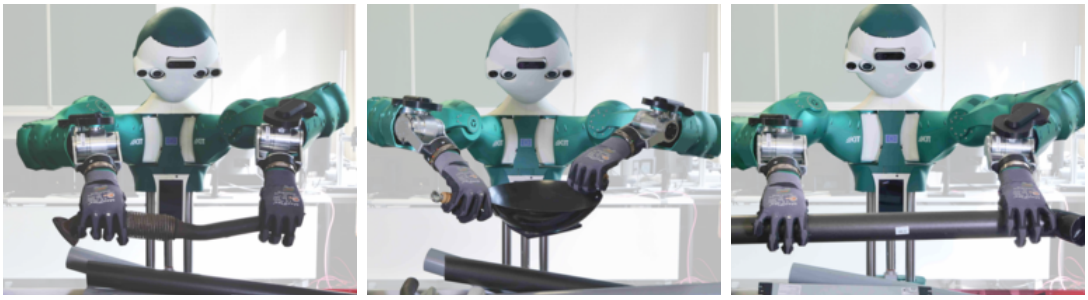


Figure 5.4.: The humanoid robot ARMAR-6 *grasping* and carrying multiple objects (exhaust, pan, and pipe) bimanually.

For the generation of a bimanual `ActionHypothesis`, a human operator selects two grasp points in the *Discovery stage* – one for the left end-effector and one for the right – in the scene via a graphical user interface. Then, grasp candidates are generated based on the local surface structure of the neighborhood of the selected points using GAE (see Section 3.2). Finally, the robot executes the grasp after the operator's approval (given during the *Selection stage* as only one `ExecutableAction` was generated), with synchronization between the arms handled by predefined keypoints (i. e., `preEES`, `execEES`, end of the `EndEffectorTraj`, and `retractEES`) and compliant control. A few examples of ARMAR-6 *grasping* diverse unknown objects bimanually can be seen in Figure 5.4.

Drawer-Opening Task

To demonstrate the relevance of MAkEable for Learning from Demonstration and the *transfer* of knowledge and experience across different *robots*, a drawer-*opening* experiment was performed, in which ARMAR-DE learned and executed a *task* that was previously taught to ARMAR-6. The primary objective was to assess the framework's ability to *transfer* learned skills between different robotic platforms, and, by doing so, showcase the beneficial impact of MAkEable on robot learning.

In this experiment, ARMAR-6 was tasked with *opening* a drawer. Although ARMAR-6 understood what it means to open a drawer (i. e., it can handle an `ActionHypothesis` with the `ActionType` "*Opening*"), it lacked the `EndEffectorTraj` to interact with the drawer. The robot utilized its prior knowledge of the drawer's handle position to derive and approach a suitable `execEES`. A human then demonstrated the motion for *opening* the drawer through kinesthetic teaching for ARMAR-6's right hand, which was stored in the robot's procedural memory. Afterwards, the trajectory was transferred (by simply copying the `EndEffectorTraj`) to ARMAR-DE, which then executed it using its left hand. This outcome underscores MAkEable's ability to abstract and generalize knowledge using its universal *task description*, which facilitates the sharing of experience across different *robots* to enhance the adaptability and *autonomy* of robotic assistants in real-world applications. Key scenes from the drawer-*opening* task can be seen in Figure 5.5.

Note: Reprinted from Pohl et al. (2024). © 2024 IEEE.

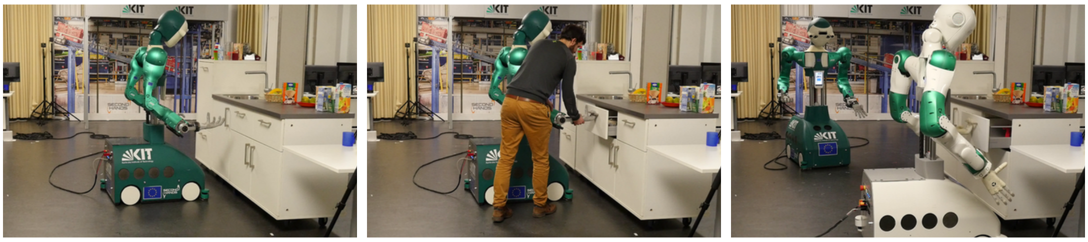
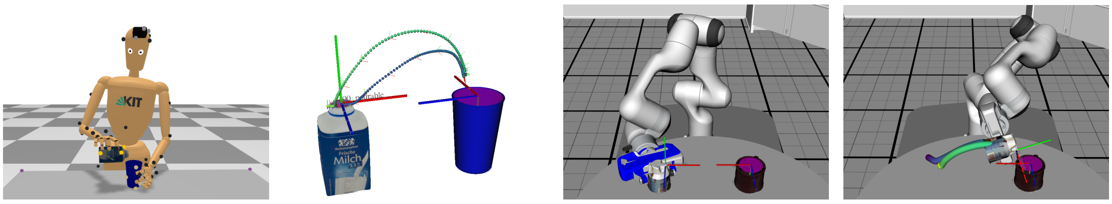


Figure 5.5.: Transfer of drawer-*opening* skill, which was learned through kinesthetic teaching, from ARMAR-6 to ARMAR-DE.

Relational Pouring in Simulation

To show that MAkEable can be used to *transfer* knowledge and experience between different *embodiments*, a simulated experiment was designed where a human motion was executed on an industrial manipulator. To this end, the *grasping* and *pouring* skills were instantiated for an Omni-Frankie robot (Haviland et al., 2022), which features a 7-DoF Franka-Emika Panda manipulator mounted on an Omron LD-60 two-wheel differential-drive base. Unlike the ARMAR humanoid robots, Omni-Frankie is equipped with a parallel-jaw gripper. Despite these differences, MAkEable's universal task and robot description enabled the autonomous generation of grasp hypotheses for the robot and its gripper in the same way as in the table-clearing experiments, allowing it to grasp and lift the object successfully.

Note: Reprinted from Pohl et al. (2024). © 2024 IEEE.



(a) Example (b) Reference motion (c) Grasped mug (d) *Pouring*

Figure 5.6.: Exemplary *pouring* motion selected from the *KIT Bimanual Manipulation Dataset* (Krebs et al., 2021) (a). Extracted reference motion of the bottle's *pourability* AAF relative to the cup's *fillability* AAF (b). OmniFrankie *grasping* a milk jug (c) and *pouring* milk into a mug using an `EndEffectorTraj` learned and adapted from human demonstrations (d).

Instead of focusing on the motion of the robot's gripper, the approach for *pouring* concentrates on the motion of two AAFs relative to one another, as inspired by Muhlig et al. (2009). This facilitates the execution of tasks that require considering two affordances. Examples include hitting a nail with a hammer (*contact* of the hammer head with the nail) or pouring water from a bottle into a glass (*pouring* out of the bottleneck and *filling* into the glass). In the latter case, the motion of the *pouring* frame of the bottle can be described within the *filling* AAF of the glass, as illustrated in Figure 5.6b. This representation can be easily transferred to new *pouring* tasks (e. g., using different objects) as long as the AAFs are included in the object description. For the simulation, VMPs are learned from the reference motion of a human from the *KIT Bimanual Manipulation Dataset* (Krebs et al., 2021), and subsequently transferred to MAkEable's universal *task description*. For the execution of the *pouring* task, an `ActionHypothesis` was generated using this knowledge for the Frankie robot, which then executed a *grasping* (Figure 5.6c) followed by a *pouring* (Figure 5.6d) skill adapted to the new objects and their poses.

When combined, the results from the experiments demonstrate MAkEable's ability to adapt manipulation knowledge and experience to various circumstances, thereby addressing Research Question 3. These observations underscore the framework's potential to generalize learned behaviors, thereby addressing the research question of how mobile manipulation knowledge and experience can be transferred across *robots*, *tasks*, and *environments*.

5.2. Affordance-based Task Planning with Large Language Models

Having introduced the highly-flexible *Executive* framework MAkEable in Section 5.1, this section will investigate the *Planning* layer of the three-tiered robot architecture. Specifically, the main focus is to demonstrate how combining a flexible *Planning* system with MAkEable can further improve the adaptability in real-world applications. The *Planning* layer is responsible for high-level decision-making and strategy. It interprets goals and tasks, generating a sequence of actions or plans to achieve those objectives. This layer typically coordinates the robot's activities by considering the current environment, future states, and potential obstacles. Recently, the integration of LLMs into the *Planning* layer (such as Brohan et al., 2023c; Liu et al., 2023a; Song et al., 2023) has led to a large improvement over conventional planners, as LLMs offer great generalization capabilities and understanding of natural language instructions. As MAkEable already demonstrated, employing an affordance-based environment representation can largely enhance the adaptability of the execution of mobile manipulation actions.

Therefore, this section will introduce AutoGPT+P, a LLM-based *Planning* system that uses the concept of affordances to increase the flexibility of the PDDL domain by describing actions in terms of the functionality of objects instead of their concrete class. This has the additional advantage of being able to handle incomplete information and, therefore, relaxing the closed-world assumption of conventional planners. Being able to adapt to the current circumstances while planning, possibly tackling solvable sub-problems first (like exploring the environment for objects) or suggesting replacements for missing objects based on their affordances, improves the success rate of AutoGPT+P for planning in real-world applications.

The following content has previously been published in the paper by Birr et al. (2024). The rest of this chapter will detail the implementation of AutoGPT+P and its implications on the *autonomy* of robots in the personal sector.

5.2.1. Motivation

Section 5.1 has demonstrated the advantages of an *Executive* framework that facilitates the *transfer* of knowledge and experience across *tasks*, *environments*, and *robots*. However, per design, MAkEable is only able to execute single actions like *grasping* and object or *placing* it. Its role as an *Executive* framework is to translate the high-level requirements for a task into the low-level instructions needed by the *Behavioral Control* layer. Therefore, planning mobile manipulation actions is of great importance for applications in the *service* and *assistance* domains, as many tasks require multiple steps to be fulfilled. For example, the simple request of "Bring me a glass of milk!" can require the execution of multiple manipulation actions. To this end, a robotic assistant might need to go to the fridge, open its door, grasp the bottle of milk, and pour the milk into a glass to bring it to a human. To create this sequence of actions from a user request is the goal of the *Planning* layer of robot architectures. Increasing the adaptability of this layer is of great importance for the overall *autonomy* of robots, as problems can arise at every step of the plan. Staying with the above example, once the robot opens the fridge, it might realize that there is no milk in it. Conventional planners would abort the plan at this point and the request of the user would remain unfulfilled. Therefore, Research Question 3 can be further addressed by combining MAkEable with a flexible *Planning* system that can capitalize on MAkEable's abilities to *transfer* knowledge and experience.

Natural language interaction is essential for improving the usability and efficiency of robots, especially in situations where they have to coexist or cooperate with humans. Studies have demonstrated that natural language commands offer an intuitive and effective means for humans to communicate with robots Liu and Zhang, 2019. Recently, LLMs have shown great promise for improving the capabilities and the adaptability of the *Planning* layer (see Section 2.3.2 and Table 2.5), as they excel in understanding and generalizing natural language tasks. However, they struggle to translate instructions directly into executable plans for robotic tasks. This limitation is mainly due to their restricted reasoning abilities (Valmeekam et al., 2022), which hinder their effectiveness in handling the complexities of task planning in dynamic environments (e. g., Brohan et al., 2023c or Valmeekam et al., 2024). Recent efforts, such as LLM+P (Liu et al., 2023a), have sought to improve LLMs' planning capabilities by combining them with classical planners. However, these systems are constrained by the closed-world assumption, meaning they can only generate plans if all necessary objects are present. Additionally, they lack automated error correction and are susceptible to contradictory goal definitions, further limiting their applicability in real-world scenarios.

To address these limitations and improve the adaptability of the *Planning* layer to unforeseen circumstances, AutoGPT+P, a system designed to enable robots to execute tasks based on natural language commands, even when some objects required for the task are missing from the immediate environment, is introduced. AutoGPT+P enhances robots' ability to dynamically respond to such constraints by searching for missing objects, proposing alternatives, or progressing towards sub-goals. By employing an affordance-based environment representation, AutoGPT+P can dynamically deduce viable actions within a given scene, facilitating the formation of a plan to achieve the user's objective. For example, if a user requests a glass of milk but no glass is detected, AutoGPT+P can propose using a cup instead. Additionally, it has been shown that the combination of parametric and non-parametric memories, so-called Retrieval-Augmented Generation (RAG, Lewis et al., 2020), leads to improved content generation results for LLMs. Therefore, AutoGPT+P extends the LLM+P approach by incorporating automated semantic and syntactic error correction and dynamic planning domain generation based on the robot's capabilities by connecting it to a non-parametric, cognitive memory architecture (Peller-Konrad et al., 2023), similar to RAG. This allows the robot to seek human assistance when encountering tasks beyond its capabilities, such as opening a milk box. Therefore, AutoGPT+P ensures that the robot can fulfill the user's request despite environmental limitations, thereby addressing Research Question 3 by increasing the system's adaptability and *autonomy*.

The main contributions of AutoGPT+P are as follows: (i) A novel affordance-based scene representation that combines object detection with an Object Affordance Mapping (OAM) automatically generated using ChatGPT. (ii) A task planning approach based on the established OAM and an LLM-based tool selection to generate plans, partial plans, and explore alternatives in case of missing objects needed to achieve a task goal specified by the user in natural language. (iii) Real-world validation experiments with the ARMAR humanoid robots demonstrating that a combination with MAkEable can improve the adaptability of task *execution*.

In Section 5.2.2, the AutoGPT+P framework and its affordance-based environment representation will be introduced. Subsequently, Section 5.2.3 will give an insight into the evaluation of the system and the positive impact of AutoGPT+P to the adaptability of the execution of mobile manipulation actions, thereby linking the approach to the main objective of this thesis.

5.2.2. Affordance-based Planning in Unstructured Environments

AutoGPT+P's approach for task planning consists of two stages: First, an affordance-based scene representation is extracted from visual perception using off-the-shelf object detection and the OAM. Second, the AutoGPT+P feedback loop is used to iteratively improve the PDDL domain and problem description in order to solve the user-specified task despite incomplete knowledge. Both steps depend largely on the generalization capabilities and the natural language understanding of LLMs. In the following, the details of both stages will be further explained. Additionally, both steps rely heavily on the concept of affordances, as it facilitates actions in the planning domain to be defined by the functionality of the objects involved and not their semantic class, allowing for a more generic planning approach (Lörken and Hertzberg, 2008).

Scene Representation

The affordance-based scene representation of AutoGPT+P establishes a foundation for relaxing the closed-world assumption of the PDDL planning problem. By representing the interaction possibilities of an agent ζ in a scene S, they allow for symbolic planning using Conceptual Equivalence Classes (CECs, Varadarajan and Vincze, 2011), i. e., sets of objects that share the same functionalities, instead of the semantic class. This improves the generality of the planning domain Δ and facilitates the suggestion of alternatives if an object is missing. To this end, a scene $S \in \mathbb{S}$ is represented as a set of object-affordance pair p_i, i. e., $S = \{p_1 \ldots p_n\}$.

For the use in AutoGPT+P's scene representation, the *representationalist* view of affordances by Şahin et al. (2007) was adopted. This means that, similar to the concept of an action hypothesis in Section 4.1 and Section 5.1, the agent's capabilities are not taken into account when observing the scene. Instead, the semantic information about the functionality of an object that is encoded in its class is used to assign affordances. To this end, objects in an image are first identified using a conventional object detector, and subsequently, an offline-generated OAM is used to map the object to its affordances.

In contrast to Varadarajan and Vincze (2011), where the CEC was defined using multiple ontologies and grasp datasets by categorizing objects with respect to their affordances, the OAM uses the advanced commonsense knowledge (see e. g., Kandpal et al., 2022) of LLMs to query the affordances that a class has directly. Therefore, the OAM (i. e., mapping from objects to affordances) represents the inverse of CECs (i. e., affordance to object mapping).

The OAM was created using multiple, simple, binary and atomic queries (e. g., "Can a typical <*object_class*> be used to contain fluids" for *fillability*) per affordance investigating the functionalities of an object class. If the LLM answered all of these questions with "yes", the object class was assigned the corresponding affordance. An ablation study on the effectiveness of different querying strategies for the creation of the OAM can be found in Appendix C.1.

Tool-based Architecture

The tool-based architecture of AutoGPT+P is designed to generate plans from user commands by defining a PDDL goal state Ω through the use of LLMs and iteratively updating the robot's memory, so that the definition of the domain Δ from the memory allows a conventional planner to solve the problem Ξ. The main planning loop queries an LLM to select the appropriate tool based on the current scene state and prior knowledge. This iterative process continues until a final plan is found or a threshold for maximum iteration is reached, as illustrated in Figure 5.7.

Note: Adapted from Birr et al. (2024).

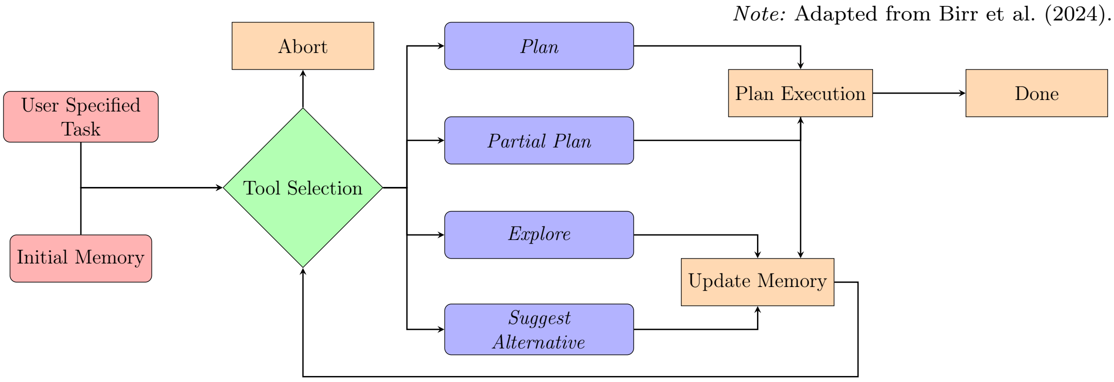


Figure 5.7.: Overview of the AutoGPT+P tool selection process presented in Section 5.2.2.

The planning task is formally defined as generating an action sequence or plan $P = (\alpha_1, \ldots, \alpha_n)$ given a scene $S \in \mathbb{S}$, object relations $R \in \mathbb{R}_S$, explorable locations $L \in \mathbb{L}$, and a user-specified task $\lambda \in \Lambda$ in natural language. The actions $\alpha \in \mathrm{A}$ are defined as capabilities $c \in C_\zeta$ executed by an agent ζ. A capability is a symbolic representation that specifies the parameters, logical preconditions, and effects of an action and is derived from the available skills in the robot's procedural memory at run-time. The scene S can be updated by exploring new locations location $l \in L$ and adding the object-affordance pairs p, which are detected in an image I taken at that location.

AutoGPT+P's approach to adapt to incomplete information and varying circumstances (see Research Question 3) is based on a hybrid approach combining the *Step-by-Step Autoregressive Plan Generation* and *LLM with Planner* paradigms (see Section 2.3.2). Specifically, it tries to relax the closed-world assumption by iteratively improving the amount of knowledge in the memory until the planning problem is solvable by a conventional planner. To this end, AutoGPT+P is centered around a tool-based architecture and planning loop that chooses the best possible tool for the current situation out of four possible tools:

Plan: Given the current knowledge in the memory, generate a planning problem for the user-specified task that a conventional planner can solve.

Partial Plan: Given the current knowledge in the memory, generate a planning problem that a conventional planner can solve, which addresses the user-specified task as much as possible under the current circumstances.

Explore: Move the robot to an unexplored location $l \in L$, extract the object-affordance pairs p from the camera image I and update the scene S.

Suggest Alternative: Suggest an alternative for a missing object that is crucial for the realization of a plan. The alternative object should be in the CEC for the object regarding its utility for the task.

A tool is selected based on the user prompt and current memory state, which includes the affordance-based scene representation, object relations, explorable locations, agent locations, instruction history, known alternatives, and the most recent plan. After selecting a tool, it is executed, and its results are written into the memory. This process is repeated until a final plan is reached or a maximum number of iterations is reached.

If an explicitly requested object is not available, AutoGPT+P can leverage the affordances involved in the user-specified task λ to suggest a replacement. This is done using a handcrafted *Chain-of-Thought* process (Wei et al., 2023): First, the LLM is queried to list the affordances of the missing object that are relevant to λ. Subsequently, the most relevant of these affordances in the scene is found using a heuristic (i. e., taking the least common affordance). Finally, the LLM is tasked to find the object that is most similar to the missing object regarding this affordance. If no replacement object could be found (e. g., no object has all the relevant affordances or the LLM returns an object not in the scene), the LLM is asked to find a replacement without any additional reasoning as a fallback.

Plan Generation

Inspired by Liu et al. (2023a), AutoGPT+P dynamically generates a PDDL domain $\Delta = (\Theta, \Upsilon, \mathrm{A})$ and problem $\Xi = (\Gamma, \Psi, \Omega)$ from the affordance-based scene representation S and user-specified task λ. The goal of the *Plan* tool is, therefore, to generate the goal state Ω in PDDL syntax with a LLM using the domain Δ and the initial state Γ, which are derived from the current state of the memory. Subsequently, the generated goal state is checked for semantic and syntactic errors and, if necessary, corrected by the LLM. Finally, the generated domain Δ and problem Ξ are given to a conventional planer for solving. An overview of the *Plan* tool can be seen in Figure 5.8. The procedure for the *Partial*

Plan is very similar, however it allows explicitly for incomplete goal states to be generated by the LLM, so that a plan for sub-goals can be created.

Note: Reprinted from Birr et al. (2024).

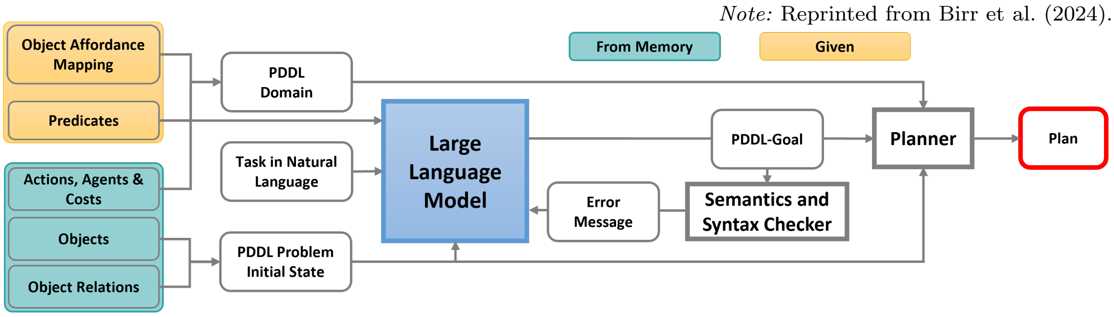


Figure 5.8.: Overview of the *Plan* tool. Rounded boxes represent the input and output of the components, which are represented as rectangles.

In PDDL, the types Θ are hierarchically structured by defining sub-types of a given type, with three top-level types for AutoGPT+P: *object*, *agent*, and *location*. To allow for interactions such as handing over an object, the *agent* is also a subtype of *location*. The type hierarchy is built by first declaring all affordances in $\mathbb{A}$ as subtypes of *objects* and then defining all objects that have this affordance as a subtype of it in turn. This is done by reversing the OAM to map affordances to object classes in the scene. For *agents*, all capabilities in C_ζ are defined as subtypes, and each agent with a specific capability c is further categorized under the type for c. In this way, human-robot collaboration is enabled by defining different *agent* types, such as robot and human, and dynamically assigning costs to them based on user preferences, influencing which agent performs an action α. The initial state Γ of the problem Ξ is defined by adding each object instance with its type and specifying current agent locations. The goal state Ω can then be determined using a LLM that has access to the initial state and domain definition. Using affordances simplifies the domain by allowing a single logical action for all objects with the same affordance (i. e., using CECs instead of semantic classes). This reduces the complexity of the domain and the search time for the planner, making the planning process more efficient and scalable.

Gou et al. (2023) demonstrated that conversational agents using LLMs can self-correct when provided with expressive error messages from external programs. Therefore, an automated check for syntactic and semantic errors in the goal state is performed. Syntactic errors, such as incorrect use of parentheses or invalid predicates, can be identified by matching predicate names and object types with those in the domain and initial state. Semantic errors involve the logical inconsistency of multiple predicates being true simultaneously, as when an object cannot be in two places at once, e.g., `(on apple table)` and `(on apple counter)`. Unlike Chen et al. (2023b), where semantic errors are tied to action sequences, this work defines them based on the feasibility of predicate co-occurrence. Semantic conditions are expressed in predicate logic and checked using Prolog by transforming the goal state into its Disjunctive Normal Form and mapping sub-states to Prolog predicates. If no sub-state satisfies all semantic conditions, the error message for the sub-state with the fewest errors is returned to the LLM for self-correction.

5.2.3. Experiments

To demonstrate the advantages of a combination of a flexible *Planning* system like AutoGPT+P with MAkEable, several experiments were performed on the humanoid robots of the ARMAR family. For all experiments, Fast Downward (Helmert, 2006) with a time limit of 300 seconds was used as the planner.

In addition to the real-world experiments, Birr et al. (2024) performed a series of ablation studies in simulation to evaluate the efficacy of the parts of AutoGPT+P. For the sake of completeness, these results are listed in Appendix C. The OAM evaluations show that using logical combinations of simple yes/no questions achieves the highest accuracy for generating the OAM. For the *Plan* tool, the results indicate better performance compared to standard *LLM as Planner* implementations, particularly in long-horizon and embodied tasks. While having no negative effect on the *Plan* tool in AutoGPT+P, the inclusion of affordances was found to reduce planning effectiveness due to information overload for the *LLM as Planner* baselines. However, the affordance-based approach for alternative suggestion outperforms a naive approach in identifying suitable replacements for missing objects. Error correction experiments showed that self-correction mechanisms led to a noticeable increase in success rates, especially in complex scenarios requiring commonsense reasoning. Lastly, the integration of all tools in complex tasks demonstrated AutoGPT+P's ability to effectively manage tool selection.

