# VISUAL 6D OBJECT POSE ESTIMATION AND TRACKING

By

LINFANG ZHENG

A thesis submitted to
the University of Birmingham
for the degree of
DOCTOR OF PHILOSOPHY


Human-Centred Visual Learning Group
School of Computer Science
College of Engineering and Physical Sciences
University of Birmingham
April 2024

## ABSTRACT

Visual 6D object pose estimation and tracking is crucial for enabling machines to understand and interact with the three-dimensional world. Despite significant advancements, challenges persist in this field. Existing methods often struggle in complex real-world scenarios, particularly with symmetric and textureless objects under occlusion. Category-level methods also face limitations in generalizability, especially with complex-shaped objects and noisy environments. Additionally, there is a significant gap in effective methods for category-level object pose refinement, which is crucial for achieving high-precision pose information with previously unseen objects. To address these challenges, this thesis proposes three approaches. First, an instance-level object pose tracking method is introduced, leveraging temporal information with augmented autoencoder-based reconstruction to enhance robustness to symmetric and textureless objects under occlusion. Second, a hybrid scope feature extraction layer (HS-layer) is presented for category-level object pose estimation. The HS layer encodes translation and scale information, perceives geometric structural information for handling complex-shaped objects with robustness to outliers. Lastly, a method that combines latent geometric feature extraction and learnable affine transformation is proposed to address shape discrepancy issues in category-level object pose refinement, improving pose refinement accuracy and generalizability. Extensive experiments validate the effectiveness of these approaches in advancing practical applications of visual 6D object pose estimation and tracking. Additionally, all the proposed methods exhibit real-time performance, crucial for real-life applications.

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my perfect supervisors, in alphabetical order, Dr. Aleš Leonardis, Dr. Hyung Jin Chang, and Dr. Wei Zhang. Their guidance and support have been invaluable throughout my PhD journey. Aleš' insightful questions have challenged me to delve deeper into the core of my research, while Hyung Jin's meticulous attention to detail has greatly enhanced the quality of my work. Wei's guidance and encouragement have been a constant source of reassurance, especially during moments of uncertainty. I consider myself incredibly fortunate to have had the opportunity to work under their supervision.

I am also indebted to Dr. Hua Chen for his valuable suggestions and advice on my research and paper writing skills. My heartfelt thanks also go to my co-authors, Chen Wang, Yinghan Sun, Tze Ho Elden Tse, Zhongqun Zhang, Esha Dasgupta, and Nora Horanyi, with whom I have collaborated closely on various projects. Their contributions have been instrumental in the success of our joint efforts.

I would like to acknowledge the members of my Research Student Monitoring Group, Dr. Iain Styles and Dr. Jinming Duan, for their guidance and support in monitoring my research progress. I am also grateful to the school administration team for their assistance in ensuring that my research stayed on track.

Special thanks are due to Dr. Xinxin Guo for her unwavering support and encouragement during challenging times. I am also grateful to my friends, Xianhui Shao, Huan He, and Haotian Fan, for their constant encouragement and companionship throughout my PhD journey.

I would like to extend my thanks to all my labmates, whose support and friendship made the

journey of studying for a PhD much more joyful. Among them, I would like to offer a special thank you to Yuan Wang for her invaluable assistance in various aspects of lab issues. Her support has been instrumental in overcoming many challenges during my research.

Last but not least, I would like to thank my family for their full support and encouragement throughout this journey.

# PUBLICATIONS

## Publications included in thesis

Zheng, Linfang, Aleš Leonardis, Tze Ho Elden Tse, Nora Horanyi, Hua Chen, Wei Zhang, and Hyung Jin Chang (2022). "TP-AE: Temporally Primed 6D Object Pose Tracking with Auto-Encoders". In: *2022 IEEE International Conference on Robotics and Automation (ICRA)*.

Zheng, Linfang, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang (June 2024). "GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pose Refinement". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10693–10703.

Zheng, Linfang, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang (June 2023). "HS-Pose: Hybrid Scope Feature Extraction for Category-Level Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17163–17173.

## Other publications

Chen, Hua, Linfang Zheng, and Wei Zhang (2020). "Optimal Control Inspired Q-Learning for Switched Linear Systems". In: *2020 American Control Conference (ACC)*, pp. 4003–4010. DOI: 10.23919/ACC45564.2020.9147818.

Horanyi, Nora, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang (June 2023). "Where Are They Looking in the 3D Space?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2678–2687.

