Investigating Scene Understanding for Robotic Grasping: From Pose Estimation to Explainable AI

by

Emily Zhixuan Zeng

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2023

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

The following papers are used in this thesis. I was co-author with major contributions to the design, analysis, writing, and editing.

Yuhao Chen, E Zhixuan Zeng, Maximilian Gilles, and Alexander Wong. Metagraspnet_v0: A large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis. In *Journal of Computational Vision and Imaging Systems*, 2021. This paper is incorporated in Chapter 3. I designed the automatic labelling of scene layout annotations along with major contributions to the design of the difficulty labels and layout-weighted evaluation metric, and was responsible for a significant amount of writing and editing, especially related to those sections.

E Zhixuan Zeng, Yuhao Chen, and Alexander Wong. Investigating use of keypoints for object pose recognition. In *Journal of Computational Vision and Imaging Systems*, 2022. This paper is incorporated in Chapter 4. I designed and implemented all experiments described in the paper, and was responsible for the majority of writing and editing.

E Zhixuan Zeng, Yuhao Chen, and Alexander Wong. Shapeshift: Superquadric-based object pose estimation for robotic grasping. In *WICV workshop*, 2023. This paper is incorporated in Chapter 5. I designed and implemented all experiments described in the paper, and was responsible for the majority of writing and editing.

Yuhao Chen, Hayden Gunraj, E Zhixuan Zeng, Robbie Meyer, Maximilian Gilles, and Alexander Wong. Mmrnet: Improving reliability for multimodal object detection and segmentation for bin picking via multimodal redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 68–77, 2023. This paper is incorporated in Chapter 6. My contribution is limited to designing the Multimodal Consistency score used for uncertainty estimation, as well as experimental results related to that uncertainty score. I was also responsible for most of the writing and editing for sections related to that uncertainty score.

E Zhixuan Zeng, Hayden Gunraj, Sheldon Fernandez, and Alexander Wong. Explaining explainability: Towards deeper actionable insights into deep learning through second-order explainability. In *XAI4CV workshop*, 2023. This paper is incorporated in Chapter 7. My main contribution is the analysis of the second order explainability results, proving the improvement in segmentation performance based on explainability results, as well as most of the writing and editing.

Abstract

In the rapidly evolving field of robotics, the ability to accurately grasp and manipulate objects—known as robotic grasping—is a cornerstone of autonomous operation. This capability is pivotal across a multitude of applications, from industrial manufacturing automation to supply chain management, and is a key determinant of a robot's ability to interact effectively with its environment. Central to this capability is the concept of scene understanding, a complex task that involves interpreting the robot's environment to facilitate decision-making and action planning. This thesis presents a comprehensive exploration of scene understanding for robotic grasping, with a particular emphasis on pose estimation, a critical aspect of scene understanding.

Pose estimation, the process of determining the position and orientation of objects within the robot's environment, is a crucial component of robotic grasping. It provides the robot with the necessary spatial information about the objects in the scene, enabling it to plan and execute grasping actions effectively. However, many current pose estimation methods provide relative pose compared to a 3D model, which lacks descriptiveness without referencing the 3D model. This thesis explores the use of keypoints and superquadrics as more general and descriptive representations of an object's pose. These novel approaches address the limitations of traditional methods and significantly enhance the generalizability and descriptiveness of pose estimation, thereby improving the overall effectiveness of robotic grasping.

In addition to pose estimation, this thesis briefly touches upon the importance of uncertainty estimation and explainable AI in the context of robotic grasping. It introduces the concept of multimodal consistency for uncertainty estimation, providing a reliable measure of uncertainty that can enhance decision-making in human-in-the-loop situations. Furthermore, it explores the realm of explainable AI, presenting a method for gaining deeper insights into deep learning models, thereby enhancing their transparency and interpretability.

In summary, this thesis presents a comprehensive approach to scene understanding for robotic grasping, with a particular emphasis on pose estimation. It addresses key challenges and advances the state of the art in this critical area of robotics research. The research is structured around five published papers, each contributing to a unique aspect of the overall study.

Acknowledgements

I would like to thank all the people who made this thesis possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This introductory chapter provides a high-level overview of the entire thesis. First, in Section 1.1, an overview of robotic grasping and scene understanding will be given. Then, in Section 1.2, the problem of pose estimation and the need for better scene understanding will be discussed. Finally, in Section 1.3, the main scientific contributions and outline of the thesis will be presented.

1.1 Motivation

The field of robotics has seen significant advancements in recent years, with robots becoming increasingly prevalent in a variety of sectors including manufacturing, healthcare, and logistics. A key capability that underpins many of these applications is robotic grasping - the ability of a robot to accurately identify, grasp, and manipulate objects in its environment. This is a fundamental requirement for autonomous operation and has the potential to revolutionize many industries by automating tasks that were previously labor-intensive or challenging for humans. However, despite the progress made, robotic grasping remains a complex problem due to the diversity and unpredictability of real-world environments.

A key aspect of robotic grasping is scene understanding, which involves the interpretation of the robot's environment to facilitate decision-making and action planning. Scene understanding encompasses several tasks, including object detection, segmentation, and pose estimation. Among these, pose estimation plays a crucial role as it provides the robot with the necessary spatial information about the objects in the scene, enabling it to plan and execute grasping actions effectively.

Scene understanding is a critical aspect of robotic grasping, providing the robot with the necessary information to interact effectively with its environment. This understanding is crucial at all stages of the grasping process: before, during, and after the grasp.

Before the Grasp: Prior to executing a grasp, the robot must identify and localize objects, especially in applications like bin picking where items are randomly oriented and placed. Additionally, understanding the scene can help the robot determine the optimal order for grasping objects, a crucial factor in cluttered environments.

During the Grasp: As the robot executes the grasp, it must navigate safely within its environment. This requires the ability to identify potential obstacles and plan a path that avoids them, a task made more challenging by the close proximity and random arrangement of objects in applications like bin picking.

After the Grasp: Once an object is grasped, the robot often needs to manipulate it, such as placing it in a specific location or orienting it for a task. These manipulation tasks require knowledge of not only the grasped object's pose but also the spatial layout and properties of other objects in the scene. This understanding can help the robot determine where to place the grasped object without disturbing other objects or causing instability.

In essence, scene understanding is a cornerstone of effective robotic grasping, influencing every stage of the process. However, achieving this understanding is a complex task, requiring the integration of various techniques and methodologies. This thesis aims to address this challenge, exploring innovative approaches to enhance scene understanding and, consequently, improve the efficiency and reliability of robotic grasping. The following sections will delve into the specific problems this thesis aim to solve and the contributions made towards this goal.

1.2 Overview of Problem

The task of robotic grasping is inherently complex, involving a multitude of sub-tasks that each present their own unique challenges. One of the most critical of these sub-tasks is pose estimation, the process of determining the position and orientation of objects within the robot's environment. Accurate pose estimation is crucial for successful robotic grasping, as it provides the spatial information necessary for the robot to plan and execute its grasping actions.

However, many current pose estimation methods provide relative pose compared to a 3D model. This approach, while useful in certain contexts, lacks descriptiveness without

referencing the 3D model, limiting its applicability and generalizability. This limitation underscores the need for more general descriptors.

Keypoints, distinctive features of an object that can be reliably detected and tracked across different views, offer the advantage of defining important features and locations independent of having to reference a 3D model, and can generalize to an entire class of objects. However, keypoints also present their own set of challenges. They are difficult to consistently define for arbitrary objects, especially those with uniform or repetitive patterns. Issues related to symmetry and proximity of objects can further complicate the pose estimation process, leading to confusion and inaccuracies.

Beyond pose estimation, uncertainty estimation is another challenge in the field of robotic grasping. In real-world applications, it is often necessary to have a human-in-the-loop for situations where the model is uncertain. However, current methods for estimating uncertainty in robotic grasping are often inadequate, lacking the ability to provide reliable and interpretable measures of uncertainty.

Understanding the model and the dataset is another critical aspect of robotic grasping. With the increasing complexity of models used in robotic grasping, there is a growing need for methods that can provide deeper insights into the model's behavior and the characteristics of the dataset. This is particularly important in human-in-the-loop situations, where the ability to provide actionable insights can significantly improve the effectiveness of the human-robot collaboration.

In summary, the central issue this thesis aims to address is the enhancement of scene understanding for robotic grasping. This involves addressing challenges in pose estimation, uncertainty estimation, and understanding of the model and the dataset. Each of these aspects presents its own unique challenges. Addressing these challenges is not only crucial for advancing the field of robotic grasping, but also sets the stage for the contributions of this thesis, which are aimed at improving the scene understanding model, an essential component of the overall robotics grasping pipeline. These contributions will be outlined in the following section.

1.3 Thesis Contributions & Outline

This thesis makes several contributions to the field of robotic grasping and scene understanding, each addressing a unique aspect of the problem statement outlined in Section 1.2. These contributions are encapsulated in the five chapters that follow this introduction,

with each chapter corresponding to a published paper that delves into a specific aspect of the overall study.

Contribution 1: MetaGraspNet v0 (Chapter 3)

The first contribution of this thesis is the creation of MetaGraspNet v0, a large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis. This dataset provides a comprehensive resource for training and evaluating models for robotic grasping, offering a diverse range of objects and grasp scenarios. The creation of MetaGraspNet v0 addresses the need for high-quality, diverse datasets in the field of robotic grasping, providing a foundation for further research and development. This dataset also provides scene layout labels, emphasizing the importance of scene understanding in a robotics grasping scenario. Subsequent chapters will use the dataset introduced in this chapter for deeper exploration of various challenges.

Contribution 2: Keypoint-based Pose Estimation (Chapter 4)

Chapter 4 investigates the use of keypoints for object pose recognition. This approach offers a more general and descriptive representation of an object's pose compared to traditional methods that provide relative pose compared to a 3D model. The use of semantically important keypoints for object pose estimation presents a way to directly gain an understanding of where important features of an object are located without having to reference an exact 3D model, enabling more effective robotic manipulation and understanding of object affordances. However, the implementation of keypoints also presents its own set of challenges, which are discussed in detail in this chapter. The exploration of keypoint-based pose estimation contributes to the ongoing efforts to improve the accuracy and generalizability of pose estimation methods.

Contribution 3: Superquadric-based Pose Estimation (Chapter 5)

To address the limitations of keypoint-based methods, Chapter 5 introduces ShapeShift, a superquadric-based object pose estimation method for robotic grasping. Superquadrics provide a more flexible and robust representation of object shape, enhancing the ability of the robot to understand and interact with its environment. Through predicting the poses of shape primitives used to represent object parts, the method is able to directly predict useful geometric properties of the scene for use in robotic grasping.

Contribution 4: Multimodal Consistency for Uncertainty Estimation (Chapter 6)

To address the need for reliable and interpretable measures of uncertainty in robotic grasping, thus enhancing the robot's ability to make informed decisions and facilitating effective human-in-the-loop collaboration, Chapter 6 presents MMRNet. Through a gated

fusion module and a dynamic ensemble learning strategy, MMRNet trains for multimodal redundancy, enabling reliable prediction even when only one modality of input is present. Through looking at the consistency in model behaviour with different input modalities enabled, the multimodal consistency (MC) introduced here can be used as an estimation for model uncertainty.

Contribution 5: Second-order Explainable AI (Chapter 7)

Finally, Chapter 7 delves into the realm of explainable AI, presenting a method for gaining deeper actionable insights into deep learning through second-order explainability. This approach not only helps understand the model better but also aids in human-in-the-loop situations by providing actionable insights. The exploration of second-order explainable AI contributes to the ongoing efforts to make AI models more transparent and interpretable, enhancing their usability and trustworthiness in real-world applications such as robotics grasping.

In summary, this thesis presents a comprehensive approach to scene understanding for robotic grasping, addressing key challenges and advancing the state-of-the-art in this critical area of robotics research. The following chapters will delve into each of these contributions in detail, providing a thorough exploration of the methodologies, experiments, and results associated with each aspect of the study.

Chapter 2

Background

This chapter provides an overview of the foundational concepts and prior works that underpin the research presented in this thesis. We begin by discussing prior work for direct predictions of robotic grasping and the challenges associated with it. We then delve into the various methodologies and techniques that have been proposed to address these challenges, focusing on improvements for scene understanding including keypoints, and pose estimation. After discussing some disadvantages of those works, we then explore the use of geometric primitives and superquadrics representations in robotic grasping. The chapter concludes by looking at the need for uncertainty estimation and Explainable AI in a robotics grasping task.

2.1 Robotic grasping

Robotic grasping is a fundamental capability for robots, enabling them to interact with and manipulate objects in their environment. The end goal of robotic grasping is not just to pick up an object but to do so in a manner that is efficient, safe, and suitable for subsequent tasks. This section reviews the key advancements in the domain of grasp predictions.

2.1.1 Jaw Grippers

Jaw grippers are among the most common tools used in robotic grasping. They function by clamping down on an object, much like a human hand. Several works have focused

on improving the efficiency of jaw grippers. There exists a multitude of prior works for direct robotics grasp prediction using jaw grippers. Jiang et al. [6] introduced an efficient grasping approach from RGBD images, utilizing a new rectangle representation paradigm. This method has been used by many subsequent works, such as Kumra et al. [7], who applied deep convolutional neural networks for robotic grasp detection. The field has seen continuous advancements with contributions like GraspNet by Asif et al. [8], GraspNet-1Billion by Fang et al. [9], and Dex-Net 2.0 by Mahler et al. [10]. Yan et al. [11] explored learning 6-DOF grasping interaction via deep geometry-aware 3D representations, while Zhang et al. [12] focused on ROI-based robotic grasp detection for object overlapping scenes. More recent works include Contact-GraspNet by Sundermeyer et al. [13] and a real-time, generative grasp synthesis approach by Morrison et al. [14].

2.1.2 Suction Grasping

Suction grasping involves using a vacuum to pick up objects. This method is particularly useful for objects that are difficult to grasp with jaw grippers. Suction grasping has also been extensively studied, with works like SuctionNet-1Billion by Cao et al. [15] and a CNN-based grasp planning method by Zhang et al. [16] for random picking of unknown objects with a vacuum gripper.

2.1.3 Challenges and Advantages

Despite the progress, direct grasp prediction suffers from several issues. A separate model is often needed for each type of grasper, and the models generally predict many possible grasp points for each given object. This makes it difficult to integrate for later manipulation task reasoning, such as navigating around other objects or considering fragile areas of the object. However, direct grasp prediction does have some advantages, such as being object-agnostic and capable of picking up any general object.

2.1.4 Spatial relationship reasoning

Some researchers have attempted to integrate reasoning between object relationships, such as manipulation order. Works like REGRAD by Zhang et al. [17] and Graph-Based Visual Manipulation Relationship Reasoning Network by Zuo et al. [18] have explored this direction. However, these outputs usually only simple graphs about the occlusion

or superposition relationship and may be difficult to reason about more complex spatial relationships, such as path planning for navigating an object around another object.

2.2 Pose estimation

Pose estimation [19] offers a way to have more information about objects in a scene, including their orientation and spatial relation to each other. There has been extensive work in this area, ranging from early works using template-based and context descriptor-based approaches [20–23] to more modern techniques utilizing deep learning.

2.2.1 Datasets

Several datasets have been introduced to facilitate research in pose estimation. PASCAL3D+ [24] augments 12 rigid categories, including cars, planes, sofas, and so forth, of the PASCAL VOC 2021 [25] with 3D annotations. 3d meshes are labeled according to their semantic keypoint locations on an image. There are multiple meshes available for each category but none are guaranteed to be an exact match to the image.

In contrast, datasets used in the BOP challenges [26–28] are geared directly towards smaller objects that may be used for robotic grasping. These objects generally match directly with the 3D models used in labelling, and are labeled by the transformation necessary to move the provided 3d mesh or CAD model to the right location and orientation on the image rather than by any keypoints. BOP datasets contain both synthetic and real images. However, not all objects in the scene are guaranteed to be labeled with the corresponding pose. Recent iterations of the BOP challenge evaluates on the VIVO task (varying number of instances of a varying number of objects), while previous iterations evaluated on the SISO task (single instance of a single object). Some major datasets included within the challenges includes LineMOD [29], the updated linemod-occluded dataset [21], YCB-V [30, 31], T-LESSS [32], HOPE [33], and HomebrewedDB [34].

Knowing what datasets the methods are being evaluated on can often be informative about some of the problems it attempts to solve. For example, the T-LESS [32] dataset has a significant number symmetrical objects which may make poses ambiguous, and can have repeated instances of each class per image, unlike with LineMOD [29] and YCB-V [31].

2.2.2 Direct 6DOF prediction

Direct methods go directly from input image to pose estimation without any intermediate representation or keypoint matching. These methods have the advantage of generally being fully differentiable and also tend to deal better with occlusion out of the box.

PoseCNN [31] is a model introduced with the YCB-V dataset, and pose is predicted through localizing its center in the image, predicting its distance from the camera, and and a direct regression for the rotation quaternerion. SSD-6D [35] extends the SSD object detection architecture (as suggested in the name) to cover the 6D pose space, generating an additional score for possible viewpoints and rotation. EfficientPose [36] similarily expands upon the EfficientDet architecture [37]. Other techniques include using iterative refinement [38–40], implicit learning [41], segmentation mask [42], and multiple view points [43].

However, since the output of direct pose estimation methods are only the location and rotation matrix of the object, it only becomes useful for robotics grasping when a corresponding 3D model is referenced. This not only creates further computational requirement on the system, it also necessitates a close match with a 3D mesh or point-cloud, which makes the data collection process for new objects significantly more difficult than other tasks like object detection or segmentation.

2.2.3 Keypoints based methods

The use of semantic keypoints is common for the task of human pose estimation [44–47] but can be difficult to translate well into the object pose estimation domain. There are several works exploring the advantages and challenges in this approach, using a number of different keypoints definitions. These methods would use the point correspondence outputs as an intermediate representation to be taken as input for a secondary stage, either RANSAC-based [48] or another network prediction head [49–51] for the final pose prediction.

Box corner points representation

Some methods simply takes the 8 corners of the orientated 3D bounding boxes [52–55] or interpolated bounding box [7] as the keypoint labels. Even though such methods utilize keypoints as intermediate representations, they do not provide much further geometric information compared to direct prediction methods beyond a rough spatial occupancy.

Furthest point sampling

Other methods have looked at sampling a set number of points on the object surface which are furthest away from each other as a way to identify keypoints, known as furthest point sampling (FPS). This was first introduced in Pvnet [56] which argued that the floating points of the 3D bounding box corners would be more difficult to find due to being disconnected from the object itself. PVN3D followed by predicting 3D keypoints rather than 2d [57], ffb6d [58] uses SIFT-FPS to distinctive texture in 2D, to then resample using FPS.

Furthest point sampling has proven very popular in many subsequent works [49, 50, 59–63] due to being fully automatic and can be computed on any point cloud. However, this also means that each keypoint does not contain any information on the geometric or semantic properties of the surface region around it.

Dense correspondence maps

While the prior two keypoints representations can be thought of as sparse point correspondences, a more popular method uses dense correspondences between the image and its 3D model. One significant reason is due a higher sensitivity to occlusion in sparse keypoint representation methods [64]. Some methods handle this through predicting each pixel's relationship to its sparse keypoint locations [57, 60] while others focuses on augmentation techniques [62] or explicitly modeling self-occlusion [61].

Dense correspondence methods [51, 64–71] predicts a mapping of each pixel or point cloud vertex, making each individual correspondence prediction less important for the final pose calculation. These methods have proven very robust, with GDR-Net [51] in particular dominating the 2022 BOP challenge [28], as variations of the model won 8 out of 11 awards.

Semantically important keypoints

Other works have elected to use semantically important keypoints as an intermediate representation. Pavlakos *et al.* [72] uses the pre-defined points from the PASCAL3D+ dataset [24], while Kundu *et al.* [73] uses the same dataset but also samples a denser skeleton of keypoints between the originally structurally important points. Merrill *et al.* [74] used the YCB-V dataset [31] and annotated the 3D objects for structurally important segments.

Semantically important keypoints does not have any advantages in terms of accuracy in the pose estimation task compared to other representation methods. Dense keypoints

have shown to be more robust to occlusion compared to sparse keypoints methods [64]. In addition compared to other sparse methods, requiring each keypoint to be of semantic importance makes the definition of keypoint placement significantly more difficult to label and maintain consistency between different object classes.

However, such keypoint definition has shown some advantages in the downstream task of robotics manipulation. Specifically, semantically important keypoints enables task-specific grasps [75] as well as understanding object affordances during robotic manipulation [76].

In addition, although semantically keypoints struggle with inter-category consistency, it is able to maintain intra-category consistency as they're based off semantic details rather than a particular 3D model. As such, these methods are able to applied to category-level pose estimation rather than instance-level. Other methods targeting this specific problem will also be addressed in the following section.