To evaluate the feasibility of using AutoGPT+P with MAkEable in mobile manipulation under realistic conditions, several experiments were conducted on the humanoid robots ARMAR-6 and ARMAR-DE. To facilitate the integration with the robotic platforms, several controlled assumptions were made. Predefined object models were used for manipulation tasks, including *grasping*, *placing*, and *pouring*. The robot operated within a fully known environment model, with fixed locations for navigation. However, the system dynamically detected the positions of all manipulable objects. For the detection of liquids within containers, a predefined liquid type was assumed in each container; however, to prevent potential damage to the robot, real liquids were not used during the experiments. Object relations were inferred based on the affordances of related objects, and the spatial relationships between object poses were estimated using a fine-tuned version of MegaPose (Labbé et al., 2022). Object detection was performed using the *yolov5* object detector (Jocher et al., 2022), which was fine-tuned on a predefined object set as described in Younes and Asfour, 2024. For grasping and placing MAkEable (see Section 5.1) was used, while for pouring tasks, the affordance keypoint detection method from Gao et al. (2023) was employed to identify the opening of the source container. It was assumed that the target container was symmetric, so aligning the source container's keypoint above the center of the target container was deemed sufficient. The experimental tasks comprised four types: *pick-and-place*, *handover*, *pouring*, and *wiping* tasks. These tasks varied in the degree of required human-robot collaboration, with the robot executing *pick-and-place* and *wiping* tasks autonomously but requiring human assistance to open liquid containers during *pouring* tasks. In *handover* tasks, both the robot and the human contributed equally to task completion.

The experimental results demonstrate the proficiency of the AutoGPT+P system in generating executable plans for robotic tasks. Out of 20 real-life scenarios, the system successfully planned 15. Analysis of failure cases identified that most errors were due to false positive object detections or the robot's inability to accurately grasp the target object. This shows AutoGPT+P's ability to adapt to varying circumstances under realistic conditions. By combining a flexible *Planning* framework, which facilitates the generation of plans that can adjust to incomplete or missing information and task specifications (thereby relaxing the closed-world assumption of conventional planners), with an *Executive* framework that focuses on transferability of skills like MAkEable, this section addresses Research Question 3 and contributes to the main objective of this thesis.

5.3. Conclusion

This chapter investigated how the adaptability of the *execution* of mobile manipulation skills in *unstructured environments* can be improved by combining a *Executive* framework focusing on the *transfer* of knowledge and experience with a flexible, affordance-based *Planning* system. By providing methods for adapting to the current conditions and requirements of the task, this combination increases the *task generality* while decreasing the amount of *task-specific knowledge* required for real-world *service* and *assistance* applications.

Section 5.1 presented the memory-centered and affordance-based task execution framework for transferable mobile manipulation skills, a modular *Executive* framework designed to unify the autonomous manipulation of both known and unknown objects across *environments*, *tasks*, and *robots*. It introduced an affordance-based *task description* that facilitates the generalization of manipulation knowledge. The memory-centered architecture of MAkEable enables unified internal and external communication and improves the introspection of the framework. The subdivision of the manipulation cycle into five distinct *stages* ensures the separation of concerns and makes the behavior of MAkEable easily adaptable to the circumstances and requirements of the current task. By bundling different versions of the *stages* into high-level skills, an easily accessible user interface through the memory is created. The integration in a cognitive memory system enhances contextual awareness, enabling the use of common knowledge in manipulation tasks and facilitating learning from both successes and failures across different agents.

In Section 5.2, a flexible *Planning* system, called AutoGPT+P, for the generation of mobile manipulation plans using a LLM-based hybrid architecture was introduced. By means of an Object Affordance Mapping, it leverages the general knowledge encoded in LLMs to create Conceptual Equivalence Classes that can be used to simplify the planning domain and suggest alternatives for missing objects necessary to fulfill the user-specified task. Combined with an iterative procedure centered around the four tools *Plan*, *Partial Plan*, *Explore*, *Suggest Alternative* to update the memory of the robot, this facilitates the generation of a successful plan despite the initial knowledge about the scene being incomplete. Additionally, the autonomously derived domain of AutoGPT+P supports human-robot collaboration by definition and the subsequently generated goal state is checked for semantic and syntactic errors before the plan is generated through a conventional planner.

Through extensive real-world and simulated experiments of varying use cases, MAkEable demonstrated its adaptability by changing the behavior of the *stages* depending on the current *task*, *environment*, and *robot*. Multiple realistic experiments on the humanoid robots ARMAR-6 and ARMAR-DE, like *grasping* known and unknown objects, *placing* objects depending on their common locations, or *opening* a drawer by learning a trajectory through Learning from Demonstration, showed the various applications of the framework to mobile manipulation. By supporting multiple *autonomy* levels, including full autonomy, semi-autonomy, and teleoperation, as well as the generation and execution of both uni- and bimanual actions, MAkEable can adapt to the various requirements and constraints on general-purpose robots when operating in real-world scenarios. The framework's capability to transfer knowledge and experience across *tasks*, *environments*, and *robot* is beneficial for the adaptability of robots in the real world where they have to cope with changing environments and circumstances. Furthermore, the validation experiments of a combination of AutoGPT+P with MAkEable on the humanoid ARMAR robots confirmed that the generated plans can be successfully executed on a robot when integrated with a flexible *Executive* framework that can transform the symbolic plans to the sub-symbolic representations required for execution on a real robot. This supports the claim that MAkEable and AutoGPT+P complement each other to form an overarching and adaptable task execution framework for mobile manipulation that can handle and adjust to changing circumstances and various real-world situations. Therefore, by addressing the Research Question 3 through adapting the execution of mobile manipulation skills, MAkEable and AutoGPT+P contribute to the main objective of this thesis.

6. Conclusion

The stated main objective of this thesis is to increase the degree of *autonomy* of robotic assistants in *unstructured environments* in order to facilitate their deployment to applications in different domains. To exhaust their full transformative potential in these environments, robots need to be able to handle a variety of tasks, deal with *uncertainties* and incomplete information, and adapt to changing requirements and conditions. Therefore, research in mobile manipulation tries to increase the *task generality* of autonomous robots while decreasing the amount of task-specific knowledge required.

The concept of affordances from cognitive psychology is utilized to represent environments in terms of their provided functionality instead of the explicit objects therein. The *representationalist* view of affordances synergizes well with a *discriminative* approach to mobile manipulation, as it facilitates the separation of robot-agnostic and robot-specific functionalities and can be mapped to the three stages of manipulation: *discovery*, *selection*, and *execution*.

Consequently, this thesis explicitly targets three *core capabilities* – one corresponding to each stage of the *discriminative* approach – that should be improved. The versatility of the *discovery* of interaction possibilities with the *unstructured environment* directly influences the robot's ability to perform a large number of tasks, thereby improving its *task generality*. Enhancing the reliability of the *selected* actions by accounting for *uncertainty* in visual perception and proprioception makes HRI more secure and fosters trust and acceptance of the robotic assistants. Finally, making the *execution* of mobile manipulation tasks more adaptable to external influences and situational circumstances decreases the amount of a priori, task-specific knowledge required to complete the robot's duties.

To improve the core capabilities of robotic assistants in mobile manipulation, this thesis investigates three research questions (Research Questions 1 to 3) by proposing three contributions (Contributions 1 to 3). To this end, each chapter addresses a single research question by detailing a contribution consisting of previously published papers.

6.1. Summary

Chapter 2 introduces the state of the art in *discriminative* mobile manipulation. Therefore, it is split into three sections, each dealing with one of the substages of *discovery*, *selection*, and *execution*. Due to the relevance to the experimental validations of the contributions of this thesis, this chapter focuses on works in the context of *grasping*. Therefore, Section 2.1 reviews work related to the *discovery* of grasp candidates for similar and unknown objects. Subsequently, Section 2.2 categorizes approaches for the *selection* of grasp candidates based on different *quality metrics*. Finally, in Section 2.3 focuses on the *Executive* and *Planning* layers of the three-tiered robot architecture for the *execution* of *grasping* and mobile manipulation skills.

Chapter 3 presents and discusses approaches for improving the versatility of the *discovery* of action hypotheses in *unstructured environments*. To this end, it concentrates on grasp *discovery* for similar and unknown objects. In Section 3.1, the Multi-feature Implicit Model (MIMO, Cai et al., 2024) is introduced, which is a neural fields-based network that facilitates the intra-class transfer of poses. Combined with Visual Imitation Learning, the Multi-feature Implicit Model can be used for task-oriented grasping and rearrangement of similar objects. However, as the Multi-feature Implicit Model does not facilitate the grasp synthesis without any prior knowledge of the objects involved, a Geometry-based Action Extraction (GAE, Pohl and Asfour, 2022) based on the local surface structure of point clouds is presented in Section 3.2 for grasping and manipulation of unknown objects.

Chapter 4 is concerned with increasing the reliability of the *selected* action candidates. It focuses on probabilistic approaches that enhance the robustness of grasping and manipulation by accounting for perceptual and proprioceptive *uncertainties*. In Section 4.1 a combination of a UKF and HMM is used by the Probabilistic Action Extraction and Fusion (PAEF, Pohl and Asfour, 2022) to spatiotemporally track the state of action hypotheses across multiple observations of a scene in order to estimate the existence certainty and covariance of the pose of an abstract affordance frame connected to the action. Subsequently, the Uncertainty-Aware Sensitivity Optimization (UASO, Baek et al., 2022) from Section 4.2 uses these estimates in combination with other *uncertainty-affected* grasp metrics to optimize a grasp score based on the *sensitivities* of the metrics towards the success rate of grasp executions in a dataset.

Chapter 5 is comprised of the contributions regarding the adaptability of the *execution* phase. To this end, a Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills (MAkEable, Pohl et al., 2024) is introduced in Section 5.1. This is an *Executive* framework that explicitly facilitates the *transfer* of knowledge, experience, and skills across *tasks*, *environments*, and *robots* by creating a universal, affordance-based *task description*. As a complementary system for MAkEable in the *Planning* layer, AutoGPT+P (Birr et al., 2024), a hybrid task planning approach based on LLMs, was presented in Section 5.2. AutoGPT+P relaxes the closed-world assumption of conventional planners by using an affordance-based scene representation for simplifying the PDDL planning domain and suggesting alternatives for missing objects in the scene.

6.2. Contributions

The three research questions formulated in Section 1.1.2 provide the foundation and structure of this thesis by defining the scope of the scientific framework, which was addressed in Chapters 3 to 5. Therefore, they will be summarized and addressed hereafter.

Research Question 1 is investigating how flexible action hypotheses can be extracted from the visual perception of unstructured environments. The versatility of action *discovery* is particularly important for real-world applications of general-purpose robots as it ensures their usability in a broad range of tasks. The Multi-feature Implicit Model and the Geometry-based Action Extraction are the two components of Contribution 1 that address this research question detailed in Chapter 3. The improved spatial features of the Multi-feature Implicit Model facilitate the task-oriented grasping and rearrangement of similar objects using only partial views. Validation experiments on the humanoid robots ARMAR-6 and ARMAR-DE proved that the visual imitation learning approach in combination with the Multi-feature Implicit Model can facilitate the versatile manipulations in *unstructured environments*. In the case of unknown objects, the box-emptying and table-clearing experiments with the humanoid robot ARMAR-6 demonstrated the improved flexibility of grasp candidates generated by the Geometry-based Action Extraction when compared to a baseline method based on scene segmentation and OOBBs. Here, the success rate of grasping could be increased by almost 10% when using the Geometry-based Action Extraction. Collectively, these experiments demonstrate the increased versatility of action *discovery* when employing the methods developed in this thesis.

Research Question 2 considers the impact *uncertainty* has on the reliability of robotic assistants in realistic scenarios. Specifically, Contribution 2 investigates the benefit of incorporating probabilistic methods into the grasp *selection* process in order to improve the reliability of grasping and manipulation in Chapter 4. The evaluation of the Probabilistic Action Extraction and Fusion with ARMAR-6 showed that by spatiotemporally fusing action observations over multiple sequential camera images, the success rate of grasping in *unstructured environments* could be improved by almost 5% compared to the Geometry-based Action Extraction. This showed that the filtered pose of the abstract affordance frame is indeed more robust to noise in the visual perception. Additionally, by tracking the state of an action hypothesis with an HMM, a measure for the confidence in the existence of an interaction possibility can be established. As a consequence, the Uncertainty-Aware Sensitivity Optimization can use this existence certainty in combination with other grasp metrics to optimize a grasp score. In the experiments with ARMAR-6 the grasp *selection* using the optimized grasp score was able to improve the success rate of the executed grasps by more than 40% compared to randomly selected grasps. This supports the claim that integrating *uncertainty* measures in the grasp *selection* process can vastly improve the reliability grasping and manipulation.

Research Question 3 is concerned with the *execution* of actions in *unstructured environments* under changing conditions and incomplete information. As robots will eventually encounter new situations and the requirements for their application will evolve over time, the adaptability of task *execution* is fundamental for their continued *autonomy* in these scenarios. To this end, Contribution 3 examines the *Executive* and *Planning* layers of the three-tiered robot architecture paradigm for mobile manipulation software frameworks. The transferability of knowledge, experience, and skills across *tasks*, *environments*, and *robots* that MAkEable provides promotes flexible manipulation skills that lay the groundwork for the *Planning* layer. This adaptability has been demonstrated in multiple real-world validation experiments with the robots of the humanoid ARMAR family, including *pick-and-place* of known and unknown objects, bimanual semi-autonomous *grasping*, and learning from demonstration for a drawer-*opening* task. Subsequently, AutoGPT+P could capitalize on the transferable manipulation skills of MAkEable to prove its capabilities in multiple real-world mobile manipulation scenarios with ARMAR-6 and ARMAR-DE. Consequently, the combination of these two frameworks improves the adaptability of task *Execution* in real-world applications.

6.3. Outlook and Future Work

This thesis has contributed multiple improvements for the versatility, reliability, and adaptability of robotic assistants in *unstructured environments*, leading to more *autonomy* for real-world applications in the personal sector. Hereafter, a short outlook to future work centered around this thesis' contributions is given.

The Multi-feature Implicit Model has shown great potential in transferring geometrical knowledge across instances of the same class to improve grasping and rearrangement of similar objects, while the Geometry-based Action Extraction has improved the grasp synthesis for unknown objects. This leads to greater versatility of mobile manipulation in *unstructured environments* by decreasing the amount of *task-specific knowledge* needed. In the future, this could be further enhanced by combining both approaches to *transfer* geometrical knowledge independent of semantic classes, thereby utilizing prior knowledge for grasping geometrically consistent structures in unknown objects.

The integration of probabilistic methods in the *selection* process to account for perceptual and proprioceptive *uncertainties* had a significant impact on the reliability of grasping and manipulation. The combination of the Probabilistic Action Extraction and Fusion with the Uncertainty-Aware Sensitivity Optimization allowed for the estimation of noise in the visual perception of a scene and facilitated the optimization of a grasp score for choosing the best action in a situation. The analysis of the chosen grasp metric showed that the height of a grasp had the most influence on the outcome. Therefore, additional metrics should be investigated to find a set of informative values that might also apply to the execution of different actions than *grasping*. Furthermore, the combination of the Uncertainty-Aware Sensitivity Optimization with a cognitive memory architecture (e. g., Peller-Konrad et al., 2023) could enable lifelong learning of the sensitivities to continuously improve the *selection* process.

The combination of the memory-centered and affordance-based task execution framework MAkEable and the LLM-based planning framework AutoGPT+P increases the adaptability of mobile manipulation tasks by introducing more flexibility in the *Executive* and *Planning* layers, respectively, utilizing a top-down communication. However, the three-tiered robot architecture also intends for feedback to be given from the lower to the higher layers. Such feedback could potentially improve the functionality of the *Executive* layer in *unstructured environments* by facilitating a more reactive approach to manipulation. Similarly, incorporating more feedback for the *Planning* layer could help AutoGPT+P's fault tolerance and further improve the *autonomy* of robots.

Appendices

A. Affordances

The theory of affordances, initially proposed by Gibson (1966, 1979), suggests to represent the environment in terms of properties or possibilities that it offers to an organism. Therefore, affordances are a relationship between the organism and its environment, highlighting how an organism perceives its environment in terms of what it can do with it. In their view on Gibson's definition, Turvey (1992) define affordances as *dispositions* of the environment that become *actualized* if they combine with their counterpart. Disagreeing with this purely environment-centric view, Chemero (2003) and Stoffregen (2003) extend the *gibsonian* understanding of affordances by proposing to define affordances not as properties of the environment or the organism individually but as relations between the abilities of animals and features of the environment. Şahin et al. (2007) conclude that these different representations of affordances are caused by differing perspectives on where to place them. Accordingly, they argue that there exist three – not one – different perspectives: the (i) *agent perspective* (the agent realizes it has the affordance of interacting once it sees the object), the (ii) *environmental perspective* (the object offers the interaction potential to the agent), and the (iii) *observer perspective* (an external observer would say that the object-agent system affords this interaction). Subsequently, Şahin et al. define their own formalization as the tuple of $(effect, (entity, behavior))$, meaning that an agent can generate the *effect* by applying the *behavior* to the *entity*. This approach was coined by Chemero and Turvey (2007) as the *representationalist* view on affordances. They use hyperset theory to compare the *representationalist* and *gibsonian* formulations and argue that even though *gibsonian* systems are complex, they can still be modeled computationally and offer a more accurate representation of perception and action in natural systems.

Following the *representationalist* view while focusing on a developmental context for robotics, Montesano et al. (2008) formalize affordances using Bayesian Networks as the probabilistic relationships between actions, objects, and the resulting effects. These relationships are learned by the robot through exploration and interaction with the environment, using sensory inputs and motor outputs. In their framework, the robot does not inherently know what actions are possible; instead, it must learn these possibilities through interaction. In a complementary approach, Krüger et al. propose Object Action Complexes (OACs, Krüger et al., 2011) as grounded abstractions that link sensorimotor experiences with symbolic representations, thereby providing a hierarchical structure for autonomous cognitive robots to learn and refine their interactions with the environment. OACs extend the concept of affordances by formalizing not only the potential actions an object offers but also integrating prediction, execution, and learning processes to support adaptive behavior in cognitive systems. Centering around Conceptual Equivalence Classes (CECs, Varadarajan and Vincze, 2011), the Affordance Network (AfNet, Varadarajan and Vincze, 2013) and its extension to domestic robotics, Affordance Network Ontology for Robotics (AfRob, Varadarajan and Vincze, 2013), use the k-TR theory to visual perception to formalize affordances using structured ontologies. While AfNet employs the Recognition by Component Affordances and k-TR theories to define a broader database of affordance features for various objects, AfRob uses so-called afbits and affordance filtrations for scalable, real-time object recognition in robots. The formalization of affordances using Dempster-Shafer belief functions from Kaiser et al. (2018) allows for the integration and hierarchical composition of evidence from various sources, enhancing the robot's ability to detect and reason about action possibilities in complex environments. This approach facilitates the consistent fusion and propagation of affordance-related evidence, supporting more robust and adaptable robotic behaviors. Based on the affordance representation of Montesano et al., Moldovan et al. (2018) introduce relational affordances by considering the interactions and spatial relationships between multiple objects, rather than treating objects in isolation. This approach uses probabilistic programming to manage the complexity and uncertainty in multi-object scenarios, allowing for more generalized and effective manipulation in robotic systems.

A more detailed view about affordances and their application to robotics can be found in multiple reviews regarding this topic. Investigating the relation of perceiving objects, identifying the actions on them, and estimating the outcome or effect of applying the action, the works of Ardón et al. (2020, 2021) categorize the related work regarding their reliance on prior knowledge (similar to e. g., "known", "similar", "unknown" objects) and how this object-action-effect relationship is established. Another general review of affordances for mobile manipulation by Yamanobe et al. (2017) focuses on object recognition and grasping, manipulation, and planning. Additionally, it introduces a cloud database that accumulates various data related to manipulation tasks.

B. Ablation Studies for MIMO

To assess the proposed task-oriented grasp generation framework based on MIMO, multiple experiments and ablation studies across various manipulation tasks were conducted. By doing so, MIMO was compared against state-of-the-art methods, including Neural Descriptor Field (NDF, Simeonov et al., 2022), Relational-Neural Descriptor Field (R-NDF, Simeonov et al., 2023), and Neural Interaction Field and Template (NIFT, Huang et al., 2023). This evaluation assesses the versatility of the MIMO-based grasping and rearrangement framework and its effectiveness in generating grasps for similar objects. More details, evaluation videos, and source code are available via the project page[1].

Disclaimer

The experiments and results of this section were not contributed by this thesis and are only listed for completeness and better understanding. The original content is taken from:
Cai, Yichen, Jianfeng Gao, **Christoph Pohl**, and Tamim Asfour (2024). "Visual Imitation Learning of Task-Oriented Object Grasping and Rearrangement". In: *Proc. of the 2024 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*. International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, UAE: IEEE/RSJ, accepted for publication

B.1. Training and Experimental Setup

The training of MIMO employs a multi-task loss function that combines the loss functions of each feature branch through a weighted sum. To avoid the challenge of manually adjusting these weights, homoscedastic uncertainty (Kendall et al., 2018) is introduced for each branch. This approach models the likelihood as a Gaussian distribution where the output is the mean and the uncertainty is the variance. The total loss function, defined as $\mathcal{L} = \sum_{i=1}^{4}(e^{-s^i}\mathcal{L}_i(\mathbf{W_i}) + s_i)$, combines binary cross-entropy loss for occupancy, clamped L1 loss for signed distance, and L1 losses

[1] https://sites.google.com/view/mimo4

for ESCF and CDD branches. By minimizing this loss function with respect to the model weights $\mathbf{W_i}$ and uncertainty s_i, the training process ensures balanced and effective learning without manual tuning.

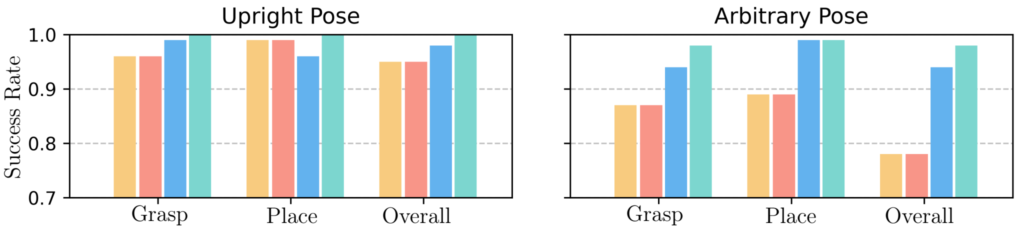
MIMO can be trained entirely without human annotation of data. The general procedure is very similar to that of NIFT: The training process of NIFT begins with the creation of a large dataset in a simulated environment containing 100,000 point clouds of various objects in random poses and scales. Spatial points are uniformly sampled around these objects, and their spherical function features are computed to provide ground truth for network training. The neural network, designed to predict the SCF for each point, is trained using an L1 loss for SCF and binary cross-entropy loss for occupancy prediction, optimized with the Adam optimizer over 50 epochs. In theory, both NIFT and NDF provide datasets that could be used for training; however, two issues prohibit their use for MIMO: (i) The bottom of meshes of the "bottle" class from NDF is hollowed out, which influences the shape reconstruction quality, and (ii) the scaling of meshes is non-uniform, leading to wrong labels for SCF and signed distance. Therefore, a new dataset consisting of watertight meshes with rendered point clouds for each mesh is generated from the ShapeNet dataset (Chang et al., 2015) using the methodology described in Stutz and Geiger (2018). Training NIFT and MIMO is conducted using the new dataset on a single NVIDIA A100 GPU, employing the pre-trained weights of NDF and R-NDF as provided by the original authors.

For the experiments in simulation, three different settings were considered: (S1) ten demonstrations with four viewpoints, where the point cloud is fused from four depth cameras positioned at the corners of the table; (S2) a single demonstration with four viewpoints, using the same camera positions, and (S3) a single demonstration with a single viewpoint, ensuring visibility of the mug handle and bottle opening The evaluation utilized Basis Point Set (BPS) for all models, distinguishing between upright (U) and arbitrary (A) initial object poses. The success of the task was determined by the grasp success (object grasped without dropping) and placement success (object correctly placed in the target pose).

The performance of MIMO is first evaluated against various state-of-the-art approaches. To demonstrate the effectiveness of the novel ESCF and CDD features in MIMO (denoted *MIMO4*), additional evaluation results are provided for a variant (denoted *MIMO3*) with three branches in the decoder to predict occupancy, signed distance, and SCF separately. Additionally, the results of *MIMO4* without shape reconstruction (denoted *MIMO4−*) are evaluated to investigate the effect of this step on grasp candidate generation.

B.2. Comparison with NDF

To benchmark the proposed approach, a comparison with NDF was conducted using a simulation environment. The evaluation included three pick-and-place tasks: (T1) picking a mug by the rim and placing it on a rack by the handle, (T2) picking a bowl and placing it on a shelf, and (T3) picking a bottle from the side and placing it on a shelf. Each task was performed 100 times under settings (S1) and (S3).



Note: Adapted from Cai et al. (2024). © 2024 IEEE.

Figure B.1.: Success rate of the pick-and-place tasks (T1)-(T3) with unseen objects under setting (S1) for models NDF(🟨), NIFT(🟥), *MIMO3*(🟦), and *MIMO4*(🟩), respectively.

As shown in Figure B.1, all approaches achieve high success rates for tasks (T1)-(T3) in setting (S1). *MIMO4* consistently achieves the best results, with *MIMO3* slightly lower. The overall success rates of *MIMO4* drop by only 2% in arbitrary poses compared to upright poses, demonstrating superior SE(3)-equivariance.

Table B.1.: Unseen object pick-and-place success rate with setting (S3) (single viewpoint, single demonstration).