Sun, Yinghan, Linfang Zheng, Hua Chen, and Wei Zhang (2023). *Multi-Resolution Planar Region Extraction for Uneven Terrains*. arXiv: 2311.12562 [cs.CV].

Zhang, Zhongqun, Wei Chen, Linfang Zheng, Aleš Leonardis, and Hyung Jin Chang (2023). "Trans6D: Transformer-Based 6D Object Pose Estimation and Refinement". In: *Computer Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, pp. 112–128. ISBN: 978-3-031-25085-9.

# Contents

# List of Figures

# List of Tables

# Chapter One

# Introduction

*Visual 6D object pose estimation and tracking* is a fundamental field of computer vision, where we aim to design algorithms for machines to perceive and understand the three-dimensional (3D) world around them. This process involves determining both the position (where an object is located) and orientation (how it is rotated) of the target object in 3D space using visual information, as illustrated in Figure 1.1. Additionally, for methods dealing with objects with unknown sizes, the object's 3D size of its tight bounding box is also estimated. The visual data is usually obtained from perception sensors such as RGB cameras, depth cameras, RGB-Depth (RGB-D) cameras, or Lidars.

While we may not always be aware of it, humans regularly rely on the 6D poses of objects in their daily lives. Understanding an object's 6D pose is immensely beneficial, aiding in tasks such as grasping objects, avoiding collisions, navigating surroundings, and interacting with tools. Similarly, for machines, comprehending object poses is crucial for effective and efficient interaction and manipulation within their environments, making it a key area of research for applications ranging from robotics to augmented reality systems.

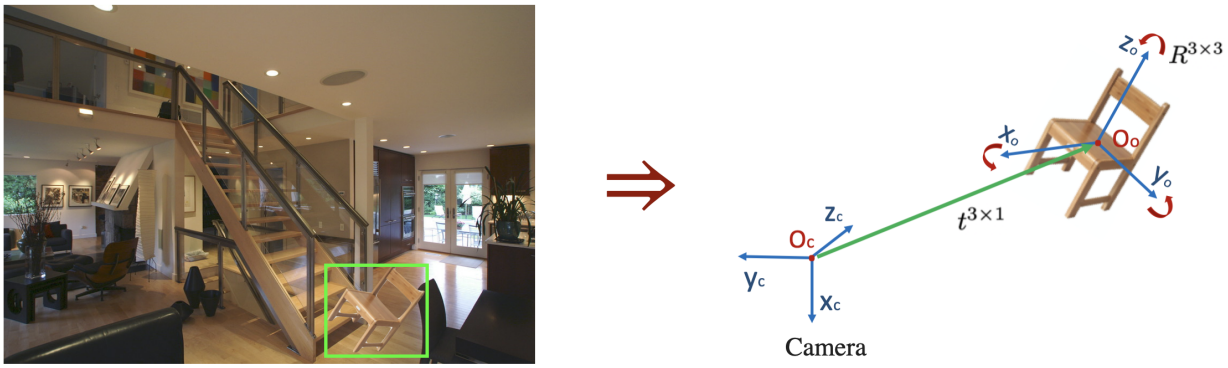



Figure 1.1: Illustration of 6D object pose estimation task

Today, machines that utilize 6D object poses are commonplace in our lives. For example, Augmented Reality (AR) depends on accurate object poses to improve user experiences. In AR, digital information is overlaid onto the real world, and knowing object poses ensures that virtual objects blend seamlessly into the environment. In manufacturing and industrial settings, industrial automation machines use object poses for tasks such as quality control, sorting, and automated assembly. Autonomous vehicles and drones rely on object pose information for navigation and localization, aiding them in understanding their position relative to objects in the environment and avoiding collisions. Robots utilize this information to plan and execute actions such as grasping, manipulation, and assembly effectively. Object pose information is also valuable in surveillance, human-computer interaction, and various other tracking applications for monitoring objects as they move through a scene.

Despite the significant enhancements brought by the integration of 6D object pose estimation and tracking into various machines and systems, notable limitations in existing 6D pose estimators and trackers remain. Current methods can struggle to accurately estimate object pose in challenging situations, especially when the target object exhibits symmetric, with insufficient surface textures, and is partially occluded. Another challenge is the algorithms' ability to generalize to objects not encountered during training, as well as their resistance to noise. Everyday objects often have shapes that differ from those used in training, and acquiring low-noise data typically

demands expensive sensors, limiting their deployment in everyday consumer devices. Moreover, there is a lack of effective refinement methods to improve the accuracy of pose estimation for previously unseen objects, which is essential for high-precision applications or when initial pose estimations are noisy. Improving the robustness and generalization capabilities of pose estimation algorithms is crucial for advancing the field and enabling wider use of machines that rely on 6D object pose information.