2.2.4 Beyond instance-level pose estimation

Beyond instance-level pose estimation is the task of category-level pose estimation. Here, there is no one-to-one match between the input image and a reference 3D model, and the model must instead generalize across the intra-category variations. Sparse semantic keypoints can do this through maintaining a consistent intra-category keypoint definition, but other methods [77–86] were also introduced to specifically target this problem.

Pose can also be defined as a relative pose from another image or shape. Rather than trained for a specific finite set of 3D objects, these techniques look at an arbitrary object in a reference pose, and attempt to predict the transformation needed for another input pose. These works [87–92] are generally referred to as one-shot pose estimation methods.

However, due to large intra-category variation, it is very difficult to use category-level pose information for robotic grasping and manipulation, as there is little information on the physical structure of the object. One-shot methods still may require 3D models [87, 89], or, when photos or video based, would either generate a point cloud that needs to be referenced for grasping prediction [90, 91], or would only predict the pose, and there would be no structural information that can be passed on to the robotics grasping pipeline [88, 92].

2.3 Primitive shapes

Another way to represent 3D shapes while directly predicting geometric properties of the object is to represent an object as a composition of shape primitives. This concept of an

object as a composition of primitive parts has been thought about by researchers very early on [93]. This sort of 3D shape approximation offers advantages over 3D reconstruction, including directly describing the predicted geometric features and object or part segmentation without further processing, less computational intensity, and less noise, but will not exactly describe the 3D scene.

Such shape abstractions include using a collection of shape primitive categories such as cylinders, rings, cuboid, sticks, and so forth [94–97], parametric surface patches [98], generalized sweeps with cuboids or cylinders [7, 99, 100], cuboids [101–106], ellipsoidal structures [107], category-specific structures [108, 109], as well as superquadrics.

Superquadrics are 3D structures defined through a simple implicit function. Through changing the shape parameters of the function, it can become cubes, spheres, cylindars, and intermediate shapes in between. The concept was first introduced in 1981 by Barr [110], Solina and Bajcsy [111] and Jakli [112] offers a more detailed introduction to these parameterized structures.

Many works explore fitting superquadrics to objects. However, prior works have primarily focused on approximating point clouds or 3D meshes [113–116], or a single object without complex background or occlusions [117]. Other works have also learned to predict the pose and shape of objects exactly matching the primitive shape rather than real objects approximated using primitive shapes [118, 119]

In addition, superquadrics have shown to be useful in learning grasping behaviour [120–122]. However, these works did not use any deep learning methods to compute the superquadric pose directly. Makhal *et al.* [120] used Principal Component Analysis (PCA) on the depth data, Vezzani *et al.* [121] used a classifier to determine the discrete category of superquadric (parallelepipeds, cylinders, spheres) before continuing with classical optimization methods, and Wu *et al.* [122] utilized the non-deep-learning optimization described by Liu *et al.* [114].

One constraint of the superquadric representation is that, in its basic form, it is restricted to symmetrical and convex parts. This means that, to describe a single concave shape, most superquadric fitting techniques elect to fill the object with many thin blobs of superquadrics. A solution would be to add further complexity to the superquadric definition through the sue of deformable superquadrics.

2.4 Uncertainty Estimation

In situations where a human is actively interacting with a robot, communication regarding uncertainty is highly important [123, 124]. However, using only the softmax output of most computer vision models is insufficient for an estimation of the model's certainty, as that output can become overinflated even in very uncertain, out of domain situations [125].

There are two separate types of uncertainty that needs to be modeled: aleatoric and epistemic [126, 127]. Aleatoric uncertainty is also commonly known as data uncertainty, and is caused by inherent noise inside the data, such as poor sensor quality. This type of uncertainty cannot be reduced through more data or better training. Epistemic uncertainty is also called model uncertainty, and is produced due poor training data in that particular group. Aleatoric uncertainty is more important in large data situations, where epistemic uncertainty can be largely resolved through having enough data samples, while epistemic uncertainty is highly important for safety-critical applications, sparse training data, and when there may be expectations of distribution shift from the training data [128].

Popular ways of estimating uncertainty include bayesian inference [125, 129–134], ensemble approaches [135–140], and explicit prediction approaches [141, 141–145].

Bayesian inference approaches perform multiple forward passes of a single model, where each pass produces a different output. This can be achieved through dropout [125] or through a bayesian neural network [146, 147]. The variation within the samples of the output can be used to approximate the model uncertainty.

Ensemble approaches are similar except instead of using a single model, outputs of multiple deterministic models are sampled.

Finally, explicit prediction approaches seeks to directly predict an uncertainty score for each input as an additional output of the model. This is computationally less intensive than the other two methods, requiring only a single forward pass rather than sampling many outputs, but is highly sensitive to the training process.

2.5 Explainable AI

In order to better understand the model's prediction, we can also explore the realm of model explainability, or explainable AI.

A significant number of publications focuses on identifying salient areas of the input. These generally come in the form of heat-maps identifying important areas in the input

image [148–150]. These can be classified as local approaches and generally aid in understanding a model's reasoning behind a single image.

A different approach seeks to provide a more global understanding of the model through probing at the layer within the model itself, such as through generating or finding images which would most maximally activate particular layers [151–153].

Both approaches require a heavy amount of human interpretation of results and are highly qualitative. Local approaches only provide information on the single image input provided, and usually one must manually go through the heat map results to identify potential patterns of interest that may be present in the dataset. Global approaches promise data agnostic results that offer a more comprehensive view of how the model may react in different circumstances. However, the highly abstract visualizations common in these methods are not generally understandable by a human viewer without significant, case-by-case investigation.

Chapter 3

MetaGraspNet_v0: A Large-Scale Benchmark Dataset for Vision-driven Robotic Grasping via Physics-based Metaverse Synthesis

This chapter delves into the creation and application of MetaGraspNet v0, a large-scale benchmark dataset designed to enhance the capabilities of vision-driven robotic grasping. Developed through physics-based metaverse synthesis, MetaGraspNet v0 offers a diverse range of scenarios with 100,000 images and 25 different object types. The chapter will discuss the unique features of the dataset, such as object layout labels and layout-based difficulty levels. This chapter will explore how these features contribute to a more nuanced understanding of the scene, thereby facilitating more effective robotic grasping. The insights gained from this dataset will serve as a stepping stone towards addressing the broader challenges in the field of robotic grasping. The MetaGraspNet benchmark dataset will be available open-source on Kaggle [1], with the first phase consisting of detailed object detection, segmentation, layout annotations, and a script for layout-weighted performance metric (https://github.com/y2863/MetaGraspNet).

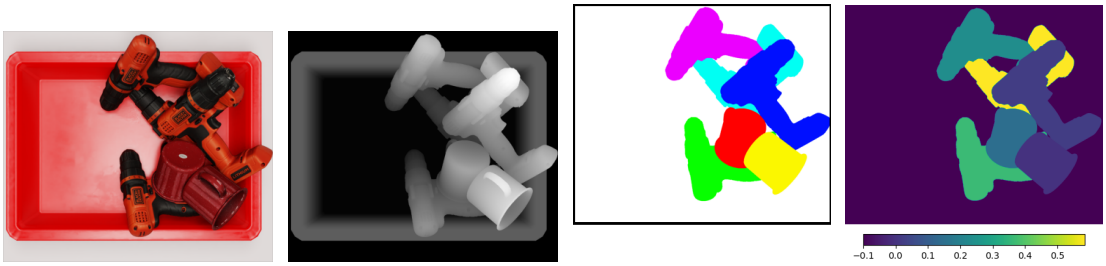
[1] https://www.kaggle.com/metagrasp/metagraspnetdifficulty1-easy, https://www.kaggle.com/metagrasp/metagraspnetdifficulty2-medium, https://www.kaggle.com/metagrasp/metagraspnetdifficulty3-hard1, https://www.kaggle.com/metagrasp/metagraspnetdifficulty4-hard2, https://www.kaggle.com/metagrasp/metagraspnetdifficulty5-very-hard

3.1 Introduction

There has been increasing interest in smart factories powered by robotics systems to tackle repetitive, laborious tasks such as handing and sorting objects or managing the material flow. One particular impactful yet challenging task in robotics-powered smart factory applications is robotic grasping which involves using robotic arms to grasp objects. A common robotic grasping scenario found in production system or warehouses includes moving a specific object from one bin to another (order picking). The seemingly simple task for human is quite complex for the robots to perform, requiring a variety of computer vision tasks such as object detection, segmentation, grasp prediction, pick planning, etc. While significant progress has been made in the leveraging of machine learning strategies for robotic grasping [9, 13, 14, 154], particularly with deep learning, a very big challenge in tackling this problem is the need for large-scale, high-quality RGBD datasets that cover a wide diversity of scenarios and permutations (e.g., different combination of objects, different ordering and orientation of objects, different ways of stacking objects.).

Many existing grasping datasets [8–12, 31] provide large-scale and high image quality data, but they have simple and similar environment settings, such as objects are placed in a common way without stacking. Another important attribute current large grasping datasets lack is environment layout on how the objects are positioned and stacked, especially in a cluttered environment. In scenarios where robotic arms are required to pick a specific item from a cluttered scene, picking an obstructed object before removing obstacles could lead to significant damage as the objects covering it are forced out of the way. With environment layout labels, pick planning can be trained more intelligently to avoid object damages [18, 155, 156]. However, there are only a few datasets [155–157] providing layout labels. Most of the datasets [155, 156] have limited training value as they lack in data size, depth information, as well as segmentation labels. In addition, objects should be picked sequentially. Occluded objects will be revealed once top objects are picked. Thus, not all the objects in a scene are equally important, and top objects are more important to evaluate. An object detection and segmentation metric weighted according the environment layout would better reflect the performance of a model for a robotic grasping task.

Motivated to tackle this big, diverse data problem, we are inspired by the recent rise in the concept of metaverse, which has greatly closed the gap between virtual worlds and the physical world. In particular, metaverses allow us to create digital twins of real-world manufacturing scenarios and enter these metaverses to virtually create different scenarios from which large volumes of high quality data can be generated for training models. this chapter presents MetaGraspNet: a large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis. This dataset contains 100,000 RGBD images,



(a) (b) (c) (d)

Figure 3.1: Example data in MetaGraspNet benchmark dataset: a) RGB image. b) Depth image. c) Instance annotations. d) Occlusion percentage annotations. Objects are marked with their occlusion percentage, while background is marked with a value of -0.1 .

11,000 scenes, and 25 classes of objects. The dataset is split into 5 difficulties to evaluate object detection and segmentation model performance in different grasping scenarios. In addition, this chapter propose a new layout-weighted performance metric alongside the MetaGraspNet benchmark dataset for evaluating object detection and segmentation performance in a manner that is more appropriate for robotic grasp applications compared to existing general-purpose performance metrics. The proposed MetaGraspNet benchmark dataset will be available in an open-source form on Kaggle [158], with the first phase consisting of detailed object detection, segmentation, layout annotations, and a script for layout-weighted performance metric (https://github.com/y2863/MetaGraspNet).

3.2 Dataset

3.2.1 Physics-based Data Synthesis in the Metaverse

Manually capturing object detection datasets in real-world robotic bin picking environments is intractable in most practical scenarios for a number of key reasons. First of all, this manual capturing process involves repeatedly setting up physical grasping environment, physically placing different objects into the physical environment, recording images with sensors, and removing objects. As such, this process is very time consuming, laborious, and unscalable in most practical scenarios as it requires the entire capture process to be manually repeated for each environment and scenario. Second, the manual process of placing different objects into different layouts by a human operator also means that the way the objects are arranged in three dimensional space often does not reflect how objects are

physically dropped together into a pile during the material handling processes in real-world warehouses or industry related scenarios. Third, manually labeling the sensor data is very time consuming, therefore static, and cannot keep up with the emerging demand for data needed for training deep neural networks. In [9], Fang *et al.* have come up with an intuitive way to overcome the enormous labeling effort for each viewpoint by mounting the camera to the robotic end effector and recording the relative movement between image frames, however their approach still needs precise initial manual annotations for each scene and is restricted to known objects and physical environments with a robotic manipulator. Image synthesis approaches that generated images based on randomizing object counts, poses, and positions can be used to cut down the data collection time significantly, but creates unrealistic or physically impossible layouts where an object can overlap with other objects in the same spatial location. As such, the effectiveness of training deep learning models using images generated in this fashion can be very limited for real-world deployment scenarios. Therefore, a way to generated large-scale benchmark datasets with highly diverse environments and layout permutations for vision-driven robotic grasping in a scalable yet realistic manner is highly desired.

Motivated by this, this chapter take inspiration from the recent rise of metaverses, which are highly immersive virtual environments that facilitates for significant interaction. The significant advancements in metaverses have significantly closed the gap between virtual worlds and the physical world, particularly in physics-based metaverse creation platforms such as the Nvidia Omniverse [159].

To create the proposed MetaGraspNet benchmark dataset, we leverage Nvidia Omniverse to create photorealistic, physics-driven digital twins of different real-world manufacturing scenarios. Within these digital twins, we then randomly drop objects under different environment configurations and let the objects interact through physics simulation to ensure the object layouts as captured within the MetaGraspNet dataset are realistic and physically accurate (as shown in Figure 3.2). Performing the data capturing process in such realistic manufacturing digital twin metaverses enables us to greatly scale in data quantity and diversity beyond what is possible with real-world manual capturing approaches in a very efficient and effective manner, but also allow us to obtain high quality, realistic data that mimics real-world physical scenarios well beyond what is possible with image synthesis approaches.

3.2.2 Object Layout Label

In addition to the typical semantic mask labels, we propose three more labels to characterize the object layouts.

Figure 3.2: Physics-based data synthesis: Items are dropped into the photorealistic, physics-driven metaverse digital twin of different manufacturing scenarios.

The first label, occlusion percentage describes the percentage area of each object being occluded. This score provides an indirect measure for each object on their relationship with other objects in the layout. This score is calculated as the percent of pixels removed in the instance segmentation mask compared to the total number of pixels of the object if all other objects in the image are removed.

The second label is a matrix storing the relation between each pair of objects, providing a comprehensive layout representation. To construct the relation matrix, we define three types of relationship for a pair of objects A and B. If A is occluding B, we define the relationship between (A, B) as positive, with a numerical value of 1. If A is occluded by B, we define the relationship between (A, B) as negative, with a numerical value of -1. If A and B have no direct relationship or $A = B$, we define the relationship between (A, B) as neutral, with a numerical value of 0. Based on these definitions, for a layout with N objects, we create a relation matrix with NxN elements, where element (i, j) in the matrix is the relationship between object i and object j.

The third label is aiming to provide a simpler layout description in line with the robotic grasping task. For each object in the environment, we want to answer to the following question. How many other objects are on top of the current object that need to be moved

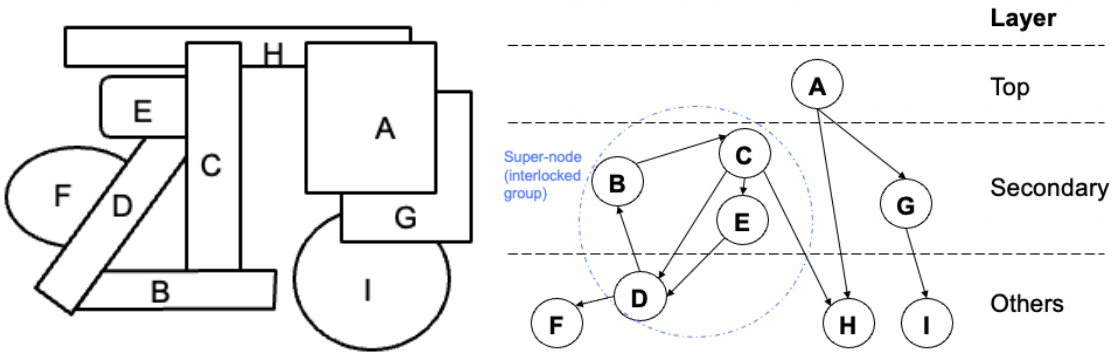


Figure 3.3: Example of how a graph of objects is categorized into the different layers (Top, secondary, others). Each graph edge represents an instance of an object covering the other.

away before picking? To better understand the order in which objects must be grasped, we create a directed graph to represent each layout. Each node represents an object in the layout and each edge represents an obstruction relationship where the parent object is covering the child. From this representation, we can see what objects and how many objects are occluding the same object. As robots pick objects sequentially, occluded objects will be revealed entirely once the objects on top of them are picked. Therefore, it is not necessary to evaluate occluded objects that are at the bottom of the scene. Sometimes, the occlusion between objects is small enough that can be ignored, so objects occluded by only a single other objects are also important to be evaluated. Given this, we categorize each object in a layout into 3 different layers. Top layer contains objects that are clear of any obstructions. Secondary layer includes objects that are covered by only a single other object. Others layer includes the rest of the objects. In some cases, there could be groups of interlocked objects. Interlocked objects that are being directly covered by only one object would be considered to be within the secondary layer. An example of a environment of objects from the top down view and the resulting graph can be seen in Figure 3.3.

3.2.3 Layout-based Difficulty Levels

While per category object detection metrics can measure performance on specific categories of objects, a difficulty rating for the overall environment would better allow us to eliminate hard examples and better understand how the model would perform under different environment conditions.

(a) (b) (c) (d) (e)

Figure 3.4: Example images from 5 difficulty levels in MetaGraspNet benchmark dataset. a) level 1: minimal occlusion. b) level 2: some occlusion. c) level 3: incomplete objects (scissors is crosscut by the banana). d) level 4: multiple instances of the same object class. e) level 5 includes all difficult characteristics.

We label images according to 5 different levels of difficulty. Those levels are defined by 4 different characteristics: Number of layers, occlusion percentage, instance completeness, and class uniqueness. Instance completeness refers to if a single object instance is visually crosscut into multiple segments due to occlusion. In such case, we refer to this kind of objects as incomplete objects, and refer to objects without visual crosscuts as complete objects. Incomplete objects often result in an object over-detection or over-segmentation, and thus it is a good characteristic to test object detection and segmentation models. Class uniqueness refers to if all objects in an image belong to different categories, or are visually distinct from each other. This characteristic is to evaluate object detection and segmentation models on distinguishing objects with similar visual features while clustered.

The first two levels of difficulty will be primarily concerned with understanding how a model deals with different levels of occlusion and layers. The layer limit for level 1 difficulty is set to 1, and the occlusion limit is set to 5% empirically. The next Three levels are primarily concerned with measuring the model's ability to correctly label object instances. Level 3 includes incomplete objects in an image, and level 4 includes non-unique objects. Level 5 includes both incomplete as well as non-unique objects. Table 3.1 describes all the difficulty levels, and Figure 3.4 shows images from each difficulty level.

3.2.4 Dataset Details

The proposed MetaGraspNet benchmark dataset contains 100,000 images, with 11,000 different scenes and 25 different household objects whose 3D models are provided by the

Level	layer limit	occlusion limit	complete object	unique class
1	2	5%	✓	✓
2	N/A	N/A	✓	✓
3	N/A	N/A		✓
4	N/A	N/A	✓	
5	N/A	N/A		

Table 3.1: Table description for difficulty levels

Yale-CMU-Berkeley Object and Model Set [160]. The objects are placed in a red plastic box in the metaverse environment, and represents an universal small load carrier which is used in many intralogistics use-cases. A scene is a single arrangement of objects in the bin and multiple images are taken of that scene at various viewpoints.

3.3 Layout-weighted Evaluation Metric

As discussed in Section 3.2.2, not all objects are of the same importance in a grasping task. Top and secondary layer objects have a priority to be picked, while picking the rest of objects requires moving away top and secondary layer objects. Therefore, our proposed metric focuses on evaluating top and secondary layer objects. Besides evaluating model performance for top and secondary objects separately, we propose a layout-weighted metric which considers the model performance on both top and secondary layer objects. In particular, we use objects' percentage of unoccluded area to weigh objects' evaluation score in each grasping scene. Occlusion percentage measures the maximum percentage area an object could be in contact with other objects, which indirectly measures how much disturbance moving the object can cause to other objects. The less disturbance an object creates, the more likely it is picked first, and thus the more important it is in a scene.

Given the occlusion percentage of an object, p, let the object's weight be $w = 1 - p$. We consider a scene $S = \{o_i | i \in [1, n]\}$ containing n objects, where each object o_i is indexed by i. Let A contain all indices for top layer objects, and B contain all indices for secondary layer objects. Let the evaluation score for object o_i be v_i. This score can be produced by any standard object detection and segmentation metric such as average precision, and intersection over union (IoU). Then the per-scene layout-weighted score V_S for the scene S is calculated as in (3.1):

$$V_S = \sum_{i \in \{A+B\}} \frac{w_i}{\sum_{j \in \{T+S\}} w_j} v_i \tag{3.1}$$

Objects from others layers are not considered during the evaluation. Once we compute all the per-scene layout-weighted score, we take the mean as our layout-weighted score.