		Mug (T1)			Bowl (T2)			Bottle (T3)			Mean		
		Grasp	Place	Overall	Grasp	Place	Overall	Grasp	Place	Overall	Grasp	Place	Overall
Upr. Pose U	NDF	0.95	0.73	0.72	0.89	0.93	0.84	0.90	0.69	0.65	0.91	0.78	0.74
	NIFT	0.99	0.92	0.92	0.98	**1.00**	0.98	0.96	0.94	0.90	0.98	0.95	0.93
	MIMO3	**1.00**	0.92	0.92	0.99	**1.00**	**0.99**	0.92	0.93	0.91	0.97	0.95	0.94
	MIMO4−	0.99	0.92	0.92	0.98	0.98	0.97	0.94	0.64	0.62	0.97	0.85	0.84
	MIMO4	**1.00**	**0.98**	**0.98**	**1.00**	0.99	**0.99**	**0.97**	**0.97**	**0.95**	**0.99**	**0.98**	**0.97**
Arb. Pose A	NDF	0.53	0.58	0.34	0.76	0.80	0.64	0.42	0.91	0.40	0.57	0.76	0.46
	NIFT	0.46	0.90	0.42	0.96	0.88	0.87	0.38	0.93	0.37	0.60	0.90	0.55
	MIMO3	0.86	0.94	0.80	0.94	**0.99**	0.94	0.77	0.87	0.71	0.86	0.93	0.82
	MIMO4−	0.53	0.96	0.50	0.97	0.95	0.94	0.67	0.52	0.50	0.72	0.81	0.65
	MIMO4	**0.92**	**0.97**	**0.90**	**0.98**	0.97	**0.95**	**0.95**	**0.97**	**0.93**	**0.95**	**0.97**	**0.93**

Note: Reprinted from Cai et al. (2024). © 2024 IEEE.

In contrast, Table B.1 shows that *MIMO4* significantly outperforms others in setting (S3), particularly in tasks (T1) and (T2) with arbitrary object poses. *MIMO3* and NIFT perform comparably to *MIMO4* only in the placing phase of

(T2) involving bowls. This is due to the partially-observed point cloud of bowls with large openings, making it easier to distinguish their orientation compared to mugs and bottles in (T1) and (T2).

Comparing the average success rates, methods incorporating shape reconstruction (*MIMO3*, *MIMO4*) outperform those without it (NDF, NIFT, *MIMO4−*), verifying the effectiveness of shape reconstruction. Interestingly, NIFT achieves a higher success rate in task (T3) for placing, likely because it is trained on bottles hollowed at the bottom, making it easier to distinguish the top and bottom.

Note: Adapted from Cai et al. (2024). © 2024 IEEE.

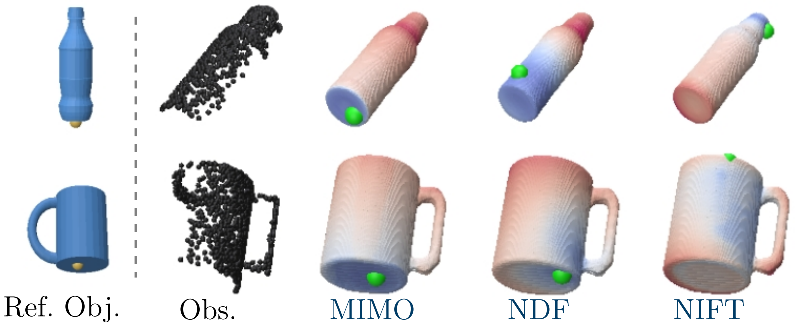


Figure B.2.: Point correspondence and shape similarity measure using point descriptors from partially-observed point clouds (●). Given a point on a reference object, the novel object mesh is colorized based on the L1 distance of point descriptors to the reference point, where blue means more similar and marks the most similar points (●).

As shown in Figure B.2, NDF and NIFT often fail to distinguish the top and bottom of bottles and mug handles, causing low success rates in (T1) and (T3) with arbitrary poses. MIMO's descriptor field is more informative, achieving accurate pose transfer and higher success rates.

In Figure B.3, the angle error between the object's upright direction and gravity is computed at the target pose for bowls and bottles in (T2) and (T3). A smaller angle indicates more precise placement. *MIMO4* has the smallest average angle error and variance across all tasks, further validating the superiority of the novel neural descriptors.

B.3. Comparison with R-NDF

In addition to the comparison to NDF, the effectiveness of the approach for task-oriented object rearrangement with respect to object relations was evaluated using the simulation environments from R-NDF, focusing on three specific tasks:

Note: Adapted from Cai et al. (2024). © 2024 IEEE.

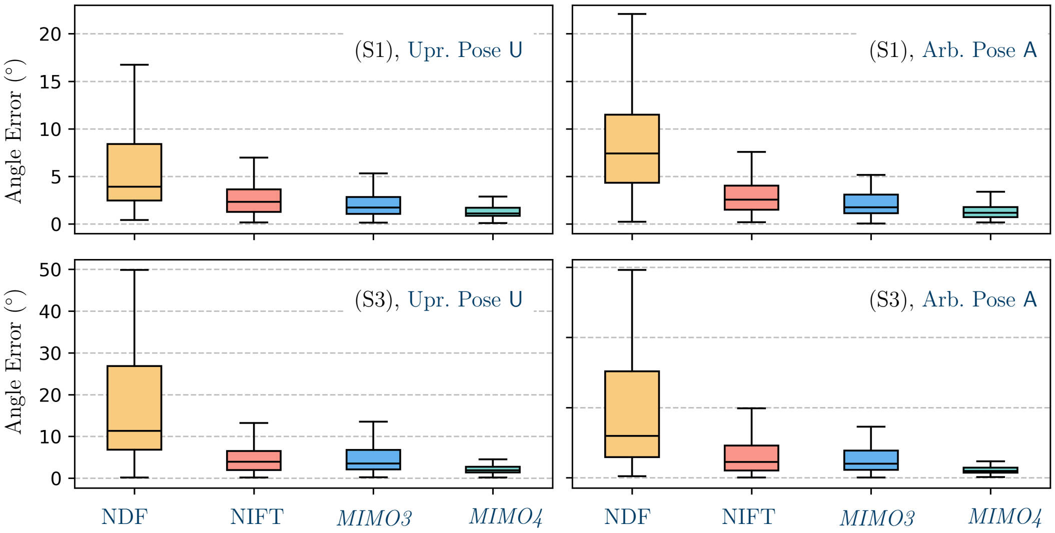


Figure B.3.: Angle error of bowls and bottles. Colors as Figure B.1.

(T4) hanging a mug on a hook, (T5) placing a bowl on a mug, and (T6) placing a bottle in a container. These tasks were tested under the three settings (S1), (S2), and (S3), concentrating solely on the target configurations of the objects and disregarding the grasp procedure. A task was deemed successful if the source object was placed on the target object without falling or exerting excessive force. Each task was conducted over 100 trials to compute success rates.

Table B.2.: Success rates of unseen object rearrangement. U and A stand for upright and arbitrary poses, respectively.

		(T4)		(T5)		(T6)		Mean	
	Models	U	A	U	A	U	A	U	A
(S1)	R-NDF	0.71	0.55	0.75	0.75	0.80	0.54	0.75	0.61
	MIMO3	**0.91**	**0.87**	**0.92**	**0.91**	0.84	0.85	**0.89**	0.88
	MIMO4	0.88	0.85	0.91	0.89	**0.87**	**0.93**	**0.89**	**0.89**
(S2)	R-NDF	0.56	0.53	0.64	0.61	0.12	0.18	0.44	0.44
	MIMO3	0.89	0.89	**0.90**	**0.88**	0.85	0.87	0.88	0.88
	MIMO4	**0.92**	**0.92**	**0.90**	0.87	**0.91**	**0.93**	**0.91**	**0.92**
(S3)	R-NDF	0.29	0.21	0.10	0.13	0.16	0.07	0.18	0.14
	MIMO3	0.85	0.85	0.88	0.87	0.72	0.70	0.82	0.81
	MIMO4	**0.89**	**0.86**	**0.90**	**0.88**	**0.90**	**0.83**	**0.90**	**0.86**

Note: Reprinted from Cai et al. (2024). © 2024 IEEE.

The results, as detailed in Table B.2, indicate that *MIMO4* and *MIMO3* performed equally well in setting (S1) with a success rate of approximately 89%. In settings (S2) and (S3), *MIMO4* significantly outperformed R-NDF by about 48% and 70%, respectively. The performance of *MIMO3* declined in these settings, demonstrating the effectiveness of the novel ESCF and CDD features in the partly shared decoder of MIMO.

B.4. Evaluation of the MIMO-based Grasping Framework

To evaluate the task-oriented grasp generation approach, additional experiments were conducted in simulation. Four specific tasks (equivalent to the tasks (E1) - (E4)) were defined for these experiments: (T7) grasping a mug at its rim and placing it upright, (T8) grasping a mug at its handle and pouring into a bowl, (T9) grasping a bottle at its neck and placing it upright, and (T10) grasping a bottle at its body and pouring it into a bowl. Object poses were randomly initialized, ensuring the visibility of mug handles. For all experiments, *MIMO4* was used to reconstruct object shapes from the partially-observed point cloud. The grasp poses are sampled from the GMM, transferred to the observed objects, and evaluated by the grasp evaluator (see Section 3.1.3). If the estimated success probability dropped below 0.9, the grasp pose was optimized with a learning rate of 10^{-3}.

Table B.3.: The success rates of unseen object grasping (G) and rearrangement (R).

	(T7)		(T8)		(T9)		(T10)		Mean	
Models	G	R	G	R	G	R	G	R	G	R
NIFT	0.80	0.62	0.92	0.80	0.86	0.08	0.92	0.68	0.88	0.55
MIMO4	**0.94**	**0.88**	**0.96**	**0.94**	**0.90**	**0.80**	**0.98**	**0.88**	**0.95**	**0.88**

Note: Reprinted from Cai et al. (2024). © 2024 IEEE.

The simulation experiments were executed using Isaac Gym and compared *MIMO4* against NIFT as a baseline. Each task was executed 50 times in (S3). As shown in Table B.3, *MIMO4* outperforms NIFT in all tasks, particularly excelling in (T9), with a success rate improvement of about 72%. As in previous experiments, the baseline approach struggled to differentiate between the top and bottom of the bottle, leading to failures in placement. In contrast, the MIMO-based task-oriented grasping framework achieved higher success rates, benefiting from the reconstructed shape and enhanced descriptor space, with an average success rate of 95% for grasping, including challenging side grasps at the mug handle.

C. Ablation Studies for AutoGPT+P

In order to investigate the behavior and benefits of the single components of AutoGPT+P (Section 5.2), multiple ablation studies in simulation were conducted. Appendix C.1 investigates the three different querying strategies used for generating the OAM. Furthermore, Appendix C.2 compares the efficacy of different planning paradigms on the SayCan instruction set. Subsequently, Appendix C.3 demonstrates the ability of the automatic error correction to improve the planning success rate of AutoGPT+P. Appendix C.4 compares the *Suggest Alternative* tool against a naive suggestion using LLMs. Finally, Appendix C.5 evaluates the efficacy of the entire AutoGPT+P planning loop and demonstrates its ability to correctly select tools for the circumstances.

Disclaimer

The experiments and results of this section were not contributed by this thesis and are only listed for completeness and better understanding. The original content is taken from:
Timo Birr, **Christoph Pohl**, Abdelrahman Younes, and Tamim Asfour (2024). “AutoGPT+P: Affordance-based Task Planning with Large Language Models”. In: *Proceedings of Robotics: Science and Systems*. Robotics: Science and Systems. Vol. 20. Delft, Netherlands

C.1. Object-Affordance Mapping

To assess the Object Affordance Mappings (OAMs), the key metrics[1] are precision (prec), recall (rec), and the F1-score (Chinchor, 1992; Van Rijsbergen, 1977; equivalent to DICE Dice, 1945), defined as follows:

$$\text{prec} = \frac{TP}{TP + FP}, \quad \text{rec} = \frac{TP}{TP + FN}, \quad \text{F1} = 2 \times \frac{\text{prec} \times \text{rec}}{\text{prec} + \text{rec}}$$

[1] All these metrics range from 0 to 1, with higher values indicating better performance.

where:

- TP (True Positives) represents the number of object-affordance pairs p correctly identified, i.e., those present in both the ground truth (GT) and the detected set.
- FP (False Positives) represents the number of object-affordance pairs that were detected but do not appear in the GT.
- FN (False Negatives) denotes the number of object-affordance pairs that are present in the GT but were not detected.

To generate the OAM, the reproduction abilities of LLMs when prompted with simple questions are utilized. Three different querying strategies were distinguished when generating the OAM:

1. The *List-Affordance strategy* involves querying the LLM for the affordances of each object directly and providing a list of affordances with descriptions. This method is efficient regarding token usage and speed but may lack accuracy due to potential ambiguities in affordance descriptions.
2. The *Yes/No-Questions strategy* improves accuracy by querying the LLM with binary questions about each affordance. This method allows for precise definitions of affordances, reducing ambiguity. However, it requires more tokens and time, as each affordance must be queried individually, making it a more resource-intensive approach.
3. The *Yes/No-Questions + Logical Combinations strategy* further enhances accuracy by breaking down complex queries into atomic yes/no questions. This method ensures that each query is simple and unambiguous, although it consumes even more tokens and time. It is the most accurate approach but also the most resource-intensive.

An independent training set of 30 object classes was used to optimize the prompts. The evaluation was conducted using a test set of 70 new object classes, each annotated with their corresponding affordances. These annotations were also utilized to evaluate the *Plan* and *Suggest Alternative* tools' functionalities. The F1-score was calculated for the three querying strategies for affordances and analyzed using a set of 40 affordances, partially derived from Varadarajan and Vincze, 2012, 2013.

As shown in Table C.1, *GPT-4* generally outperforms *GPT-3* across most strategies. The data suggests that the most effective approach combines yes/no questions with logical reasoning. However, despite the high accuracy achieved, the *uncertainty* in affordance estimation remains a consideration for future research.

Note: Reprinted from Birr et al. (2024).

GPT	List-Affordances			Yes/No			Logical		
	prec	rec	F1	prec	rec	F1	prec	rec	F1
3	0.31	0.49	0.38	0.70	0.78	0.74	0.78	0.85	0.81
4	0.59	0.67	0.62	0.78	**0.95**	0.86	**0.87**	0.91	**0.89**

Table C.1.: Comparison of OAM methods using different versions of the Generative Pre-trained Transformer (GPT) on the proposed set of affordances with the best values for precision, recall, and F1-score in bold

C.2. Planning Tool

The evaluation involved a series of scenarios, each defined by a user-specified task, a formal goal state, and the scene's specifications, including object relations, and explorable locations. The effectiveness of the generated plans was assessed by simulating their execution using Prolog and determining whether the resulting scene state met the desired goal state.

The *Plan* tool was evaluated against the provided code for SayCan on its instruction set and two additional *LLM as Planner* implementations. The first implementation generated plans based on a textual representation of the initial state and the user-specified task. The second implementation included additional affordance information about objects in the prompt to examine the impact of affordances on planning performance.

In Table C.2, the results of the evaluation on SayCan's instruction set can be seen. The proposed method outperformed the naive *LLM as Planner* implementations for both *GPT-3* and *GPT-4*, aligning with findings from previous studies. However, adding affordance information in the prompt reduced performance, likely due to information overload. Compared to SayCan, the method showed superior results when using *GPT-4*, particularly in scenarios involving embodiment and long-horizon tasks, although it performed worse with *GPT-3*. However, as SayCan is based on PaLM, the difference in LLMs employed makes it impossible to directly compare both approaches.

The results indicate that while *GPT-4* handles explicit tasks well, it struggles with tasks requiring nuanced interpretation of the user's intentions. Contextual understanding remains crucial, and the system should seek clarification when the goal is not clearly stated.

Note: Reprinted from Birr et al. (2024).

Instruction Family	SayCan (plan)	*GPT-3* As Planner	*GPT-3* As Planner+A	*GPT-4* As Planner	*GPT-4* As Planner+A	AutoGPT+P (*GPT-3*)	AutoGPT+P (*GPT-3*, Auto)	AutoGPT+P (*GPT-4*)	AutoGPT+P (*GPT-4*, Auto)
NL Primitive	0.93	0.47	0.53	0.93	0.93	0.73	0.73	**1.00**	**1.00**
NL Verb	0.60	0.00	0.00	0.67	0.87	0.27	0.33	0.93	**1.00**
NL Noun	0.93	0.13	0.07	0.26	0.20	0.27	0.40	0.93	**1.00**
Structured	**0.93**	0.20	0.13	0.87	0.60	0.00	0.13	0.20	**0.93**
Embodiment	0.64	0.09	0.00	0.55	0.55	0.64	0.64	0.82	**1.00**
Crowd-Sourced	0.73	0.13	0.07	**0.93**	0.73	0.27	0.33	0.73	**0.93**
Long-Horizon	0.73	0.00	0.00	0.40	0.33	0.20	0.33	0.80	**1.00**
Drawer	**1.00**	0.00	0.00	0.00	0.00	0.66	0.33	**1.00**	**1.00**
Average	0.81	0.14	0.12	0.66	0.59	0.34	0.40	0.78	**0.98**

Table C.2.: Ablation results of the planning success rate with the *Plan Tool* with different versions of GPT and with (Auto) or without automatic self-corrections on the SayCan instruction set. *GPT-X as Planner* refers to the naive baseline of using the LLM directly as the planner, *GPT-X as Planner+A* refers to the same planner with additional context information about affordances added to the prompt.

C.3. Error Correction

To identify limitations in the reasoning capabilities of a LLM in understanding user intentions, five subsets of scenarios were created, each containing 30 prompts with diverse tasks, such as cutting, heating, and moving objects. The subsets *Simple Task*, *Simple Goal*, *Complex Scene*, *Complex Goal*, *Knowledge*, and *Implicit* vary in complexity and the necessity for commonsense knowledge or implicit understanding.

The *Plan* tool's performance was evaluated across these subsets. *Simple Task* and *Simple Goal* involved straightforward requests, while *Complex Scene* and *Complex Goal* introduced higher complexity through more objects and logical subgoal connections. *Knowledge* required commonsense reasoning, and *Implicit* involved indirect user intentions. Additionally, the impact of scene complexity on planning time was assessed, alongside the role of self-correction mechanisms.

Note: Reprinted from Birr et al. (2024).

Subset	*GPT-3*		*GPT-3* Auto		*GPT-4*		*GPT-4* Auto	
	success	min	success	min	success	min	success	min
Simple Task	0.70	0.63	0.70	0.63	0.97	0.97	**1.00**	**1.00**
Simple Goal	0.63	0.60	0.90	0.83	**1.00**	**0.97**	**1.00**	0.93
Complex Scene	0.17	0.13	0.77	0.53	0.93	0.87	**0.97**	**0.93**
Complex Goal	0.23	0.17	0.33	0.23	**0.87**	0.70	**0.87**	**0.73**
Knowledge	0.10	0.10	0.10	0.10	0.53	0.53	**0.57**	**0.57**
Implicit	0.10	0.10	0.13	0.13	0.43	0.40	**0.47**	**0.43**
Average	0.32	0.29	0.49	0.42	0.79	0.74	**0.81**	**0.77**

Table C.3.: Ablation results of planning with the *Plan* tool with different versions of GPT and with or without automatic self-corrections (Auto) on AutoGPT+P's own instruction set. Success refers to the success rate, whereas min refers to the rate of plans that had the minimal length possible for the given goal.

As can be seen in Table C.3, the *Plan* tool performed reliably on simple tasks but struggled with more complex and vague tasks, especially in the *Knowledge* and *Implicit* subsets. *GPT-3* faced difficulties with vague instructions, while *GPT-4* showed better performance, often rendering self-correction unnecessary. The self-correction improved success rates slightly, more so for *GPT-3* than *GPT-4*. In the SayCan instruction set (see Table C.2), self-correction led to notable improvements, particularly in handling structured language (i. e., the *Structured* instruction set).

The experiment revealed significant increases in planning time as scene complexity grew. For the *Simple Goal* subset, where scenes involved around 30 objects, the average planning time was 8.4 seconds with *GPT-3* and 28.0 seconds with *GPT-4*. However, in the *Complex Scene* subset, which increased the object count to 100, the planning time rose sharply to 31.6 seconds for *GPT-3* and 59.4 seconds for *GPT-4*. Despite the minimal increase in LLM inference time – from 2.7 to 2.8 seconds for *GPT-3* and from 19.1 to 21.8 seconds for *GPT-4* – the overall planning time was primarily impacted by the Fast Downward planner.

This demonstrated the planner's effectiveness in handling simple tasks and exposed its limitations with complex and implicit instructions. *GPT-4* demonstrated strong performance, minimizing the need for self-correction, while scene complexity primarily affected planning time. The results suggest that enhancing error messaging and addressing planner inefficiencies could further improve performance.

C.4. Alternative Suggestion

The *Suggest Alternative* tool experiment compares AutoGPT+P's approach (explained in Section 5.2.2) with a naive alternative suggestion method. The naive method relies on the LLM to identify a suitable replacement for a missing object within a scene without incorporating any additional reasoning processes. The performance of both methods was evaluated across 30 predefined scenarios, each characterized by a missing object, a user-specified task, a set of objects present in the scene, and a list of permissible alternative objects. The task is deemed successful if the method identifies one of the allowed alternatives. The scenarios are categorized into three levels of difficulty based on the number of objects in the scene: simple (5 objects), medium (20 objects), and complex (70 objects). This reflects the hypothesis that increased scene complexity, in terms of object quantity, poses a greater challenge for accurately identifying the missing object.

Note: Reprinted from Birr et al. (2024).

	GPT-3		*GPT-4*	
	Naive	AutoGPT+P	Naive	AutoGPT+P
simple	0.73	0.87	**0.90**	**0.90**
medium	0.63	**0.90**	0.70	0.83
complex	0.33	0.80	0.67	**0.80**

Table C.4.: Comparison of the success rate of the *Suggest Alternative* tool with a naive approach. The best values for the success rate are in bold.

The experimental results, shown in Table C.4, reveal a trend where the accuracy of both methods diminishes as the complexity of the scene increases. The naive approach shows a significant decline in accuracy, from 0.73 to 0.33 with *GPT-3* and from 0.9 to 0.67 with *GPT-4*, as the number of objects increases from 5 to 70. In contrast, AutoGPT+P's approach demonstrates greater resilience, with accuracy only slightly decreasing from 0.9 to 0.8 with *GPT-3* and from 0.87 to 0.8 with *GPT-4* under the same conditions. Notably, the difference in performance between *GPT-3* and *GPT-4* is minimal when using the novel method. This consistent accuracy is attributed to AutoGPT+P's utilization of a directed Chain-of-Thought process, which guides the LLM through the reasoning needed for object replacement, thereby reducing the likelihood of incorrect suggestions.

C.5. AutoGPT+P

The evaluation of the overall planning approach and tool selection process of AutoGPT+P was conducted using five distinct scenario sets, each consisting of 30 scenarios designed to rigorously assess the system's performance across various tasks. Four of these sets were focused on evaluating the system's interaction with individual tools, while the fifth set required the integration of all tools to complete more complex tasks. Specifically, the *Plan* subset included randomly selected scenarios from previous evaluations, whereas the *Explore* and *Partial Plan* subsets were constructed with entirely new scenarios. The *Explore* scenarios provided partial hints about the location of objects, such as "Bring me the cucumber from the fridge," to evaluate the system's exploration capabilities. The *Suggest Alternative* and *Combined* sets shared the same scenarios, with the latter set involving initial location exploration to assess tool selection and combination strategies. An additional metric, referred to as *minimal tools*, was incorporated to determine the optimal number of tools required for each scenario, aiming to quantify the efficiency of the tool selection process in achieving successful outcomes.

The experimental results from Table C.5 demonstrate that AutoGPT+P is effective in selecting the appropriate tools for task completion, with performance metrics indicating that the introduction of a prior tool selection process does not diminish its effectiveness in planning-based scenarios. However, including exploration tasks resulted in a slight decrease in success rates, primarily due to premature planning before all relevant locations were fully explored. The *Suggest Alternative* set exhibited a similar reduction in success rate, attributable to invalid alternative suggestions, with an overall success rate 0.07 lower in the *Combined* set, emphasizing

Note: Reprinted from Birr et al. (2024).

Subset	*GPT-3*			*GPT-4*		
	success	minimal	minimal tools	success	minimal	minimal tools
Plan	0.53	0.50	0.30	**0.87**	**0.80**	**0.80**
Partial Plan	0.37	0.20	**0.23**	**0.83**	**0.67**	0.13
Explore	0.10	0.03	0.00	**0.77**	**0.23**	**0.63**
Suggest Alternative	0.13	0.13	0.03	**0.77**	**0.53**	**0.73**
Combined	0.13	0.07	0.10	**0.70**	**0.53**	**0.47**
Average	0.25	0.19	0.13	**0.79**	**0.55**	**0.55**

Table C.5.: Evaluation of AutoGPT+P in the metrics success rate, minimal plan length, and minimal tool usage rate comparing *GPT-3* to *GPT-4*. Best values are written in bold.

the challenge of integrating exploration with tool selection. Additionally, the analysis highlighted a trend towards minimal tool usage when only one tool was required, although the *Partial Plan* set occasionally exhibited inefficient tool selection, often cycling through the *Suggest Alternative* tool. It was observed that when provided with hints, the tool selection process effectively identified the correct location, whereas the absence of clues led to random and sometimes illogical explorations. In comparison, *GPT-3* displayed suboptimal tool selection, particularly in scenarios requiring exploration or alternative suggestions, with a tendency to randomly select tools, resulting in a markedly lower success rate. Overall, while the tool selection process of AutoGPT+P shows promise in addressing tasks with missing objects and partially unexplored environments, challenges remain, particularly in avoiding premature planning and improving the accuracy of alternative suggestions.

D. Tools and Resources

The following tables provide an overview of the tools and resources utilized in the course of this research. Their inclusion serves to ensure transparency regarding the methodologies and technologies employed, aligning with the principles of the *German Research Foundation*'s (DFG) Code of Conduct for Good Research Practice[1].