In the upcoming sections, I will first discuss the motivation behind my research, highlighting the practical applications of visual 6D object pose estimation and tracking. I will then outline the key qualities necessary for these methods to meet real-world demands. Afterward, I will present three significant challenges in this field, and explain how I tackled these challenges in my work. Finally, I'll clearly outline what my research has contributed and give an overview of the entire thesis.

## 1.1 Motivation

The motivation for research in visual 6D object pose estimation and tracking arises from its profound impact on a wide array of practical applications, spanning across industries and playing a crucial role in advancing technology and enhancing daily life experiences.

**Autonomous Driving** One of the most prominent areas benefiting from 6D pose estimation is autonomous driving (Kothari et al., 2017; Hoque et al., 2023). As autonomous vehicles become more prevalent, the need for precise object detection and tracking becomes increasingly critical. Accurate 6D pose estimation enables vehicles to detect and react to objects and obstacles in their surroundings, ensuring safe navigation through complex environments like bustling city streets and fast-paced highways. By providing vehicles with the ability to precisely determine the position

and orientation of objects, such as fallen debris or other vehicles, 6D pose estimation enhances the safety and reliability of autonomous driving systems.

**Augmented Reality (AR)** In the realm of augmented reality (AR), accurate object pose estimation is essential for creating immersive and interactive experiences (Y. Su, Rambach, et al., 2019). AR applications, such as games, retail, and healthcare experiences, rely on 6D pose estimation to seamlessly integrate virtual elements into the real world. This technology has the potential to revolutionize industries by offering innovative and engaging experiences. For example, in AR games like Pokémon GO, precise pose estimation ensures that virtual creatures interact realistically with the environment. Similarly, in retail, companies use AR applications to allow customers to visualize products in their own space before making a purchase decision. In healthcare, AR is used for medical training and surgical planning, enabling precise overlay of virtual anatomy models onto the surgical environment.

**Robotics** The field of robotics is another area where 6D object pose estimation plays a crucial role (Collet, Martinez, and Srinivasa, 2011; J. Liu et al., 2023). Robots equipped with accurate pose estimation capabilities can perform a wide range of tasks with precision and flexibility. From household chores to healthcare and entertainment, robots can interact effectively with their environment by leveraging object poses. For example, in household settings, robots can use object pose estimation to pick up and move objects, assisting with tasks such as cleaning or organizing. In healthcare, robots can support medical professionals by preparing surgical tools or transporting medical supplies. Additionally, in entertainment, robots can engage in interactive games such as Treasure Hunts with humans, providing new forms of entertainment and engagement.

**Manufacturing** Utilizing 6D object pose estimation in manufacturing provides multiple benefits (Liang et al., 2021), particularly in assembly processes and quality control. Precise pose

estimation allows for visual inspections of component positions and orientations before assembly, ensuring precise assembly with minimal errors. This results in reduced component damage rates and higher-quality outputs. Moreover, in electronics manufacturing, machines using visual 6D object pose estimation can accurately place tiny components on circuit boards, thereby enhancing product reliability. Quality control processes also benefit from pose estimation, as it allows for inspections of component alignments to ensure they meet quality standards.

**Warehouse Automation** In the context of warehouse automation, 6D pose estimation is essential for efficient and accurate object manipulation (D.-C. Hoang, Stoyanov, and Lilienthal, 2019; V.-D. Vu et al., 2024). As e-commerce continues to grow, automated warehouse solutions rely on robots to pick, pack, and transport items. Accurate pose estimation ensures that robots can grasp and manipulate objects reliably, even in densely packed environments. By enhancing pose estimation algorithms, warehouse operations can be optimized, reducing errors and increasing efficiency in logistics processes.

**Security and Surveillance** 6D object pose estimation plays a vital role in security and surveillance applications by accurately tracking the pose of objects in the environment (Vidanage, Fernando, Abeywardhana, et al., 2023). Surveillance systems equipped with 6D pose estimation capabilities can enhance situational awareness and improve threat detection. For example, in public spaces, surveillance cameras can use pose estimation to track the movements of suspicious objects, such as unattended vehicles, alerting security personnel to potential threats.