3.4 Conclusion

In this chapter, we proposed MetaGraspNet: a large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis. This dataset contains 100,000 RGBD images, 11,000 scenes, and 25 classes of objects. The proposed MetaGraspNet benchmark dataset consists of detailed object detection, segmentation, layout annotations, and a script for layout-weighted performance metric. We presented 5 difficulties to evaluate model performance in different grasping scenarios. Moreover, we proposed a new layout-weighted performance metric to evaluate object detection and segmentation performance in a manner that is more appropriate for robotic grasp applications. Subsequent chapters will then use the dataset provided here to explore the problem of scene understanding for robotic grasping.

Chapter 4

Investigating Use of Keypoints for Object Pose Recognition

This chapter delves into the exploration of keypoints for object pose recognition, a critical aspect of scene understanding for robotic grasping. The chapter is based on a study that demonstrates the feasibility of a pose estimation network based on detecting semantically important keypoints on the MetagraspNet dataset, which contains heavy occlusion and greater scene complexity. The chapter discusses various challenges in using semantically important keypoints as a way to perform object pose estimation. These challenges include maintaining consistent keypoint definition, as well as dealing with heavy occlusion and similar visual features. The chapter also presents experimental results and discusses the impact of occlusion and similar keypoints on the performance of the pose estimation model. The insights gained from this study contribute to the ongoing efforts to improve the accuracy and generalizability of pose estimation methods, thereby enhancing scene understanding for robotic grasping. The chapter is structured to first introduce the concept and importance of keypoints in Section 4.1, then delve into the implementation details in Section 4.2, followed by a presentation of experimental results in Section 4.3, and finally a discussion of the findings and their implications in Section 4.4.

4.1 Introduction

In recent years, there has been a rising interest in the use of robotic systems to handle object manipulation tasks in scenarios ranging from manufacturing to domestic settings.

Figure 4.1: An example of keypoints detection results for drills.

One important hurdle to performing automated object manipulation is estimating accurate object pose. Object pose provides important knowledge both before, during, and after a grasp action. Before a grasp, pose information can allow the robot to target different parts of the object depending on its task. During a grasp, pose information is vital for moving objects through space without collision, as well as to operate tools that are being grasped. Finally, it can be important to place an object in the correct orientation when releasing a grasp.

Object pose estimation methods can be categorised as direct pose estimation methods or Perspective-n-Point(PnP) methods. Direct pose estimation predict the 3D rotation and translation matrix of an object relative to a reference pose, such as an exact 3D model of the object in question [31, 35, 38, 42]. Perspective-n-Point(PnP)/RANSAC methods use an intermediate representation that is used to match up with the 3D model pose. Those representations can be categorized as either dense or sparse representations. Dense PnP methods [51, 61, 64–66] predicts a correspondence to a reference model for each input point. In contrast, sparse methods only predicts a limited number of correspondences or keypoints, typically in the range of 5-20. These keypoints may be defined as the corners of the 3D object bounding box [52, 53] or a set of points defined relative to the object surface [55, 57, 58, 62, 72, 76, 161]. These surface keypoints are most similar to pose estimation methods used in human pose estimation.

Direct pose estimation and dense PnP pose estimation methods both have a heavy reliance on exact 3D reference models. Such 3D models are costly to collect and makes it difficult to expand the dataset to new objects. Sparse keypoints are more flexible, and do not require exact matches with a reference mode. The Pascal3D+ dataset [162] for example uses semantic keypoints based on a limited number of 3D models to represent the pose of a much wider variety of real world examples that do not match exactly.

Furthermore, because direct and dense PnP methods are informative only relative to a reference model, the resultant prediction contains little intrinsic information. In contrast, surface keypoints may directly inform us of part locations or surface properties. For instance, a particular keypoint can be defined as the "tip" of a object, while another may be the "end" of the handle. This direct description without the need to refer back to a reference model can decrease computation requirement during inference, removing the need to load a 3D CAD model into the system for each predicted object.

Previous work using semantically important sparse keypoints for robotic grasping are very limited in the complexity of its environment and the variety of available object classes. There is often very little occlusion, and only two or three object categories are considered.

In this work, we investigate the feasibility of using semantically important keypoints in object pose estimation in complex, clustered environments. We train a heatmap-based keypoint detection model on the MetagraspNet dataset [163]. Our model is evaluated based on both pose estimation performance as well as 2D keypoint similarity scores. Through experiments, we conclude that the model performance is heavily impacted by occlusion and similar nearby points.

4.2 Implementation Details

The keypoint detection network is implemented using mmpose [164] with a ResNet [165] backbone and a heatmap-based keypoint predictor head based off of Simple Baseline 2D [46]. A separate head is used for each prediction class. An example of keypoint detection results for the drill class can be seen in Figure 4.1.

The model is trained on the synthetic MetagraspNet Dataset [163]. Each object category is labelled with keypoints containing unique ids at semantically important locations on the object surface. Boxes, for example, are labeled with 8 keypoints on each corner. Some example classes are displayed in Figure 4.2.

The 2D keypoints detected on an image can then be converted into an estimated pose through solving for the rotation and translation matrices to minimize the reprojection error

Figure 4.2: Example keypoint labels for various classes.

from 3D-2D point correspondences using the RANSAC algorithm. Example pose results can be seen in Figure 4.3.

4.2.1 Keypoints Definition

The most difficult component of using keypoints for object pose recognition is defining keypoint placement. A popular option is to use furthest point sampling [57], where a fixed number of keypoints are sampled evenly around a 3D shape. However, such a sampling method offers no semantic information.

Instead, a method similar to Manuelli *et al.* [76] is preferable, where keypoints are defined in areas significant to the object, such as the top, bottom, and handle. Such keypoints offer important semantic information, but can be more difficult to implement for both model prediction and in defining where keypoints are placed.

During model prediction, semantically important keypoints can be less robust to occlusion due to being fewer in number. It can also be difficult to ensure that they are located in areas that are visually distinct.

During keypoint definition, it is difficult to define a generalized pattern for placing keypoints, especially for more complex objects like scissors or drills. The complexity and uniqueness of their geometric shapes makes such objects less likely to fit into any previous

pattern of keypoint placement, or can even conflict. For example, corners areimportant in boxes but labelling all corners of a more complex shape that contains many corners may not be necessary or even desired.

Another challenge is symmetry. Symmetrical objects with theoretically different poses may visually be exactly identical. Human pose estimation only needs to handle bi-radial symmetry, but objects with more than a single plane of symmetry are very common. Pose estimation specific metrics, such as those in [27] often take into account symmetry in the final loss calculations, where visually identical poses are not punished, while [76] defines keypoints on the line of symmetry itself to reduce ambiguity. However, important object properties such as edges or corners often lies outside that line or plane of symmetry, and makes it difficult to limit keypoints to only along such axis. A box, for example, is intuitively defined with keypoints at the corners. However, this definition causes ambiguity for flipped or rotated poses where the particular keypoints may not necessarily match.

Our implementation defines keypoints along axis of symmetry, such as one on center-top, and another on center-bottom, but also on important features such as corners of boxes or edges of cups. Examples of keypoints on various object classes can be seen in Figure 4.2.

4.3 Experimental Results

4.3.1 Metrics

We evaluate the pose prediction on the average distance (ADD) [20] metric :

$$e_{\text{ADD}} = \frac{1}{m} \sum_{x \in \mathcal{M}} ||(Rx + T) - (\tilde{R}x + \tilde{T})|| \tag{4.1}$$

Where $\mathcal{M}$ is the set of 3D model points, m is the number of points, R and T are the ground truth rotation and translation matrices, and $\tilde{R}$ and $\tilde{T}$ are the estimated rotation and translation matrices. This can be computed in both world coordinates using only the estimated pose matrices, as well as after projecting both poses into image coordinates based on the camera's intrinsic matrix. This metric can be converted into an accuracy score by the following equation:

$$\text{Accuracy}_{\text{ADD}} = \frac{1}{n} \sum_{e_{\text{ADD}} \in \mathcal{N}} \begin{cases} 1 & \text{if } e_{\text{ADD}} < t \\ 0 & \text{otherwise} \end{cases} \tag{4.2}$$

Where $\mathcal{N}$ is the set of predictions, n is the number of predictions, and t is the distance threshold.

We also look at the object keypoint similarity (OKS) score for the 2D keypoints based on the COCO evaluation metrics:

$$\text{OKS} = \sum_i \left[exp\left(\frac{-d_i^2}{2s^2 k_i^2} \right) \right] \tag{4.3}$$

where d_i is the euclidean distance between detected and ground truth keypoint, s_i is object scale, and k_i is a per-keypoint constant to control falloff. Precision and recall metrics are computed with OKS at 0.5 threshold. Results can be seen in Table 4.1.

4.3.2 Pose

After solving for rotation and translation matrices using RANSAC, pose estimation example results can be seen in Figure 4.3.

Our model is trained on cereal box and drill classes. The accuracy curve for different ADD score threshold can be seen in Figure 4.4.

4.4 Discussion

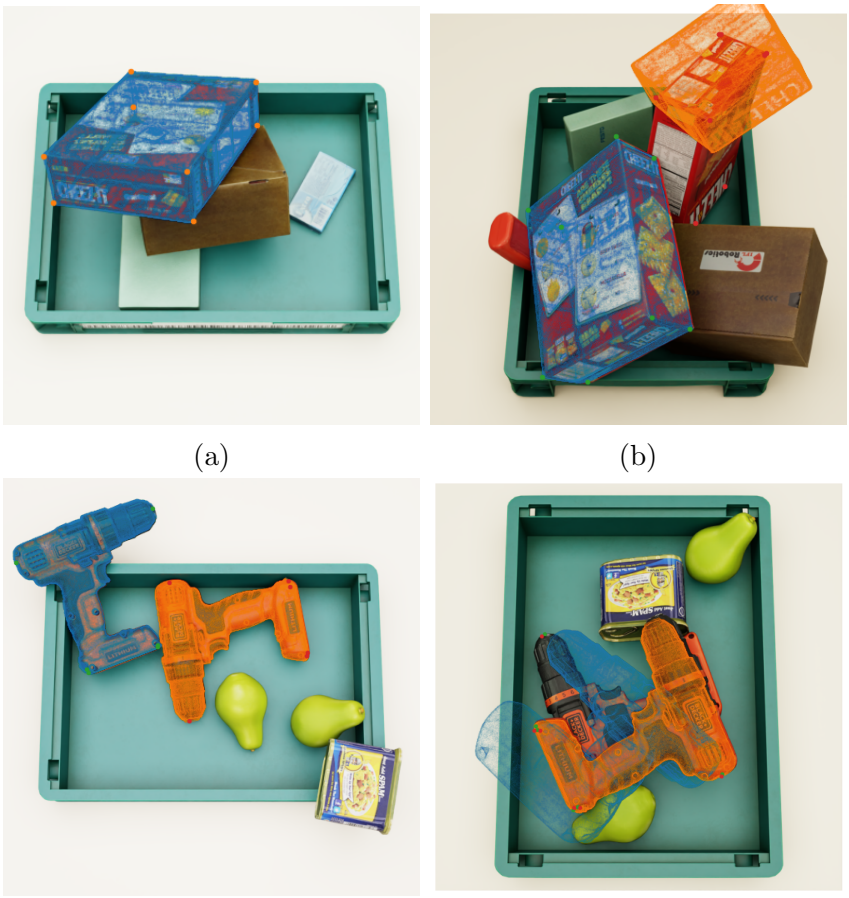
4.4.1 Occlusion

The average precision and recall scores are heavily impacted by occlusion, as shown in Table 4.1.

Table 4.1: Average precision and recall performance for occluded and visible keypoints in drills

	AP	AR
all	83.1	87.2
occluded points	69.4	78.8
visible points	86.7	89.4

When a keypoint is occluded and not visible from the camera,, there are no longer local features for the model to identify, making it significantly more difficult to predict



(a) (b)

(c) (d)

Figure 4.3: Visualization of pose results.

that keypoint's location. This issue has also been pointed out by Pavlakos *et al.* [72]. Chen *et al.* [62] works to improve this weakness through heavy augmentation, the lack of robustness to occlusion is a key reason for the popularity of dense PnP and direct pose estimation methods compared to sparse PnP methods.

Occlusion is especially problematic with semantically important keypoints. Often, the number of defined points are fewer than those defined in automatically sampled methods, and thus the loss of one point more heavily impacts the pose prediction. An example of this can be seen in Figure 4.3b and 4.3d.

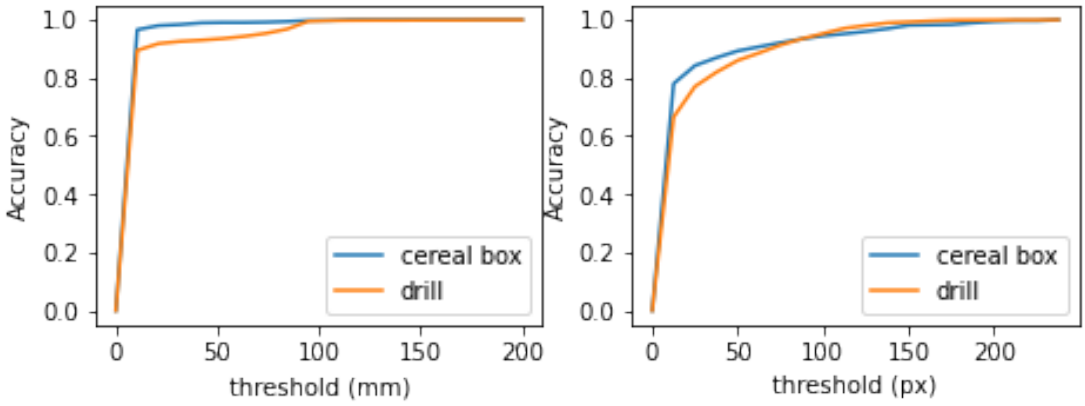


Figure 4.4: ADD accuracy scores given different thresholds for world coordinates (left) and projected image coordinates (right).

4.4.2 Similar keypoints

The model can be easily confused by keypoints where the local features are very similar to each other. This can happen in two different scenarios: between keypoints on the same object as well as if a similar object is within the bounding box.

An example of confusion between similar keypoints shown in Figure 4.5, where keypoints 0, 1, 2, and 3 are misclassified with each other.

We demonstrate this phenomenon through calculating the percentage of keypoint observations which contain more than one peak above the detection threshold of 0.3. As seen in Table 4.2, 15 to 24 percent of predictions for keypoints 0-3 are confused between different possible locations. This compares to the around 5% probability for keypoints 4 and 5, which are more visually unique.

In comparison, cracker boxes have more visual graphics that aid in distinguishing between the different keypoints. An example of a cracker box can be seen in figure 4.6.

Those visual graphics result decrease the likelihood of there being multiple peaks in the heatmap prediction in comparison to keypoints 0-3 on drills, but are still less unique than keypoints 4 and 5.

Using this multiple-peaks probability, we can iteratively improve the location of various keypoints to more visually distinct locations through identifying keypoint locations to minimize this value. However, it can be difficult to do so while maintaining their semantic significance.

(a) keypoint 0 (b) keypoint 1 (c) keypoint 2

(d) keypoint 3 (e) keypoint 4 (f) keypoint 5

Figure 4.5: Visualization of heatmap for predicting a keypoint for example in Figure 4.1, overlayed on the original image.

This confusion between keypoints is sometimes caused by symmetry. Human pose estimation deals with symmetry pre-defined pairs of key-points that can be flipped with each other (eg. left-ear, right-ear) when the image itself is flipped as a data augmentation. However, there are often more than a single axis of symmetry in objects, and it may be non-obvious which keypoints to flip.

Similar to Manuelli *et al.* [76], we primarily define keypoints on the line of symmetry itself to reduce ambiguity. However, important object properties such as edges or corners often lies outside that line or plane of symmetry, and makes it difficult to limit keypoints to only along such axis. A box, for example, is intuitively defined with keypoints at the corners.

The second scenario which causes confusion between similar keypoints is the presence of nearby objects of the same class. A great example of this can be seen in Figure 4.3d. The bounding box for the bottom drill (blue model, green keypoints) overlaps heavily with the top drill. When predicting the keypoints, only keypoint 4 (see Figure 4.5e for location) was correctly predicted on the bottom drill. The other points were all placed on the top drill. This suggests that the model is only focusing on the local features in the near vicinity of the keypoint location rather than the overall object. A comparison of ADD accuracy scores for images containing multiple objects of a given class vs those with a single object of a given class can be seen in Figure 4.7, with the former case performing notably worse.

Table 4.2: Probability that each keypoint would contain multiple detected peaks above the 0.3 confidence threshold for drills.

keypoint id	probability of multiple peaks (%)
0	22.1
1	20.3
2	15.6
3	23.7
4	5.8
5	4.2

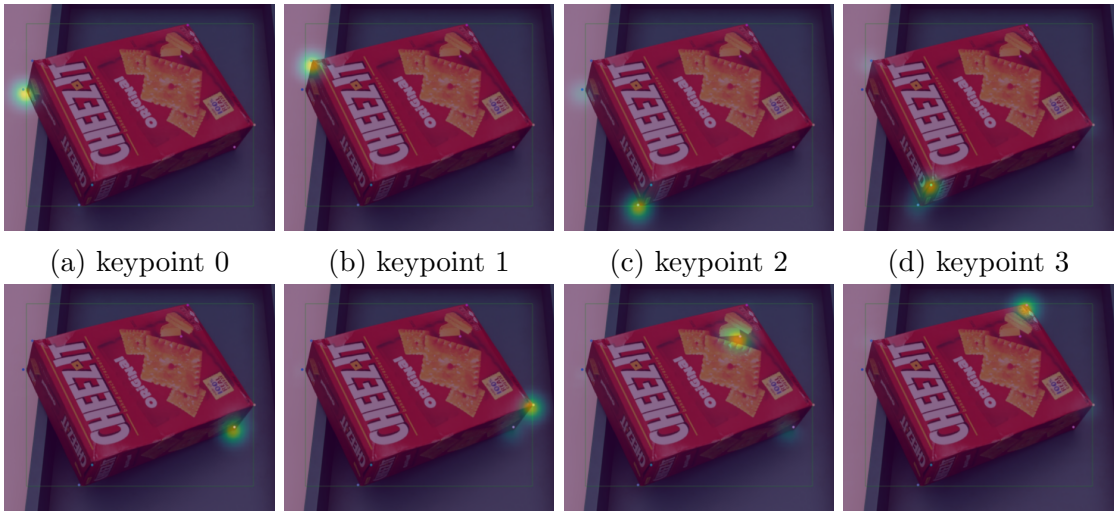
Table 4.3: Probability that each keypoint would contain multiple detected peaks above the 0.3 confidence threshold for cracker boxes.

keypoint id	probability of multiple peaks (%)
0	7.9
1	9.6
2	6.8
3	7.2
4	11.1
5	13.2
6	16.2
7	15.8

4.5 Summary

Through training a heatmap-based pose detection model on the MetagraspNet dataset, we demonstrate that semantically important keypoints can be an effective way to estimate the pose of objects. However, the complexity of the available object classes results in keypoint definitions that do not have a consistent pattern, leading to ambiguities on where to place them on new objects.

Furthermore, the complex scenes with high level of occlusion, sometimes with similar objects on top of each other, present in the MetagraspNet dataset reveal challenges that were not well observed in previous implementations. In particular, the model struggles heavily with occluded keypoints as well as with the presence of points with similar visual features nearby. Those similar local features may be present on the same object, or on other objects close-by. We demonstrate this issue through observing that the probability



(a) keypoint 0 (b) keypoint 1 (c) keypoint 2 (d) keypoint 3

(e) keypoint 4 (f) keypoint 5 (g) keypoint 6 (h) keypoint 7

Figure 4.6: Visualization of heatmap predictions for a cracker box.

of multiple peaks in the heat-map prediction varies based for different keypoints based on the local visual uniqueness of that point (as shown in Tables 4.2 and 4.3). We further demonstrate this problem through the decrease in ADD performance when multiple objects of the same class are present in the same object (as shown in Figure 4.7).

Future work may use the multi-peak probability to identify poor keypoints and iteratively improve the keypoint definition to ensure easy recognizably. In addition, we may also explore alternative ways to represent object poses, such as with categorical keypoints (as opposed to keypoints with unique ids), as well as volumetric primitives, so that the representation is less arbitrarily defined, and enable easier integration with the downstream robotics grasping pipeline.