D.1. Generative Tools and Websites

Tool	**Usage**
ChatGPT	Generation of TikZ figures
	Analysis and summarization of papers
	Structuring
	LaTeX troubleshooting
	Text reformulation
GPT-API	Analysis and summarization of papers
	Structuring of related work
	Reformulation of papers
Grammarly Premium	Grammar analysis and correction
	Text reformulation
DeepL	Translation of text

Website	**Usage**
SciSummary	Analysis of papers
	Summarization of papers
Semantic Scholar	Semantic search for papers
Google Scholar	Search for papers
Thesaurus	Search for synonyms
Overleaf	Writing and compiling LaTeX documents
StackExchange	Help with solving technical problems
Linguee	Translation of words

[1] https://www.dfg.de/en/basics-topics/basics-and-principles-of-funding/good-scientific-practice

D.2. Copyright Notice

In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of Karlsruhe Institute of Technology's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to `http://www.ieee.org/publications_standards/publications/rights/rights_link.html` to learn how to obtain a License from RightsLink. If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.

List of Figures

List of Tables

Glossary

Large Language Model (LLM)

Familiy of neural networks that show great capabilities for generating text

Interpretable Data Format (IDF, Peller-Konrad et al., 2023)

data format of the cognitive memory architecture of ArmarX

Human-Robot Interaction (HRI)

the field of study concerning itself with the design of systems that facilitate communication and collaboration between humans and robots

Virtual Reality (VR)

a computer-generated simulation of a three-dimensional environment that can be interacted with in a seemingly real or physical way using special electronic equipment

Grasp Wrench Space (GWS)

a mathematical representation of all possible forces and torques that can be applied by a robotic gripper to an object

Inverse Kinematics (IK)

a computational method used to determine the joint angles needed to place the end-effector of a robotic arm at a desired position and orientation in space

Behavior Tree (BT)

a hierarchical model used to control the decision-making process of autonomous agents by organizing actions and conditions in a tree structure

Robot Operating System (ROS, Quigley et al., 2009)

a flexible framework for writing robot software that provides tools and libraries for building, simulating, and controlling robotic systems

Finite State Machine (FSM)

a computational model consisting of a finite number of states, transitions between those states, and actions, used to design both computer programs and sequential logic circuits

Learning from Demonstration (LfD)

a technique in robotics where a system learns to perform tasks by observing and mimicking human actions

Random Sample Consensus (RANSAC)

a robust statistical method used to estimate parameters of a mathematical model from a subset of inliers within a dataset containing outliers

Probability Density Function (PDF)

a function that describes the likelihood of a continuous random variable taking on a particular value within a given range

Kullback-Leibler (KL)

see D_{KL}

Dynamic Movement Primitive (DMP)

a framework for encoding and reproducing complex motor behaviors in robots through a combination of learned and adaptable movement patterns

Inertial Measurement Unit (IMU)

a device that measures and reports a body's specific force, angular rate, and sometimes the magnetic field surrounding the body, often used in navigation and motion tracking systems

Markov Decision Process (MDP)

a mathematical framework used for modeling decision-making in situations where outcomes are partly random and partly under the control of a decision-maker

Contributor Role Taxonomy (CRediT, `https://credit.niso.org/`)

A standardized framework used to describe and acknowledge the various roles and contributions of individuals in collaborative projects.

Affordance Template (AT, Hart et al., 2014)

Executive framework for mobile manipulation

Height Accumulated Features (HAF, Fischinger and Vincze, 2012)

visual feature based on a height map of unknown objects

Symmetry Height Accumulated Features (SHAF, Fischinger et al., 2013)

improved version of HAF

instantaneous Task Specification using Constraints (iTaSC, De Schutter et al., 2007)

a method for defining robot tasks in real-time by applying specific constraints to guide the robot's actions and behaviors

Layered Architecture for Autonomous Interactive Robots (LAAIR, Jiang et al., 2018)
a design framework that organizes the control and decision-making processes of autonomous robots into hierarchical layers, each responsible for different levels of abstraction and functionality

Neural Network (NN)
a computational model inspired by the human brain, consisting of interconnected nodes (neurons) that process information in layers to recognize patterns and make decisions

Deep Neural Network (DNN)
a type of artificial NN with multiple layers between the input and output layers, enabling the modeling of complex patterns and representations in data

Convolutional Neural Network (CNN)
a type of DNN designed to process and analyze visual data by using convolutional layers to automatically and adaptively learn spatial hierarchies of features from input images

Variational Auto-Encoder (VAE)
a type of generative model that learns to encode input data into a latent space and then decodes it back to the original data distribution while incorporating probabilistic elements

Conditional Variational Auto-Encoder (CVAE)
a type of neural network that combines VAE with conditional inputs to generate data samples conditioned on specific attributes or labels

Mixture Density Network (MDN)
a neural network model that predicts the parameters of a mixture of probability distributions, allowing for the modeling of complex, multimodal output distributions

Long-Short-Term Memory (LSTM)
a type of recurrent neural network architecture designed to effectively capture and utilize long-range dependencies in sequential data through its unique gating mechanisms

Multi-Layer Perceptron (MLP)
a type of NN composed of multiple layers of nodes, where each layer is fully connected to the next one and used for tasks such as classification and regression

Gaussian Process Latent Variable Model (GP-LVM)

a probabilistic model that uses Gaussian processes to capture the underlying structure of high-dimensional data by mapping it to a lower-dimensional latent space

Pathways Language Model (PaLM, Chowdhery et al., 2023)

a LLM designed to understand and generate human language by leveraging multiple pathways for processing information and improving efficiency and accuracy

Retrieval-Augmented Generation (RAG, Lewis et al., 2020)

a technique that combines information retrieval with text generation in LLMs to produce more accurate and contextually relevant responses by leveraging external knowledge sources

Abstract Affordance Frame (AAF)

abstract coordinate frame, relative to the Darboux frame at any point on an object's surface, that is connected to an affordance in which actions can be defined relative to the pose of the frame; equivalent to the Local Curvature Frame of Pohl and Asfour (2022)

Oriented Bounding Box (OBB)

a bounding parallelepiped whose faces and edges are not parallel to the basis vectors of the frame in which they're defined.

Object-Oriented Bounding Box (OOBB)

a OBB that is oriented according to the object it bounds

Unscented Kalman Filter (UKF, Wan and Van Der Merwe, 2000)

an extension of the Kalman filter that works well for non-linear models

Hidden Markov Model (HMM)

a statistical model used to represent systems with hidden states, where the system transitions between these states with certain probabilities and generates observable outputs based on the current hidden state, see e. g., Rabiner (1990)

Continuous Density Hidden Markov Model (CDHMM)

a type of HMM that uses continuous probability density functions to model the observation probabilities for each state

Point Cloud Library (PCL, Rusu and Cousins, 2011)

an open-source software framework designed for processing and analyzing 3D point cloud data, enabling tasks such as filtering, feature estimation, surface reconstruction, and object recognition

Probabilistic Action Extraction and Fusion (PAEF, Pohl and Asfour, 2022)

probabilistic approach for the extraction of affordance-based manipulation actions

Geometry-based Action Extraction (GAE, Pohl and Asfour, 2022)

surface patch-based action extraction

Generative Pre-trained Transformer (GPT)

type of LLM and a prominent example generative artificial intelligence developed by OpenAI

Multi-feature Implicit Model (MIMO, Cai et al., 2024)

neural network for transferring points across intra-class object models

Neural Descriptor Field (NDF, Simeonov et al., 2022)

a representation technique that encodes spatial information using neural networks to describe the geometry and appearance of 3D objects or scenes as SE(3)-equivariant point and pose descriptors

Relational-Neural Descriptor Field (R-NDF, Simeonov et al., 2023)

a version of NDF that relaxes the limitation of one object being known and fixed by manually selecting keypoints and associated local frames in task-relevant regions

Neural Interaction Field and Template (NIFT, Huang et al., 2023)

a descriptive and robust interaction representation of object manipulations to facilitate imitation learning leveraging the SCF

Space Coverage Feature (SCF, Zhao et al., 2016)

a descriptor of spatial relations between object surfaces that encodes the geometry of the open space around objects

Gaussian Mixture Model (GMM)

a probabilistic model that represents a distribution of data points as a combination of multiple Gaussian distributions, each with its own mean and variance

Extended Space Coverage Feature (ESCF, Cai et al., 2024)

extended version of the SCF using coefficients of spherical harmonics expansion across all orders and degrees

Closest Distance Direction (CDD, Cai et al., 2024)

novel feature improving the direction awareness of the MIMO

Basis Point Set (BPS, Prokudin et al., 2019)

a method for representing 3D shapes by a set of basis points

Visual Imitation Learning (VIL)

a machine learning technique where robots learn to perform tasks by observing and mimicking human actions through visual inputs

Karlsruhe Institute of Technology (KIT)

university of the city of Karlsruhe

Anthropomorphic Multi-Armed Robot (ARMAR, Asfour et al., 1999; Asfour et al., 2017)

humanoid robot family developed at the KIT

Degree of Freedom (DoF)

a parameter that defines the number of independent movements or variables a system or mechanism can have

Tool Center Point (TCP)

the specific point on a robotic tool or end-effector that is used as a reference for positioning and orientation in a robotic system.

Via-point Movement Primitive (VMP, Zhou et al., 2019)

a form of DMP in which the motion can be diverted through via-points

Uncertainty-Aware Sensitivity Optimization (UASO, Baek et al., 2022)

probabilistic approach to improving grasp success rates through sensitivity optimization of uncertainty-affected metrics

Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills (MAkEable, Pohl et al., 2024)

transferable Task Description and Execution Framework

Object Affordance Mapping (OAM)

mapping of affordances to class types of objects

Object Affordance Detection (OAD)

the process of detecting affordances in an image

Natural Language Processing (NLP)

a field of artificial intelligence that focuses on the interaction between computers and humans through natural language.

Planning Domain Definition Language (PDDL)

a formal language used to specify the components and constraints of planning problems in artificial intelligence

Conceptual Equivalence Class (CEC, Varadarajan and Vincze, 2011)

sets of objects that are interchangeable based on their functional affordances, which refer to the potential actions that the objects can support

Disjunctive Normal Form (DNF)

a standardization of a logical formula in Boolean algebra where the formula is expressed as an OR of ANDs, with each AND term consisting of literals

Affordance Network (AfNet, Varadarajan and Vincze, 2013)

affordance-based framework for cognitive object recognition

Affordance Network Ontology for Robotics (AfRob, Varadarajan and Vincze, 2013)

extension of AfNet to domestic robotics

Recognition by Component Affordances (RBCA, Varadarajan, 2011)

extension of the Recognition by Component theory designed for the use with affordances

Object Action Complex (OAC, Krüger et al., 2011)

formalized entities that represent and operationalize the interaction between objects and actions, integrating prediction, execution, and learning to enable cognitive systems to adapt and reason about their environment

tertiary sector

the segment of the economy that provides services rather than goods, including industries such as healthcare and nursing, service and hospitality, and domestic services

AutoGPT+P (Birr et al., 2024)

planning system using a LLM and an affordance-based scene representation to solve planning tasks based on user-specified tasks in natural language

affordance (Gibson, 1966, 1979)

the possibility of an action on an object or environment based on various properties such as shape, weight, stability etc. For example, a chair affords *sitting*, but it does not afford *rolling*.

Local Curvature Frame

See AAF

Darboux frame

a moving orthonormal coordinate system along a curve on a surface, consisting of the tangent vector to the curve, the normal vector to the surface, and the binormal vector orthogonal to both. In the special case of the curve being the principal curve of the surface, it consists of the normal vector and the two Principal Curvature Directions.

Principal Curvature Direction

the direction along a surface at a given point where the curvature is either maximized or minimized, providing critical information about the surface's geometric properties

known object

an object whose properties, dimensions, and characteristics are fully identified and understood within the context of a given robotic system or application

similar object

an object that shares common characteristics or features with another one from a specific class of objects (e. g., cups, bottles, etc.), making them comparable in certain aspects

unknown object

an object about which no specific prior information (like meshes, shapes, weight, features, etc.) is known, sometimes not even that it is an object

statechart

a visual representation of a system's states and the transitions between them, often used to model the behavior of complex systems in software and robotics

Jacobian

a matrix of all first-order partial derivatives of a vector-valued function, used to describe the rate of change of the function with respect to its variables

quadric

a hypersurface (of dimension D) in a $(D+1)$-dimensional space, and defined as the zero set of an irreducible polynomial of degree two in $D+1$ variables

recursive Bayesian estimation

a method for updating the probability estimate for a hypothesis as more evidence or information becomes available, using Bayes' theorem in a recursive manner

L1 distance

a measure of the distance between two points in some space, calculated as the sum of the absolute differences of their coordinates

visual perception (see e.g., Chapter 3, Kragic and Vincze, 2009)

the process by which an entity interprets visual information and stimuli from the environment to form a coherent representation of the surroundings

proprioception

ability of a system to sense its own position, movement, and orientation in space, allowing it to coordinate and control its actions accurately

three-tiered robot architecture (Bonasso, 1991; Firby, 1989)

a hierarchical software design paradigm for robots that separates functionality into three distinct layers: *Behavioral Control*, *Executive*, and *Planning*

Strategy (Gamma et al., 1993)
behavioral software design pattern that lets the program choose different algorithms at runtime without changing the code that uses them by keeping the algorithms separate and interchangeable

closed-world assumption
presumption that anything not explicitly known or stated within the system's knowledge base is considered false or does not exist

object detector
a system, network, or algorithm designed to identify and locate objects within an image or video frame, usually returning the type object and its bounding box

Prolog
a high-level programming language associated with artificial intelligence and computational linguistics, known for its use of logic and rules to solve problems through pattern matching and automated reasoning

representationalist **(Şahin et al., 2007)**
a view on Gibson's theory of affordances

gibsonian **(Chemero, 2003; Chemero and Turvey, 2007)**
a view on Gibson's theory of affordances

Fast Downward (Helmert, 2006)
a planning system that employs heuristic search techniques to solve automated planning problems efficiently

ARMAR-6 (Asfour et al., 2019)
6th iteration of the ARMAR humanoid robot family; intended for the assistance in maintenance tasks

ARMAR-III (Asfour et al., 2006)
3rd iteration of the ARMAR humanoid robot family; developed for real-world applications in domestic environments

ARMAR-DE
newer, updated version of ARMAR-6 with stronger motors and 4-DoFs hands

ArmarX (Vahrenkamp et al., 2015)
robot software framework of the ARMAR humanoid robot family

end-effector
the component of a robotic arm designed to interact with the environment, performing tasks such as gripping, welding, or sensing.

long-term memory

a type of memory system responsible for storing information over extended periods, allowing for the retention and retrieval of knowledge and experiences

working memory

a cognitive system responsible for temporarily holding and processing information necessary for complex tasks such as reasoning, learning, and comprehension

prior knowledge

information or understanding that is already known before encountering new data or experiences, often used to inform decision-making or analysis.

ChatGPT

A variant of the GPT model developed by OpenAI, designed for conversational tasks.

LLM+P (Liu et al., 2023a)

A planning approach that integrates LLMs with classical planners to improve planning capabilities

SayCan (Brohan et al., 2023c)

a task planning system that integrates language models with physical actions to enable robots to understand and execute complex tasks based on natural language instructions

procedural memory (see e.g., Peller-Konrad et al., 2023)

a type of long-term memory responsible for the storage and retrieval of motor skills and actions

Robotic Transformer (Brohan et al., 2023a,b; O'Neill et al., 2024)

type of multi-modal foundational model trained on large datasets for robotic mobile manipulation

Isaac Gym (Makoviychuk et al., 2021)

a high-performance robotics simulation environment developed by NVIDIA, designed to facilitate large-scale training and testing of robotic systems using GPU acceleration

adaptability

the ability of a robotic system to deal with changes in the circumstances or requirements for task execution; one of the core capabilities of this thesis

versatility

the ability of a system or component to handle a wide range of tasks or functions efficiently and effectively and adapt to changes in its environment or conditions; one of the core capabilities of this thesis

reliability

the likelihood that a robotic system will perform its required functions under stated conditions for a specified period of time; one of the core capabilities of this thesis

Symbols

Name	Symbol	Description
Time step	t	a single instance in time, corresponding to e. g., a specific image from a camera
Action observation	$\mathbf{A}$	observation of a candidate for a mobile manipulation action for a single time step connected to all possible affordances at that point
Action hypothesis	$\bar{\mathbf{A}}$	spatio-temporally coherent state of a mobile manipulation action, corresponding to multiple fused action observations
Principal curvatures	$\kappa_{\pm}$	the maximum and minimum values of the curvature as expressed by the eigenvalues of the shape operator at a given point of a surface
Principal directions	$\lambda_{\pm}$	the directions of the minimum and maximum principal curvatures tangential to the surface
Direction of curvature	$\mathbf{k}_{\lambda_{\pm}}$	direction of curvature in Euclidean space
Affordance	a	see affordance
Surface normal	$\mathbf{n}$	normal vector to a surface at a specific point
Covariance	$\boldsymbol{\Sigma}$	a square matrix that summarizes the covariances between multiple variables, with each element indicating the degree to which two variables change together
Position	$\mathbf{t}$	coordinates in $\mathbb{R}^3$ representing the translation from the origin
Position covariance	$\boldsymbol{\Sigma}_{\mathbf{t}}$	covariance of the position
Mean position	$\bar{\mathbf{t}}$	mean value of the position

Name	Symbol	Description
Orientation	$\mathbf{R}$	specific direction or alignment of an object or system in relation to a reference point or coordinate system
Orientation covariance	$\mathbf{\Sigma_R}$	covariance of the orientation
Mean orientation	$\bar{\mathbf{R}}$	mean value of the orientation
Pose	$\mathbf{T}$	combined position and orientation
Pose covariance	$\mathbf{\Sigma_T}$	covariance of the pose
Mean pose	$\bar{\mathbf{T}}$	mean value of the pose
Special orthogonal group	$\mathrm{SO}(3)$	group of all 3x3 orthogonal matrices with determinant 1, representing rotations in three-dimensional space
Special euclidean group	$\mathrm{SE}(3)$	comprises all rigid body transformations in three-dimensional space, combining rotations and translations
Euclidean space	$\mathbb{R}^3$	three-dimensional space of all ordered triples of real numbers, representing points with three coordinates
Lie group	$\mathcal{G}$	a mathematical structure that combines elements of group theory and differential geometry, where the group's elements form a smooth manifold, and the group operations (multiplication and inversion) are smooth functions
Manifold	$\mathcal{M}$	a space that, around every point, looks like a flat, Euclidean space (like a plane or higher-dimensional equivalent) when viewed up close, making it possible to use methods of calculus within these small regions
PDF	p	a function that describes the likelihood of a continuous random variable taking on a particular value within a given range
Correspondence	C	binary random variable representing whether or not two poses correspond to each other

Name	**Symbol**	**Description**
Correspondence likelihood	$p(C\|\mathbf{R}, \mathbf{t})$	likelihood that the observed action $\mathbf{A}$ at position $\mathbf{t}$ and orientation $\mathbf{R}$ corresponds to the hypothesis $\bar{\mathbf{A}}$
Existence certainty	p_E^a	the likelihood that an action observation exists at its associated pose after multiple temporally-distinct observations
Point	$\mathbf{Y}$	point on a Lie group
Mean point	$\bar{\mathbf{Y}}$	average of points on a Lie group
Point covariance	$\mathbf{\Sigma}_{\mathbf{Y}}$	covariance of the mean points on a Lie group
Tangent space	$\mathcal{T}_{\bar{\mathbf{Y}}}\mathcal{M}$	tangent space of a manifold at the mean point $\bar{\mathbf{Y}}$
Tangent	$\boldsymbol{\tau}$	vector in the tangent space of a Lie group $\mathcal{G}$
First fundamental form	I	captures the inner product of tangent vectors, providing information on metric properties like lengths and angles of a parametric surface
Second fundamental form	II	describes the curvature of a parametric surface, describing how it bends by incorporating the derivative of the unit normal vector
Second fundamental form coefficients	L, M, N	coefficients of the second fundamental form
First fundamental form coefficients	E, F, G	coefficients of the first fundamental form
Parametric surface	$\mathbf{r}$	two dimensional parametric surface in $3D$ space
Gaussian curvature	K	a measure of the intrinsic curvature of a surface at a point, calculated as the product of the principal curvatures at that point
Supervoxel	V	a cluster of points with similar properties in a point cloud
Color	$\mathbf{c}$	a vector of the RGB color values of a point
Existing state	S_1	initial state of the HMM, representing that the $\bar{\mathbf{A}}$ exists
Non-exisiting state	S_2	state of the HMM, representing that the $\bar{\mathbf{A}}$ does not exists
HMM state	λ	state of the HMM

Name	Symbol	Description
State transition probabilities	$\mathbf{A}$	state transition probability matrix, representing the probabilities of transitioning from one state to another in the HMM
Observation probabilities	$\mathbf{B}$	observation probability matrix, which defines the probability of observing a particular output given a specific state
Initial state probabilities	π	initial state probability vector, indicating the probabilities of starting in each possible state when the process begins
Identity transition	a_{11}	probability of transitioning from state S_1 to S_1, i. e., staying in the same state
True positive observation	b_{22}	probability of observing the second type of observation given that the system is in S_2
Point cloud	$\mathbf{P}$	perceived point cloud of an object, as seen from a single view
Point cloud of object A	$\mathbf{P}_A$	point cloud of object A
Reconstructed point cloud of object A	$\mathbf{P}_A^r$	reconstructed point cloud of object A
Canonical point cloud	$\mathbf{P}_S^c$	canonical point cloud of source object
Observed point cloud	$\mathbf{P}_S^o$	observed point cloud of source object
Demonstrated point cloud	$\mathbf{P}_S^d$	point cloud from human demonstration of a task in a VIL setup
Point	$\mathbf{x}$	single point in a point cloud $\mathbf{P}$
Point set	$\mathbf{X}$	set of points $\mathbf{x}$ in a point cloud $\mathbf{P}$
Point descriptor	$\mathbf{z}$	point descriptor to measure geometric similarity obtained from the activation layers of the partly-shared decoder for ESCF and CDD
Pose descriptor	$\mathbf{Z}$	concatenation of the point descriptors of a set of points around an object

Name	Symbol	Description
Pose descriptor of object B with respect to object A	${}^{A}\mathbf{Z}_{B}$	pose descriptor encoding the pose $\mathbf{T}$ of $\mathcal{O}_B$ with respect to the point cloud of object A $\mathbf{P}_A$
Reference pose descriptor	${}^{A}\hat{\mathbf{Z}}_{B}$	reference pose descriptor of object B with respect to object A obtained from e. g., VIL
Vector Neurons-PointNet encoder	$\epsilon(\mathbf{P})$	shared encoder of MIMO
Arbitrary pose	A	experimental setting where the objects are initially in an arbitrary pose in the air
Upright pose	U	experimental setting where the objects are initially in an upright pose on a table
MIMO4	*MIMO4*	complete version of MIMO, using all feature branches
MIMO4-	*MIMO4*−	version of MIMO, using all feature branches but no shape completion
MIMO3	*MIMO3*	version of MIMO, with three branches in the decoder to predict occupancy, signed distance, and SCF separately
Demonstrated grasp pose	$\mathbf{T}_g^d$	pose of the grasp demonstrated by a human in a VIL setup
Set of task-agnostic grasp poses	$\{\,\mathbf{T}_g^a\,\}$	set of task-agnostic grasp candidates generated on the canonical point cloud
Set of task-relevant grasp poses	$\{\,\mathbf{T}_g^r\,\}$	set of task-relevant grasp candidates generated by MIMO for the source object
Set of successful task-relevant grasp poses	$\{\,\bar{\mathbf{T}}_g^r\,\}$	set of task-relevant grasp candidates that were executed successfully in simulation
Sampled grasp pose	$\hat{\mathbf{T}}_g$	sampled grasp candidates generated by the GMM trained in simulation for the source object
Optimized grasp pose	$\mathbf{T}_g^*$	grasp pose that has been optimized using the grasp evaluation network
Grasp pose	$\mathbf{T}_g$	end-effector pose of a grasp candidate

Name	Symbol	Description
Transferred grasp pose	$\tilde{\mathbf{T}}_g$	the sampled grasp pose after being transferred to the observed point cloud $\mathbf{P}^o_S$
Success probability	$p_S(\tilde{\mathbf{T}}_g)$	success probability of a transferred grasp pose $\tilde{\mathbf{T}}_g$
Object A	$\mathcal{O}_A$	object space for class A
Object B	$\mathcal{O}_B$	object space for class B
Source object	$\mathcal{O}_S$	source object, i. e., the object being grasped
Target object	$\mathcal{O}_T$	target object, i. e., the object that sets a reference frame for placing the source object $\mathcal{O}_S$
Grasping time step	t_g	time step of the video demonstration where the source object $\mathcal{O}_S$ is grasped
Last time step	t_T	last time step of the video demonstration
Occupancy features	Φ_{occ}	output features of the occupancy feature branch of MIMO
Occupancy features	Φ_{escf}	output features of the ESCF feature branch of MIMO
Occupancy features	Φ_{sdf}	output features of the signed distance feature branch of MIMO
Occupancy features	Φ_{cdd}	output features of the CDD feature branch of MIMO
KL divergence	D_{KL}	a measure of how one probability distribution diverges from a second, expected probability distribution
Grasp score	z	A scalar value used to select the most likely successful grasp
Grasp	g	An attempt by a robotic system to hold or manipulate an object
Grasp metric	m	A specific measure used to evaluate the quality of a grasp
Mean value of a grasp metric	μ	average value of a specific grasp metric
Standard deviation of a grasp metric	σ	measure of the amount of variation or dispersion of a grasp metric
Successful grasp	g^s	A grasp that successfully holds or manipulates an object

Name	Symbol	Description
Failed grasp	g^{f}	A grasp that fails to hold or manipulate an object
Success rate	r^{s}	The ratio of successful grasps to the total number of executed grasps
Total number of executed grasps	g^{tot}	The total number of grasps attempted by the robotic system
Global weighting factor	f^{glob}	global score for a grasp metric obtained from the KL divergence of the successful and failed grasp attempts
Local weighting factor	f^{loc}	local score for a grasp metric representing the likelihood of belonging to the set of successful grasps
Total weighting factor	f^{tot}	the combined score consisting of global weighting factor and local weighting factor for a grasp
Functional model	y	A model used to derive the ranking score for grasp selection
Manipulability	a	grasp metric representing extended manipulability score
Support relation	s	grasp metric representing probabilistic support relations
Distance to center	d	grasp metric representing the distance of a grasp candidate from the center of a point cloud segment
Height	h	grasp metric representing the height of a grasp candidate
Action hypothesis	`ActionHypothesis`	a class that represents a hypothesis of an action
Action type	`ActionType`	a class that represents the type of an action
End-effector trajectory	`EndEffectorTraj`	a class that represents a finger-TCP-trajectory of the end-effector
Affordance	`Affordance`	a class that represents an affordance
Executable action	`ExecutableAction`	a class that represents an action that has all necessary information to be executed
Unimanual action	`Unimanual`	a class that contains all necessary information for a single end-effector