In conclusion, the applications of 6D object pose estimation and tracking are vast and impactful, ranging from enhancing the safety and efficiency of autonomous vehicles to revolutionizing industries like AR, robotics, warehouse automation, manufacturing, and security and surveillance. These applications highlight the importance of continued research and development in the field of 6D object pose estimation, as it continues to shape the future of technology and improve

the quality of life for people around the world.

## 1.2 Key qualities for real-world pose estimation and tracking

In light of these applications, it raises the crucial question: what qualities should an ideal 6D pose estimator or tracker possess to meet the demands of real-life applications? To answer this question, we examine several key characteristics that are essential for practical deployment in a wide range of scenarios.

**Accuracy.** Accurately estimating the 6D pose of an object is fundamental for many applications. High accuracy ensures that the system can make precise decisions and actions based on the object's pose information.

**Robustness.** Real-world environments are often cluttered, under varying lighting conditions, and objects of interest may be partially or fully occluded. A robust pose estimator or tracker should be able to handle clutters and illumination changes, and still provide accurate pose estimates even when parts of the object are not visible.

**Noise resistance.** Visual data from sensors can be noisy, especially in everyday environments. A robust pose estimator should be able to handle noise in the data and provide accurate pose estimates even in the presence of such noise. A pose estimator that can achieve high accuracy and robustness without requiring expensive sensors or hardware is more likely to be widely adopted.

**Efficiency/Real-time performance.** Efficiency is crucial for real-time applications. A pose estimator should be able to process data and estimate poses quickly enough to be useful in real-time

scenarios, such as robotics or augmented reality.

**Generalization.** The ability to generalize to new environments, objects, and lighting conditions is important for a pose estimator. A system that can adapt to new scenarios without the need for extensive retraining is more likely to be deployed in real-world applications.

**Data efficiency.** Some methods require large amounts of labeled training data, which can be expensive and time-consuming to collect and annotate.

In conclusion, the ideal 6D pose estimator or tracker for real-life applications should exhibit high accuracy, robustness to clutter and occlusions, noise resistance, efficiency in processing and real-time performance, and the ability to generalize to new environments without extensive retraining. These qualities are essential for practical deployment across various scenarios, ensuring that the system can effectively perceive and interact with its surroundings in real-world applications. My approaches in this thesis also focused on meeting these qualities, making them more suitable for real-world applications.

## 1.3 Outstanding problems

Despite several decades of development in 6D object pose estimation and tracking, there are still significant limitations in this field that hinder its practical deployment in real-world scenarios. Building on the discussion of essential qualities for real-world pose estimation, this section presents three outstanding problems that are essential for realistic applications.

**Robustness to symmetry, texturelessness, and occlusion.** The first challenge is maintaining robustness in the presence of symmetric and textureless objects under occlusion. Objects with

symmetric shapes, lack of surface texture, or partial occlusion pose difficulties for pose estimation systems. Existing methods often address these challenges individually, leading to failures when all three are present simultaneously (Cao, Sheikh, and Banerjee, 2016; Oberweger, Rad, and Vincent Lepetit, 2018). For instance, as shown in Figure 1.2, in environments like kitchens or manufacturing plants, objects may exhibit all three challenges concurrently, necessitating robust algorithms capable of handling such scenarios. Addressing this issue is crucial for ensuring algorithm stability in practical applications. Therefore, the question arises: Is there an effective approach that can simultaneously tackle all three challenges?

Figure 1.2: A daily life scenario where symmetric and textureless objects are under occlusion

**Generalization to previously unseen objects, especially for complex-shaped objects.** Beyond handling the aforementioned complex scenarios, a significant challenge is the generalization problem, where algorithms struggle to manage objects not encountered during training, particularly when encountering complex-shaped objects. Instance-level methods rely on pre-existing CAD models, limiting their applicability to a small set of objects. Category-level methods aim to generalize from known objects to new ones within a category, leveraging the shape similarity between the objects. However, existing geometric-based methods are sensitive to noise and limited to simple-shaped objects (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021; Di et al., 2022). As

the common features among objects are essential for category-level methods, the question arises: which features are crucial for category-level object pose estimation, especially for complex-shaped objects and noise robustness?