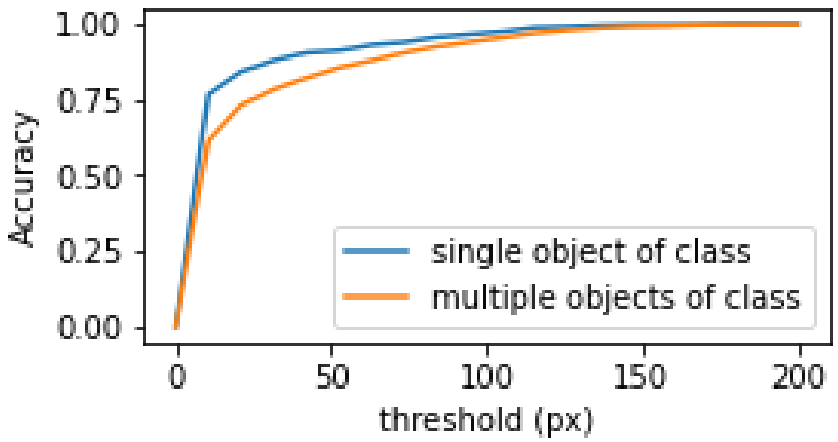


Figure 4.7: ADD accuracy scores given different thresholds in projected image coordinates for images with multiple objects of the same class compared to images with only a single object of a given class.

Chapter 5

ShapeShift: Superquadric-based Pose Estimation for Enhanced Scene Understanding in Robotic Grasping

This chapter introduces ShapeShift, a superquadric-based object pose estimation method that offers a flexible approach to understanding the scene for robotic grasping. By predicting an object's pose relative to a primitive shape fitted to the object, ShapeShift provides intrinsic descriptiveness and information about the 3D object without the need for a reference 3D model. This approach enhances the robot's ability to interact with its environment, thereby contributing to more effective robotic grasping. The chapter will detail the development of ShapeShift, its underlying principles, and its application in the context of scene understanding for robotic grasping.

5.1 Introduction

Object pose estimation is a crucial task in robotics, enabling precise manipulation of objects in the environment. However, a common challenge faced by current object pose estimation techniques is their heavy reliance on a reference 3D object. Adding new object categories into the model incurs a substantial expense, as it necessitates accurate 3D scans. Furthermore, these models have limited generalizability and are confined to a small set of objects. In practice, referring back to the 3D model to obtain useful information about grasp points and positions can limit the effectiveness of these techniques in robotic grasping applications.

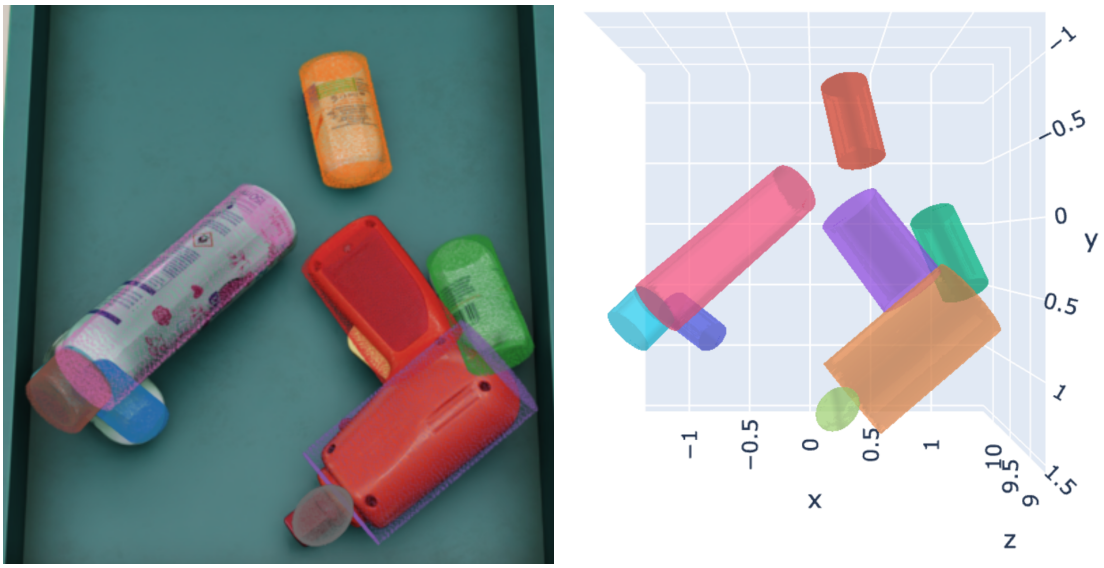


Figure 5.1: An example of ShapeShift on a typical scene.

Keypoint-based methods for object pose estimation, such as those described in [75] offer intrinsic descriptiveness about the 3D object without the need to reference a 3D model. However, these methods can be arbitrary in their definition and lack consistency between objects. Such methods deal poorly with multiple degrees of symmetry and heavy occlusion.

To overcome these challenges, this paper proposes ShapeShift, a framework for object pose estimation based on primitive shapes. By fitting a primitive shape to an object, the proposed approach provides intrinsic descriptiveness and information about the 3D object without relying on a 3D model. In this paper, we specifically utilize superquadrics, which are three-dimensional shapes described by a mathematical equation that has been used to simplify shape representation in previous works [114, 117, 119]. The proposed approach predicts the object's pose in reference to a predicted primitive shape fitted to the object. This not only provides intrinsic descriptiveness but also enables generalization to arbitrary geometric shapes not present in the training set, making it a promising solution to the challenges faced by current object pose estimation techniques.

5.2 Method

The proposed ShapeShift framework can be described as follows. In the first phase, primitive shapes are fitted to each geometric part of an object using superquadrics. In the second phase, the superquadrics fits are leveraged as "ground truth" for a superquadric-guided direct regression network to directly predict pose and shape information.

5.2.1 Phase 1: Superquadric fitting

Superquadrics, described by Equation 5.2 are characterized by parameters in Equation 5.1.

$$\theta = \{\epsilon_1, \epsilon_2, a_x, a_y, a_z, R, t\} \tag{5.1}$$

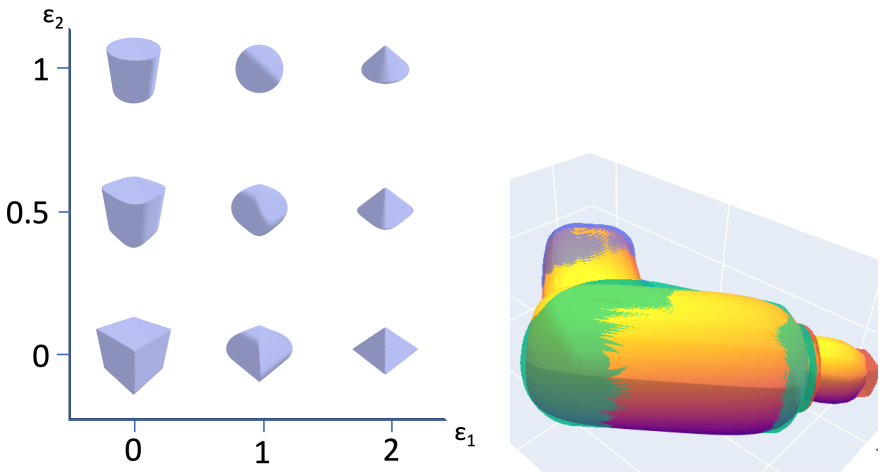
a_z are the scale, ϵ_1 and ϵ_2 define the shape of the surface, R defines the rotation matrix, and t is translation.

$$F(x) = \left(\left(\frac{x}{a_x} \right)^{\frac{2}{\epsilon_2}} + \left(\frac{y}{a_y} \right)^{\frac{2}{\epsilon_2}} \right)^{\frac{\epsilon_2}{\epsilon_1}} + \left(\frac{z}{a_z} \right)^{\frac{2}{\epsilon_1}} \tag{5.2}$$

The equation represents the surface implicitly, returning $F(x) = 1$ for points on the surface, $F(x) < 1$ for points inside the object, and $F(x) > 1$ for points outside. Through varying the parameters in (5.1), different 3D shapes can be expressed. Examples of different superquadrics are shown in Figure 5.2a.

The first phase of ShapeShift leverages the superquadric fitting technique from Liu [114] to fit superquadrics to parts. This primarily serves to provide ground truths for superquadric poses, shapes, and scales for Phase 2: Superquadric-guided Pose Estimation described Section 5.2.2. After each 3D model is fitted with a collection of superquadrics, as seen in Figure 5.2b, each image can then be labelled with the superquadric pose, replacing the original object pose.

It is possible to go without Phase 1, and instead compare the superquadric prediction in Phase 2 directly with the original object, either through comparing point cloud differences or through an implicit loss similar to [119]. However, that would require significantly more compute per loss calculation, and would likely be more difficult to regress to compared to a pre-labeled superquadric pose.



(a) (b)

Figure 5.2: Superquadrics shape space and when fitted to an object.

Regardless of whether superquadrics are explicitly fit to the 3D models for ground truth labels, or compared directly to the original objects, there arises a problem of similar superquadrics characterized by different parameters, which was explored by Liu in [114].

Recall that superquadrics are characterized by shape, scale, rotation, and translation parameters (5.1). Axis mismatch similarity occurs when $\epsilon_1 \approx \epsilon_2$, and similar superquadrics can be achieved through reassigning the principal axis to either x or y, then applying a corresponding rotation. In Equation 5.3, θ_1^c and θ_2^c shows the parameters for the two possible similar superquadrics.

$$\begin{aligned} \theta_1^c &= \{\epsilon_2, \epsilon_1, a_y, a_z, a_x, [r_2, r_3, r_1], t\} \\ \theta_2^c &= \{\epsilon_2, \epsilon_1, a_z, a_x, a_y, [r_3, r_1, r_2], t\} \end{aligned} \tag{5.3}$$

where $[r_1, r_2, r_3]$ are orthonormal column vectors of R, the rotation matrix

There also exists duality similarity when $a_x \approx a_y$ and $\epsilon \in [0, 2]$. The new parameter for this similar superquadric is shown in Equation 5.4

$$\theta_3^c = \{\epsilon_1, 2 - \epsilon_2, s \cdot \bar{a}, s \cdot \bar{a}, a_z, R \cdot R_z(\pi/4), t\} \tag{5.4}$$

where

$$
\begin{aligned}
s &= \begin{cases} ((1-\sqrt{2})\epsilon_2 + \sqrt{2}) & \text{if } \epsilon_2 < 1 \\ (\sqrt{2}/2 - 1)\epsilon_2 + 2 - \sqrt{2}/2 & \text{otherwise} \end{cases} \\
\bar{a} &= (a_x + a_y)/2
\end{aligned} \tag{5.5}
$$

Duality similarity exists because, as ϵ_2 becomes greater than 1, the shape that it represents becomes the same as when $\epsilon_2 = 2 - \epsilon_2$, or the same as the shapes mirrored about $\epsilon_2 = 1$. The only difference is that the shapes are rotated by 45° about the z axis, and the x and y scales are different, while a_z remains the same. $R_z(\pi/4)$ in equation 5.4 represents this rotation about the z axis, while s in equation 5.5 represents the scale change.

To avoid cases of duality similarity, we maintain ϵ_2 within the range $[0, 1]$, while redefining the scale as a warp transformation rather than directly in the implicit equation in 5.2. Additionally, the rotation can be transformed into a new rotation using the following approach:

$$
\begin{aligned}
S &= \begin{bmatrix} a_x * s, & a_y * s, & a_z \end{bmatrix}^T \mathbf{I} \\
S_{warp} &= R_d^{-1} S R_d \\
R_{new} &= R R_d
\end{aligned} \tag{5.6}
$$

where $R_d = R_z(\pi/4)$ is a rotation of $\pi/4$ about the z axis

Instead of a 3x3 transformation matrix for S, it can be further expanded to a pure scale and a shear transformation. Reversely, S and R can be combined into a single 3x3 transformation matrix (S).

Other situation of similar superquadrics are solved through defining a series of discrete and continuous symmetrical transformations for the loss calculation in Phase 2.

5.2.2 Phase 2: Superquadric-guided Pose Estimation

The second phase of ShapeShift involves superquadric-guided pose estimation to directly predict pose and shape. The proposed architecture (see Figure 5.3) uses depth images as input and extends upon [51] in several ways. To reduce search dimensions, superquadrics were discretized based on shape parameters and treated as different object categories. This also allows for easily defining symmetry (discrete and continuous) for error calculations.

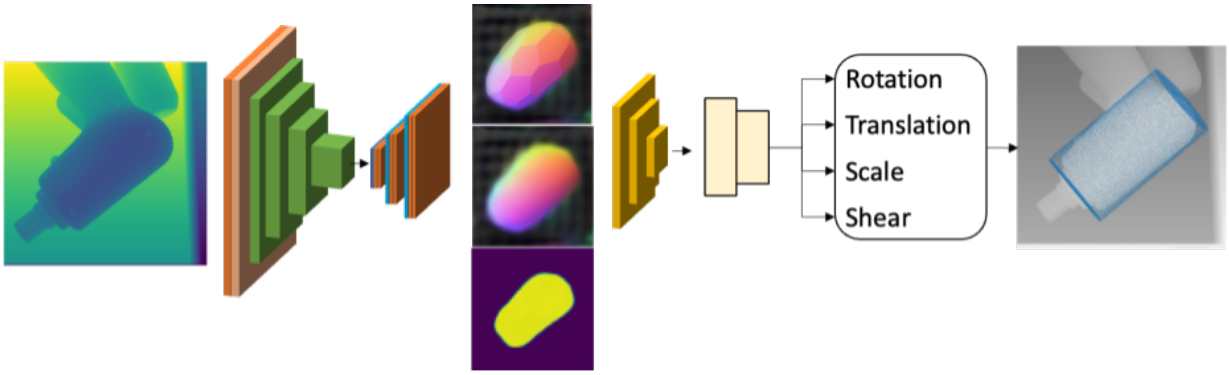


Figure 5.3: The proposed superquadric-guided pose estimation architecture extends upon [51] by introducing a new scale and shear head. The main difference is defining the 3D correspondences in reference to a superquadric shape rather than the 3D model of the full object.

To generate an intermediate representation, un-scaled primitive shapes were sampled using furthest point sampling. Additionally, the proposed architecture introduces additional scale and shear heads. An additional head can be added in the future to predict shape parameter offsets.

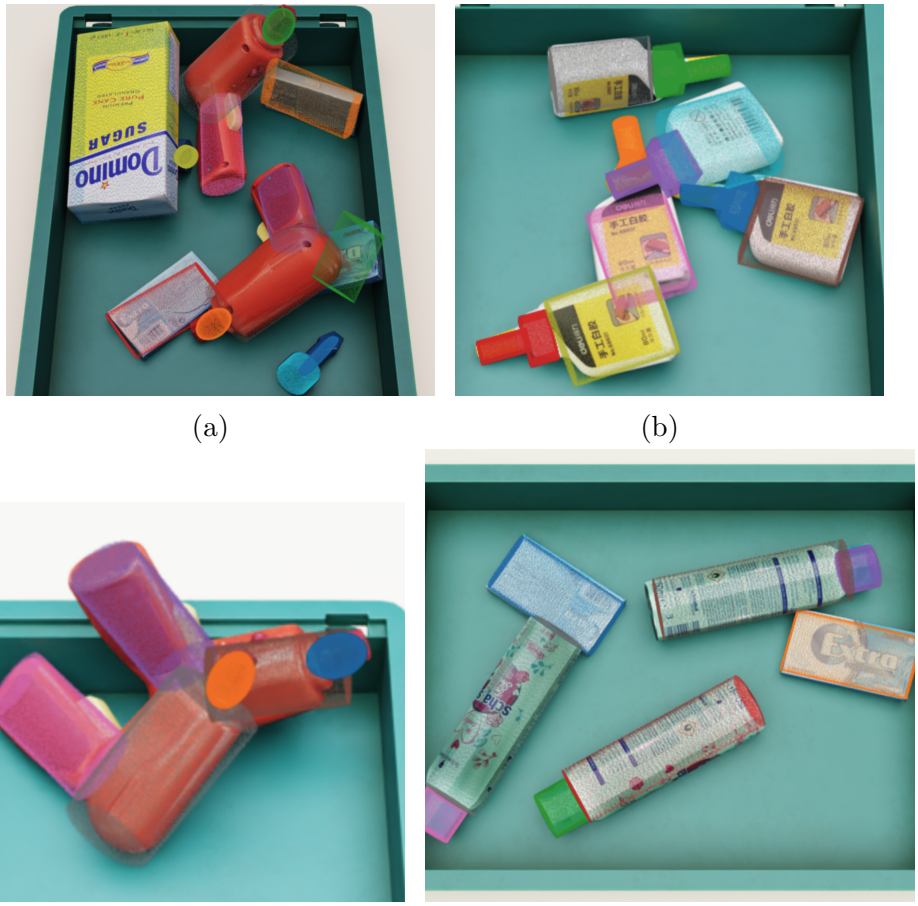
5.3 Experimental results and discussion

We conducted experiments on a subset of the MetaGraspNet benchmark dataset [163] containing 54 object categories of objects that are composable using multiple superquadrics. We evaluated the performance of our proposed ShapeShift framework for object pose estimation using the Maximum Symmetry-Aware Surface Distance (MSSD) and Maximum Symmetry-Aware Projection Distance (MSPD) metrics [27]. For evaluation, we combined rotation, scale, and shear into a single transformation matrix. The accuracy scores based on both metrics are shown in Figure 5.5.

During experimentation, we observed three challenges.

First, the method's performance was lower for underrepresented shape categories, as shown in Figure 5.5.

Second, shear is underrepresented in the ground truth, but there exists a head dedicated to predicting it, causing a number of false predictions. False predictions of shear caused projected error to be higher than 3D surface distance error for $\epsilon = (0, 0)$ shapes.



(a) (b)

(c) (d)

Figure 5.4: Superquadrics pose and shape prediction examples.

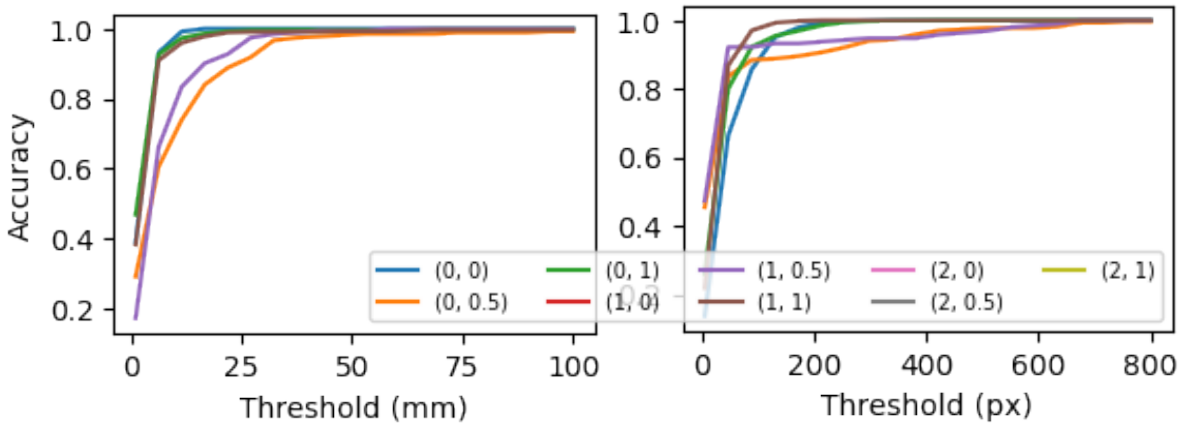


Figure 5.5: MSSD (left) and MSPD (right) based accuracy scores over different thresholds on discrete superquadric shapes

Finally, the proposed method is robust to partial occlusion on objects composed of multiple primitive shapes. As illustrated in Figures 5.4b, 5.4c, parts of the object that are not occluded can still be properly predicted, even while the method struggles on heavy occlusion. Further work exploring data augmentation methods and an addition amodal mask prediction head should help with overall occluded object performance.

5.3.1 Summary and future work

In this paper, we introduced ShapeShift, a superquadric-based framework for object pose estimation is proposed that predicts an object's pose relative to a primitive shape fitted to the object. This approach provides intrinsic descriptiveness and information about the 3D object without relying on a 3D model. The approach was further tested on the MetaGraspNet benchmark dataset, and has demonstrated the ability to approximate the shapes present in the image. Future work would focus on optimizing performance in cases of occlusion, dealing with the imbalanced distribution of shape types, adding an additional head to predict precise shape parameters, evaluating performance on novel objects not seen in the training set, and evaluating performance on a full robotics grasping pipeline.

Chapter 6

MMRNet: Multimodal Consistency for Reliable Uncertainty Estimation

This chapter presents MMRNet, a novel method for improving the reliability of multimodal object detection and segmentation for bin picking. By introducing the concept of multimodal consistency (MC) for uncertainty estimation, MMRNet aims to provide reliable and interpretable measures of uncertainty, which is critical in robotic grasping involving human-in-the-loop scenarios. The chapter will discuss the development of MMRNet, the concept of MC, and the impact of these advancements on robotic grasping.

6.1 Introduction

Global labor shortages and the need for resilient supply chains has accelerated companies' upgrades to industry 4.0 and introduced a range of technologies such as big data, cloud computing, internet of things (IoT), robotics, and artificial intelligence (AI) into production systems. With warehouses and manufacturing units becoming smart environments, a crucial objective is to develop an autonomous flow of both material and information, and robotic bin picking plays an essential role in this task.