Name	Symbol	Description
Executed action	`ExecutedAction`	a class that contains all information from an execution of an `ExecutableAction`
Date	`DateTime`	a class representing some point in time
Framed pose	`FramedPose`	a pose that is connected to a certain frame or coordinate system
Handedness	`Handedness`	a class that represents the handedness (e. g., left or right) of an end-effector
End-effector state	`EndEffectorState`	a class that represents the state of the end-effector at certain point of execution
Execution pose	`execEES`	a class member that represents the pose of the end-effector at the start of the execution of an `ExecutableAction`
Pre-pose	`preEES`	a class member that represents a safe pose before the `execEES`
Retract pose	`retractEES`	a class member that represents a safe pose after an `ExecutableAction` has been executed
Approach info	`ApproachInfo`	a class representing additional information about the approach direction
Object info	`ObjectInfo`	a class representing additional information about the target object
Planning info	`PlanningInfo`	a class representing additional information that influences planning trajectories (arm or platform)
Robot description	`RobotDescription`	a class representing a description of the robot used
Scene	S	an observation of the environment through visual perception
Set of scenes	$\mathbb{S}$	set of all possibly observable scenes
Object relations	R	an observation of object relations in the current scene
Set of object relations	$\mathbb{R}_S$	set of all possible object relations in the scene
Space of natural language	Λ	space of all possible instructions in natural language
User-specified task	λ	task description in natural language

Name	Symbol	Description
Explorable locations	L	explorable locations in a scene
Set of all explorable locations	$\mathbb{L}$	set of all possible explorable locations in a scene
Location	l	specific location in a scene
Plan	P	sequence of actions needed to fulfill a task
Action	α	executed capability by an agent
Actions	A	specification of the possible operations that can change the state of the world, including the conditions under which these operations can be performed (preconditions) and the effects that result from performing them
Set of capabilities	C_ζ	set of all possible actions that an agent could execute
Capability	c	ability of an agent to perform an action in a scene
Agent	ζ	actor in a plan
Argument	ρ	parameter for a capability of an agent
Image	I	image of a scene at a specific location
Object-affordance pair	p	mapping of an object to its affordances
Object type	o	type of an object
Set of object types	$\mathbb{O}$	set of all possible object types
Object instance number	k	instance number of an object
Natural numbers	$\mathbb{N}_0$	set of natural numbers including 0
Set of affordances	$\mathbb{A}$	set of all possible affordances
Goal state	Ω	goal state of the PDDL problem
Domain	Δ	domain of the PDDL problem
Initial state	Γ	initial state of the PDDL problem
Problem	Ξ	state description of the PDDL problem
Objects	Ψ	set of objects that are in the scene

Name	Symbol	Description
Predicate	Υ	set of predicates of the PDDL problem, i. e., logical statements that describe the properties or relations between objects in the domain
Type	Θ	categories or classifications of objects that can exist in the domain, allowing for the organization and specification of different objects that actions can interact with
GPT-4	$GPT\text{-}4$	4th iteration of the GPT architecture; version `GPT-4-0613` was used for experiments in this thesis
GPT-3	$GPT\text{-}3$	3rd iteration of the GPT architecture; version `GPT-3.5-turbo-0613` was used for experiments in this thesis

Bibliography

Aertbelien, Erwin and Joris De Schutter (2014). “eTaSL/eTC: A Constraint-Based Task Specification Language and Robot Controller Using Expression Graphs”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014). Chicago, IL, USA: IEEE, pp. 1540–1546 (cit. on p. 43).

Ala, Rajeshkanna, Dong Hwan Kim, Sung Yul Shin, Changhwan Kim, and Sung-Kee Park (2015). “A 3D-grasp Synthesis Algorithm to Grasp Unknown Objects Based on Graspable Boundary and Convex Segments”. In: *Information Sciences* 295, pp. 91–106 (cit. on pp. 23, 24).

Almeida, Luís and Plinio Moreno (2021). “Uncertainty and Heuristics for Underactuated Hands: Grasp Pose Selection Based on the Ppotential Grasp Robustness Metric”. In: *SN Applied Sciences* 3.6, p. 681 (cit. on pp. 31, 34).

Arapi, Visar, Yujie Zhang, Giuseppe Averta, Manuel G. Catalano, Daniela Rus, Cosimo Della Santina, and Matteo Bianchi (2020). “To Grasp or Not to Grasp: An End-to-End Deep-Learning Approach for Predicting Grasping Failures in Soft Hands”. In: *2020 3rd IEEE International Conference on Soft Robotics (RoboSoft)*. 2020 3rd IEEE International Conference on Soft Robotics (RoboSoft), pp. 653–660 (cit. on pp. 30, 31).

Ardón, Paola, Èric Pairet, Katrin S. Lohan, Subramanian Ramamoorthy, and Ronald P. A. Petrick (2020). *Affordances in Robotic Tasks – A Survey*. URL: `http://arxiv.org/abs/2004.07400`. Pre-published (cit. on p. 134).

Ardón, Paola, Èric Pairet, Katrin S. Lohan, Subramanian Ramamoorthy, and Ronald P. A. Petrick (2021). “Building Affordance Relations for Robotic Agents - A Review” (cit. on p. 134).

Asfour, T., K. Berns, and R. Dillmann (1999). “The Humanoid Robot ARMAR”. In: *The Second International Symposium in HUmanoid RObots (HURO’99)*. Tokyo, Japan, pp. 174–180 (cit. on pp. 39, 162).

Asfour, T., K. Regenstein, P. Azad, J. Schröder, N. Vahrenkamp, and R. Dillmann (2006). “ARMAR-III: An Integrated Humanoid Platform for Sensory-Motor Control”. In: *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*. Most Influential Paper Award Finalist, IEEE-RAS International

Conference on Humanoid Robots 20th anniversary. Genova, Italy, pp. 169–175 (cit. on pp. 39, 165).

Asfour, Tamim, Rüdiger Dillmann, Nikolaus Vahrenkamp, Martin Do, Mirko Wächter, Christian Mandery, Peter Kaiser, Manfred Kröhnert, and Markus Grotz (2017). "The Karlsruhe ARMAR Humanoid Robot Family". In: *Humanoid Robotics: A Reference*. Ed. by Ambarish Goswami and Prahlad Vadakkepat. Dordrecht: Springer Netherlands, pp. 1–32 (cit. on pp. 39, 162).

Asfour, Tamim, Mirko Wächter, Lukas Kaul, Samuel Rader, Pascal Weiner, Simon Ottenhaus, Raphael Grimm, You Zhou, Markus Grotz, and Fabian Paus (2019). "ARMAR-6: A High-Performance Humanoid for Human-Robot Collaboration in Real World Scenarios". In: *IEEE Robotics & Automation Magazine* 26.4, pp. 108–121 (cit. on pp. 12, 39, 165).

Asif, Umar, Mohammed Bennamoun, and Ferdous Sohel (2014). "Model-Free Segmentation and Grasp Selection of Unknown Stacked Objects". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, pp. 659–674 (cit. on pp. 31, 34).

Baek, Woo Jeong, Christoph Pohl, Philipp Pelcz, Torsten Kroger, and Tamim Asfour (2022). "Improving Humanoid Grasp Success Rate Based on Uncertainty-Aware Metrics and Sensitivity Optimization". In: *IEEE-RAS International Conference on Humanoid Robots*. Vol. 2022-Novem, pp. 786–793 (cit. on pp. 10, 13, 31, 38, 90–92, 96–98, 126, 162).

Bagnell, J. Andrew, Felipe Cavalcanti, Lei Cui, Thomas Galluzzo, Martial Hebert, Moslem Kazemi, Matthew Klingensmith, Jacqueline Libby, Tian Yu Liu, Nancy Pollard, Mihail Pivtoraiko, Jean-Sebastien Valois, and Ranqi Zhu (2012). "An Integrated System for Autonomous Robotics Manipulation". In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012). Vilamoura-Algarve, Portugal: IEEE, pp. 2955–2962 (cit. on p. 43).

Barad, Kuldeep R., Andrej Orsula, Antoine Richard, Jan Dentler, Miguel Olivares-Mendez, and Carol Martinez (2023). *GraspLDM: Generative 6-DoF Grasp Synthesis Using Latent Diffusion Models*. URL: http://arxiv.org/abs/2312.11243 (visited on 02/07/2024). Pre-published (cit. on pp. 23, 28).

Baressi Šegota, Sandi, Nikola Andelic, Zlatan Car, and Mario Šercer (2022). "Prediction of Robot Grasp Robustness Using Artificial Intelligence Algorithms". In: *Tehnicki vjesnik - Technical Gazette* 29.1 (cit. on p. 31).

Bärmann, Leonard, Rainer Kartmann, Fabian Peller-Konrad, Jan Niehues, Alex Waibel, and Tamim Asfour (2024). "Incremental Learning of Humanoid Robot

Behavior from Natural Interaction and Large Language Models" (cit. on pp. 53, 55).

Beer, Jenay M, Arthur D Fisk, and Wendy A Rogers (2014). "Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction". In: *Journal of Human-Robot Interaction* 3.2, p. 74 (cit. on p. 3).

Beetz, Michael, Raja Chatila, Joachim Hertzberg, and Federico Pecora (2016). "AI Reasoning Methods for Robotics". In: *Springer Handbook of Robotics*. Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 329–356 (cit. on pp. 3–5).

Billard, Aude and Danica Kragic (2019). "Trends and Challenges in Robot Manipulation". In: *Science* 364.6446, eaat8414 (cit. on pp. 2, 5).

Billard, Aude G., Sylvain Calinon, and Rüdiger Dillmann (2016). "Learning from Humans". In: *Springer Handbook of Robotics*. Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 1995–2014 (cit. on p. 5).

Bilyea, A., N. Seth, S. Nesathurai, and H.A. Abdullah (2017). "Robotic Assistants in Personal Care: A Scoping Review". In: *Medical Engineering & Physics* 49, pp. 1–6 (cit. on pp. 1, 2, 4).

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP, and OIML (2008). *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement*. Joint Committee for Guides in Metrology, JCGM 100:2008 (cit. on p. 93).

Birr, Timo, Christoph Pohl, and Tamim Asfour (2022). "Oriented Surface Reachability Maps for Robot Placement". In: *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 3357–3363 (cit. on p. 10).

Birr, Timo, Christoph Pohl, Abdelrahman Younes, and Tamim Asfour (2024). "AutoGPT+P: Affordance-based Task Planning with Large Language Models". In: *Proceedings of Robotics: Science and Systems*. Robotics: Science and Systems. Vol. 20. Delft, Netherlands (cit. on pp. 10, 14, 51–53, 113, 117, 119, 120, 127, 143–146, 148, 163).

Biza, Ondrej, Skye Thompson, Kishore Reddy Pagidi, Abhinav Kumar, Elise van der Pol, Robin Walters, Thomas Kipf, Jan-Willem van de Meent, Lawson LS Wong, and Robert Platt (2023). "One-Shot Imitation Learning via Interaction Warping" (cit. on p. 59).

Bohg, Jeannette, Matthew Johnson-Roberson, Beatriz Leon, Javier Felip, Xavi Gratal, Niklas Bergstrom, Danica Kragic, and Antonio Morales (2011). "Mind the Gap - Robotic Grasping under Incomplete Observation". In: *2011 IEEE International Conference on Robotics and Automation*. 2011 IEEE International

Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, pp. 686–693 (cit. on pp. 22, 23).

Bohg, Jeannette, Antonio Morales, Tamim Asfour, and Danica Kragic (2014). "Data-Driven Grasp Synthesis - A Survey". In: *IEEE Transactions on Robotics* 30.2, pp. 289–309 (cit. on pp. 17, 29).

Bohg, Jeannette, Kai Welke, Beatriz León, Martin Do, Dan Song, Walter Wohlkinger, Marianna Madry, Aitor Aldóma, Markus Przybylski, Tamim Asfour, Higinio Martí, Danica Kragic, Antonio Morales, and Markus Vincze (2012). "Task-Based Grasp Adaptation on a Humanoid Robot". In: *IFAC Proceedings Volumes* 45.22, pp. 779–786 (cit. on pp. 19, 20).

Bohren, Jonathan and Steve Cousins (2010). "The SMACH High-Level Executive [ROS News]". In: *IEEE Robotics & Automation Magazine* 17.4, pp. 18–20 (cit. on p. 41).

Bonasso, R. Peter (1991). "Integrating Reaction Plans and Layered Competences through Synchronous Control". In: *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 2*. IJCAI'91. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 1225–1231 (cit. on pp. 38, 164).

Borghesan, Gianni, Erwin Aertbelien, and Joris De Schutter (2014). "Constraint-and Synergy-Based Specification of Manipulation Tasks". In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 2014 IEEE International Conference on Robotics and Automation (ICRA). Hong Kong, China: IEEE, pp. 397–402 (cit. on pp. 44, 46, 54, 104).

Brock, Oliver, Jaeheung Park, and Marc Toussaint (2016). "Mobility and Manipulation". In: *Springer Handbook of Robotics*. Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 1007–1036 (cit. on pp. 2, 3, 6, 17, 90, 103).

Brohan, Anthony, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, Tianli Ding, Danny Driess, Avinava Dubey, Chelsea Finn, Pete Florence, Chuyuan Fu, Montse Gonzalez Arenas, Keerthana Gopalakrishnan, Kehang Han, Karol Hausman, Alexander Herzog, Jasmine Hsu, Brian Ichter, Alex Irpan, Nikhil Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Lisa Lee, Tsang-Wei Edward Lee, Sergey Levine, Yao Lu, Henryk Michalewski, Igor Mordatch, Karl Pertsch, Kanishka Rao, Krista Reymann, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Pierre Sermanet, Jaspiar Singh, Anikait Singh, Radu Soricut, Huong Tran, Vincent Vanhoucke, Quan Vuong, Ayzaan Wahid, Stefan Welker, Paul Wohlhart, Jialin Wu, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich (2023a). *RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic*

Control. URL: `http://arxiv.org/abs/2307.15818` (visited on 08/30/2024). Pre-published (cit. on pp. 39, 166).

Brohan, Anthony, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J. Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich (2023b). *RT-1: Robotics Transformer for Real-World Control at Scale.* URL: `http://arxiv.org/abs/2212.06817` (visited on 08/30/2024). Pre-published (cit. on pp. 39, 166).

Brohan, Anthony, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu (2023c). “Do As I Can, Not As I Say: Grounding Language in Robotic Affordances”. In: *Proceedings of the 6th Conference on Robot Learning.* Conference on Robot Learning. Ed. by Karen Liu, Dana Kulic, and Jeff Ichnowski. Vol. 205. Proceedings of Machine Learning Research. PMLR, pp. 287–318 (cit. on pp. 41, 52, 53, 55, 113, 114, 166).

Brossard, Martin, Axel Barrau, and Silvere Bonnabel (2020). “A Code for Unscented Kalman Filtering on Manifolds (UKF-M)”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA).* IEEE, pp. 5701–5708 (cit. on p. 86).

Brossard, Martin, Silvere Bonnabel, and Jean-Philippe Condomines (2017). “Unscented Kalman Filtering on Lie Groups”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* Vol. 2017-Septe. IEEE, pp. 2485–2491 (cit. on pp. 81, 86).

Burgess-Limerick, Ben, Chris Lehnert, Jurgen Leitner, and Peter Corke (2022). *An Architecture for Reactive Mobile Manipulation On-The-Move.* URL: `http://arxiv.org/abs/2212.06991` (visited on 02/06/2024). Pre-published (cit. on pp. 44, 46).

Cai, Yichen, Jianfeng Gao, Christoph Pohl, and Tamim Asfour (2024). "Visual Imitation Learning of Task-Oriented Object Grasping and Rearrangement". In: *Proc. of the 2024 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems.* International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, UAE: IEEE/RSJ, accepted for publication (cit. on pp. 10, 11, 19, 29, 58, 61, 63–65, 126, 137–140, 161).

Cavalli, Luca, Gian Pietro, and M. Matteucci (2019). "Towards Affordance Prediction with Vision via Task Oriented Grasp Quality Metrics". In: *ArXiv* (cit. on pp. 31, 33).

Chang, Angel X, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. (2015). "ShapeNet: An Information-rich 3D Model Repository" (cit. on p. 136).

Chemero, Anthony (2003). "An Outline of a Theory of Affordances". In: *Ecological Psychology* 15.2, pp. 181–195 (cit. on pp. 133, 165).

Chemero, Anthony and Michael T. Turvey (2007). "Gibsonian Affordances for Roboticists". In: *Adaptive Behavior* 15.4, pp. 473–480 (cit. on pp. 133, 165).

Chen, Dong, Vincent Dietrich, Ziyuan Liu, and Georg von Wichert (2018). "A Probabilistic Framework for Uncertainty-Aware High-Accuracy Precision Grasping of Unknown Objects". In: *Journal of Intelligent & Robotic Systems* 90.1-2, pp. 19–43 (cit. on pp. 31, 34).

Chen, Dong, Vincent Dietrich, and Georg Von Wichert (2016). "Precision Grasping Based on Probabilistic Models of Unknown Objects". In: *2016 IEEE International Conference on Robotics and Automation (ICRA).* 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm, Sweden: IEEE, pp. 2044–2051 (cit. on pp. 23, 24).

Chen, Junting, Yao Mu, Qiaojun Yu, Tianming Wei, Silang Wu, Zhecheng Yuan, Zhixuan Liang, Chao Yang, Kaipeng Zhang, Wenqi Shao, Yu Qiao, Huazhe Xu, Mingyu Ding, and Ping Luo (2024). *RoboScript: Code Generation for Free-Form Manipulation Tasks across Real and Simulation.* URL: `http://arxiv.org/abs/2402.14623` (visited on 02/27/2024). Pre-published (cit. on pp. 44, 50).

Chen, Tiffany L., Matei Ciocarlie, Steve Cousins, Phillip M. Grice, Kelsey Hawkins, Kaijen Hsiao, Charles C. Kemp, Chih-Hung King, Daniel A. Lazewatsky, Adam E. Leeper, Hai Nguyen, Andreas Paepcke, Caroline Pantofaru, William

D. Smart, and Leila Takayama (2013). “Robots for Humanity: Using Assistive Robotics to Empower People with Disabilities”. In: *IEEE Robotics & Automation Magazine* 20.1, pp. 30–39 (cit. on p. 4).

Chen, Wenkai, Hongzhuo Liang, Zhaopeng Chen, Fuchun Sun, and Jianwei Zhang (2022). “Improving Object Grasp Performance via Transformer-Based Sparse Shape Completion”. In: *Journal of Intelligent & Robotic Systems* 104.3, p. 45 (cit. on pp. 19, 20).

Chen, Yiye, Yunzhi Lin, Ruinian Xu, and Patricio A. Vela (2023a). “Keypoint-GraspNet: Keypoint-based 6-DoF Grasp Generation from the Monocular RGB-D Input”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA).* 2023 IEEE International Conference on Robotics and Automation (ICRA), pp. 7988–7995 (cit. on pp. 19, 21).

Chen, Yongchao, Jacob Arkin, Yang Zhang, Nicholas Roy, and Chuchu Fan (2023b). “AutoTAMP: Autoregressive Task and Motion Planning with LLMs as Translators and Checkers” (cit. on pp. 53, 54, 120).

Chen, Zibo, Zhixuan Liu, Shangjin Xie, and Wei-Shi Zheng (2023c). “Grasp Region Exploration for 7-DoF Robotic Grasping in Cluttered Scenes”. In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, MI, USA: IEEE, pp. 3169–3175 (cit. on pp. 23, 25).

Cheng, Hu, Danny Ho, and Max Q.-H. Meng (2020). “High Accuracy and Efficiency Grasp Pose Detection Scheme with Dense Predictions”. In: *2020 IEEE International Conference on Robotics and Automation (ICRA).* 2020 IEEE International Conference on Robotics and Automation (ICRA). Paris, France: IEEE, pp. 3604–3610 (cit. on pp. 23, 26).

Cheng, Hu, Yingying Wang, and Max Q.-H. Meng (2022). “A Robot Grasping System With Single-Stage Anchor-Free Deep Grasp Detector”. In: *IEEE Transactions on Instrumentation and Measurement* 71, pp. 1–12 (cit. on pp. 23, 26).

Chiang, Ai-Hsuan and Silvana Trimi (2020). “Impacts of Service Robots on Service Quality”. In: *Service Business* 14.3, pp. 439–459 (cit. on pp. 1, 4).

Chinchor, Nancy (1992). “MUC-4 Evaluation Metrics”. In: *Proceedings of the 4th Conference on Message Understanding - MUC4 ’92.* The 4th Conference. McLean, Virginia: Association for Computational Linguistics, p. 22 (cit. on p. 141).

Chowdhery, Aakanksha, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua

Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, B. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, M. Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, S. Ghemawat, Sunipa Dev, H. Michalewski, Xavier García, Vedant Misra, Kevin Robinson, L. Fedus, Denny Zhou, Daphne Ippolito, D. Luan, Hyeontaek Lim, Barret Zoph, A. Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, T. S. Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, K. Meier-Hellstern, D. Eck, J. Dean, Slav Petrov, and Noah Fiedel (2023). “PaLM: Scaling Language Modeling with Pathways”. In: *Journal of Machine Learning Research* 24.240, pp. 1–113 (cit. on pp. 54, 160).

Ciocarlie, Matei, Kaijen Hsiao, Edward Gil Jones, Sachin Chitta, Radu Bogdan Rusu, and Ioan A. Şucan (2014). “Towards Reliable Grasping and Manipulation in Household Environments”. In: *Experimental Robotics: The 12th International Symposium on Experimental Robotics*. Ed. by Oussama Khatib, Vijay Kumar, and Gaurav Sukhatme. Berlin, Heidelberg: Springer, pp. 241–252 (cit. on p. 4).

Danielczuk, Michael, Ashwin Balakrishna, Daniel S. Brown, Shivin Devgon, and Ken Goldberg (2020). *Exploratory Grasping: Asymptotically Optimal Algorithms for Grasping Challenging Polyhedral Objects*. URL: `http://arxiv.org/abs/2011.05632` (visited on 02/22/2024). Pre-published (cit. on pp. 23, 27).

De Graaf, Maartje M. A., Somaya Ben Allouch, and Jan A. G. M. Van Dijk (2019). “Why Would I Use This in My Home? A Model of Domestic Social Robot Acceptance”. In: *Human–Computer Interaction* 34.2, pp. 115–173 (cit. on pp. 2, 5).

De Schutter, Joris, Tinne De Laet, Johan Rutgeerts, Wilm Decré, Ruben Smits, Erwin Aertbeliën, Kasper Claes, and Herman Bruyninckx (2007). “Constraint-Based Task Specification and Estimation for Sensor-Based Robot Systems in the Presence of Geometric Uncertainty”. In: *The International Journal of Robotics Research* 26.5, pp. 433–455 (cit. on pp. 42, 158).

De Farias, Cristiana, Brahim Tamadazte, Rustam Stolkin, and Naresh Marturi (2022). “Grasp Transfer for Deformable Objects by Functional Map Correspondence”. Version 1. In: (cit. on p. 18).

DeGol, Joseph, Aadeel Akhtar, Bhargava Manja, and Timothy Bretl (2016). “Automatic Grasp Selection Using a Camera in a Hand Prosthesis”. In: *Annual International Conference of the IEEE Engineering in Medicine and Biology So-*

ciety. IEEE Engineering in Medicine and Biology Society. Annual International Conference 2016, pp. 431–434 (cit. on pp. 31, 36).

De Jong, Michiel, Kevin Zhang, Aaron M Roth, Travers Rhodes, Robin Schmucker, Chenghui Zhou, and Sofia Ferreira (2018). “Towards a Robust Interactive and Learning Social Robot”. In: (cit. on p. 4).

Deng, Congyue, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas (2021). “Vector Neurons: A General Framework for SO(3)-Equivariant Networks”. In: *Intl. Conf. on Computer Vision (ICCV)*. Intl. Conf. on Computer Vision (ICCV), pp. 12200–12209 (cit. on p. 60).

Deng, Zhen, Ge Gao, Simone Frintrop, Fuchun Sun, Changshui Zhang, and Jianwei Zhang (2019). “Attention Based Visual Analysis for Fast Grasp Planning With a Multi-Fingered Robotic Hand”. In: *Frontiers in Neurorobotics* 13, p. 60 (cit. on pp. 23, 26).

Detry, Renaud, Jeremie Papon, and Larry Matthies (2017). “Task-Oriented Grasping with Semantic and Geometric Scene Understanding”. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3266–3273 (cit. on pp. 19, 21).

Dice, Lee R. (1945). “Measures of the Amount of Ecologic Association between Species”. In: *Ecology* 26.3, pp. 297–302 (cit. on p. 141).

Ding, Yan, Xiaohan Zhang, Saeid Amiri, Nieqing Cao, Hao Yang, Andy Kaminski, Chad Esselink, and Shiqi Zhang (2023). “Integrating Action Knowledge and LLMs for Task Planning and Situation Handling in Open Worlds” (cit. on pp. 53, 54).

Dömel, Andreas, Simon Kriegel, Michael Kaßecker, Manuel Brucker, Tim Bodenmüller, and Michael Suppa (2017). “Toward Fully Autonomous Mobile Manipulation for Industrial Environments”. In: *International Journal of Advanced Robotic Systems* 14.4, p. 172988141771858 (cit. on pp. 44, 49, 54, 104).

Driess, Danny, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence (2023). “PaLM-E: An Embodied Multimodal Language Model” (cit. on pp. 53–55).