**Lack of effective methods for category-level object pose refinement.** When moving into the realistic application of category-level methods, a key challenge is the lack of effective methods for object pose refinement. While initial pose estimation is crucial, refining this estimate for high-precision poses is equally important, particularly in scenarios like AR medical training. Traditional methods compare the observed object with its CAD models for refinement, but this isn't feasible for category-level refinement where CAD models are unavailable. This gap in research is evident, with only one existing method, CATRE (X. Liu et al., 2022), proposed before my work. CATRE refines object poses by aligning the observed object's point cloud with a category-level shape prior, but it struggles to capture detailed geometric relationships essential for accurate refinement. This raises the question: Can we improve pose refinement by leveraging geometric relationships between the shape prior and the target object?

These outstanding problems highlight the need for advancements in 6D object pose estimation and tracking algorithms to meet the demands of real-world applications. Addressing these challenges will significantly enhance the practicality and effectiveness of pose estimation systems, paving the way for their widespread adoption across various industries.

## 1.4 My approaches

I proposed three approaches to address each of the aforementioned outstanding problems.

**Approach 1: tackling prevalent real-world challenges** The first approach aims to improve object pose estimation in challenging scenarios where objects exhibit symmetry, lack texture, and are partially occluded. While template-matching methods are effective for handling symmetry and textureless objects, they often struggle with occluded objects. Understanding an object's historical motion can aid in pose estimation, particularly when it's occluded. To address the challenges posed by complex scenarios, I propose a *Temporal-Primed object pose tracking framework with Auto-Encoders (TP-AE)*. This framework comprises a prior pose prediction module and a temporally primed pose estimation module. The prior pose prediction module uses historical pose data to predict the most probable pose for the current frame. The temporally primed pose estimation module adjusts the prior pose based on the current image observation. This adjustment involves embedding the prior pose into the observed image using a pose-image fusion procedure. Subsequently, the fused image is input into an auto-encoder for high-quality object reconstruction, even in cases of significant occlusion. Finally, the object pose is determined using a dynamic visibility-guided template-matching procedure that utilizes the auto-encoder's latent features. This approach strategically employs object motion information to guide latent feature extraction and matching, enabling real-time and accurate tracking of partially occluded and textureless objects with symmetric shapes.

**Approach 2: enhancing generalizability to category-level objects.** The second approach focuses on enhancing algorithm generalizability to category-level objects, crucial for real-world applications where accurate estimation of unseen objects is necessary without relying on CAD models. Existing 3D graph covolution (3D-GC) based methods often struggle to extract essential translation and scale information necessary for precise pose estimation. Moreover, their reliance on local geometric structures limits their effectiveness to simple-shaped objects and makes them more susceptible to noise. To overcome these challenges, I propose a novel network structure called the Hybrid Scope Feature Extraction Layer (HS-layer) for extracting latent features essential for

category-level object pose estimation. The HS-layer can encode translation and scale information and perceive outlier-robust local-global geometric information. This is achieved by integrating a *feature-distance-based receptive field construction procedure* and an *Outlier-Robust Feature Extraction Layer (ORL)* with 3D-GC, along with adding a parallel path for scale and translation encoding. The HS-layer is then used to build a real-time category-level pose estimation structure called HS-Pose. This model showcases substantial enhancements in accuracy, demonstrating its ability to handle complex-shaped objects while also exhibiting robustness against outliers.

**Approach 3: achieving high-precision in category-level applications.** The third approach targets high-precision category-level applications, where accurately comparing the pose between a category-level shape prior and the actual object is crucial. To achieve this, I propose leveraging the geometric relationships that exist between the shape prior and the target object, as objects within a category often share certain geometric similarities. I begin by incorporating the HS-layer into the feature extraction process, enabling the extraction of both local and global geometric features. This step is essential for gaining a comprehensive understanding of the object's geometry, which is crucial for achieving precise pose estimation. I then apply learnable affine transformations to align the extracted features, addressing geometric discrepancies between observed objects and shape priors. Additionally, I propose a mechanism to merge information from observed data and shape priors, enhancing the algorithm's ability to detect pose errors and refine object poses. By incorporating these strategies, the algorithm demonstrates improved performance in category-level object pose refinement, making it more reliable and effective in high-precision applications for real-world scenarios.