Robotic bin picking has been an active area of research for many decades given the complexity of the task, ranging from joint control and trajectory planning [166] to object identification [167] and grasp detection [15].. In particular, we examine the object detection and segmentation task in autonomous bin picking. Different from object detection and segmentation in other areas such as autonomous driving, robotic vision system works in

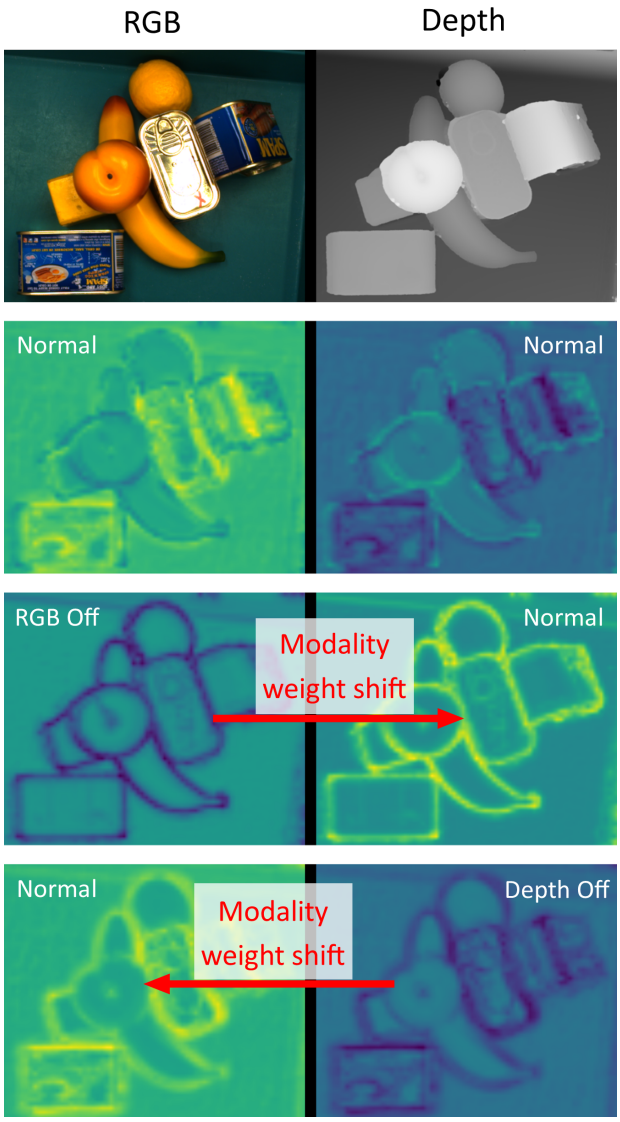


Figure 6.1: The dynamic modality weight shifting of our network ensures a reliable overall performance when a modality is missing. Row 2-4 heatmaps describe the average gate weights of each modality at a single feature scale. Yellow indicates high weight, dark purple indicates low weight.

environments that are very close to the camera, dealing with heavy occlusions, shadows, dense object layouts, and complex stacking relations. It plays an essential role in a robot's

perception system.

Deep neural networks have been proven effective for object detection and segmentation [168, 169]. But, deploying such systems in robotic picking applications is challenging due to the many sources of uncertainty present in practical scenarios. Real-world bin scenes may consist of a wide variety of unknown or occluded items arranged in an infinite number of poses and illuminated with variable lighting conditions. In addition to the variability of real-world bin scenes, errors in the camera system can make a computer vision system unreliable. Camera sensors are prone to noise and can fail in various situations such as specular reflections (missing values), black areas (missing depth), overexposure, blur, and artifacts. In practice, commercial systems are expected to run 24/7 to be feasible, which increases the risk of imaging sensor failures compared to research environments. If not accounted for, sensor failures can lead to wrongly commissioned orders and in the worst case to product and hardware damages, leading to expensive recall campaigns or production downtime. Therefore, vision systems capable of handling uncertain inputs and producing reliable predictions under sensor errors are critical to creating fail-safe applications.

One approach to creating fault-tolerant object detection and segmentation systems is to introduce system duplication, where portions of the system are duplicated to allow the system to continue to operate despite failures of its constituent parts. This approach assumes that failures are caused by either input sensor failures or computational failures. However, duplication may not provide fault-tolerance in situations where the system is operating correctly but its sensors are unable to adequately measure the inputs. For example, a camera may fail to adequately image a piece of glass due to its transparency, and so the use of a second identical camera cannot address this issue. In addition, deep neural networks as a data-driven approach are designed to capture feature distributions of the input dataset. A simple duplication of these networks will not detect features that are not in the training distribution. Instead, we add image data from depth sensor as an additional modality to capture object feature characteristics from a different perspective. More specifically, depth data has very simple texture yet rich geometric features, that are more transferable to unseen objects than RGB data.

A good system duplication design duplicates components that are more likely to fail, preventing any disruption in the information flow from the system input to output. Non-data-driven methods have well defined explicit logic to control the information flow. In comparison, deep learning system learns the input and output mapping through high-dimensional implicit feature representations. A typical deep learning model encodes input information through a backbone network into a high-dimensional latent representation, and downstream tasks use the representation to predict low-dimensional outputs. Consequently, a large amount of information is lost during the dimensionality reduction of

downstream tasks. However, in robotic bin-picking, unseen items may contain highly complex image characteristics that require both RGB and depth to work collaboratively. For example, RGB backbones are better at detecting transparent objects and depth backbones are better at detecting dark objects. A pair of eyeglasses with black frame will require the RGB backbone to focus on glass parts while the depth backbone to focus on the frame for a complete detection and segmentation. With the reduction of dimensionality, a simple result aggregation on two low quality detections will create another low quality detection. Additional result merging networks or explicit merging logic will introduce errors and instabilities into the system. An effective modality fusion technique that will dynamically fuse modality features with limited loss of information is therefore greatly desired. In addition, modality features merging may introduce dependencies between them, causing unexpected model behavior when one of the modality feature is absent. We tackle this problem with a multimodal redundancy framework consists of two key techniques: 1) we use a multi-scale soft-gating mechanism to make the network learn to weigh and combine features between modalities dynamically, and 2) we use a dynamic ensemble learning strategy to train the sub-system independently and collaboratively in an alternating fashion. With this framework, only one modality needs to be present for the model to operate.

Finally, we propose a novel multimodal consistency (MC) score as a more objective reliability indicator for the system output based on the overlaps of detected bounding boxes and segmentation masks. This can be used as an indicator for model uncertainty on individual predictions, as well as model reliability on particular datasets.

Through experiments, we demonstrate that in an event of missing modality, our MMR-Net provides a much more reliable performance compared to baseline models. When depth is removed, our network's performance drop is within 1% where other models have a performance drop greater than 6%. When RGB is removed, our network's performance drop is within 11% where other models have a performance drop greater than 80%. Furthermore, we demonstrate that our MC score is a more reliable indicator for output confidence during inference compared to the often overly-confident confidence scores. We summarize our contribution as the following:

- A multimodal redundancy framework consisting of a multi-scale soft-gating feature fusion module and a dynamic ensemble learning strategy allowing trained sub-systems to operate both independently and collaboratively.
- A multimodal consistency score to describe the reliability of the system output.

6.2 Related Work

Reliability study for deep learning-based systems: Deep learning-based methods are data-driven, encoding the decision making process through continuous latent vectors, which makes the model behavior hard to predict and fix. Only a few of the studies focus on the reliability aspect of the deep learning-based systems. In [170], Santhanam *et al.* list differences between traditional and deep learning-based software systems and discuss the challenges involved in the development of reliable deep learning-based systems. In [171], Xu *et al.* study the reliability of object detection systems in autonomous driving. In [172], dos Santos *et al.* study the relationship between reliability and GPU precision (half, single, and double) for object detection tasks. Other reliability related work can be found in model uncertainty estimation [173]. To the best of our knowledge, none of the work investigates reliability or uncertainty for multimodal applications, in particular for robotic bin picking.

Multimodal Data Fusion: Multimodal learning [174–178] has been rigorously studied. In multimodal learning, there are three types of data fusion: early fusion, intermediate fusion, and late fusion. Each corresponds to merging information at input, intermediate, and output stage respectively. Early fusion involves combining and pre-processing inputs. A simple example is replacing the blue channel of RGB with depth channel [7]. Late fusion merges the low-dimensional output of all networks. For example, Simonyan *et al.* [179] combine spatial and temporal network output with i) averaging, and ii) linear Support Vector Machine [180]. Early fusion and late fusion are simpler to implement but have a lower dimensional representation compared to the intermediate fusion. Intermediate fusion involves merging high-dimensional feature vectors. Common intermediate fusion includes concatenation [174], and weighted summation [167]. Recently, more advanced techniques are developed to dynamically merge the modalities. In [181], Wang *et al.* propose a feature channel exchange technique based on Batch Normalization's [182] scaling factor to dynamically fuse the modalities. In [183], Cao *et al.* propose to replace the basic convolution operator with Shapeconv to achieve RGB and depth fusion at the basic operator level. In [184], Xue *et al.* focus on the efficiency aspect of multimodal learning and propose a hard gating function which outputs an one-hot encoded vector to select modalities. In robotic grasping, Back *et al.* [167] take the weighted summation approach and propose a multi-scale feature fusion module by applying a 1x1 convolutional layer to the feature layers before passing them into a feature pyramid network (FPN) [185].

The aforementioned works are designed to optimize the overall network performance but at the same time introduce dependencies among modality features, which are extremely vulnerable in case of an abnormal event, such as an input sensor failure. In this paper, we

address the multimodal fusion strategy from the system reliability perspective, where our goal is to design a simple yet effective network architecture that enables sub-modal systems to work independently as well as collaboratively to increase the overall system reliability.

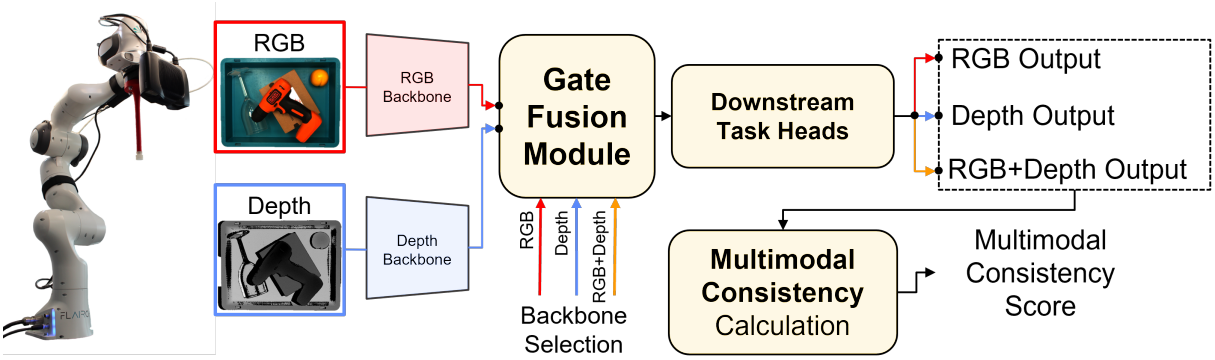


Figure 6.2: Block diagram of our multimodal redundancy framework. Gate fusion module allows simple switching between modalities. Trained with dynamic ensemble learning, our system is able to use both modalities independently (RGB or depth output) as well as collaboratively (RGB+depth output). A multimodal consistency score is computed at the end to indicate the reliability of the output.

Ensemble learning: Ensemble learning typically involves training multiple weak learners and aggregating their predictions to improve predictive performance [186]. One of the simplest approaches to construct ensembles is bagging [187], where weak learners are trained on randomly-sampled subsets of a dataset and subsequently have their predictions combined via averaging or voting techniques [186]. Instead of aggregating predictions directly, one may also use a meta-learner which considers the input data as well as each weak learner's predictions in order to make a final prediction, a technique known as stacking [188]. Boosting [189] is another common approach where weak learners are added sequentially and leverage the previous learner's mistakes to re-weight training samples, effectively attempting to correct the previous learner's mistakes.

While ensemble learning has long been a common technique in classical machine learning, it can be expensive to apply to deep learning due to the increased computational complexity and training time of deep neural networks. Of particular relevance to this work is the application of ensemble learning to multimodal deep learning problems. In multimodal problems, the data distributions typically differ significantly between modalities and thus may violate the assumptions of certain ensembling techniques [190]. Nevertheless, ensemble methods have been applied to a variety of multimodal problems [190–193]. For example, Menon *et al.* [191] trained modality-specific convolutional neural networks on

three different magnetic resonance imaging modalities and combined the models' predictions via majority voting. In [193], Zhou *et al.* used a stacking-based approach to combine the outputs of neural networks trained on text, audio, and video inputs, thereby reducing noise and inter-modality conflicts.

Rather than combining multiple models with a typical ensembling strategy, in this work we consider a *dynamic* ensemble where multiple unimodal systems are dynamically fused into a single network. This network is capable of both unimodal operation using each of its inputs independently as well as multimodal operation through the fusion of the constituent unimodal systems.

6.3 Methodology

The subsequent sub-sections outline the key components of our MMRNet architecture. Firstly, we introduce a multi-scale soft gating mechanism that effectively combines information from the two modalities. Secondly, we propose a dynamic ensemble learning strategy, which, in conjunction with the multi-scale soft gating mechanism, constitutes the multimodal redundancy framework. This framework helps to remove the inter-modality dependencies. Lastly, we present the formulation of the multimodal consistency score, which serves as our system's reliability measure. We show our system block diagram in Figure 6.2.

6.3.1 Multimodal Redundancy Framework

Multi-scale Soft-Gate Feature Fusion (MSG Fusion): Fusing high-dimensional latent representation from two data distribution involves integrating information from multiple scales as well as multiple modalities. While a simple convolution as proposed in [167] can merge the information, it also constrains the information exchange between modalities to be within the same scale. The other modality's high-level features may contain crucial contextual information for localizing and segmenting objects with intricate RGB and depth features. In order to maximize the utilization of contextual information from both modalities, we concatenate the features and input them into a Feature Pyramid Network (FPN) [185]. This FPN fuses multi-scale modality features in a hierarchical manner, enabling effective contextualization. Nonetheless, this process can result in inter-modality dependencies. To address this issue, we draw inspiration from [194] and incorporate a soft gating mechanism. This mechanism enables the dynamic adjustment of feature weights

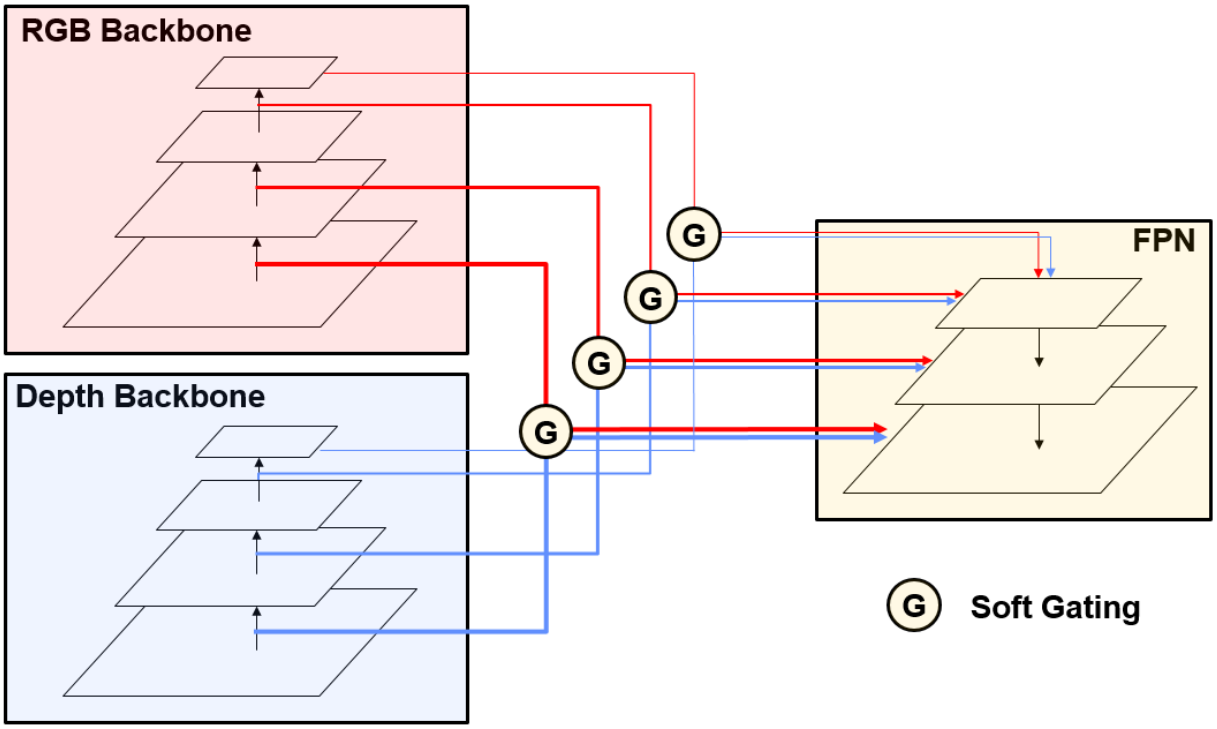


Figure 6.3: Gate fusion module fuses the multi-scale feature from each modality.

from each backbone, thereby facilitating modality feature selection that is optimized for detecting individual object classes. More importantly, this method enables the model to disentangle features from different modality backbones. We define the total number of modalities to be N and denote the jth scale feature layer in modality m as $f_{m,j}$. Features of all modalities pass through a 1x1 convolution layer G_m. The convolution layer takes N jth scale modality features with C channels and outputs one feature layer with C channels for modality m. We obtain $g_{m,j}$:

$$g_{m,j} = G_m\left(\{f_{m,j} | m \in [0, N)\}\right) \tag{6.1}$$

The output gate weight $w_{m,j}$ is calculated by:

$$w_{m,j} = \sigma(\{g_{m,j} | m \in [0, N)\}) \tag{6.2}$$

, where σ is the softmax function ensuring modality weights sum to one. Finally, the gated feature layer for scale j and modality m is updated by:

$$f_{m,j} \leftarrow f_{m,j} w_{m,j} \tag{6.3}$$

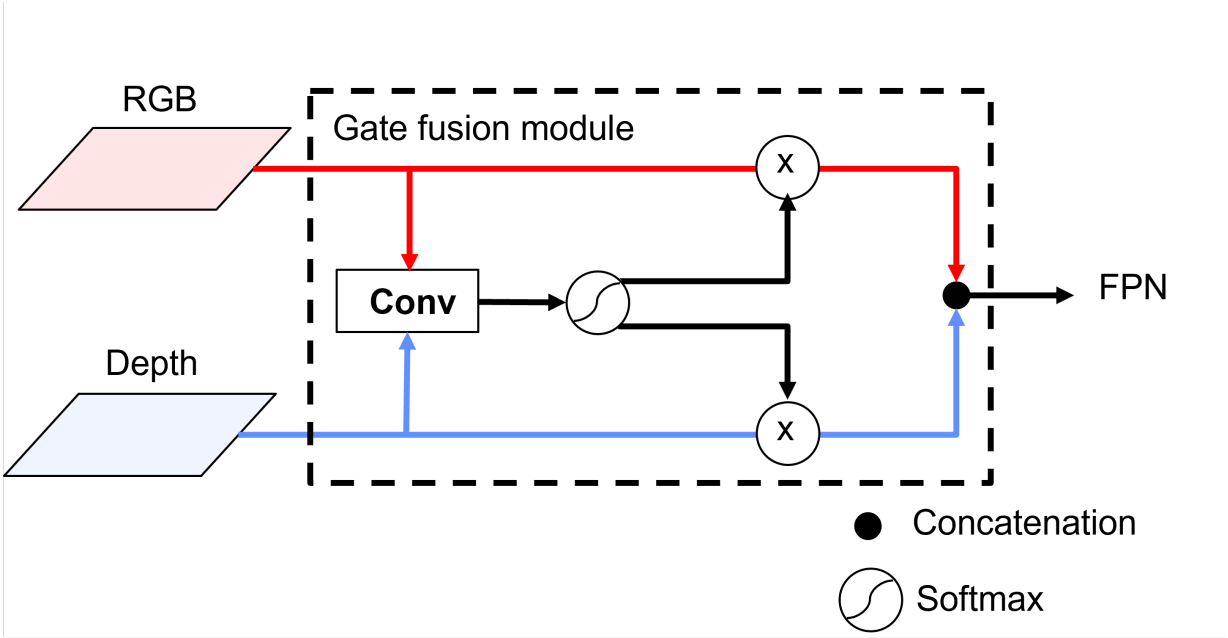


Figure 6.4: Soft gating architecture applied to every scale of feature layers.