Dune, C., E. Marchand, C. Collowet, and C. Leroux (2008). “Active Rough Shape Estimation of Unknown Objects”. In: *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems. Nice: IEEE, pp. 3622–3627 (cit. on pp. 22, 23).

Eppner, Clemens and Oliver Brock (2013). “Grasping Unknown Objects by Exploiting Shape Adaptability and Environmental Constraints”. In: *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2013). Tokyo: IEEE, pp. 4000–4006 (cit. on pp. 23, 24).

Erkan, Ayşe Naz, Oliver Kroemer, Renaud Detry, Yasemin Altun, Justus Piater, and Jan Peters (2010). “Learning Probabilistic Discriminative Models of Grasp Affordances under Limited Supervision”. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, pp. 1586–1591 (cit. on pp. 31, 35, 38).

Feil-Seifer, David and Maja J. Matarić (2009). “Toward Socially Assistive Robotics for Augmenting Interventions for Children with Autism Spectrum Disorders”. In: *Experimental Robotics*. Ed. by Oussama Khatib, Vijay Kumar, and George J. Pappas. Red. by Bruno Siciliano, Oussama Khatib, and Frans Groen. Vol. 54. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 201–210 (cit. on p. 2).

Ficuciello, F., A. Migliozzi, G. Laudante, P. Falco, and B. Siciliano (2019). “Vision-Based Grasp Learning of an Anthropomorphic Hand-Arm System in a Synergy-Based Control Framework”. In: *Science Robotics* 4.26, eaao4900 (cit. on pp. 19, 21).

Firby, Robert James (1989). “Adaptive Execution in Complex Dynamic Worlds”. PhD thesis. USA: Yale University (cit. on pp. 38, 164).

Fischinger, David, Peter Einramhof, Konstantinos Papoutsakis, Walter Wohlkinger, Peter Mayer, Paul Panek, Stefan Hofmann, Tobias Koertner, Astrid Weiss, Antonis Argyros, and Markus Vincze (2016). “Hobbit, a Care Robot Supporting Independent Living at Home: First Prototype and Lessons Learned”. In: *Robotics and Autonomous Systems* 75, pp. 60–78 (cit. on pp. 1, 2, 4).

Fischinger, David and Markus Vincze (2012). “Empty the Basket - a Shape Based Learning Approach for Grasping Piles of Unknown Objects”. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012). Vilamoura-Algarve, Portugal: IEEE, pp. 2051–2057 (cit. on pp. 23, 24, 158).

Fischinger, David, Markus Vincze, and Yun Jiang (2013). “Learning Grasps for Unknown Objects in Cluttered Scenes”. In: *2013 IEEE International Conference on Robotics and Automation*. 2013 IEEE International Conference on Robotics and Automation (ICRA). Karlsruhe, Germany: IEEE, pp. 609–616 (cit. on pp. 23, 24, 158).

Fischinger, David, Astrid Weiss, and Markus Vincze (2015). “Learning Grasps with Topographic Features”. In: *The International Journal of Robotics Research* 34.9, pp. 1167–1194 (cit. on pp. 23, 24).

Gabellieri, Chiara, Franco Angelini, Visar Arapi, Alessandro Palleschi, Manuel G. Catalano, Giorgio Grioli, Lucia Pallottino, Antonio Bicchi, Matteo Bianchi, and Manolo Garabini (2020). “Grasp It Like a Pro: Grasp of Unknown Objects With Robotic Hands Based on Skilled Human Expertise”. In: *IEEE Robotics and Automation Letters* 5.2, pp. 2808–2815 (cit. on pp. 23, 25).

Gamma, Erich, Richard Helm, Ralph Johnson, and John Vlissides (1993). “Design Patterns: Abstraction and Reuse of Object-Oriented Design”. In: *ECOOP’93—Object-Oriented Programming: 7th European Conference Kaiserslautern, Germany, July 26–30, 1993 Proceedings 7*. Springer, pp. 406–431 (cit. on pp. 107, 165).

Gao, Jianfeng, Xiaoshu Jin, Franziska Krebs, Noémie Jaquier, and Tamim Asfour (2024). *Bi-KVIL: Keypoints-based Visual Imitation Learning of Bimanual Manipulation Tasks.* URL: http://arxiv.org/abs/2403.03270 (visited on 06/14/2024). Pre-published (cit. on p. 63).

Gao, Jianfeng, Zhi Tao, Noémie Jaquier, and Tamim Asfour (2023). “K-VIL: Keypoints-based Visual Imitation Learning”. In: *IEEE Transactions on Robotics* 39.5, pp. 3888–3908 (cit. on p. 121).

Garcia, Sergio, Claudio Menghi, Patrizio Pelliccione, Thorsten Berger, and Rebekka Wohlrab (2018). “An Architecture for Decentralized, Collaborative, and Autonomous Robots”. In: *2018 IEEE International Conference on Software Architecture (ICSA)*. 2018 IEEE International Conference on Software Architecture (ICSA). Seattle, WA: IEEE, pp. 75–7509 (cit. on pp. 44, 48).

Ge, Shihao, Beiping Hou, Wen Zhu, Yuzhen Zhu, Senjian Lu, and Yangbin Zheng (2023). “Pixel-Level Collision-Free Grasp Prediction Network for Medical Test Tube Sorting on Cluttered Trays”. In: *IEEE Robotics and Automation Letters* 8.12, pp. 7897–7904 (cit. on p. 18).

Ghalamzan E., Amir M., Nikos Mavrakis, Marek Kopicki, Rustam Stolkin, and Ales Leonardis (2016). “Task-Relevant Grasp Selection: A Joint Solution to Planning Grasps and Manipulative Motion Trajectories”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, South Korea: IEEE, pp. 907–914 (cit. on pp. 31, 37).

Gibson, James J. (1966). *The Senses Considered as Perceptual Systems.* The Senses Considered as Perceptual Systems. Oxford, England: Houghton Mifflin (cit. on pp. 6, 133, 163).

Gibson, James J. (1979). “The Theory of Affordances”. In: *The Ecological Approach to Visual Perception.* Houghton Mifflin, pp. 119–137 (cit. on pp. 6, 105, 133, 163, 165).

Goins, Alex K., Ryan Carpenter, Weng-Keen Wong, and Ravi Balasubramanian (2014). “Evaluating the Efficacy of Grasp Metrics for Utilization in a Gaussian Process-based Grasp Predictor”. In: *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2014). Chicago, IL, USA: IEEE, pp. 3353–3360 (cit. on pp. 31, 37).

Gori, Ilaria, Ugo Pattacini, Vadim Tikhanoff, and Giorgio Metta (2013). “Ranking the Good Points: A Comprehensive Method for Humanoid Robots to Grasp Unknown Objects”. In: *2013 16th International Conference on Advanced Robotics (ICAR).* 2013 16th International Conference on Advanced Robotics (ICAR 2013). Montevideo, Uruguay: IEEE, pp. 1–7 (cit. on pp. 31, 34).

Gou, Zhibin, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen (2023). “CRITIC: Large Language Models Can Self-Correct with Tool-Interactive Critiquing” (cit. on p. 119).

Gravdahl, Irja, Katrine Seel, and Esten Ingar Grotli (2019). “Robotic Bin-Picking under Geometric End-Effector Constraints: Bin Placement and Grasp Selection”. In: *2019 7th International Conference on Control, Mechatronics and Automation (ICCMA).* 2019 7th International Conference on Control, Mechatronics and Automation (ICCMA). Delft, Netherlands: IEEE, pp. 197–203 (cit. on pp. 31, 34, 37).

Grimm, Raphael, Markus Grotz, Simon Ottenhaus, and Tamim Asfour (2021). “Vision-Based Robotic Pushing and Grasping for Stone Sample Collection under Computing Resource Constraints”. In: *2021 IEEE International Conference on Robotics and Automation (ICRA).* 2021 IEEE International Conference on Robotics and Automation (ICRA), pp. 6498–6504 (cit. on pp. 23, 24, 74, 87, 109).

Gualtieri, Marcus, Andreas ten Pas, Kate Saenko, and Robert Platt (2016). “High Precision Grasp Pose Detection in Dense Clutter”. In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* Vol. 2016-Novem. IEEE, pp. 598–605 (cit. on pp. 31, 36).

Guan, Lin, Karthik Valmeekam, Sarath Sreedharan, and Subbarao Kambhampati (2023). "Leveraging Pre-trained Large Language Models to Construct and Utilize World Models for Model-based Task Planning" (cit. on pp. 53, 54).

Guo, Dingkun, Yuqi Xiang, Shuqi Zhao, Xinghao Zhu, Masayoshi Tomizuka, Mingyu Ding, and Wei Zhan (2024). *PhyGrasp: Generalizing Robotic Grasping with Physics-informed Large Multimodal Models*. URL: `http://arxiv.org/abs/2402.16836` (visited on 02/29/2024). Pre-published (cit. on pp. 23, 25).

Hart, Stephen, Paul Dinh, and Kimberly A. Hambuchen (2014). "Affordance Templates for Shared Robot Control". In: *AAAI Fall Symposium - Technical Report* FS-14-01.c, pp. 81–82 (cit. on pp. 44, 50, 104, 158).

Hart, Stephen, Paul Dinh, and Kimberly A. Hambuchen (2015). "The Affordance Template ROS Package for Robot Task Programming". In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*. 2015 IEEE International Conference on Robotics and Automation (ICRA). Seattle, WA, USA: IEEE, pp. 6227–6234 (cit. on pp. 44, 50).

Hart, Stephen, Ana Huamán Quispe, Michael W. Lanighan, and Seth Gee (2022). "Generalized Affordance Templates for Mobile Manipulation". In: *2022 International Conference on Robotics and Automation (ICRA)*. 2022 IEEE International Conference on Robotics and Automation (ICRA). Philadelphia, PA, USA: IEEE, pp. 6240–6246 (cit. on pp. 44, 50, 54).

Hartshorne, Robin (2013). *Algebraic Geometry*. Vol. 52. Springer Science & Business Media (cit. on p. 69).

Haviland, Jesse, Niko Sünderhauf, and Peter Corke (2022). "A Holistic Approach to Reactive Mobile Manipulation". In: *IEEE Robotics and Automation Letters* 7.2, pp. 3122–3129 (cit. on p. 111).

Hawes, Nick, Christopher Burbridge, Ferdian Jovan, Lars Kunze, Bruno Lacerda, Lenka Mudrova, Jay Young, Jeremy Wyatt, Denise Hebesberger, Tobias Kortner, Rares Ambrus, Nils Bore, John Folkesson, Patric Jensfelt, Lucas Beyer, Alexander Hermans, Bastian Leibe, Aitor Aldoma, Thomas Faulhammer, Michael Zillich, Markus Vincze, Eris Chinellato, Muhannad Al-Omari, Paul Duckworth, Yiannis Gatsoulis, David C. Hogg, Anthony G. Cohn, Christian Dondrup, Jaime Pulido Fentanes, Tomas Krajnik, Joao M. Santos, Tom Duckett, and Marc Hanheide (2017). "The STRANDS Project: Long-Term Autonomy in Everyday Environments". In: *IEEE Robotics & Automation Magazine* 24.3, pp. 146–156 (cit. on p. 5).

Helmert, M. (2006). "The Fast Downward Planning System". In: *Journal of Artificial Intelligence Research* 26, pp. 191–246 (cit. on pp. 120, 165).

Hermann, Andreas, Zhixing Xue, Steffen W. Ruhl, and R. Dillmann (2011). “Hardware and Software Architecture of a Bimanual Mobile Manipulator for Industrial Application”. In: *2011 IEEE International Conference on Robotics and Biomimetics.* 2011 IEEE International Conference on Robotics and Biomimetics (ROBIO). Karon Beach, Thailand: IEEE, pp. 2282–2288 (cit. on pp. 44, 46, 104).

Herzog, Alexander, Peter Pastor, Mrinal Kalakrishnan, Ludovic Righetti, Tamim Asfour, and Stefan Schaal (2012). “Template-Based Learning of Grasp Selection”. In: *2012 IEEE International Conference on Robotics and Automation.* 2012 IEEE International Conference on Robotics and Automation (ICRA). St Paul, MN, USA: IEEE, pp. 2379–2384 (cit. on pp. 31, 35).

Herzog, Alexander, Peter Pastor, Mrinal Kalakrishnan, Ludovic Righetti, Jeannette Bohg, Tamim Asfour, and Stefan Schaal (2014). “Learning of Grasp Selection Based on Shape-Templates”. In: *Autonomous Robots* 36.1-2, pp. 51–65 (cit. on pp. 31, 35).

Hidalgo-Carvajal, Diego, Hanzhi Chen, Gemma C. Bettelani, Jaesug Jung, Melissa Zavaglia, Laura Busse, Abdeldjallil Naceri, Stefan Leutenegger, and Sami Haddadin (2023). *Anthropomorphic Grasping with Neural Object Shape Completion.* URL: `http://arxiv.org/abs/2311.02510` (visited on 04/04/2024). Pre-published (cit. on pp. 19, 20, 59).

Hoang, Dinh-Cuong, Johannes A. Stork, and Todor Stoyanov (2022). “Context-Aware Grasp Generation in Cluttered Scenes”. In: *2022 International Conference on Robotics and Automation (ICRA).* 2022 IEEE International Conference on Robotics and Automation (ICRA). Philadelphia, PA, USA: IEEE, pp. 1492–1498 (cit. on pp. 23, 25).

Houliston, Trent, Jake Fountain, Yuqing Lin, Alexandre Mendes, Mitchell Metcalfe, Josiah Walker, and Stephan K. Chalup (2016). “NUClear: A Loosely Coupled Software Architecture for Humanoid Robot Systems”. In: *Frontiers in Robotics and AI* 3 (cit. on p. 41).

Huang, Wenlong, Pieter Abbeel, Deepak Pathak, and Igor Mordatch (2022a). “Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents” (cit. on p. 53).

Huang, Wenlong, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter (2022b). “Inner Monologue: Embodied Reasoning through Planning with Language Models” (cit. on pp. 53, 55).

Huang, Zeyu, Juzhan Xu, Sisi Dai, Kai Xu, Hao Zhang, Hui Huang, and Ruizhen Hu (2023). "NIFT: Neural Interaction Field and Template for Object Manipulation". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, pp. 1875–1881 (cit. on pp. 19, 22, 59, 61, 64, 135, 161).

Huber, Andreas, Lara Lammer, Astrid Weiss, and Makcus Vincze (2014). "Designing Adaptive Roles for Socially Assistive Robots: A New Method to Reduce Technological Determinism and Role Stereotypes". In: *Journal of Human-Robot Interaction* 3.2, p. 100 (cit. on p. 5).

Iovino, Matteo, Julian Förster, Pietro Falco, Jen Jen Chung, Roland Siegwart, and Christian Smith (2023a). "On the Programming Effort Required to Generate Behavior Trees and Finite State Machines for Robotic Applications". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, pp. 5807–5813 (cit. on p. 41).

Iovino, Matteo, Jonathan Styrud, Pietro Falco, and Christian Smith (2023b). "A Framework for Learning Behavior Trees in Collaborative Robotic Applications". In: *2023 IEEE 19th International Conference on Automation Science and Engineering (CASE)*, pp. 1–8 (cit. on pp. 44, 45, 54).

Jaquier, Noémie, Michael C. Welle, Andrej Gams, Kunpeng Yao, Bernardo Fichera, Aude Billard, Aleš Ude, Asfour Tamim, and Danica Kragic (2024). "Transfer Learning in Robotics: An Upcoming Breakthrough? A Review of Promises and Challenges". In: *International Journal of Robotics Research* (cit. on pp. 3, 5, 38–41, 43, 48, 51, 54, 102, 103).

Jiang, Yuqian, Nick Walker, Minkyu Kim, Nicolas Brissonneau, Daniel S. Brown, Justin W. Hart, Scott Niekum, Luis Sentis, and Peter Stone (2018). *LAAIR: A Layered Architecture for Autonomous Interactive Robots*. URL: `http://arxiv.org/abs/1811.03563` (visited on 02/28/2024). Pre-published (cit. on pp. 44, 45, 159).

Jiang, Zhenyu, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu (2021). *Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations*. URL: `http://arxiv.org/abs/2104.01542` (visited on 04/04/2024). Pre-published (cit. on pp. 23, 27).

Jocher, Glenn, Ayush Chaurasia, Alex Stoken, Jirka Borovec, Yonghye Kwon, Kalen Michael, Jiacong Fang, Zeng Yifu, Colin Wong, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, UnglvKitDe, Victor Sonck, tkianai, Piotr Skalski, Adam Hogan, Dhruv Nair, Max Strobel, and Mrinal Jain (2022).

Yolov5: V7.0 - YOLOv5 SOTA Realtime Instance Segmentation. Zenodo (cit. on p. 121).

Kaiser, Peter, Eren E. Aksoy, Markus Grotz, Dimitrios Kanoulas, Nikos G. Tsagarakis, and Tamim Asfour (2017). “Experimental Evaluation of a Perceptual Pipeline for Hierarchical Affordance Extraction”. In: *International Symposium on Experimental Robotics (ISER).* Ed. by Dana Kulić, Yoshihiko Nakamura, Oussama Khatib, and Gentiane Venture. Springer International Publishing, pp. 136–146 (cit. on p. 67).

Kaiser, Peter and Tamim Asfour (2018). “Autonomous Detection and Experimental Validation of Affordances”. In: *IEEE Robotics and Automation Letters (RA-L)* 3.3, pp. 1949–1956 (cit. on pp. 67, 81).

Kaiser, Peter, Markus Grotz, Fabian Paus, and Tamim Asfour (2018). “Towards the Formalization of Affordances as Dempster-Shafer Belief Functions”. In: *1st International Workshop on Computational Models of Affordance in Robotics, Robotics Science and Systems (RSS)* (cit. on pp. 81, 134).

Kaiser, Peter, Dimitrios Kanoulas, Markus Grotz, Luca Muratore, Alessio Rocchi, Enrico Mingo Hoffman, Nikos G. Tsagarakis, and Tamim Asfour (2016). “An Affordance-Based Pilot Interface for High-Level Control of Humanoid Robots in Supervised Autonomy”. In: *IEEE/RAS International Conference on Humanoid Robots (Humanoids).* Cancún, Mexico, pp. 621–628 (cit. on p. 71).

Kandpal, Nikhil, H. Deng, Adam Roberts, Eric Wallace, and Colin Raffel (2022). “Large Language Models Struggle to Learn Long-Tail Knowledge”. In: *International Conference on Machine Learning* (cit. on p. 116).

Kappler, Daniel, Jeannette Bohg, and Stefan Schaal (2015). “Leveraging Big Data for Grasp Planning”. In: *2015 IEEE International Conference on Robotics and Automation (ICRA).* 2015 IEEE International Conference on Robotics and Automation (ICRA). Seattle, WA, USA: IEEE, pp. 4304–4311 (cit. on pp. 31, 36).

Kappler, Daniel, Stefan Schaal, and Jeannette Bohg (2016). “Optimizing for What Matters: The Top Grasp Hypothesis”. In: *2016 IEEE International Conference on Robotics and Automation (ICRA).* 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm: IEEE, pp. 2167–2174 (cit. on pp. 31, 36).

Kasaei, Hamidreza and Mohammadreza Kasaei (2024). *Harnessing the Synergy between Pushing, Grasping, and Throwing to Enhance Object Manipulation in Cluttered Scenarios.* URL: `http://arxiv.org/abs/2402.16045` (visited on 02/29/2024). Pre-published (cit. on pp. 44, 49, 54).

Keleştemur, Tarik, Naoki Yokoyama, Joanne Truong, Anas Abou Allaban, and Taşkin Padir (2019). “System Architecture for Autonomous Mobile Manipulation of Everyday Objects in Domestic Environments”. In: *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. PETRA ’19: The 12th PErvasive Technologies Related to Assistive Environments Conference. Rhodes Greece: ACM, pp. 264–269 (cit. on pp. 44, 46, 104).

Kendall, Alex, Yarin Gal, and Roberto Cipolla (2018). “Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics”. In: *Conf. on Computer Vision and Pattern Recognition (CVPR)*. Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 7482–7491 (cit. on p. 135).

Kent, David and Russell Toris (2018). “Adaptive Autonomous Grasp Selection via Pairwise Ranking”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, pp. 2971–2976 (cit. on pp. 31, 34, 37).

Kerr, Justin, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik (2023). “LERF: Language Embedded Radiance Fields”. In: *Intl. Conf. on Computer Vision (ICCV)*. Intl. Conf. on Computer Vision (ICCV), pp. 19729–19739 (cit. on p. 59).

Kirillov, Alexander, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick (2023). “Segment Anything”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, pp. 3992–4003 (cit. on p. 63).

Kobayashi, Shoshichi and Katsumi Nomizu (1996). *Foundations of Differential Geometry*. Vol. 61. 2. John Wiley & Sons. 296 pp. (cit. on p. 69).

Konrad, Anna, John McDonald, and Rudi Villing (2022). “VGQ-CNN: Moving Beyond Fixed Cameras and Top-Grasps for Grasp Quality Prediction”. In: *2022 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8 (cit. on pp. 31, 36).

Kopicki, Marek, Dominik Belter, and Jeremy L. Wyatt (2019). *Learning Better Generative Models for Dexterous, Single-View Grasping of Novel Objects*. URL: `http://arxiv.org/abs/1907.06053` (visited on 02/22/2024). Pre-published (cit. on pp. 23, 28).

Kopicki, Marek, Renaud Detry, Maxime Adjigble, Rustam Stolkin, Ales Leonardis, and Jeremy L. Wyatt (2016). “One-Shot Learning and Generation of Dexterous

Grasps for Novel Objects". In: *The International Journal of Robotics Research* 35.8, pp. 959–976 (cit. on pp. 23, 28).

Kortenkamp, David, Reid Simmons, and Davide Brugali (2016). "Robotic Systems Architectures and Programming". In: *Springer Handbook of Robotics*. Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 283–306 (cit. on pp. 2, 38, 39, 102).

Koubaa, Anis, Mohamed-Foued Sriti, Yasir Javed, Maram Alajlan, Basit Qureshi, Fatma Ellouze, and Abdelrahman Mahmoud (2016). "Turtlebot at Office: A Service-Oriented Software Architecture for Personal Assistant Robots Using ROS". In: *2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. 2016 International Conference on Autonomous Robot Systems and Competitions (ICARSC). Bragança, Portugal: IEEE, pp. 270–276 (cit. on pp. 44, 47).

Kraft, Dirk, Renaud Detry, Nicolas Pugeault, Emre Başeski, Justus Piater, and Norbert Krüger (2009). "Learning Objects and Grasp Affordances through Autonomous Exploration". In: *Computer Vision Systems*. Ed. by Mario Fritz, Bernt Schiele, and Justus H. Piater. Red. by David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Madhu Sudan, Demetri Terzopoulos, Doug Tygar, Moshe Y. Vardi, and Gerhard Weikum. Vol. 5815. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 235–244 (cit. on pp. 22, 23).

Kragic, D. and M. Vincze (2009). "Vision for Robotics". In: *Foundations and Trends in Robotics* 1.1, pp. 1–78 (cit. on pp. 18, 164).

Krebs, Franziska, Andre Meixner, Isabel Patzer, and Tamim Asfour (2021). "The KIT Bimanual Manipulation Dataset". In: *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*. Best Oral Paper Award Finalist. Munich, Germany, pp. 499–506 (cit. on pp. 112, 152).

Krug, Robert, Achim J. Lilienthal, Danica Kragic, and Yasemin Bekiroglu (2016). "Analytic Grasp Success Prediction with Tactile Feedback". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 2016 IEEE International Conference on Robotics and Automation (ICRA). Stockholm: IEEE, pp. 165–171 (cit. on pp. 31, 32, 37).

Krüger, Norbert, Christopher Geib, Justus Piater, Ronald Petrick, Mark Steedman, Florentin Wörgötter, Aleš Ude, Tamim Asfour, Dirk Kraft, Damir Omrčen, Alejandro Agostini, and Rüdiger Dillmann (2011). "Object–Action Complexes: Grounded Abstractions of Sensory–Motor Processes". In: *Robotics and Autonomous Systems* 59.10, pp. 740–757 (cit. on pp. 134, 163).

Kumar, Rajesh and Sudipto Mukherjee (2022). “Algorithmic Selection of Preferred Grasp Poses Using Manipulability Ellipsoid Forms”. In: *Journal of Mechanisms and Robotics* 14.051006 (cit. on pp. 30, 31, 37).

Kurenkov, Andrey, Viraj Mehta, Jingwei Ji, Animesh Garg, and S. Savarese (2017). “Towards Grasp Transfer Using Shape Deformation”. In: (cit. on pp. 19, 20).

Labbé, Yann, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, Dieter Fox, and Josef Sivic (2022). “MegaPose: 6D Pose Estimation of Novel Objects via Render & Compare” (cit. on p. 121).

Lee, In (2021). “Service Robots: A Systematic Literature Review”. In: *Electronics* 10.21, p. 2658 (cit. on p. 1).

Lee, Taeyoung (2018). “Bayesian Attitude Estimation with the Matrix Fisher Distribution on SO(3)”. In: *IEEE Transactions on Automatic Control* 63.10, pp. 3377–3392 (cit. on p. 86).

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. (2020). “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems* 33, pp. 9459–9474 (cit. on pp. 115, 160).

Li, Samuel, Sarthak Bhagat, Joseph Campbell, Yaqi Xie, Woojun Kim, Katia Sycara, and Simon Stepputtis (2024). *ShapeGrasp: Zero-Shot Task-Oriented Grasping with Large Language Models through Geometric Decomposition.* URL: `http://arxiv.org/abs/2403.18062` (visited on 04/02/2024). Pre-published (cit. on pp. 19, 21).

Li, Yiming, Wei Wei, Daheng Li, Peng Wang, Wanyi Li, and Jun Zhong (2022). “HGC-Net: Deep Anthropomorphic Hand Grasping in Clutter”. In: *2022 International Conference on Robotics and Automation (ICRA).* 2022 International Conference on Robotics and Automation (ICRA), pp. 714–720 (cit. on pp. 23, 25).

Liang, Jacky, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng (2023). “Code as Policies: Language Model Programs for Embodied Control” (cit. on p. 53).