## 1.5 Contributions

In this thesis, I present three significant contributions to the field of visual 6D object pose estimation and tracking, with a focus on their practical applicability in real-life scenarios:

**Novel Approach for Challenging Scenarios:** I introduce a novel approach to tackle real life complex situations where three main challenges co-exist: symmetry, texturelessness, and occlusion. By leveraging temporal information and integrating object motion information into the latent feature extraction and matching procedure, I demonstrate a significant enhancement in the accuracy of real-time object pose tracking. This approach is particularly effective for objects with ambiguous surface features, those moving with non-constant velocity, or those under occlusion (Chapter 3).

**Generalizability for Category-Level Objects:** To enhance the generalizability of object pose estimation algorithms for category-level objects—essential for applications requiring the handling of numerous previously unseen objects without specific CAD models—I introduce a hybrid scope feature extraction layer (HS-layer) within my network architecture. The HS-layer effectively extracts features in different scopes, showing great perceptiveness to geometric structure and robustness to outliers. This significantly improves the accuracy and noise resistance of category-level pose estimation, enabling the algorithm to handle complex-shaped objects and operate in real time (Chapter 4).

**Category-Level 6D Object Pose Refinement:** I advance the refinement of category-level 6D object poses, crucial for high-precision applications, by introducing a novel architecture that specifically addresses shape variation issues. My approach effectively leverages the geometric relationship between shape priors and observed point clouds. Through applying learnable affine trans-

formations and employing a unique merging mechanism, it enhances the algorithm's capacity to refine object poses and detect pose errors, resulting in consistent performance improvements and enhanced generalization abilities (Chapter 5).

These contributions collectively advance the state-of-the-art in visual 6D object pose estimation and tracking, making algorithms more reliable and effective in real-world scenarios, and laying a foundation for their widespread use in various practical applications.

## 1.6 Structure

This thesis is structured as follows:

**Chapter 1** serves as an introduction, which outlines the significance of research in visual 6D object pose estimation and tracking, highlights the outstanding problems in this field, provides an overview of how each problem is addressed, and clarifies the contributions of this work.

**Chapter 2** provides background information on visual 6D object pose estimation and tracking, including a historical overview and the categorization of existing methods. It also reviews state-of-the-art methods relevant to the focused scenarios.

**Chapter 3** presents a real-time instance-level 6D object pose tracking framework designed to address challenges in real-life scenarios, such as symmetry, texturelessness, and occlusion. This chapter demonstrates how temporal information is used and combined with auto-encoder-based reconstruction to enhance the capability to tackle the challenges. Experimental results are provided to validate the effectiveness of the approach.

**Chapter 4** explains the enhancement of generalizability of object pose estimation for category-level objects. The chapter discusses the limitations of existing state-of-the-art methods and demonstrates how a network architecture was designed to address them by leveraging hybrid-scope features. A category-level object pose estimation framework is provided in this chapter. The effectiveness of the approach is demonstrated through extensive experiments.

**Chapter 5** presents a novel architecture for the category-level object pose refinement task. This chapter describes how geometric feature extraction is combined with learnable affine transformations to address shape variation issues. It also introduces a specially designed merging mechanism to efficiently merge diverse data sources. Extensive experiments are conducted to demonstrate the effectiveness of the proposed framework.

**Chapter 6** discusses the limitations of the proposed methods and outlines potential future work. It examines the challenges and constraints faced by the developed algorithms, such as their limitation to long-term object tracking, computational consumption, and the need for object detectors. The chapter also suggests possible directions for future research to address these limitations and enhance the applicability of the proposed methods.

**Chapter 7** concludes this thesis by summarizing the contributions of this work.

# Chapter Two

# Overview of 6D Object Pose Estimation and Tracking Research

The field of visual 6D object pose estimation and tracking has undergone significant evolution over the past few decades, with recent years experiencing a notable surge in interest and innovation. Initially, researchers relied on the availability of the target object's 3D model as prior information for pose estimation and tracking (also known as the *instance-level method*). This method typically involves two main stages: first, extracting features from visual input, and then recovering the object's pose by matching these features with the 3D model through template-matching or feature-matching strategies. One of the foundational works in this area is attributed to Roberts, who calculated the perspective projection matrix as a solution of a linear system, establishing the relationship between feature points on an object and their corresponding image projections (Roberts, 1963). However, early methods like these were limited to objects with simple shapes and were sensitive to noise in correspondence points.