We show the gate fusion module architecture in Figure 6.3 and Figure 6.4.

Dynamic Ensemble Learning Strategy: Although the proposed soft gating mechanism enables dynamic re-weighting of features extracted from each input modality, it does not inherently allow for modalities to be used independently. Ideally, the network would be capable of operating reasonably using a single input modality, with each additional modality providing improved performance or reliability. This accounts for the practical scenario where the sensor used to capture an input modality fails, forcing the system to leverage its other inputs.

Classic ensemble approaches combine weak models according to their standalone performance by a simple discrete process such as weighted sum [186]. In comparison, our gating module allows more dynamic interaction since information can be exchanged across different modalities and scales with respect to different input items, poses, and scene layouts. Instead of independently training each modality model and combine them with a weighted sum, we propose a novel dynamic ensemble learning strategy to train multimodal deep learning models, allowing for different modalities to be used both collaboratively and independently.

Specifically, in each training iteration we randomly select one of the possible input conditions: both inputs, RGB-only input, or depth-only input. In the unimodal conditions,

we force the system to make predictions with only one of its usual inputs in order to encourage rich features to be extracted from both modalities. This training scheme prevents the model from learning to rely heavily on a single modality while simultaneously allowing the model to learn how to combine data from both modalities.

6.3.2 Multimodal Consistency Score

Existing object detection and segmentation networks contain a confidence score calculated by the softmax classifier for each detection. The reliability information in this score is somewhat subjective as it is estimated from the same network. Instead, we leverage the multimodal property of our model. In an ideal scenario, if we train a separate model for each modality, all models would converge to produce the same output describing the same object in the physical space. Less reliable models will produce results that deviated from the ground truth. Models trained with different modalities capture distinct feature distributions and characteristics such as textures and geometries. We assume the output deviation between the modalities is very different from each other. If the network output is reliable, then the outputs between modalities are well-aligned. This can be measured by the percentage overlap between output bounding boxes as well as the segmentation masks. Based on this assumption, we argue that the more deviations between the modalities there are, the less certain the output is. To estimate the deviation, we use Intersection Over Union (IOU). It is a ratio between the intersection of the two modalities and their union and can be applied to boxes as well as masks. Given a pair of detection/segmentation output x_0 and x_1. Each represents a set of pixels. x_0 and x_1 can either be a pair of boxes or a pair of masks. Then, IOU can be calculated by:

$$IOU(x_0, x_1) = \frac{|x_0 \cap x_1|}{|x_0 \cup x_1|} \tag{6.4}$$

Where $|.|$ is a function that computes the number of pixels for the given input. When deviation becomes larger, IOU will be smaller. When there is less deviation, IOU will be larger and close to 1. This behavior captures well the output alignment between modalities. When comparing results in object detection/segmentation, object matching is involved. There can be multiple detections for one object, so we average the IOU score for all related detections associated with this object. For simple annotation, we call the two models being compared source and target. Source results are matched to the target results. Let the set of all n_s detections in source be $D_s = \{d_{s,l} | l \in [0, n_s)\}$. We compute the IOU of all items in D_s associated with the kth target detection $d_{t,k}$, and obtain a set of IOUs $I_{D_s,d_{t,k}} = \{IOU(d_{s,i}, d_{t,k}) | d_{s,i} \in D_s\}$. We define objects with IOU lower than 30%, a typical

threshold value used in the Non-Maximum Suppression (NMS) step in object detection networks such as [168], as non-matched and remove them. The updated IOU set is

$$I^{'}_{D_s,d_{t,k}} = \left\{a | a \in I_{D_s,d_{t,k}}, a > 0.3\right\} \tag{6.5}$$

Next, we compute the average IOU $A(D_s, d_{t,k})$ for source detections D_s and target detection $d_{t,k}$:

$$A(D_s, d_{t,k}) = mean(I^{'}_{D_s,d_{t,k}}) \tag{6.6}$$

We further compute the mean IOU for all n_t detections in target D_t:

$$mIOU(D_s, D_t) = mean(\{A(D_s, d_{t,k}) | k \in [0, n_t)\}) \tag{6.7}$$

We extend this mIOU score to describe the alignment between our network output and all the modalities. We name this score multimodal consistency (MC) score. MC can be used to describe the alignment of one modality or multiple modalities in a multimodal system. Let D_o to be the network output with n_o number of detections, and D_m to be the network output using only modality m. The MC score S_m for a single modality m is calculated by:

$$S_m = mIOU(D_m, D_o) \tag{6.8}$$

The MC score S for all modalities is computed by:

$$S = mean(\{A(D_m, d_{o,k}) | m \in [0, N), k \in [0, n_o)\}) \tag{6.9}$$

The higher the MC score is, the more reliable the system is. A score of 100% means all modalities predict the exact same output, and the system is very reliable. A score of 0% means each modality predicts a different output, and the system is unreliable.

6.4 Experiments

6.4.1 Dataset and Implementation Details

Among robotic grasping datasets, the MetaGraspNet dataset [163] provides large-scale, high-resolution simulated RGB and depth data as well as real-world data from an industry-grade sensor system. In addition, the dataset contains 82 objects and has a novel object set for testing. We divide the real dataset into train, validation, test, and test novel. We first exclude all scenes with novel objects, adding them to a separate novel test data split. Then we split the rest of the real dataset into 80% train, 10% validation, and 10% test.

Due to the unique characteristics of each modality, we normalize and pre-process RGB inputs and depth inputs differently. We use the standard mean variance normalization for RGB inputs and we apply min-max normalization per scene for depth inputs, where depth values are min-max normalized to $[0, 1]$. We further flip the depth values to make 0 as the depth of the background and 1 as the closes point to the camera. With this value flip, background values are aligned to be 0 in each scene. In addition, this normalization added a data augmentation to the dataset as it stretches and compress object shapes in depth, allowing a fully utilized depth range where every depth value is used by an object.

Near objects' edges, reflective surfaces, and transparent surfaces, there are often undefined values caused by a lack of signal returning to the depth sensor. As a pre-processing step, we apply image inpainting [195] to the depth images to replace any invalid values.

We use a classic object detection and segmentation network Mask-RCNN [168] with ResNet50 [196] backbone as our baseline. All the networks in our experiment are initialized by the same ImageNet [197] pretrained weights. We train all models with the same training configuration in terms of batch size, training epoch, and optimizer. We pretrain all the models on the simulated dataset of MetaGraspNet, and finetune on the real dataset. We report the performance of our method on the real test set with bounding box mean average precision (box mAP) and segmentation mask mean average precision (mask mAP).

6.4.2 Results and Discussions

MC Score: The model exhibits a steady decline in MC score between the training set, test set, and test-novel set as seen in Table 6.1. This shows that the MC score decreases accordingly the more out-of-distribution a dataset is, correlating well with theoretically how reliable the model will be on each dataset. The MC score also differs dramatically between objects of different classes. Objects with poor MC scores include disinfection bottle, glass bottle, cables in transparent bag, eyeglasses, and so forth, while boxes, cups, cables (not in plastic bags), and pears have higher MC scores, as seen in Table 6.2. This shows the ability of the MC score to identify challenging objects in the dataset. We also compute the MC score against only RGB input or only depth input. Some objects exhibit a significant difference in MC score between those two options in Table 6.3. The starkest contrast appears when the object's material properties result in significant noise in one sensor, such as when transparent or reflective objects causes errors in the depth sensor. Through this, we can identify when the model is highly reliant on a particular sensor for its predictions. Eyeglasses, for example, performs both poorly overall (Table 6.2, Figure 6.5d), and relies heavily on RGB input due to the transparency and reflection of its glass component.

Data Split	MC Score - Box	MC Score - Mask
train	91.4	92.7
test	82.9	84.7
test-novel	73.4	73.7

Table 6.1: Class-agnostic MC scores from different data splits

		MC Score (%)	
	Class	Box	Mask
non-novel objects	disinfection bottle	60.5	62.6
	glass bottle	76.9	64.7
	cups d	96.0	96.5
novel objects	cables in transparent bag	68.4	61.0
	eyeglasses	73.6	63.7
	pear	77.6	91.4

Table 6.2: MC scores for different object classes

	Mask MC Score (%)	
Class	RGB only	Depth only
power drill	68.4	73.3
wineglass	89.8	76.5
eyeglasses	69.6	58.0

Table 6.3: Comparing MC score using only RGB or depth

Examples of object-level MC scores for segmentation mask detections can be seen in in Figure 6.5. Note that the confidence score predictions for each object remains at an inflated 0.999 for all four examples, while the MC score shows a greater distinction between the objects depending on difficulty. This is especially true in Figure 6.5d, where the model outputs a poor detection, but with high confidence. This, supported by previous dataset-level results, shows that the MC score is a better indicator for model reliability and uncertainty compared to the confidence score.

6.5 Conclusion

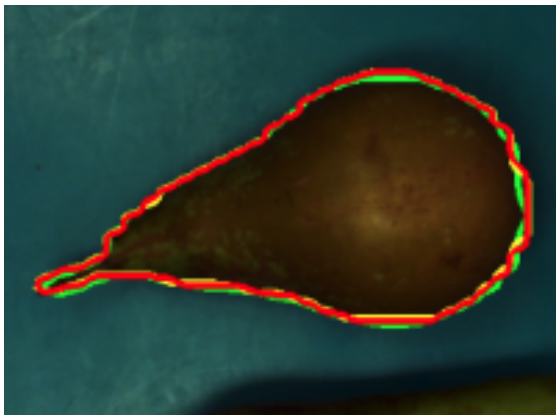
This paper has addressed the crucial aspect of reliability in deep learning-based computer vision systems for robotic grasping through the introduction of a multimodal redundancy framework called MMRNet. Specifically, we have achieved multimodal redundancy by leveraging a multi-scale soft-gate feature fusion and dynamic ensemble learning strategy to train modality models both independently and collaboratively. Additionally, we have proposed a multimodal consistency score as a reliable indicator of network output certainty. The results demonstrate that our MMRNet delivers robust performance in the event of a modality input failure, and that the MC score serves as a well-suited output reliability indicator that is independent of the network's confidence score. Through such methods, a computer vision system would be able to effectively warn a human supervisor regarding uncertain scenarios for tasks such as robotics grasping.

Acknowledgements

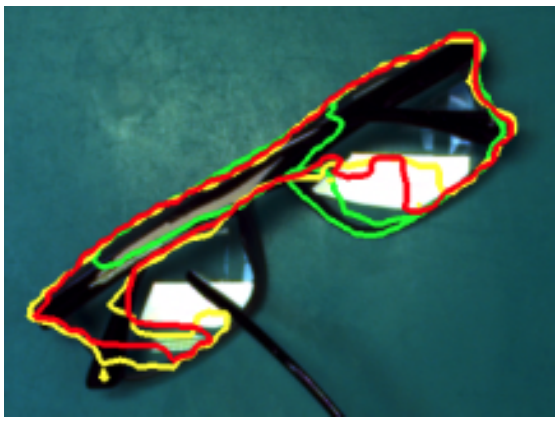
This work was supported by the National Research Council Canada (NRC) and German Federal Ministry for Economic Affairs and Climate Action (BMWK) under grant 01MJ21007B. We also extend our appreciation to Festo and DarwinAI for their valuable input and assistance throughout the project.

(a) Train set object
MC score: 0.964
Confidence score: 0.999

(b) Test set object
MC score: 0.778
Confidence score: 0.999



(c) Test-novel set object
MC score: 0.966
Confidence score: 0.999



(d) Test-novel set object
MC score: 0.633
Confidence score: 0.999

Figure 6.5: Examples of object level MC scores. Gate fusion output is marked in red contour. RGB and depth only are marked in yellow and green contours respectively. In Figure 6.5b, some predictions identified the object as separate boxes, decreasing the MC score.

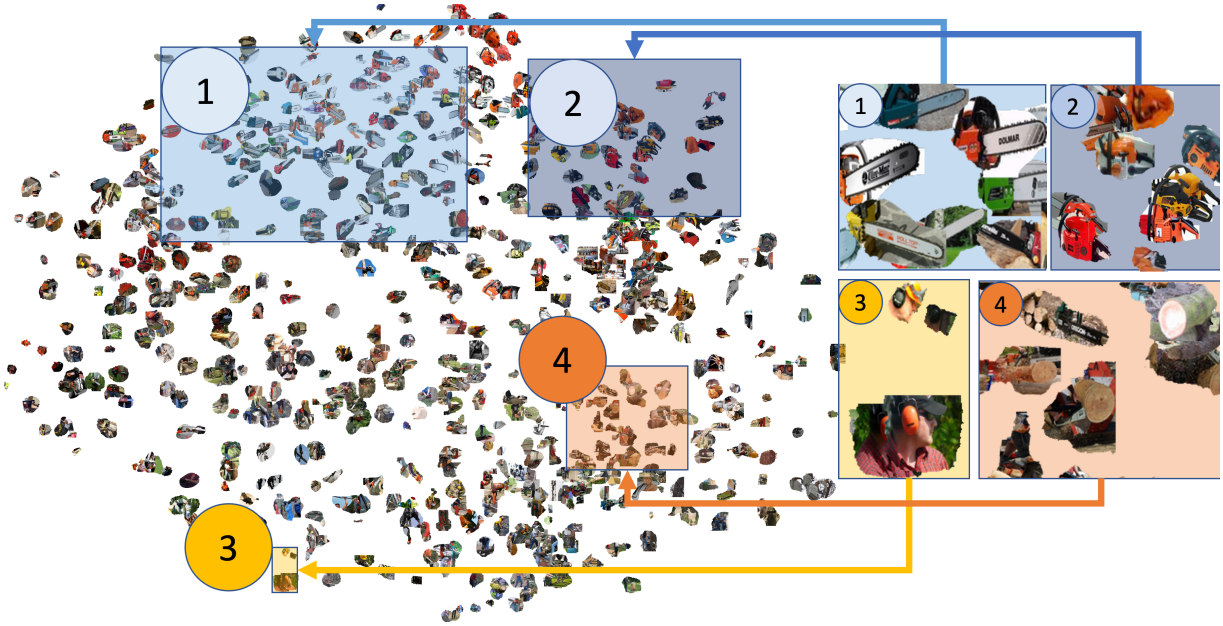
Chapter 7

Second-Order Explainable AI: Unveiling Actionable Insights for Enhanced Scene Understanding in Robotic Grasping

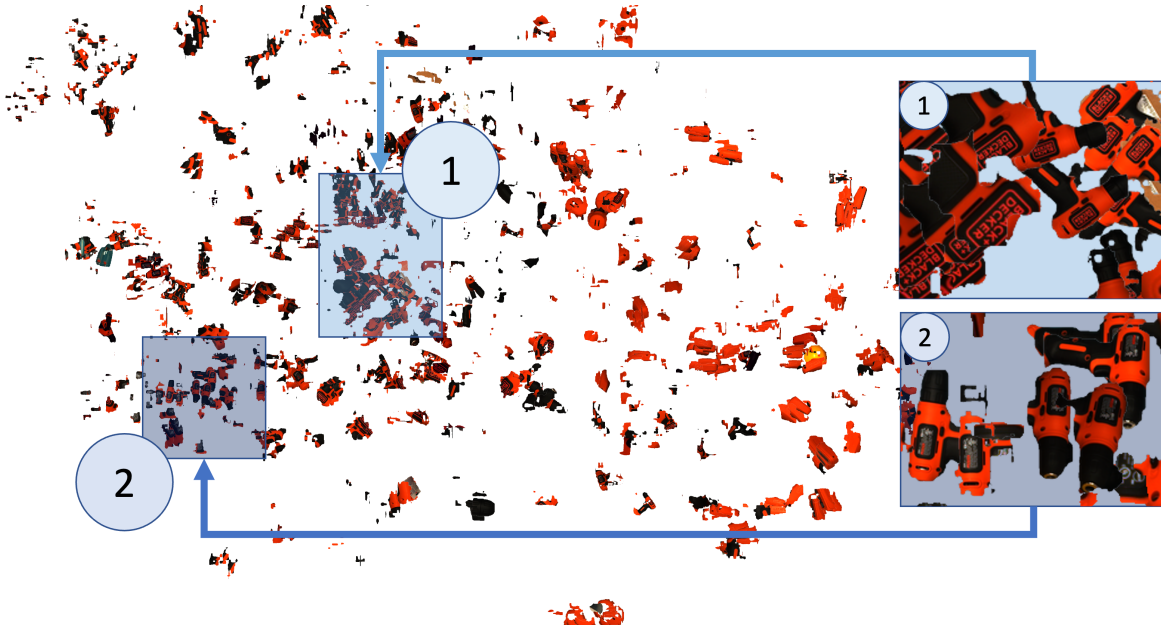
This chapter delves into the realm of second-order explainable AI (SOXAI), a concept that extends explainable AI from the instance level to the dataset level. SOXAI provides a higher-level interpretation of a deep neural network's behavior, allowing us to "explain the explainability" for actionable insights. In the context of scene understanding for robotic grasping, these insights can help uncover biases in the model's decision-making process or in the training data, which can then guide improvements in the training framework. This chapter will detail the development and application of SOXAI, demonstrating how it can enhance the efficiency and reliability of robotic grasping by providing deeper insights into the model's behavior and the characteristics of the dataset.

7.1 Introduction

Although quantitative performance metrics such as accuracy are essential indicators of a deep neural network's performance, they do not offer insights into the decision-making process. To fill this gap in the performance analysis, explainable AI (XAI) can facilitate the auditing of model behaviour. This auditing helps ensure that the decisions are based



(a)



(b)

Figure 7.1: SOXAI visualizations of a classification model on chainsaws 7.1a and a segmentation model on hand drills 7.1b. Different regions show groupings of related quantitative explanations via first-order XAI, with significance discussed in Section 7.3.

on relevant visual indicators. Additionally, it can uncover potential biases in the training data, which may then be used to guide improvements to the training framework.

First-order explainability techniques such as Grad-CAM [198], integrated/expected gradients [199, 200], LIME [201], GSInquire [202], and SHAP [203] yield per-instance visualizations of explanations. However, reviewing these visualizations can be time-consuming, particularly for large-scale datasets with multiple classes or high intra-class variability. In addition, human biases can impact manual review.

In this work, we explore the concept of second-order explainable AI (SOXAI) [204] for obtaining actionable insights and demonstrate, for the first time, that such insights can be used to enhance model performance. SOXAI extends XAI from the instance level to the dataset level to enable the auditing of the model and dataset during development. Rather than relying on manual reviews of visual explanations to explore patterns in a model's decision-making behavior, SOXAI seeks to automatically unveil these patterns through the analysis of the relationships between quantitative explanations. This expedites the identification of the shared visual concepts utilized by a model during inference and can uncover apparent model and dataset biases. Furthermore, this improves transparency by uncovering problematic patterns that exist among a groupings of examples in the dataset, which can adversely impact the model's decision-making process. In essence, SOXAI enables us to "explain the explainability" by providing higher-level interpretations of model behaviour for actionable insights.

7.2 Methods

The concept of SOXAI takes first-order instance-level quantitative explanations of samples in a dataset and groups similar embeddings of these explanations to generate a user-friendly visualization that enables the uncovering of patterns among different groupings of data to unveil trends.

Here, we employ GSInquire [202] to generate first-order quantitative explanations of a neural network's decision-making process across a dataset. GSInquire examines the network's activation signals in response to the input image and employs them to identify critical features within the sample that quantitatively led to the network's decision.

7.2.1 Second-order explainability

Second-order explainability is treated as an embedding problem: given an image I and the corresponding quantitative explanation α for the trained model M, we define the n^{th} element of the embedding $f : (I, \alpha) \to \mathbb{R}^N$ as:

$$f(I, \alpha)_n = \frac{\sum_{i=1}^{H} \sum_{j=1}^{W} M(I)_{ijn} \alpha_{ij}}{\sum_{i=1}^{H} \sum_{j=1}^{W} \alpha_{ij}}, \tag{7.1}$$

producing an N-dimensional vector embedding from the regions of I weighted by α. Notably, M is truncated such that its output is a convolutional feature map of size $H \times W \times N$, and α is resized to $H \times W$ to match. Equation 7.1 ignores regions not identified as critical and only considers regions with higher weighting score provided by α – in essence, f performs a weighted average of $M(I)$ with weights α.