Liang, Jacky, Mohit Sharma, Alex LaGrassa, Shivam Vats, Saumya Saxena, and Oliver Kroemer (2022). “Search-Based Task Planning with Learned Skill Effect Models for Lifelong Robotic Manipulation”. In: *2022 International Conference on Robotics and Automation (ICRA)*, pp. 6351–6357 (cit. on pp. 44, 45).

Lin, Kevin, Christopher Agia, Toki Migimatsu, Marco Pavone, and Jeannette Bohg (2023). “Text2Motion: From Natural Language Instructions to Feasible Plans” (cit. on pp. 52, 53, 55).

Lin, Kevin, Lijuan Wang, and Zicheng Liu (2021). “Mesh Graphormer”. In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021 IEEE/CVF International Conference on Computer Vision (ICCV). Montreal, QC, Canada: IEEE, pp. 12919–12928 (cit. on p. 63).

Lin, Yun and Yu Sun (2015). “Grasp Planning to Maximize Task Coverage”. In: *The International Journal of Robotics Research* 34.9, pp. 1195–1210 (cit. on pp. 31, 32, 37).

Liu, Bo, Yuqian Jiang, Xiaohan Zhang, Qiang Liu, Shiqi Zhang, Joydeep Biswas, and Peter Stone (2023a). *LLM+P: Empowering Large Language Models with Optimal Planning Proficiency*. URL: `http://arxiv.org/abs/2304.11477` (visited on 04/03/2024). Pre-published (cit. on pp. 51, 53, 54, 113, 114, 118, 166).

Liu, Peiqi, Yaswanth Orru, Chris Paxton, Nur Muhammad Mahi Shafiullah, and Lerrel Pinto (2024a). *OK-Robot: What Really Matters in Integrating Open-Knowledge Models for Robotics*. URL: `http://arxiv.org/abs/2401.12202` (visited on 02/22/2024). Pre-published (cit. on pp. 44, 47).

Liu, Rui and Xiaoli Zhang (2019). “A Review of Methodologies for Natural-Language-Facilitated Human–Robot Cooperation”. In: *International Journal of Advanced Robotic Systems* 16.3, p. 1729881419851402 (cit. on p. 114).

Liu, Shilong, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang (2023b). *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. URL: `http://arxiv.org/abs/2303.05499` (visited on 04/29/2024). Pre-published (cit. on p. 63).

Liu, Wenhai, Weiming Wang, Yang You, Teng Xue, Zhenyu Pan, Jin Qi, and Jie Hu (2022). “Robotic Picking in Dense Clutter via Domain Invariant Learning from Synthetic Dense Cluttered Rendering”. In: *Robotics and Autonomous Systems* 147, p. 103901 (cit. on pp. 23, 25, 29).

Liu, Yuchen, Luigi Palmieri, Sebastian Koch, Ilche Georgievski, and Marco Aiello (2024b). “DELTA: Decomposed Efficient Long-Term Robot Task Planning Using Large Language Models” (cit. on p. 54).

Logothetis, Michalis, George C. Karras, Shahab Heshmati-Alamdari, Panagiotis Vlantis, and Kostas J. Kyriakopoulos (2018). “A Model Predictive Control Approach for Vision-Based Object Grasping via Mobile Manipulator”. In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid: IEEE, pp. 1–6 (cit. on p. 18).

Lörken, Christopher and Joachim Hertzberg (2008). "Grounding Planning Operators by Affordances". In: *International Conference on Cognitive Systems (CogSys)*, pp. 79–84 (cit. on p. 116).

Lu, Qingkai, Mark Van Der Merwe, Balakumar Sundaralingam, and Tucker Hermans (2020). "Multifingered Grasp Planning via Inference in Deep Neural Networks: Outperforming Sampling by Learning Differentiable Models". In: *IEEE Robotics & Automation Magazine* 27.2, pp. 55–65 (cit. on pp. 31, 37).

Madry, Marianna, Dan Song, and Danica Kragic (2012). "From Object Categories to Grasp Transfer Using Probabilistic Reasoning". In: *2012 IEEE International Conference on Robotics and Automation*. 2012 IEEE International Conference on Robotics and Automation (ICRA). St Paul, MN, USA: IEEE, pp. 1716–1723 (cit. on pp. 19, 20).

Mahler, Jeffrey, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg (2017). *Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics*. URL: `http://arxiv.org/abs/1703.09312` (visited on 02/12/2024). Prepublished (cit. on pp. 23, 27).

Makoviychuk, Viktor, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State (2021). "Isaac Gym: High Performance GPU Based Physics Simulation for Robot Learning". In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Ed. by J. Vanschoren and S. Yeung. Vol. 1 (cit. on pp. 63, 166).

Martinez-Martin, Ester and Angel P. del Pobil (2018). "Personal Robot Assistants for Elderly Care: An Overview". In: *Personal Assistants: Emerging Computational Technologies*. Ed. by Angelo Costa, Vicente Julian, and Paulo Novais. Cham: Springer International Publishing, pp. 77–91 (cit. on pp. 2, 4).

Martins, Diogo, Sara Aldhaheri, Pavel Kopanev, Èric Pairet, Paola Ardón, and Alirio Sá (2023). "An Action-Level Assistant for Robotic Manipulation: User Experience and Performance". In: *2023 XIII Brazilian Symposium on Computing Systems Engineering (SBESC)*. 2023 XIII Brazilian Symposium on Computing Systems Engineering (SBESC). Porto Alegre, Brazil: IEEE, pp. 1–6 (cit. on pp. 44, 49, 104).

Marton, Z, D Pangercic, N Blodow, J Kleinehellefort, and M Beetz (2010). "General 3D Modelling of Novel Objects from a Single View". In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*. 2010 IEEE/RSJ

International Conference on Intelligent Robots and Systems (IROS 2010). Taipei: IEEE, pp. 3700–3705 (cit. on pp. 23, 24).

Mavrakis, Nikos, E. Amir M. Ghalamzan, and Rustam Stolkin (2017). "Safe Robotic Grasping: Minimum Impact-Force Grasp Selection". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Vancouver, BC: IEEE, pp. 4034–4041 (cit. on pp. 31, 32).

Mayr, Matthias, Francesco Rovida, and Volker Krueger (2023). "SkiROS2: A Skill-Based Robot Control Platform for ROS". In: *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Detroit, MI, USA: IEEE, pp. 6273–6280 (cit. on pp. 44, 45).

Mescheder, Lars, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger (2019). "Occupancy Networks: Learning 3D Reconstruction in Function Space". In: *Conf. on Computer Vision and Pattern Recognition (CVPR).* Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4460–4470 (cit. on p. 61).

Miller, David P. and Illah Nourbakhsh (2016). "Robotics for Education". In: *Springer Handbook of Robotics.* Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 2115–2134 (cit. on p. 1).

Mnyussiwalla, H., P. Seguin, P. Vulliez, and J. P. Gazeau (2022). "Evaluation and Selection of Grasp Quality Criteria for Dexterous Manipulation". In: *Journal of Intelligent & Robotic Systems* 104.2, p. 20 (cit. on pp. 31, 32).

Moldovan, Bogdan, Plinio Moreno, Davide Nitti, José Santos-Victor, and Luc De Raedt (2018). "Relational Affordances for Multiple-Object Manipulation". In: *Autonomous Robots* 42.1, pp. 19–44 (cit. on p. 134).

Montesano, Luis, Manuel Lopes, Alexandre Bernardino, and JosÉ Santos-Victor (2008). "Learning Object Affordances: From Sensory–Motor Coordination to Imitation". In: *IEEE Transactions on Robotics* 24.1, pp. 15–26 (cit. on pp. 133, 134).

Morales, A., E. Chinellato, A.H. Fagg, and A.P. Del Pobil (2003). "Experimental Prediction of the Performance of Grasp Tasks from Visual Features". In: *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453).* 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vol. 3. Las Vegas, NV, USA: IEEE, pp. 3423–3428 (cit. on pp. 31, 35).

Mosbach, Malte and Sven Behnke (2024). *Grasp Anything: Combining Teacher-Augmented Policy Gradient Learning with Instance Segmentation to Grasp*

Arbitrary Objects. URL: `http://arxiv.org/abs/2403.10187` (visited on 04/02/2024). Pre-published (cit. on pp. 23, 27, 29).

Muhlig, Manuel, Michael Gienger, Sven Hellbach, Jochen J. Steil, and Christian Goerick (2009). “Task-Level Imitation Learning Using Variance-based Movement Optimization”. In: *2009 IEEE International Conference on Robotics and Automation.* 2009 IEEE International Conference on Robotics and Automation, pp. 1177–1184 (cit. on p. 112).

Mullen Jr., James F. and Dinesh Manocha (2024). *Towards Robots That Know When They Need Help: Affordance-Based Uncertainty for Large Language Model Planners.* URL: `http://arxiv.org/abs/2403.13198` (visited on 04/03/2024). Pre-published (cit. on p. 42).

Murphy, Jamie, Charles Hofacker, and Ulrike Gretzel (2017). “Dawning of the Age of Robots in Hospitality and Tourism: Challenges for Teaching and Research”. In: *European Journal of Tourism Research* 15, pp. 104–111 (cit. on p. 1).

Nadon, Felix and Pierre Payeur (2020). “Grasp Selection for In-Hand Robotic Manipulation of Non-Rigid Objects with Shape Control”. In: *2020 IEEE International Systems Conference (SysCon).* 2020 IEEE International Systems Conference (SysCon). Montreal, QC, Canada: IEEE, pp. 1–8 (cit. on pp. 31, 34, 37).

Nam, Changjoo, Seokjun Lee, Jeongho Lee, Sang Hun Cheong, Dong Hwan Kim, Changhwan Kim, Incheol Kim, and Sung-Kee Park (2020). “A Software Architecture for Service Robots Manipulating Objects in Human Environments”. In: *IEEE Access* 8, pp. 117900–117920 (cit. on pp. 44, 49).

Nebot, Patricio and Enric Cervera (2007). “An Integrated Agent-Based Software Architecture for Mobile and Manipulator Systems”. In: *Robotica* 25.2, pp. 213–220 (cit. on pp. 44, 47).

Newbury, Rhys, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, Dieter Fox, and Akansel Cosgun (2023). *Deep Learning Approaches to Grasp Synthesis: A Review.* URL: `http://arxiv.org/abs/2207.02556` (visited on 02/12/2024). Pre-published (cit. on pp. 18, 22, 29).

Ni, Peiyuan, Wenguang Zhang, Xiaoxiao Zhu, and Qixin Cao (2021). “Learning an End-to-End Spatial Grasp Generation and Refinement Algorithm from Simulation”. In: *Machine Vision and Applications* 32.1, p. 10 (cit. on pp. 23, 25).

O’Neill, Abby et al. (2024). “Open X-Embodiment: Robotic Learning Datasets and RT-X Models : Open X-Embodiment Collaboration0”. In: *2024 IEEE International Conference on Robotics and Automation (ICRA).* 2024 IEEE

International Conference on Robotics and Automation (ICRA), pp. 6892–6903 (cit. on pp. 39, 166).

Ohneberg, Christoph, Nicole Stöbich, Angelika Warmbein, Ivanka Rathgeber, Amrei Christin Mehler-Klamt, Uli Fischer, and Inge Eberl (2023). “Assistive Robotic Systems in Nursing Care: A Scoping Review”. In: *BMC Nursing* 22.1, p. 72 (cit. on pp. 2, 4).

Paikan, Ali, David Schiebener, Mirko Wachter, Tamim Asfour, Giorgio Metta, and Lorenzo Natale (2015). “Transferring Object Grasping Knowledge and Skill across Different Robotic Platforms”. In: *2015 International Conference on Advanced Robotics (ICAR).* 2015 International Conference on Advanced Robotics (ICAR). Istanbul, Turkey: IEEE, pp. 498–503 (cit. on pp. 44, 47).

Pane, Yudha, Erwin Aertbelien, Joris De Schutter, and Wilm Decre (2020). “Skill-Based Programming Framework for Composable Reactive Robot Behaviors”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, NV, USA: IEEE, pp. 7087–7094 (cit. on pp. 44, 45).

Papon, Jeremie, Alexey Abramov, Markus Schoeler, and Florentin Worgotter (2013). “Voxel Cloud Connectivity Segmentation - Supervoxels for Point Clouds”. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, pp. 2027–2034 (cit. on p. 71).

Pardi, Tommaso, Ghalamzan E. Amir, Valerio Ortenzi, and Rustam Stolkin (2021). “Optimal Grasp Selection, and Control for Stabilising a Grasped Object, with Respect to Slippage and External Forces”. In: *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids).* 2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids). Munich, Germany: IEEE, pp. 429–436 (cit. on pp. 31, 33).

Park, Jeong Joon, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove (2019). “DeepSDF: Learning Continuous Signed Distance Functions for Shape Representation”. In: *Conf. on Computer Vision and Pattern Recognition (CVPR).* Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 165–174 (cit. on p. 61).

Patrikalakis, Nicholas M. and Takashi Maekawa (2010). *Shape Interrogation for Computer Aided Design and Manufacturing.* Berlin, Heidelberg: Springer Berlin Heidelberg. 1-408 (cit. on pp. 69, 72).

Patten, Timothy, Kiru Park, and Markus Vincze (2020). “DGCM-Net: Dense Geometrical Correspondence Matching Network for Incremental Experience-based Robotic Grasping". In: *Frontiers in Robotics and AI* 7 (cit. on pp. 23, 28).

Paus, Fabian and Tamim Asfour (2020). "Probabilistic Representation of Objects and Their Support Relations". In: *International Symposium on Experimental Robotics (ISER)*. Springer International Publishing, pp. 510–519 (cit. on p. 95).

Peller-Konrad, Fabian, Rainer Kartmann, Christian R. G. Dreher, Andre Meixner, Fabian Reister, Markus Grotz, and Tamim Asfour (2023). "A Memory System of a Robot Cognitive Architecture and Its Implementation in ArmarX". In: *Robotics and Autonomous Systems* 164, p. 104415 (cit. on pp. 14, 103–105, 107, 108, 115, 129, 152, 157, 166).

Pineau, Joelle, Michael Montemerlo, Martha Pollack, Nicholas Roy, and Sebastian Thrun (2003). "Towards Robotic Assistants in Nursing Homes: Challenges and Results". In: *Robotics and Autonomous Systems* 42.3-4, pp. 271–281 (cit. on pp. 2, 4).

Player, Timothy R., Dongsik Chang, Li Fuxin, and Geoffrey A. Hollinger (2023). "Real-Time Generative Grasping with Spatio-temporal Sparse Convolution". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, pp. 7981–7987 (cit. on pp. 23, 25).

Pohl, Christoph and Tamim Asfour (2022). "Probabilistic Spatio-Temporal Fusion of Affordances for Grasping and Manipulation". In: *IEEE Robotics and Automation Letters* 7.2, pp. 3226–3233 (cit. on pp. 10–12, 23, 29, 38, 66, 68, 69, 72, 75, 76, 80, 81, 87, 89, 126, 160, 161).

Pohl, Christoph, Patrick Hegemann, Byungchul An, Markus Grotz, and Tamim Asfour (2022). "Humanoid Robotic System for Grasping and Manipulation in Decontamination Tasks: Humanoides Robotersystem Für Das Greifen Und Manipulieren Bei Dekontaminierungsaufgaben". In: *at - Automatisierungstechnik* 70.10, pp. 850–858 (cit. on pp. 10, 108).

Pohl, Christoph, Kevin Hitzler, Raphael Grimm, Antonio Zea, Uwe D. Hanebeck, and Tamim Asfour (2020). "Affordance-Based Grasping and Manipulation in Real World Applications". In: *Proc. of the 2020 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*. International Conference on Intelligent Robots and Systems (IROS). Las Vegas, NV, USA: IEEE/RSJ, pp. 9569–9576 (cit. on pp. 10, 90).

Pohl, Christoph, Fabian Reister, Fabian Peller-Konrad, and Tamim Asfour (2024). "MAkEable: Memory-centered and Affordance-based Task Execution Framework for Transferable Mobile Manipulation Skills". In: *Proc. of the 2024 IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems*. International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, UAE: IEEE/RSJ, accepted for publication (cit. on pp. 10, 14, 44, 103, 105, 106, 109–112, 127, 162).

Popović, Mila, Dirk Kraft, Leon Bodenhagen, Emre Başeski, Nicolas Pugeault, Danica Kragic, Tamim Asfour, and Norbert Krüger (2010). “A Strategy for Grasping Unknown Objects Based on Co-Planarity and Colour Information”. In: *Robotics and Autonomous Systems* 58.5, pp. 551–565 (cit. on pp. 23, 24).

Prassler, Erwin, Mario E. Munich, Paolo Pirjanian, and Kazuhiro Kosuge (2016). “Domestic Robotics”. In: *Springer Handbook of Robotics*. Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 1729–1758 (cit. on p. 2).

Prokudin, Sergey, Christoph Lassner, and Javier Romero (2019). “Efficient Learning on Point Clouds with Basis Point Sets”. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). Seoul, Korea (South): IEEE, pp. 3072–3081 (cit. on pp. 61, 161).

Prorok, Amanda, Matthew Malencia, Luca Carlone, Gaurav S. Sukhatme, Brian M. Sadler, and Vijay Kumar (2021). *Beyond Robustness: A Taxonomy of Approaches towards Resilient Multi-Robot Systems*. URL: `http://arxiv.org/abs/2109.12343` (visited on 04/17/2024). Pre-published (cit. on p. 4).

Qian, Jianing, Thomas Weng, Luxin Zhang, Brian Okorn, and David Held (2020). “Cloth Region Segmentation for Robust Grasp Selection”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Las Vegas, NV, USA: IEEE, pp. 9553–9560 (cit. on pp. 31, 36).

Qin, Ran, Haoxiang Ma, Boyang Gao, and Di Huang (2023). “RGB-D Grasp Detection via Depth Guided Learning with Cross-modal Attention”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8003–8009 (cit. on pp. 23, 26).

Quigley, Morgan, Brian Gerkey, Ken Conley, Josh Faust, Tully Foote, Jeremy Leibs, Eric Berger, Rob Wheeler, and Andrew Ng (2009). “ROS: An Open-Source Robot Operating System”. In: *Proc. of the IEEE Intl. Conf. on Robotics and Automation (ICRA) Workshop on Open Source Robotics*. IEEE Intl. Conf. on Robotics and Automation (ICRA). Kobe, Japan: IEEE (cit. on pp. 41, 157).

Quispe, Ana Huaman, Heni Ben Amor, and Henrik I. Christensen (2016). “Combining Arm and Hand Metrics for Sensible Grasp Selection”. In: *2016 IEEE International Conference on Automation Science and Engineering (CASE)*.

2016 IEEE International Conference on Automation Science and Engineering (CASE). Fort Worth, TX, USA: IEEE, pp. 1170–1176 (cit. on pp. 31, 33).

Rabiner, Lawrence R. (1990). “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”. In: *Readings in Speech Recognition.* Elsevier, pp. 267–296 (cit. on pp. 84, 160).

Rana, Krishan, Jesse Haviland, Sourav Garg, Jad Abou-Chakra, Ian Reid, and Niko Suenderhauf (2023). “SayPlan: Grounding Large Language Models Using 3D Scene Graphs for Scalable Task Planning” (cit. on pp. 52, 53, 55).

Rao, Deepak, Quoc V. Le, Thanathorn Phoka, Morgan Quigley, Attawith Sudsang, and Andrew Y. Ng (2010). “Grasping Novel Objects with Depth Segmentation”. In: *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems.* 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2578–2585 (cit. on pp. 23, 24).

Rashid, Adam, Satvik Sharma, Chung Min Kim, Justin Kerr, Lawrence Yunliang Chen, Angjoo Kanazawa, and Ken Goldberg (2023). “Language Embedded Radiance Fields for Zero-Shot Task-Oriented Grasping”. In: *Conference on Robot Learning (CoRL).* Conference on Robot Learning (CoRL), pp. 178–200 (cit. on p. 59).

Ren, Yi, Zhehua Zhou, Ziwei Xu, Yang Yang, Guangyao Zhai, Marion Leibold, Fenglei Ni, Zhengyou Zhang, Martin Buss, and Yu Zheng (2024). “Enabling Versatility and Dexterity of the Dual-Arm Manipulators: A General Framework Toward Universal Cooperative Manipulation”. In: *IEEE Transactions on Robotics* 40, pp. 2024–2045 (cit. on pp. 44, 47).

Rodriguez, Diego, Corbin Cogswell, Seongyong Koo, and Sven Behnke (2018). *Transferring Grasping Skills to Novel Instances by Latent Space Non-Rigid Registration.* URL: `http://arxiv.org/abs/1809.05353` (visited on 02/12/2024). Pre-published (cit. on pp. 19, 20).

Rodriguez-Gonzalez, Carmen Guadalupe, Ana Herranz-Alonso, Vicente Escudero-Vilaplana, Maria Aranzazu Ais-Larisgoitia, Irene Iglesias-Peinado, and Maria Sanjurjo-Saez (2019). “Robotic Dispensing Improves Patient Safety, Inventory Management, and Staff Satisfaction in an Outpatient Hospital Pharmacy”. In: *Journal of Evaluation in Clinical Practice* 25.1, pp. 28–35 (cit. on pp. 2, 4).

Rohanimanesh, Khashayar, Jake Metzger, William Richards, and Aviv Tamar (2023). “Online Tool Selection with Learned Grasp Prediction Models”. In: *2023 IEEE International Conference on Robotics and Automation (ICRA).* 2023 IEEE International Conference on Robotics and Automation (ICRA). London, United Kingdom: IEEE, pp. 5844–5850 (cit. on pp. 31, 37, 38).

Rovida, Francesco and Volker Kruger (2015). "Design and Development of a Software Architecture for Autonomous Mobile Manipulators in Industrial Environments". In: *2015 IEEE International Conference on Industrial Technology (ICIT).* 2015 IEEE International Conference on Industrial Technology (ICIT). Seville: IEEE, pp. 3288–3295 (cit. on pp. 44, 45).

Rubert, Carlos, Daniel Kappler, Jeannette Bohg, and Antonio Morales (2018). "Grasp Success Prediction with Quality Metrics" (cit. on pp. 31, 35).

Rubert, Carlos, Daniel Kappler, Antonio Morales, Stefan Schaal, and Jeannette Bohg (2017). "On the Relevance of Grasp Metrics for Predicting Grasp Success". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS).* IEEE, pp. 265–272 (cit. on pp. 31, 35).

Ruiz, Oriol, Jan Rosell, and Mohammed Diab (2022). "Reasoning and State Monitoring for the Robust Execution of Robotic Manipulation Tasks". In: *2022 IEEE 27th International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1–4 (cit. on p. 42).

Ruiz-Celada, Oriol, Parikshit Verma, Mohammed Diab, and Jan Rosell (2022). "Automating Adaptive Execution Behaviors for Robot Manipulation". In: *IEEE Access* 10, pp. 123489–123497 (cit. on p. 41).

Rusu, Radu Bogdan and Steve Cousins (2011). "3D Is Here: Point Cloud Library (PCL)". In: *2011 IEEE International Conference on Robotics and Automation.* 2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, pp. 1–4 (cit. on pp. 71, 160).

Sabzejou, Ali, Mehdi Tale Masouleh, and Ahmad Kalhor (2023). "2D Skeleton-Based Keypoint Generation Method for Grasping Objects with Roughly Uniform Height Variation". In: *2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM).* 2023 11th RSI International Conference on Robotics and Mechatronics (ICRoM). Tehran, Iran, Islamic Republic of: IEEE, pp. 847–853 (cit. on pp. 23, 25, 29).

Şahin, Erol, Maya Çakmak, Mehmet R. Doğar, Emre Uğur, and Göktürk Üçoluk (2007). "To Afford or Not to Afford: A New Formalization of Affordances Toward Affordance-Based Robot Control". In: *Adaptive Behavior* 15.4, pp. 447–472 (cit. on pp. 6, 116, 133, 165).

Sarkisyan, Christina, Alexandr Korchemnyi, Alexey K. Kovalev, and Aleksandr I. Panov (2023). "Evaluation of Pretrained Large Language Models in Embodied Planning Tasks". In: *Artificial General Intelligence*, pp. 222–232 (cit. on pp. 51, 52).

Sawyer, Ben D., Dave B. Miller, Matthew Canham, and Waldemar Karwowski (2021). "Human Factors and Ergonomics in Design of A^3: Automation, Autonomy, and Artifical Intelligence". In: *Handbook of Human Factors and Ergonomics*. Ed. by Gavriel Salvendy and Waldemar Karwowski. 1st ed. Wiley, pp. 1385–1416 (cit. on pp. 3–5).

Saxena, Ashutosh, Justin Driemeyer, and Andrew Y. Ng (2008). "Robotic Grasping of Novel Objects Using Vision". In: *The International Journal of Robotics Research* 27.2, pp. 157–173 (cit. on pp. 23, 26).

Schiebener, David, Julian Schill, and Tamim Asfour (2012). "Discovery, Segmentation and Reactive Grasping of Unknown Objects". In: *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*. 2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012). Osaka, Japan: IEEE, pp. 71–77 (cit. on pp. 23, 24).

Schiebener, David, Andreas Schmidt, Nikolaus Vahrenkamp, and Tamim Asfour (2016). "Heuristic 3D Object Shape Completion Based on Symmetry and Scene Context". In: *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, South Korea: IEEE, pp. 74–81 (cit. on pp. 22, 23).

Schmidt, Philipp, Nikolaus Vahrenkamp, Mirko Wachter, and Tamim Asfour (2018). "Grasping of Unknown Objects Using Deep Convolutional Neural Networks Based on Depth Images". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. 2018 IEEE International Conference on Robotics and Automation (ICRA). Brisbane, QLD: IEEE, pp. 6831–6838 (cit. on pp. 23, 26, 29).

Sharif, Mohammadreza, Deniz Erdogmus, and Taskin Padir (2019). "Particle Filters vs Hidden Markov Models for Prosthetic Robot Hand Grasp Selection". In: *International Journal of Robotic Computing* 1.2, pp. 98–122 (cit. on pp. 31, 33).