Subsequent research efforts focused on enhancing feature matching robustness (Fischler and Bolles, 1981; Haralick et al., 1989) and extracting effective geometric primitives for correspondence matching, often with simplistic models and assumptions (David G. Lowe, 1991; Kosaka

and Nakazawa, 1995). However, reliance on geometric primitives (David G Lowe, 1987; Harris and Stennett, 1990) for feature extraction limited the algorithms' ability to handle complex objects and textures, leading to challenges in accurate pose estimation in realistic environments.

To address these limitations, researchers turned to data-driven methods, leveraging traditional machine learning algorithms to learn discriminative features for object pose estimation and tracking. Examples include the LineMOD (S. Hinterstoisser et al., 2011) algorithm and SURF (Bay, Tuytelaars, and Van Gool, 2006). Despite their success, these methods often relied on handcrafted features, requiring extensive hyper-parameter tuning, making them less scalable and adaptable.

To overcome these challenges, researchers embraced deep learning, using deep neural networks to extract feature representations directly from raw data, leading to improved accuracy and generalizability. Deep learning-based methods quickly became prevalent, with numerous studies (Xiang et al., 2018; Rad and Vincent Lepetit, 2017) exploring their potential to extract sparse or dense latent features for object pose estimation and tracking, and even to learn object poses directly from raw data. Additionally, integrating object pose estimation with scene understanding and semantic segmentation (K. He et al., 2017) is becoming prevalent, enabling more context-aware and intelligent object pose estimation and tracking systems.

As accuracy and reliability improved in simple scenarios, there was a growing emphasis on addressing real-world challenges. Recent efforts have led to the exploration of *category-level methods*, which estimate the pose of objects within a category without relying on specific 3D models, thereby improving generalizability to unseen objects (H. Wang et al., 2019). Concurrently, there is a significant emphasis on addressing complex real-world scenarios (Zheng, Leonardis, et al., 2022) characterized by occlusions, cluttered environments, symmetry, and textureless objects. Apart from these efforts, current trends in this field also involve integrating novel approaches such as attention mechanisms (Dosovitskiy et al., 2021), graph neural networks (Z.-H. Lin, S.-Y.

Huang, and Y.-C. F. Wang, 2020), and unsupervised learning (T. Lee, B. Lee, et al., 2021) to enhance object pose estimation and tracking methods, making them more adaptable to challenging real-world conditions.

The above developments lay the foundation for my work in visual 6D object pose estimation and tracking. In my research, I focus on developing algorithms that not only achieve high accuracy but also demonstrate practicality and generalizability for real-world applications. By tackling the key challenges in real-world scenarios, my work contributes to the continuous evolution of visual 6D object pose estimation and tracking, advancing us toward the development of practical and reliable algorithms for real-world use.

In the rest of this chapter, I first introduce the categorization of 6D object pose estimation and tracking methods based on different criteria, I then discuss the challenges in this field. Later, I review the state-of-the-art methods related to my work.

## 2.1 Categorization of 6D object pose estimation and tracking methods

Visual 6D object pose estimation and tracking methods can be categorized into different groups based on various criteria. Three main types of categorization are relevant to this work. The first categorization is based on objectives, where algorithms are grouped into instance-level and category-level methods. The second categorization is based on methodologies, where algorithms are roughly grouped into three types: pose estimation, tracking, and refinement. The third categorization is based on the underlying processing procedure of the algorithms: template-matching, correspondence-matching, and regression-based methods. Each of these categorizations provides a different perspective on the object pose estimation and tracking algorithms. I briefly introduce

them in this section.

### 2.1.1 Instance-level and category-level methods

**Instance-level methods.** Instance-level methods in visual 6D object pose estimation and tracking focus on precisely determining the pose of individual object instances. These methods typically require access to the 3D CAD model of the target object. Utilizing the prior information provided by the known 3D object model, instance-level methods often achieve higher accuracy compared to category-level methods. However, their reliance on specific object models limits their applicability to scenarios with a small number of known objects. Despite this limitation, instance-level methods are particularly effective in applications requiring precise object pose information, such as industrial assembly and object manipulation (Collet, Martinez, and Srinivasa, 2011). Over the years, research in instance-level methods has seen significant advancements, with current trends focusing on enhancing robustness in complex and challenging real-world scenarios (Y. Su, Saleh, et al., 2022).

**Category-level methods.** Category-level methods, unlike instance-level methods, aim to estimate the pose of objects belonging to a specific category (e.g., chairs, cars, bottles) rather than individual instances. These methods do not rely on specific object models, making them more generalizable and suitable for a wider range of everyday applications. While category-level methods can handle more object instances than instance-level methods, both types of methods are important. Instance-level methods serve as the foundation for category-level methods, and each has its suitable applications. Research on category-level object pose estimation has emerged more recently, with current efforts focused on improving the ability to handle intra-category shape variations and enhancing overall performance.