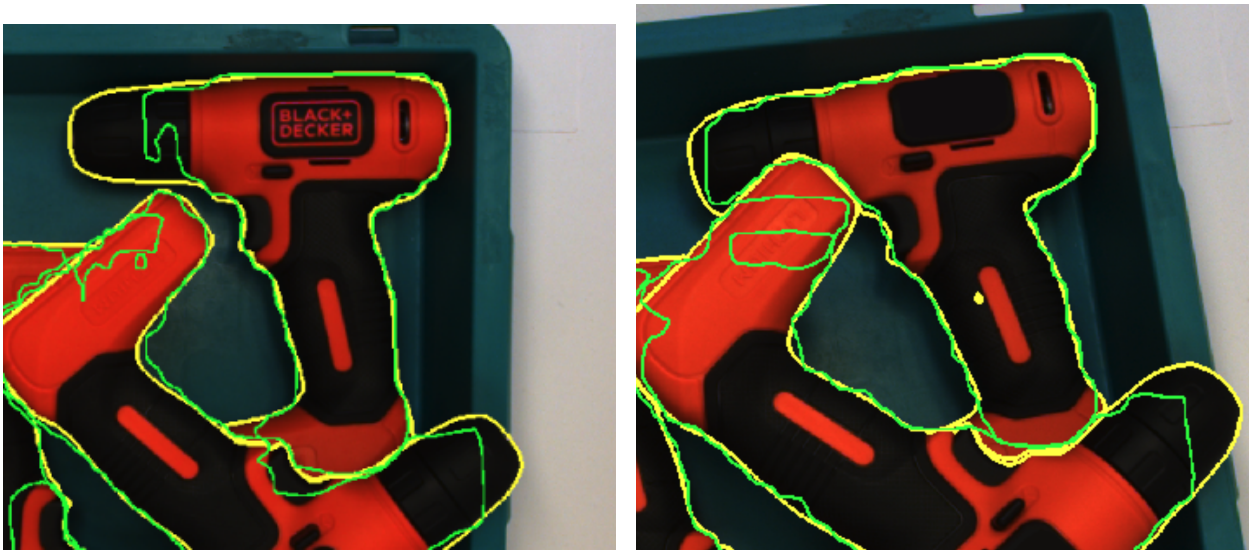
Here, we use t-distributed stochastic neighbour embedding (t-SNE) [205] to group the resulting embeddings across a full dataset [204]. In addition, embeddings were reduced to 50 dimensions via principal component analysis before applying t-SNE to map them to a 2D space for visualization.

7.3 Experimental Results and Discussion

We present two example cases of SOXAI visualization: image classification and foreground instance segmentation, discuss the actionable insights gained from each, and demonstrate how such actionable insights can be used to enhance model performance.

Chainsaw classification: To explore SOXAI for classification, we apply it to a ResNet-50 trained on ImageNet 1k [206]. An example result for the chainsaw class can be seen in Figure 7.1a, which also highlights four groupings of interest. Groupings 1 and 2 show the frontal part of chainsaws (*i.e.*, the cutting chain and guide bar) and the handle, respectively, demonstrating that the model has learned important features representing the target class. However, smaller groupings highlighted in areas 3 and 4 also reveal biases that the model has learned over time. In grouping 3, we see that the model has learned a relationship between earmuffs commonly worn when using chainsaws and the actual class prediction. Grouping 4 shows images of logs and even wooden sculptures instead of chainsaws directly.

Through the use of SOXAI, we were able to quickly identify reoccurring biases learned by the model towards objects that commonly appear in the same frame as the target class.



(a) (b)

Figure 7.2: Example of a drill with incomplete segmentation 7.2a, and with the label filled in 7.2b. Model prediction is outlined in green.

This was accomplished without the need to manually inspect each example in the validation set, as would be necessary for first-order XAI algorithms. Based on the identified biases, enhanced model performance may be achieved by better-targeted elimination of biases in future training and data collection or cleaning.

Drill segmentation: Here, we apply SOXAI to a MaskRCNN model [207] trained on the MetaGraspNet dataset [163] to detect foreground objects. As an example, we analyze the segmentation of drills, an object category not seen in the training set, chosen for its geometric and textural complexity. Figure 7.1b presents the SOXAI result, highlighting two groupings representing different faces of the drill.

The face shown in grouping 1 exhibits a high level of focus on the large logo. Since the model was not explicitly trained to recognize drills, some other foreground object must have biased it towards recognizing letters. We observe that the large logo is over-represented in the grouping, while the frontal black head of the drill is underrepresented.

To investigate further, we evaluate the prevalence of incomplete segmentations of the drill when each face is visible, such as the incomplete segmentation shown in Figure 7.2a. We find that 37% of predictions for drills with the large logo facing up are incomplete segmentations, with much of the frontal black segment missing, while only 14% of segmen-

tation predictions on the other face are incomplete.

To confirm the model's bias towards text, we mask out the logo (see Figure 7.2b), and evaluate the mAP score. We observe an increase from 0.592 to 0.618, suggesting that allowing the model to ignore its learned bias and focus on a fuller representation of the object improves its performance. These example cases demonstrate the usefulness of SOXAI for unveiling actionable insights into model biases that can be used to enhance a model's performance.

Chapter 8

Conclusion

This thesis focuses on various problems relating to robotic grasping and scene understanding, as encapsulated in the five main contributions in chapters 3 to 7. Each of these contributions corresponds to a unique aspect of the overall study and has helped to push the boundaries of what is currently possible in the field.

The creation of MetaGraspNet v0, as discussed in Chapter 3, has provided a comprehensive resource for training and evaluating models for robotic grasping. This large-scale benchmark dataset for vision-driven robotic grasping has addressed the need for high-quality, diverse datasets in the field, and contains 100,000 RGBD images, 11,000 scenes, and 25 classes of objects. The dataset includes detailed object detection, segmentation, layout annotations, and a script for a layout-weighted performance metric. Five difficulty levels were presented to evaluate model performance in different grasping scenarios. A new layout-weighted performance metric was proposed to evaluate object detection and segmentation performance in a manner that is more appropriate for robotic grasp applications. The development of this dataset has laid the groundwork for the exploration of more advanced pose estimation techniques for better scene understanding, as discussed in the subsequent two chapters.

Building upon the foundation laid by MetaGraspNet v0, Chapter 4 explored the use of keypoints for object pose recognition. This approach offers a more descriptive representation of an object's pose compared to traditional methods that provide relative pose compared to a 3D model. However, the complexity of the available object classes and the high level of occlusion in the MetaGraspNet dataset revealed challenges that were not well observed in previous implementations. In particular, the model struggles heavily with occluded keypoints as well as with the presence of points with similar visual features nearby.

Those similar local features may be present on the same object, or on other objects close-by. We demonstrate this issue through observing that the probability of multiple peaks in the heat-map prediction varies based for different keypoints based on the local visual uniqueness of that point (as shown in Tables 4.2 and 4.3). We further demonstrate this problem through the decrease in ADD performance when multiple objects of the same class are present in the same object (as shown in Figure 4.7). Despite these challenges, the exploration of keypoint-based pose estimation has contributed to ongoing efforts to improve the accuracy and generalizability of pose estimation methods. This exploration of keypoints has set the stage for the introduction of a novel pose estimation method in the next chapter.

Chapter 5 introduced ShapeShift, a superquadric-based object pose estimation method. Through representing objects as a composition of primitive shapes, this method provides a more flexible and robust representation of object geometry. It improves upon many shortcomings in the keypoints based method of the prior chapter, such as a consistent definition between different classes of objects, higher robustness to partial occlusion, and mechanisms to deal with symmetry.

Building on the advancements in scene understanding, the thesis then addressed the challenge of uncertainty estimation in robotic grasping. In real-world applications, it is often necessary to have a human-in-the-loop for situations where the model is uncertain. To do so, it is necessary to detect when a model is uncertain in its understanding of a scene.. In Chapter 6, a multimodal consistency score built on top of a multimodal redundancy framework called MMRNet was introduced to address this issue. MMRNet leverages a multi-scale soft-gate feature fusion and dynamic ensemble learning strategy to train modality models both independently and collaboratively. Through looking at the output consistency between independent and collaborative outputs of each modality, a multimodal consistency score was proposed as a reliable indicator of network output certainty. The results demonstrate that MMRNet delivers robust performance in the event of a modality input failure, and that the multimodal consistency score serves as a well-suited output reliability indicator that is independent of the network's confidence score.

Understanding the model and the dataset is another critical aspect of robotic grasping. With the increasing complexity of models used in robotic grasping, there is a growing need for methods that can provide deeper insights into the model's behavior and the characteristics of the dataset. Chapter 7 delved into the realm of explainable AI, presenting a method for gaining deeper actionable insights into deep learning through second-order explainability. This approach not only helps understand the model better but can also provide actionable insights, allowing for debugging and improvement of the model or dataset. The exploration of second-order explainable AI contributes to the ongoing efforts to make AI

models more transparent and interpretable, enhancing their usability and trustworthiness in real-world applications.

8.1 Future Work

For each major contribution of this thesis, there are several potential directions for future research. Enhancing the diversity of the MetaGraspNet v0 dataset to include more complex and challenging scenarios, such as non-rigid objects, could be a valuable area of focus. Additionally, the current difficulty scoring method is primarily based on a instance segmentation task. Comparing different difficulty scores between multiple computer vision tasks, such as pose estimation, and direct grasp prediction, may be able to provide a more holistic understanding of a scene.

Further refinement and improvement of the keypoint-based and superquadric-based pose estimation methods, particularly in handling objects with uniform or repetitive patterns and addressing issues related to symmetry and proximity of objects, could also be beneficial. Improving the performance of the superquadric-based pose estimation method on novel, out of distribution objects would also be an important step in realizing the full potential of the superquadratic-based object representation format.

The concept of multimodal consistency (MC) introduced in this thesis provides a novel approach to uncertainty estimation, and future work could focus on further refining this approach and exploring other methods for estimating uncertainty in robotic grasping. In particular, it would be important to do a more in depth study comparing different uncertainty estimation methods to this one.

Lastly, while second order explainable AI provides a very useful visualization of features grouped by their distance to each other, possible boundaries of each cluster is not marked automatically and still need to be manually identified. Further automation in detecting, marking, and even possibly describing regions and clusters of interest would be a way to provide further automated insights of a model performance.

References

[1] Yuhao Chen, E Zhixuan Zeng, Maximilian Gilles, and Alexander Wong. Meta-graspnet_v0: A large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis. In *Journal of Computational Vision and Imaging Systems*, 2021.

[2] E Zhixuan Zeng, Yuhao Chen, and Alexander Wong. Investigating use of keypoints for object pose recognition. In *Journal of Computational Vision and Imaging Systems*, 2022.

[3] E Zhixuan Zeng, Yuhao Chen, and Alexander Wong. Shapeshift: Superquadric-based object pose estimation for robotic grasping. In *WICV workshop*, 2023.

[4] Yuhao Chen, Hayden Gunraj, E Zhixuan Zeng, Robbie Meyer, Maximilian Gilles, and Alexander Wong. Mmrnet: Improving reliability for multimodal object detection and segmentation for bin picking via multimodal redundancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 68–77, 2023.

[5] E Zhixuan Zeng, Hayden Gunraj, Sheldon Fernandez, and Alexander Wong. Explaining explainability: Towards deeper actionable insights into deep learning through second-order explainability. In *XAI4CV workshop*, 2023.

[6] Yun Jiang, Stephen Moseson, and Ashutosh Saxena. Efficient grasping from rgbd images: Learning using a new rectangle representation. In *2011 IEEE International Conference on Robotics and Automation*, pages 3304–3311, 2011. doi: 10.1109/ICRA.2011.5980145.

[7] S. Kumra and C. Kanan. Robotic grasp detection using deep convolutional neural networks. In *Proceedings of the International Conference on Intelligent Robots and Systems*, pages 769–776, 2017. doi: 10.1109/IROS.2017.8202237.

[8] Umar Asif, Jianbin Tang, and Stefan Harrer. Graspnet: An efficient convolutional neural network for real-time grasp detection for low-powered devices. In *IJCAI*, volume 7, pages 4875–4882, 2018.

[9] H. Fang, , C. Wang, M. Gou, and C. Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, June 2020. URL http://openaccess.thecvf.com/content_CVPR_2020/papers/Fang_GraspNet-1Billion_A_Large-Scale_Benchmark_for_General_Object_Grasping_CVPR_2020_paper.pdf. Seattle, WA.

[10] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *robotics science and systems*, 2017. doi: 10.15607/RSS.2017.XIII.058.

[11] Xinchen Yan, Jasmined Hsu, Mohammad Khansari, Yunfei Bai, Arkanath Pathak, Abhinav Gupta, James Davidson, and Honglak Lee. Learning 6-dof grasping interaction via deep geometry-aware 3d representations. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3766–3773. IEEE, 2018.

[12] Hanbo Zhang, Xuguang Lan, Site Bai, Xinwen Zhou, Zhiqiang Tian, and Nanning Zheng. Roi-based robotic grasp detection for object overlapping scenes. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4768–4775, 2019. doi: 10.1109/IROS40897.2019.8967869.

[13] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox. Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes. In *Proceedings of the IEEE International Conference on Robotics and Automation*, May 2021. URL https://arxiv.org/abs/2103.14127. Xi'an, China.

[14] D. Morrison, P. Corke, and J. Leitner. Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach. June 2018. URL https://arxiv.org/abs/1804.05172. Pittsburgh, PA.

[15] H. Cao, H. Fang, W. Liu, and C. Lu. Suctionnet-1billion: A large-scale benchmark for suction grasping. *IEEE Robot. Automat. Lett. (RA-L)*, 6(4):8718–8725, 2021.

[16] H. Zhang, Jef Peeters, Eric Demeester, and Karel Kellens. A cnn-based grasp planning method for random picking of unknown objects with a vacuum gripper. *J. of Intell. & Robot. Syst.*, 103(4):1–19, 2021.

[17] Hanbo Zhang, Deyu Yang, Han Wang, Binglei Zhao, Xuguang Lan, and Nanning Zheng. Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter. 2021. doi: 10.48550/ARXIV.2104.14118.

[18] Guoyu Zuo, Jiayuan Tong, Hongxing Liu, Wenbai Chen, and Jianfeng Li. Graph-based visual manipulation relationship reasoning network for robotic grasping. *Frontiers in Neurorobotics*, 15:112, 2021. ISSN 1662-5218. doi: 10.3389/fnbot.2021.719731. URL https://www.frontiersin.org/article/10.3389/fnbot.2021.719731.

[19] Zhaoxin Fan, Yazhi Zhu, Yulin He, Qi Sun, Hongyan Liu, and Jun He. Deep learning on monocular object pose detection and tracking: A comprehensive overview. 2022. doi: 10.1145/3524496.

[20] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012.

[21] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pages 536–551. Springer, 2014.

[22] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. *Lecture Notes in Computer Science*, 2014. doi: 10.1007/978-3-319-10599-4_30.

[23] Anders Glent Buch, Dirk Kraft, Joni-Kristian Kamarainen, Henrik Gordon Petersen, and Norbert Krüger. Pose estimation using local structure-specific shape and appearance context. In *2013 IEEE International Conference on Robotics and Automation*, pages 2080–2087. IEEE, 2013.

[24] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE winter conference on applications of computer vision*, pages 75–82. IEEE, 2014.

[25] Mark Everingham and John Winn. The pascal visual object classes challenge 2012 (voc2012) development kit. *Pattern Analysis, Statistical Modelling and Computational Learning, Tech. Rep*, 8:5, 2011.

[26] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 19–34, 2018.

[27] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020.

[28] M. Sundermeyer, Tomás Hodan, Yann Labbé, Gu Wang, Eric Brachmann, Bertram Drost, C. Rother, and Juan E. Sala Matas. Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. *ArXiv*, 2023. doi: 10.48550/ARXIV.2302.13075.

[29] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012.

[30] B. Çalli, Aaron Walsman, Arjun Singh, S. Srinivasa, P. Abbeel, and A. Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv.org*, 2015. doi: 10.1109/MRA.2015.2448951.

[31] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[32] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017.

[33] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. *arXiv preprint arXiv:2203.05701*, 2022.

[34] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *Proceedings of*

the IEEE/CVF International Conference on Computer Vision Workshops, pages 0–0, 2019.

[35] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE international conference on computer vision*, pages 1521–1529, 2017.

[36] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach. *arXiv preprint arXiv:2011.04307*, 2020.

[37] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10781–10790, 2020.

[38] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018.

[39] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] Shun Iwase, Xingyu Liu, Rawal Khirodkar, Rio Yokota, and Kris M Kitani. Repose: Fast 6d object pose refinement via deep texture rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3303–3312, 2021.

[41] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the european conference on computer vision (ECCV)*, pages 699–715, 2018.

[42] Nuno Pereira and Luís A Alexandre. Maskedfusion: Mask-based 6d object pose estimation. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 71–78. IEEE, 2020.

[43] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020.

[44] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.

[45] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

[47] Bowen Cheng, Bin Xiao, Jingdong Wang, Honghui Shi, Thomas S Huang, and Lei Zhang. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5386–5395, 2020.

[48] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Readings in Computer Vision*, pages 726–740, 1987. URL http://dx.doi.org/10.1016/B978-0-08-051581-6.50070-2.

[49] Tuo Cao, Fei Luo, Yanping Fu, Wenxiao Zhang, Shengjie Zheng, Chunxia Xiao, Tuo Cao, Fei Luo, Yanping Fu, Wenxiao Zhang, Shengjie Zheng, and Chunxia Xiao. Dgecn: A depth-guided edge convolutional network for end-to-end 6d pose estimation. 2022. doi: 10.1109/CVPR52688.2022.00376.

[50] Pedro Castro, Tae-Kyun Kim, Pedro Castro, and Tae-Kyun Kim. Crt-6d: Fast 6d object pose estimation with cascaded refinement transformers. 2023. doi: 10.1109/WACV56688.2023.00570.

[51] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021.

[52] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017.

[53] Alexander Grabner, Peter M. Roth, and Vincent Lepetit. 3d pose estimation and 3d model retrieval for objects in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[54] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 292–301, 2018.

[55] Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Making deep heatmaps robust to partial occlusions for 3d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134, 2018.

[56] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019.

[57] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020.

[58] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3003–3013, 2021.

[59] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 431–440, 2020.

[60] Yangzheng Wu, Mohsen Zand, A. Etemad, and M. Greenspan. Vote from the center: 6 dof pose estimation in rgb-d images by radial keypoint voting. *ArXiv*, 2021.

[61] Yan Di, Fabian Manhardt, Gu Wang, Xiangyang Ji, Nassir Navab, and Federico Tombari. So-pose: Exploiting self-occlusion for direct 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12396–12405, 2021.

[62] Bo Chen, Tat-Jun Chin, and Marius Klimavicius. Occlusion-robust object pose estimation with holistic representation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2929–2939, 2022.

[63] Heng Yang and M. Pavone. Object pose estimation with statistical guarantees: Conformal keypoint detection and geometric uncertainty propagation. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2303.12246.

[64] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7678–7687, 2019.

[65] Omid Hosseini Jafari, Siva Karthik Mustikovela, Karl Pertsch, Eric Brachmann, and Carsten Rother. ipose: instance-aware 6d pose estimation of partly occluded objects. In *Asian Conference on Computer Vision*, pages 477–492. Springer, 2018.

[66] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7668–7677, 2019.

[67] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019.

[68] Tomas Hodan, Daniel Barath, and Jiri Matas. Epos: Estimating 6d pose of objects with symmetries. 2020. doi: 10.1109/CVPR42600.2020.01172.

[69] Hansheng Chen, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li. Epro-pnp: Generalized end-to-end probabilistic perspective-n-points for monocular object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2781–2790, 2022.

[70] Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, Hongsheng Li, Yan Xu, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li. Rnnpose: Recurrent 6-dof object pose refinement with robust correspondence field estimation and pose optimization. 2022. doi: 10.1109/CVPR52688.2022.01446.

[71] Yongzhi Su, Mahdi Saleh, Torben Fetzer, J. Rambach, N. Navab, Benjamin Busam, D. Stricker, and F. Tombari. Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation. *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00662.

[72] Georgios Pavlakos, Xiaowei Zhou, Aaron Chan, Konstantinos G Derpanis, and Kostas Daniilidis. 6-dof object pose from semantic keypoints. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 2011–2018. IEEE, 2017.

[73] Jogendra Nath Kundu, MV Rahul, Aditya Ganeshan, and R Venkatesh Babu. Object pose estimation from monocular image using multi-view keypoint correspondence. In *European Conference on Computer Vision*, pages 298–313. Springer, 2018.

[74] Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, Guoquan Huang, Nathaniel Merrill, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang. Symmetry and uncertainty-aware object slam for 6dof object pose estimation. 2022. doi: 10.1109/CVPR52688.2022.01448.

[75] Mark Robson and Mohan Sridharan. A keypoint-based object representation for generating task-specific grasps. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, pages 374–381. IEEE, 2022.

[76] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *The International Symposium of Robotics Research*, pages 132–157. Springer, 2019.

[77] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. 2019. doi: 10.1109/CVPR.2019.00275.

[78] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. *computer vision and pattern recognition*, 2020. doi: 10.1109/CVPR42600.2020.01199.

[79] Yang Fu and X. Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *Neural Information Processing Systems*, 2022. doi: 10.48550/ARXIV.2206.15436.

[80] Xingyu Liu, Gu Wang, Yi Li, and Xiangyang Ji. Catre: Iterative point clouds alignment for category-level object pose refinement. *European Conference on Computer Vision*, 2022. doi: 10.48550/ARXIV.2207.08082.