Si, Zilin, Zirui Zhu, Arpit Agarwal, Stuart Anderson, and Wenzhen Yuan (2022). "Grasp Stability Prediction with Sim-to-Real Transfer from Tactile Sensing". In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7809–7816 (cit. on pp. 30, 31).

Simeonov, Anthony, Yilun Du, Yen-Chen Lin, Alberto Rodriguez Garcia, Leslie Pack Kaelbling, Tomás Lozano-Pérez, and Pulkit Agrawal (2023). "SE(3)-Equivariant Relational Rearrangement with Neural Descriptor Fields". In: *Conference on Robot Learning*, pp. 835–846 (cit. on pp. 19, 22, 59, 64, 135, 161).

Simeonov, Anthony, Yilun Du, Andrea Tagliasacchi, Joshua B. Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann (2022). "Neural Descriptor

Fields: SE(3)-Equivariant Object Representations for Manipulation". In: *2022 International Conference on Robotics and Automation (ICRA)*. 2022 IEEE International Conference on Robotics and Automation (ICRA). Philadelphia, PA, USA: IEEE, pp. 6394–6400 (cit. on pp. 19, 21, 59, 62, 64, 135, 161).

Singh, Ishika, Valts Blukis, Arsalan Mousavian, Ankit Goyal, Danfei Xu, Jonathan Tremblay, Dieter Fox, Jesse Thomason, and Animesh Garg (2023). "ProgPrompt: Generating Situated Robot Task Plans Using Large Language Models". In: *Icra* (cit. on p. 53).

Sjøberg, Alexander Meyer and Olav Egeland (2021). "Lie Algebraic Unscented Kalman Filter for Pose Estimation". In: *IEEE Transactions on Automatic Control*, pp. 1–1 (cit. on pp. 81, 86).

Smarr, Cory-Ann, Tracy L. Mitzner, Jenay M. Beer, Akanksha Prakash, Tiffany L. Chen, Charles C. Kemp, and Wendy A. Rogers (2014). "Domestic Robots for Older Adults: Attitudes, Preferences, and Potential". In: *International Journal of Social Robotics* 6.2, pp. 229–247 (cit. on p. 2).

Solà, Joan, Jeremie Deray, and Dinesh Atchuthan (2018). "A Micro Lie Theory for State Estimation in Robotics". In: *arXiv* (cit. on pp. 83, 86).

Song, Chan Hee, Jiaman Wu, Clayton Washington, Brian M. Sadler, Wei-Lun Chao, and Yu Su (2023). "LLM-Planner: Few-Shot Grounded Planning for Embodied Agents with Large Language Models" (cit. on pp. 52, 53, 113).

Song, Dan, Carl Henrik Ek, Kai Huebner, and Danica Kragic (2011). "Multivariate Discretization for Bayesian Network Structure Learning in Robot Grasping". In: *2011 IEEE International Conference on Robotics and Automation*. 2011 IEEE International Conference on Robotics and Automation (ICRA). Shanghai, China: IEEE, pp. 1944–1950 (cit. on pp. 31, 36).

Song, Dan, Carl Henrik Ek, Kai Huebner, and Danica Kragic (2015). "Task-Based Robot Grasp Planning Using Probabilistic Inference". In: *IEEE Transactions on Robotics* 31.3, pp. 546–561 (cit. on pp. 31, 36).

Song, Fangjing, Zengzhi Zhao, Wei Ge, Weiwei Shang, and Shuang Cong (2018a). "Learning Optimal Grasping Posture of Multi-Fingered Dexterous Hands for Unknown Objects". In: *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. 2018 IEEE International Conference on Robotics and Biomimetics (ROBIO), pp. 2310–2315 (cit. on pp. 23, 26).

Song, Hyun Oh, Mario Fritz, Daniel Goehring, and Trevor Darrell (2016). "Learning to Detect Visual Grasp Affordance". In: *IEEE Transactions on Automation Science and Engineering* 13.2, pp. 798–809 (cit. on p. 67).

Song, Peng, Zhongqi Fu, and Ligang Liu (2018b). “Grasp Planning via Hand-Object Geometric Fitting”. In: *The Visual Computer* 34.2, pp. 257–270 (cit. on pp. 30, 31).

Spek, Andrew, Wai Ho Li, and Tom Drummond (2017). “A Fast Method For Computing Principal Curvatures From Range Images”. In: *arXiv* (cit. on pp. 69, 70).

Staroverov, Aleksei, Kirill Muravyev, Konstantin Yakovlev, and Aleksandr I. Panov (2023). “Skill Fusion in Hybrid Robotic Framework for Visual Object Goal Navigation”. In: *Robotics* 12.4, p. 104 (cit. on pp. 44, 49, 54, 104).

Stoffregen, Thomas A. (2003). “Affordances as Properties of the Animal-Environment System”. In: *Ecological Psychology* 15.2, pp. 115–134 (cit. on p. 133).

Stutz, David and Andreas Geiger (2018). “Learning 3D Shape Completion from Laser Scan Data with Weak Supervision”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, pp. 1955–1964 (cit. on p. 136).

Su, Entong, Chengzhe Jia, Yuzhe Qin, Wenxuan Zhou, Annabella Macaluso, Binghao Huang, and Xiaolong Wang (2024). *Sim2Real Manipulation on Unknown Objects with Tactile-based Reinforcement Learning*. URL: `http://arxiv.org/abs/2403.12170` (visited on 04/02/2024). Pre-published (cit. on pp. 23, 27).

Suarez, Alejandro, Rainer Kartmann, Daniel Leidner, Luca Rossini, Johann Huber, Carlos Azevedo, Marko Bjelonic, Antonio Gonzalez-Morgado, Christian Dreher, Peter Schmaus, Arturo Laurenzi, François Hélénon, Rodrigo Serra, Jean-Baptiste Mouret, Lorenz Wellhausen, Vicente Perez-Sanchez, Jianfeng Gao, Adrian Simon Bauer, Alessio De Luca, Mouad Abrini, Rui Bettencourt, Olivier Rochel, Joonho Lee, Pablo Viana, Christoph Pohl, Nesrine Batti, Diego Vedelago, Vamsi Krishna Guda, Carlos Alvarez, Fabian Reister, Werner Friedl, Corrado Burchielli, Aline Baudry, Thomas Gumpert, Luca Muratore, Philippe Gauthier, Franziska Krebs, Sebastian Jung, Lorenzo Baccelliere, Hippolyte Watrelot, Andre Meixner, Anne Köpken, Mohamed Chetouani, Florian Lay, Felix Hundhausen, Anne Reichert, Noémie Jaquier, Florian Schmidt, Marco Sewtz, Freek Stulp, Lioba Suchenwirth, Rudolph Triebl, Xuwei Wu, Begoña Arrue, Rebecca Schedl-Warpup, Marco Hutter, Serena Ivaldi, Pedro U Lima, Stéphane Doncieux, Nikos Tsagarakis, Tamim Asfour, and Alin Albu-Schäffer (2024). “Door-to-Door Parcel Delivery from Supply Point to Users Home with Heterogeneous Robot Team: euROBIN First Year Robotics Hackathon”. In: *IEEE Robotics & Automation Magazine*, accepted for publication (cit. on p. 10).

Sun, Lingfeng, Devesh K. Jha, Chiori Hori, Siddarth Jain, Radu Corcodel, Xinghao Zhu, Masayoshi Tomizuka, and Diego Romeres (2023). "Interactive Planning Using Large Language Models for Partially Observable Robotics Tasks". Version 1. In: (cit. on p. 42).

Sundermeyer, Martin, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox (2021). "Contact-GraspNet: Efficient 6-DoF Grasp Generation in Cluttered Scenes". In: *2021 IEEE International Conference on Robotics and Automation (ICRA)*. 2021 IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE, pp. 13438–13444 (cit. on pp. 23, 25, 29, 63).

Suzuki, Yosuke, Ryoya Yoshida, Tokuo Tsuji, Toshihiro Nishimura, and Tetsuyou Watanabe (2022). "Grasping Strategy for Unknown Objects Based on Real-Time Grasp-Stability Evaluation Using Proximity Sensing". In: *IEEE Robotics and Automation Letters* 7.4, pp. 8643–8650 (cit. on pp. 23, 24, 29).

Tang, Chao, Dehao Huang, Wenlong Dong, Ruinian Xu, and Hong Zhang (2024a). *FoundationGrasp: Generalizable Task-Oriented Grasping with Foundation Models*. URL: http://arxiv.org/abs/2404.10399 (visited on 04/22/2024). Pre-published (cit. on pp. 19, 21, 29).

Tang, Wei, Siang Chen, Pengwei Xie, Dingchang Hu, Wenming Yang, and Guijin Wang (2024b). *Rethinking 6-Dof Grasp Detection: A Flexible Framework for High-Quality Grasping*. URL: http://arxiv.org/abs/2403.15054 (visited on 04/02/2024). Pre-published (cit. on pp. 23, 27).

Tekden, Ahmet, Marc Peter Deisenroth, and Yasemin Bekiroglu (2023). "Grasp Transfer Based on Self-Aligning Implicit Representations of Local Surfaces". In: *IEEE Robotics and Automation Letters* 8.10, pp. 6315–6322 (cit. on pp. 19, 20).

Ten Pas, Andreas, Marcus Gualtieri, Kate Saenko, and Robert Platt (2017). *Grasp Pose Detection in Point Clouds*. URL: http://arxiv.org/abs/1706.09911 (visited on 02/16/2024). Pre-published (cit. on pp. 23, 27, 67).

Thrun, Sebastian, Wolfram Burgard, and Dieter Fox (2006). *Probabilistic Robotics*. Nachdruck. Intelligent Robotics and Autonomous Agents. Cambridge, Mass. London: MIT Press. 647 pp. (cit. on p. 81).

Tsagkas, Nikolaos, Jack Rome, Subramanian Ramamoorthy, Oisin Mac Aodha, and Chris Xiaoxuan Lu (2024). *Click to Grasp: Zero-Shot Precise Manipulation via Visual Diffusion Descriptors*. URL: http://arxiv.org/abs/2403.14526 (visited on 03/25/2024). Pre-published (cit. on pp. 19, 20).

Tuomi, Aarni, Iis P. Tussyadiah, and Jason Stienmetz (2021). "Applications and Implications of Service Robots in Hospitality". In: *Cornell Hospitality Quarterly* 62.2, pp. 232–247 (cit. on p. 1).

Turvey, M.T. (1992). “Affordances and Prospective Control: An Outline of the Ontology”. In: *Ecological Psychology* 4.3, pp. 173–187 (cit. on p. 133).

Vahrenkamp, N., M. Wächter, M. Kröhnert, K. Welke, and T. Asfour (2015). “The Robot Software Framework ArmarX”. In: *Information Technology* 57.2, pp. 99–111 (cit. on pp. 39, 104, 165).

Vahrenkamp, Nikolaus, Tamim Asfour, Giorgio Metta, Giulio Sandini, and Rüdiger Dillmann (2012). “Manipulability Analysis”. In: *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*. Osaka, Japan, pp. 568–573 (cit. on p. 95).

Vahrenkamp, Nikolaus, Leonard Westkamp, Natsuki Yamanobe, Eren E. Aksoy, and Tamim Asfour (2016). “Part-Based Grasp Planning for Familiar Objects”. In: *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*. 2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids). Cancun, Mexico: IEEE, pp. 919–925 (cit. on pp. 19, 21).

Valmeekam, Karthik, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati (2024). “On the Planning Abilities of Large Language Models: A Critical Investigation”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., pp. 75993–76005 (cit. on p. 114).

Valmeekam, Karthik, Alberto Olmo, Sarath Sreedharan, and Subbarao Kambhampati (2022). “Large Language Models Still Can’t Plan (A Benchmark for LLMs on Planning and Reasoning about Change)”. In: *NeurIPS 2022 Foundation Models for Decision Making Workshop* (cit. on pp. 51, 114).

Van der Loos, H.F. Machiel, David J. Reinkensmeyer, and Eugenio Guglielmelli (2016). “Rehabilitation and Health Care Robotics”. In: *Springer Handbook of Robotics*. Ed. by Bruno Siciliano and Oussama Khatib. Cham: Springer International Publishing, pp. 1685–1728 (cit. on pp. 3, 4).

Van Rijsbergen, C.J. (1977). “A THEORETICAL BASIS FOR THE USE OF CO-OCCURRENCE DATA IN INFORMATION RETRIEVAL”. In: *Journal of Documentation* 33.2, pp. 106–119 (cit. on p. 141).

Varadarajan, Karthik Mahesh and Markus Vincze (2011). “Affordance Based Part Recognition for Grasping and Manipulation”. In: *ICRA Workshop on Autonomous Grasping* (cit. on pp. 116, 134, 162).

Varadarajan, Karthik Mahesh and Markus Vincze (2012). “AfRob: The Affordance Network Ontology for Robots”. In: *IEEE International Conference on Intelligent Robots and Systems*. IEEE, pp. 1343–1350 (cit. on p. 142).

Varadarajan, Karthik Mahesh and Markus Vincze (2013). “AfNet: The Affordance Network”. In: *Computer Vision – ACCV 2012*. Ed. by Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu. Berlin, Heidelberg: Springer, pp. 512–523 (cit. on pp. 71, 134, 142, 163).

Varadarajan, KM (2011). “K-TR: Karmic Tabula Rasa–A Theory of Visual Perception”. In: *Conf. Intl Soc. Psychophysics.* Vol. 10, pp. 512–523 (cit. on p. 163).

Verma, Parikshit, Mohammed Diab, and Jan Rosell (2021). “Automatic Generation of Behavior Trees for the Execution of Robotic Manipulation Tasks”. In: *2021 26th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, pp. 1–4 (cit. on pp. 44, 46).

Vernon, David and Markus Vincze (2017). “Industrial Priorities for Cognitive Robotics”. In: *CEUR Workshop Proceedings.* Vol. 1855, pp. 6–9 (cit. on pp. 13, 101).

Vincze, Markus, Timothy Patten, Kiru Park, and Dominik Bauer (2020). “Learn, Detect, and Grasp Objects in Real-World Settings”. In: *e & i Elektrotechnik und Informationstechnik* 137.6, pp. 324–330 (cit. on p. 9).

Vollhardt, Ugo, Maria Makarov, Alex Caldas, Mathieu Grossard, and Pedro Rodriguez-Ayerbe (2019). “Energy-Based Stability Analysis for Grasp Selection with Compliant Multi-Fingered Hands”. In: *2019 18th European Control Conference (ECC).* 2019 18th European Control Conference (ECC). Naples, Italy: IEEE, pp. 1592–1597 (cit. on pp. 31, 32).

Wächter, Mirko, Simon Ottenhaus, Manfred Kröhnert, Nikolaus Vahrenkamp, and Tamim Asfour (2016). “The ArmarX Statechart Concept: Graphical Programing of Robot Behavior”. In: *Frontiers Robotics AI* 3 (JUN) (cit. on p. 41).

Wakabayashi, Shumpei, Shingo Kitagawa, Kento Kawaharazuka, Takayuki Murooka, Kei Okada, and Masayuki Inaba (2022). “Grasp Pose Selection Under Region Constraints for Dirty Dish Grasps Based on Inference of Grasp Success Probability through Self-Supervised Learning”. In: *2022 International Conference on Robotics and Automation (ICRA).* 2022 IEEE International Conference on Robotics and Automation (ICRA). Philadelphia, PA, USA: IEEE, pp. 8312–8318 (cit. on pp. 31, 36).

Wake, Naoki, Atsushi Kanehira, Kazuhiro Sasabuchi, Jun Takamatsu, and Katsushi Ikeuchi (2023). “ChatGPT Empowered Long-Step Robot Control in Various Environments: A Case Application”. In: *IEEE Access* 11, pp. 95060–95078 (cit. on pp. 52, 53, 55).

Wan, E.A. and Rudolph Van Der Merwe (2000). “The Unscented Kalman Filter for Nonlinear Estimation”. In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373).* IEEE, pp. 153–158 (cit. on pp. 80, 160).

Wang, Bin, Wenzhong Shi, and Zelang Miao (2015). "Confidence Analysis of Standard Deviational Ellipse and Its Extension into Higher Dimensional Euclidean Space". In: *PLOS ONE* 10.3. Ed. by Duccio Rocchini, e0118537 (cit. on p. 84).

Wang, Huaxiaoyue, Kushal Kedia, Juntao Ren, Rahma Abdullah, Atiksh Bhardwaj, Angela Chao, Kelly Y. Chen, Nathaniel Chin, Prithwish Dan, Xinyi Fan, Gonzalo Gonzalez-Pumariega, Aditya Kompella, Maximus Adrian Pace, Yash Sharma, Xiangwan Sun, Neha Sunkara, and Sanjiban Choudhury (2024). *MOSAIC: A Modular System for Assistive and Interactive Cooking.* URL: `http://arxiv.org/abs/2402.18796` (visited on 04/02/2024). Pre-published (cit. on pp. 44, 48).

Wang, Qiang, Francisco Roldan Sanchez, Robert McCarthy, David Cordova Bulens, Kevin McGuinness, Noel O'Connor, Manuel Wüthrich, Felix Widmaier, Stefan Bauer, and Stephen J. Redmond (2023a). "Dexterous Robotic Manipulation Using Deep Reinforcement Learning and Knowledge Transfer for Complex Sparse Reward-based Tasks". In: *Expert Systems* 40.6, e13205 (cit. on pp. 44, 45, 54).

Wang, Zihao, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang (2023b). "Describe, Explain, Plan and Select: Interactive Planning with Large Language Models Enables Open-World Multi-Task Agents" (cit. on p. 53).

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou (2023). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models" (cit. on p. 118).

Wei, Wei, Daheng Li, Peng Wang, Yiming Li, Wanyi Li, Yongkang Luo, and Jun Zhong (2022). *DVGG: Deep Variational Grasp Generation for Dextrous Manipulation.* URL: `http://arxiv.org/abs/2211.11154` (visited on 02/07/2024). Pre-published (cit. on pp. 23, 27).

Wen, Bowen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal (2022). *CaTGrasp: Learning Category-Level Task-Relevant Grasping in Clutter from Simulation.* URL: `http://arxiv.org/abs/2109.09163` (visited on 04/04/2024). Pre-published (cit. on pp. 19, 21, 29).

Wirtz, Jochen, Paul G. Patterson, Werner H. Kunz, Thorsten Gruber, Vinh Nhat Lu, Stefanie Paluch, and Antje Martins (2018). "Brave New World: Service Robots in the Frontline". In: *Journal of Service Management* 29.5, pp. 907–931 (cit. on pp. 3, 6).

Wu, Albert, Michelle Guo, and Karen Liu (2023a). "Learning Diverse and Physically Feasible Dexterous Grasps with Generative Model and Bilevel Optimization". In: *Proceedings of The 6th Conference on Robot Learning.* Conference on Robot Learning. PMLR, pp. 1938–1948 (cit. on pp. 23, 27, 29).

Wu, Bohan, Iretiayo Akinola, and Peter K. Allen (2019). "Pixel-Attentive Policy Gradient for Multi-Fingered Grasping in Cluttered Scenes". In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China: IEEE, pp. 1789–1796 (cit. on pp. 23, 26).

Wu, Bohan, Iretiayo Akinola, Abhi Gupta, Feng Xu, Jacob Varley, David Watkins-Valls, and Peter K. Allen (2020). "Generative Attention Learning: A "GenerAL" Framework for High-Performance Multi-Fingered Grasping in Clutter". In: *Autonomous Robots* 44.6, pp. 971–990 (cit. on pp. 44, 49).

Wu, Jimmy, Rika Antonova, Adam Kan, Marion Lepert, Andy Zeng, Shuran Song, Jeannette Bohg, Szymon Rusinkiewicz, and Thomas Funkhouser (2023b). "TidyBot: Personalized Robot Assistance with Large Language Models" (cit. on p. 53).

Wu, Rina, Tianqiang Zhu, Wanli Peng, Jinglue Hang, and Yi Sun (2023c). "Functional Grasp Transfer Across a Category of Objects From Only One Labeled Instance". In: *IEEE Robotics and Automation Letters* 8.5, pp. 2748–2755 (cit. on pp. 19, 21, 29).

Wu, Zhenyu, Ziwei Wang, Xiuwei Xu, Jiwen Lu, and Haibin Yan (2023d). "Embodied Task Planning with Large Language Models" (cit. on pp. 52, 53).

Xiao, Li and V. Kumar (2021). "Robotics for Customer Service: A Useful Complement or an Ultimate Substitute?" In: *Journal of Service Research* 24.1, pp. 9–29 (cit. on p. 1).

Xie, Yaqi, Chen Yu, Tongyao Zhu, Jinbin Bai, Ze Gong, and Harold Soh (2023). "Translating Natural Language to Planning Goals with Large-Language Models" (cit. on pp. 53, 54).

Xu, Ruinian, Fu-Jen Chu, and Patricio A Vela (2022). "GKNet: Grasp Keypoint Network for Grasp Candidates Detection". In: *The International Journal of Robotics Research* 41.4, pp. 361–389 (cit. on pp. 23, 26).

Xu, Yucheng, Mohammadreza Kasaei, Hamidreza Kasaei, and Zhibin Li (2023). "Instance-Wise Grasp Synthesis for Robotic Grasping". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1744–1750 (cit. on pp. 23, 26).

Yamanobe, Natsuki, Weiwei Wan, Ixchel G. Ramirez-Alpizar, Damien Petit, Tokuo Tsuji, Shuichi Akizuki, Manabu Hashimoto, Kazuyuki Nagata, and Kensuke Harada (2017). "A Brief Review of Affordance in Robotic Manipulation Research". In: *Advanced Robotics* 31.19-20, pp. 1086–1101 (cit. on pp. 67, 134).

Yang, Shuo and Qi Zhang (2023). "Towards Efficient Robotic Software Development by Reusing Behavior Tree Structures for Task Planning Paradigms". In: *Complex System Modeling and Simulation* 3.4, pp. 357–380 (cit. on pp. 44, 46).

Yao, Qingfeng, Jilong Wan, Shuyu Yang, Cong Wang, Linghan Meng, Qifeng Zhang, and Donglin Wang (2022). *A Transferable Legged Mobile Manipulation Framework Based on Disturbance Predictive Control.* URL: `http://arxiv.org/abs/2203.03391` (visited on 02/06/2024). Pre-published (cit. on pp. 44, 48, 54).

Yenamandra, Sriram, Arun Ramachandran, Karmesh Yadav, Austin Wang, Mukul Khanna, Theophile Gervet, Tsung-Yen Yang, Vidhi Jain, Alexander William Clegg, John Turner, Zsolt Kira, Manolis Savva, Angel Chang, Devendra Singh Chaplot, Dhruv Batra, Roozbeh Mottaghi, Yonatan Bisk, and Chris Paxton (2024). *HomeRobot: Open-Vocabulary Mobile Manipulation.* URL: `http://arxiv.org/abs/2306.11565` (visited on 02/28/2024). Pre-published (cit. on pp. 44, 47).

Yi, Jae-Bong, Taewoong Kang, Dongwoon Song, and Seung-Joon Yi (2020). "Unified Software Platform for Intelligent Home Service Robots". In: *Applied Sciences* 10.17, p. 5874 (cit. on pp. 44, 49).

Ying, Robert, Jonathan Weisz, and Peter K. Allen (2018). "Grasping with Your Brain: A Brain-Computer Interface for Fast Grasp Selection". In: *Robotics Research: Volume 1.* Ed. by Antonio Bicchi and Wolfram Burgard. Cham: Springer International Publishing, pp. 325–340 (cit. on pp. 31, 32).

Yokoyama, Naoki, Alex Clegg, Joanne Truong, Eric Undersander, Tsung-Yen Yang, Sergio Arnaud, Sehoon Ha, Dhruv Batra, and Akshara Rai (2023). *ASC: Adaptive Skill Coordination for Robotic Mobile Manipulation.* URL: `http://arxiv.org/abs/2304.00410` (visited on 02/06/2024). Pre-published (cit. on pp. 44, 46).

Younes, Abdelrahman and Tamim Asfour (2024). "KITchen: A Real-World Benchmark and Dataset for 6D Object Pose Estimation in Kitchen Environments" (cit. on p. 121).

Young, James E., Richard Hawkins, Ehud Sharlin, and Takeo Igarashi (2009). "Toward Acceptable Domestic Robots: Applying Insights from Social Psychology". In: *International Journal of Social Robotics* 1.1, pp. 95–108 (cit. on pp. 2–4).

Zhang, Hanbo, Jian Tang, Shiguang Sun, and Xuguang Lan (2022). *Robotic Grasping from Classical to Modern: A Survey.* URL: `http://arxiv.org/abs/2202.03631` (visited on 04/17/2024). Pre-published (cit. on p. 3).

Zhang, Hui, Jef Peeters, Eric Demeester, and Karel Kellens (2021). "A CNN-Based Grasp Planning Method for Random Picking of Unknown Objects with a

Vacuum Gripper”. In: *Journal of Intelligent & Robotic Systems* 103.4, p. 64 (cit. on pp. 23, 25).

Zhao, Wayne Xin, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen (2023). *A Survey of Large Language Models*. URL: `http://arxiv.org/abs/2303.18223` (visited on 08/30/2024). Pre-published (cit. on p. 51).

Zhao, Xi, Ruizhen Hu, Paul Guerrero, Niloy Mitra, and Taku Komura (2016). “Relationship Templates for Creating Scene Variations”. In: *ACM Transactions on Graphics* 35.6, pp. 1–13 (cit. on pp. 61, 161).

Zhao, Zirui, Wee Sun Lee, and David Hsu (2024). “Large Language Models as Commonsense Knowledge for Large-Scale Task Planning”. In: *Advances in Neural Information Processing Systems* 36 (cit. on pp. 52, 53, 55).

Zhou, You, Jianfeng Gao, and Tamim Asfour (2019). “Learning Via-Point Movement Primitives with Inter- and Extrapolation Capabilities”. In: *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 4301–4308 (cit. on pp. 63, 162).

Zhou, Zhehua, Jiayang Song, Kunpeng Yao, Zhan Shu, and Lei Ma (2023). “ISR-LLM: Iterative Self-Refined Large Language Model for Long-Horizon Sequential Task Planning” (cit. on pp. 52, 53).