[81] Yan Di, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, N. Navab, and F. Tombari. Gpv-pose: Category-level object pose estimation via geometry-guided point-wise voting. *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00666.

[82] Ruida Zhang, Yan Di, Zhiqiang Lou, Fabian Manhardt, N. Navab, F. Tombari, and Xiangyang Ji. Rbp-pose: Residual bounding box projection for category-level pose estimation. *ArXiv*, 2022. doi: 10.48550/ARXIV.2208.00237.

[83] Wufei Ma, Angtian Wang, A. Yuille, and Adam Kortylewski. Robust category-level 6d pose estimation with coarse-to-fine rendering of neural features. *European Conference on Computer Vision*, 2022. doi: 10.48550/ARXIV.2209.05624.

[84] Jiahao Yang, Wufei Ma, Angtian Wang, Xiaoding Yuan, A. Yuille, and Adam Kortylewski. Robust category-level 3d pose estimation from synthetic data. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2305.16124.

[85] Artur Jesslen, Guofeng Zhang, Angtian Wang, A. Yuille, and Adam Kortylewski. Robust 3d-aware object classification via discriminative render-and-compare. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2305.14668.

[86] Guanglin Li, Yifeng Li, Zhichao Ye, Qihang Zhang, Tao Kong, Zhaopeng Cui, and Guofeng Zhang. Generative category-level shape and pose estimation with semantic primitives. *Conference on Robot Learning*, 2022. doi: 10.48550/ARXIV.2210.01112.

[87] Yang Xiao, Xuchong Qiu, Pierre-Alain Langlois, Mathieu Aubry, and Renaud Marlet. Pose from shape: Deep pose estimation for arbitrary 3d objects. *arXiv preprint arXiv:1906.05105*, 2019.

[88] Yuan Liu, Yilin Wen, Sida Peng, Chu-Hsing Lin, Xiaoxiao Long, T. Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. *European Conference on Computer Vision*, 2022. doi: 10.48550/ARXIV.2204.10776.

[89] Yann Labb'e, Lucas Manuelli, Arsalan Mousavian, Stephen Tyree, Stan Birchfield, Jonathan Tremblay, Justin Carpentier, Mathieu Aubry, D. Fox, and Josef Sivic. Megapose: 6d pose estimation of novel objects via render & compare. *Conference on Robot Learning*, 2022. doi: 10.48550/ARXIV.2212.06870.

[90] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad

models. *Computer Vision and Pattern Recognition*, 2022. doi: 10.1109/CVPR52688.2022.00670.

[91] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, H. Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *ArXiv*, 2023. doi: 10.48550/ARXIV.2301.07673.

[92] Jason Y. Zhang, D. Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. *European Conference on Computer Vision*, 2022. doi: 10.48550/ARXIV.2208.05963.

[93] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 1987. doi: 10.1037/0033-295X.94.2.115.

[94] R. Rusu, Nico Blodow, Zoltán-Csaba Márton, and M. Beetz. Close-range scene segmentation and reconstruction of 3d point cloud maps for mobile manipulation in domestic environments. *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009. doi: 10.1109/IROS.2009.5354683.

[95] Yunzhi Lin, Chao Tang, Fu-Jen Chu, and Patricio A. Vela. Using synthetic data and deep networks to recognize primitive shapes for object grasping. 2020. doi: 10.1109/ICRA40945.2020.9197256.

[96] Yunzhi Lin, Chao Tang, Fu-Jen Chu, Ruinian Xu, and P. Vela. Primitive shape recognition for object grasping. *arXiv.org*, 2022.

[97] Takuya Torii and M. Hashimoto. Model-less estimation method for robot grasping parameters using 3d shape primitive approximation. *2018 IEEE 14th International Conference on Automation Science and Engineering (CASE)*, 2018. doi: 10.1109/COASE.2018.8560417.

[98] Gopal Sharma, Difan Liu, Subhransu Maji, Evangelos Kalogerakis, Siddhartha Chaudhuri, and Radomír Měch. Parsenet: A parametric surface fitting network for 3d point clouds. *european conference on computer vision*, 2020. doi: 10.1007/978-3-030-58571-6_16.

[99] Yang Zhou, K. Yin, Hui Huang, Hao Zhang, Minglun Gong, and D. Cohen-Or. Generalized cylinder decomposition. *ACM Transactions on Graphics*, 2015. doi: 10.1145/2816795.2818074.

[100] Ali Abdollahzadeh, Alejandra Sierra, and Jussi Tohka. Cylindrical shape decomposition for 3d segmentation of tubular objects. *IEEE Access*, 2021. doi: 10.1109/ACCESS.2021.3056958.

[101] Hao Jiang and Jianxiong Xiao. A linear approach to matching cuboids in rgbd images. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 2013. doi: 10.1109/CVPR.2013.282.

[102] Shubham Tulsiani, Hao Su, Leonidas J Guibas, Alexei A Efros, and Jitendra Malik. Learning shape abstractions by assembling volumetric primitives. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2635–2643, 2017.

[103] Chuhang Zou, Ersin Yumer, Jimei Yang, Duygu Ceylan, and Derek Hoiem. 3d-prnn: Generating shape primitives with recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 900–909, 2017.

[104] Chun-Yu Sun, Qian-Fang Zou, Xin Tong, and Yang Liu. Learning adaptive hierarchical cuboid abstractions of 3d shape collections. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019.

[105] Kaizhi Yang and Xuejin Chen. Unsupervised learning for cuboid shape abstraction via joint segmentation from point clouds. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021.

[106] Qian He, Desen Zhou, Bo Wan, and Xuming He. Single image 3d object estimation with primitive graph networks. 2021. doi: 10.1145/3474085.3475398.

[107] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019.

[108] Theo Deprelle, Thibault Groueix, Matthew Fisher, Vladimir Kim, Bryan Russell, and Mathieu Aubry. Learning elementary structures for 3d shape generation and matching. *Advances in Neural Information Processing Systems*, 32, 2019.

[109] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy J Mitra, and Leonidas J Guibas. Structedit: Learning structural shape variations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8859–8868, 2020.

[110] Alan H Barr. Superquadrics and angle-preserving transformations. *IEEE Computer graphics and Applications*, 1(1):11–23, 1981.

[111] F. Solina and R. Bajcsy. Recovery of parametric models from range images: The case for superquadrics with global deformations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1990. doi: 10.1109/34.44401.

[112] Aleš Jaklič, Aleš Leonardis, and Franc Solina. Segmentation and recovery of superquadrics. 2000.

[113] Despoina Paschalidou, Ali Osman Ulusoy, and Andreas Geiger. Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10344–10353, 2019.

[114] Weixiao Liu, Yuwei Wu, Sipu Ruan, and Gregory S Chirikjian. Robust and accurate superquadric recovery: a probabilistic approach. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2676–2685, 2022.

[115] Yuwei Wu, Weixiao Liu, Sipu Ruan, and Gregory S Chirikjian. Primitive-based shape abstraction via nonparametric bayesian inference. *arXiv preprint arXiv:2203.14714*, 2022.

[116] Weixiao Liu, Yuwei Wu, Sipu Ruan, and G. Chirikjian. Marching-primitives: Shape abstraction from signed distance function. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2303.13190.

[117] Despoina Paschalidou, Luc Van Gool, and Andreas Geiger. Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1060–1070, 2020.

[118] Ryo Hachiuma and Hideo Saito. Pose estimation of primitive-shaped objects from a depth image using superquadric representation. *Applied Sciences*, 2020. doi: 10.3390/APP10165442.

[119] Tim Oblak, Jaka Šircelj, Vitomir Štruc, Peter Peer, Franc Solina, and Aleš Jaklič. Learning to predict superquadric parameters from depth images with explicit and implicit supervision. *IEEE Access*, 9:1087–1102, 2021. doi: 10.1109/ACCESS.2020.3041584.

[120] Abhijit Makhal, Federico Thomas, and Alba Perez Gracia. Grasping unknown objects in clutter by superquadric representation. In *2018 Second IEEE International Conference on Robotic Computing (IRC)*, pages 292–299. IEEE, 2018.

[121] G. Vezzani, U. Pattacini, Giulia Pasquale, and L. Natale. Improving superquadric modeling and grasping with prior on object shapes. *IEEE International Conference on Robotics and Automation*, 2018. doi: 10.1109/ICRA.2018.8463161.

[122] Yuwei Wu, Weixiao Liu, Zhiyang Liu, and G. Chirikjian. Learning-free grasping of unknown objects using hidden superquadrics. *arXiv.org*, 2023. doi: 10.48550/ARXIV.2305.06591.

[123] Sebastian Thrun. Probabilistic robotics. 2005.

[124] Hugo Grimmett, Rudolph Triebel, Rohan Paul, and I. Posner. Introspective classification for robot perception. *Int. J. Robotics Res.*, 2016. doi: 10.1177/0278364915587924.

[125] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: representing model uncertainty in deep learning. 2016.

[126] Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? does it matter? *Structural safety*, 31(2):105–112, 2009.

[127] J. Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseo Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, P. Jung, R. Roscher, M. Shahzad, Wen Yang, R. Bamler, and Xiaoxiang Zhu. A survey of uncertainty in deep neural networks. *arXiv.org*, 2021.

[128] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision. 2017.

[129] Geoffrey E. Hinton and D. Camp. Keeping the neural networks simple by minimizing the description length of the weights. *COLT '93*, 1993. doi: 10.1145/168304.168306.

[130] Alex Graves. Practical variational inference for neural networks. *neural information processing systems*, 2011.

[131] Radford M. Neal. An improved acceptance procedure for the hybrid monte carlo algorithm. *Journal of Computational Physics*, 1994. doi: 10.1006/JCPH.1994.1054.

[132] Shengyang Sun, Changyou Chen, and L. Carin. Learning structured weight uncertainty in bayesian neural networks. *AISTATS*, 2017.

[133] Christopher Nemeth and Paul Fearnhead. Stochastic gradient markov chain monte carlo. *Journal of the American Statistical Association*, 2021. doi: 10.1080/01621459.2020.1847120.

[134] Jongseok Lee, Matthias Humt, Jianxiang Feng, and Rudolph Triebel. Estimating model uncertainty of neural network in sparse information form. *international conference on machine learning*, 2020.

[135] Alexander Pritzel, Charles Blundell, and Balaji Lakshminarayanan. Simple and scalable predictive uncertainty estimation using deep ensembles. *NIPS*, 2017.

[136] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. 2020.

[137] George D.C. Cavalcanti, Luiz S. Oliveira, Thiago J.M. Moura, and Guilherme V. Carvalho. Combining diversity measures for ensemble pruning. *Pattern Recognition Letters*, 2016. doi: 10.1016/J.PATREC.2016.01.029.

[138] Huaping Guo, Hongbing Liu, Ran Li, Changan Wu, Yibo Guo, and Mingliang Xu. Margin & diversity based ordering ensemble pruning. *Neurocomputing*, 275:237–246, 2018.

[139] Jakob Lindqvist, Amanda Olmin, Fredrik Lindsten, and Lennart Svensson. A general framework for ensemble distribution distillation. *international workshop on machine learning for signal processing*, 2020. doi: 10.1109/MLSP49062.2020.9231703.

[140] Waldyn G. Martinez. Ensemble pruning via quadratic margin maximization. *IEEE Access*, 2021. doi: 10.1109/ACCESS.2021.3062867.

[141] Joost van Amersfoort, Lewis Smith, Yee Whye Teh, and Yarin Gal. Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network. *international conference on machine learning*, 2020.

[142] Andrey Malinin and Mark Gales. Predictive uncertainty estimation via prior networks. *neural information processing systems*, 2018. doi: 10.17863/CAM.35237.

[143] Philipp Oberdiek, Matthias Rottmann, and Hanno Gottschalk. Classification uncertainty of deep neural networks based on gradient information. *artificial neural networks in pattern recognition*, 2018. doi: 10.1007/978-3-319-99978-4_9.

[144] Jinsol Lee and Ghassan AlRegib. Gradients as a measure of uncertainty in neural networks. 2020. doi: 10.1109/ICIP40778.2020.9190679.

[145] Maithra Raghu, Katy Blumer, Rory Abbott Sayres, Ziad Obermeyer, Robert Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. *international conference on machine learning*, 2019.

[146] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[147] David JC MacKay. A practical bayesian framework for backpropagation networks. *Neural computation*, 4(3):448–472, 1992.

[148] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[149] Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209, 2019.

[150] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

[151] Dumitru Erhan, Aaron Courville, and Yoshua Bengio. Understanding representations learned in deep architectures. 2010.

[152] Judy Borowski, Roland Simon Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. Exemplary natural images explain {cnn} activations better than state-of-the-art feature visualization. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=QO9-y8also-.

[153] Anh Nguyen, Jason Yosinski, and Jeff Clune. Understanding neural networks via feature visualization: A survey. In *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 55–76. Springer, 2019.

[154] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 3634–3642, May 2020. URL https://doi.org/10.1109/ICRA40945.2020.9197518. Paris, France.

[155] Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship network for autonomous robotics. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 118–125, 2018. doi: 10.1109/HUMANOIDS.2018.8625071.

[156] S. Panda, A. H. A. Hafez, and C. V. Jawahar. Single and multiple view support order prediction in clutter for manipulation. *Journal of Intelligent & Robotic Systems*, 83: 179–203, 2016. doi: 10.1007/s10846-015-0330-z. URL https://doi.org/10.1007/s10846-015-0330-z.

[157] H. Zhang, D. Yang, H. Wang, B. Zhao, X. Lan, J. Ding, and N. Zheng. Regrad: A large-scale relational grasp dataset for safe and object-specific robotic grasping in clutter. *https://arxiv.org/abs/2104.14118*, April 2021. URL https://arxiv.org/abs/2104.14118.

[158] kaggle. kaggle. URL https://https://www.kaggle.com/. (accessed Nov-17-2021).

[159] NVIDIA. DEVELOP WITH NVIDIA OMNIVERSE. URL https://developer.nvidia.com/nvidia-omniverse-platform. (accessed Nov-17-2021).

[160] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M. Dollar. Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set. *IEEE Robotics Automation Magazine*, 22(3):36–52, 2015. doi: 10.1109/MRA.2015.2448951.

[161] Mark Robson and Mohan Sridharan. A keypoint-based object representation for generating task-specific grasps. In *2022 IEEE International Conference on Automation Science and Engineering*, 2022.

[162] Yu Xiang, Roozbeh Mottaghi, and Silvio Savarese. Beyond pascal: A benchmark for 3d object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

[163] Yuhao Chen, Maximilian Gilles, E Zhixuan Zeng, and Alexander Wong. Metagraspnet: A large-scale benchmark dataset for vision-driven robotic grasping via physics-based metaverse synthesis. In *2022 IEEE International Conference on Automation Science and Engineering*, 2022.

[164] MMPose Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2020.

[165] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[166] Jeffrey Ichnowski, Yahav Avigal, Yi Liu, and Ken Goldberg. Gomp-fit: Grasp-optimized motion planning for fast inertial transport. *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 5255–5261, 2022.

[167] S. Back, J. Lee, T. Kim, S. Noh, R. Kang, S. Bak, and K. Lee. Unseen object amodal instance segmentation via hierarchical occlusion modeling. *arXiv preprint arXiv:2109.11103*, 2021.

[168] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2980–2988, October 2017. doi: 10.1109/iccv.2017.322. Venice, Italy.

[169] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. *arXiv preprint arXiv:1506.02640*, May 2016. URL http://arxiv.org/abs/1506.02640.

[170] P. Santhanam, E. Farchi, and V. Pankratius. Engineering reliable deep learning systems. *arXiv preprint arXiv:1910.12582*, 2019.

[171] H. Xu, J. Blanchet, M. P. Gerardo-Castro, and S. Paudel. Measuring reliability of object detection algorithms for automated driving perception tasks. In *Proceedings of the Winter Simulation Conference*, pages 1–12, 2021. doi: 10.1109/WSC52266.2021.9715295.

[172] F. F. dos Santos, P. Navaux, L. Carro, and P. Rech. Impact of reduced precision in the reliability of deep neural networks for object detection. In *IEEE European Test Symposium*, pages 1–6, 2019. doi: 10.1109/ETS.2019.8791554.

[173] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher, et al. A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*, 2021.

[174] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng. Multimodal deep learning. In *Proceedings of the International Conference on Machine Learning*, pages 689–696, June 2011. URL https://icml.cc/2011/papers/399_icmlpaper.pdf.

[175] G. Joshi, R. Walambe, and K. Kotecha. A review on explainability in multimodal deep neural nets. *IEEE Access*, 9:59800–59821, 2021. doi: 10.1109/ACCESS.2021.3070212.

[176] D. Ramachandram and G. W. Taylor. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine*, 34(6):96–108, 2017. doi: 10.1109/MSP.2017.2738401.

[177] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Gläser, F. Timm, W. Wiesbeck, and K. Dietmayer. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges. *IEEE Transactions on Intelligent Transportation Systems*, 22(3):1341–1360, 2021. doi: 10.1109/TITS.2020.2972974.

[178] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang. What makes multi-modal learning better than single (provably). In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 10944–10956, 2021. URL https://proceedings.neurips.cc/paper/2021/file/5aa3405a3f865c10f420a4a7b55cbff3-Paper.pdf.

[179] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27, 2014. URL https://proceedings.neurips.cc/paper/2014/file/00ec53c4682d36f5c4359f4ae7bd7ba1-Paper.pdf.

[180] A. J. Smola and B. Schölkopf. A tutorial on support vector regression. *Statistics and Computing*, 14(3):199–222, 2004. doi: 10.1023/B:STCO.0000.

[181] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang. Deep multimodal fusion by channel exchanging. *Proceedings of the Conference on Neural Information Processing Systems*, December 2020.

[182] S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the International Conference on Machine Learning*, pages 448–456, 2015. URL http://jmlr.org/proceedings/papers/v37/ioffe15.pdf.

[183] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 7088–7097, October 2021.

[184] Z. Xue and R. Marculescu. Dynamic multimodal fusion. *arXiv preprint arXiv:2204.00102*, 2022.

[185] T. Lin, P Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. doi: 10.1109/CVPR.2017.106. Honolulu, HI.

[186] Z. Zhou. *Machine Learning*. Springer Singapore, 08 2021. ISBN 9789811519673.

[187] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[188] David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80023-1.

[189] Y. Freund, R. E Schapire, et al. Experiments with a new boosting algorithm. In *Proceedings of the International Conference on Machine Learning*, volume 96, pages 148–156, 1996.

[190] A. Marinoni, S. Chlaily, E. Khachatrian, T. Eltoft, S. Selvakumaran, M. Girolami, and C. Jutten. Enhancing ensemble learning and transfer learning in multimodal data analysis by adaptive dimensionality reduction. *arXiv preprint arXiv:2105.03682*, 2021.

[191] S. S. Menon and K. Krishnamurthy. Multimodal ensemble deep learning to predict disruptive behavior disorders in children. *Frontiers in Neuroinformatics*, 15, 2021. ISSN 1662-5196. doi: 10.3389/fninf.2021.742807.

[192] V. Chordia and V. Kumar. Large scale multimodal classification using an ensemble of transformer models and co-attention. In *Proceedings of the SIGIR Workshop On eCommerce*, 2020.

[193] Q. Zhou, H. Liang, Z. Lin, and K. Xu. Multimodal feature fusion for video advertisements tagging via stacking ensemble. *arXiv preprint arXiv:2108.00679*, 2021.

[194] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S Huang. Free-form image inpainting with gated convolution. *arXiv preprint arXiv:1806.03589*, 2018.

[195] M. Bertalmio, A.L. Bertozzi, and G. Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–I, 2001. doi: 10.1109/CVPR.2001.990497.

[196] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, June 2016. Las Vegas, NV.

[197] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, December 2015.

[198] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017.

[199] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, pages 3319–3328. JMLR.org, 2017.

[200] Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott M. Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients. *Nature Machine Intelligence*, 3:620–631, 2021. doi: 10.1038/s42256-021-00343-w.

[201] Marco Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 97–101, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-3020. URL https://aclanthology.org/N16-3020.

[202] Zhong Qiu Lin, Mohammad Javad Shafiee, Stanislav Bochkarev, Michael St. Jules, Xiao Yu Wang, and Alexander Wong. Do explanations reflect decisions? a machine-centric strategy to quantify the performance of explainability algorithms. *arXiv preprint. arXiv:1910.07387*, 2019.

[203] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *30th International Conference on Neural Information Processing Systems (NIPS 2017)*, pages 768–4777, 2017.

[204] Hayden Gunraj, Paul Guerrier, Sheldon Fernandez, and Alexander Wong. SolderNet: Towards trustworthy visual inspection of solder joints in electronics manufacturing using explainable artificial intelligence. In *35th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI-23)*. Association for the Advancement of Artificial Intelligence (AAAI), 2023.

[205] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[206] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009.

[207] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.