### 2.1.2 Object pose estimation, tracking, and refinement Methods

**Object pose estimation methods.** Object pose estimation methods focus on determining the pose of objects in a given scene. The goal is to find the transformation that aligns the 3D model with the observed scene, thus estimating the object's pose. They determine the pose of objects in a single image or multi-view image, typically using features extracted from that image alone. These methods are typically adopted in applications that do not require temporal consistency. Object pose estimation methods serve as the foundation for object pose tracking and refinement methods, as they establish the initial pose estimate.

**Object pose tracking methods.** Object pose tracking methods, on the other hand, aim to track the pose of objects over time in a video sequence. These methods enhance object pose estimation by incorporating temporal information to maintain continuity in the estimated poses, which is crucial for applications such as autonomous driving and augmented reality (Marchand, Uchiyama, and Spindler, 2016; Kothari et al., 2017). The challenges of object pose tracking include maintaining robustness when tracking objects with complex motion (e.g., rapid and non-linear movements) and varying levels of occlusion. Tracking objects over the long term, where the object is fully occluded during tracking and reappears, is also a significant challenge in visual object pose tracking. Since object pose tracking builds upon object pose estimation methods, and category-level object pose estimation is still in its early stages, most existing works focus on instance-level object pose tracking, with only a few addressing the challenge of category-level object pose tracking (Weng et al., 2021).

**Object pose refinement methods.** Object pose refinement methods aim to enhance the accuracy of initial pose estimates. They typically take a less accurate pose estimation and the visual observation as input, refining it using a template model as additional information. The refine-

ment procedure often involves comparing the transformed template model with the visual input and estimating the pose error between the initial pose estimation and the ground truth pose (Yi Li et al., 2020). This refinement step is crucial for ensuring the reliability of pose estimates, especially in scenarios requiring high precision or where the initial estimate may be noisy or imprecise. Category-level object pose refinement is much more challenging than instance-level methods, as it requires correctly identifying the shape relationships between objects with different shapes (the template model *vs*. the target object). Research on category-level 6D object pose refinement is still in its infancy, with only one existing work introduced (X. Liu et al., 2022) before my proposed approach.

### 2.1.3 Template-matching, correspondence-matching, and end-to-end methods

**Template-matching-based methods.** Template-matching-based methods rely on comparing observed object features with a precomputed template book generated from the object's 3D model. By calculating a similarity metric, such as normalized cross-correlation, these methods search for the best match between the template book and features extracted from the observed image (V. N. Nguyen et al., 2022) for pose estimation. Template-matching methods are straightforward and suitable for textureless objects, handling rotational ambiguity. However, they can be sensitive to illumination, background clutter, and occlusion, as the extracted features tend to vary from templates in these cases. Recent advances have focused on improving robustness to these challenges (Sundermeyer, Durner, et al., 2020; Zheng, Leonardis, et al., 2022).

**Correspondence-matching based methods.** Correspondence-matching methods, also known as feature-based methods, typically employ a two-stage structure for object pose recovery: first, they identify key features (e.g., sparse or dense keypoints) in both the 3D object model and the visual

input, and then they establish correspondences between these features using matching algorithms such as Perspective-n-point (PnP) (Fischler and Bolles, 1981). Correspondence-matching-based methods are prominent in visual 6D object pose estimation and tracking, often achieving higher accuracy than template-matching-based and end-to-end methods. However, they can be prone to errors with mismatched correspondences and may struggle with texture-less objects due to a lack of distinguishable local features. Recent research in this area has focused on two main paths: first, exploring deep neural network-based feature matching methods for enhanced robustness to correspondence errors; and second, enhancing the quality of latent feature extraction (Hansheng Chen et al., 2022).

**End-to-End methods.** End-to-end methods, also called direct methods, directly regress the object's pose from input image data without explicitly identifying features or establishing correspondences. These methods arise with the surge of deep learning and typically leverage deep neural networks to learn complex mappings from visual inputs to poses. End-to-end methods are often preferred for applications that require real-time speed and show impressive performance in scenarios with complex backgrounds. However, they can be challenging to train and require large amounts of annotated data. Current research works also explored adding auxiliary modules (*e.g.*, shape reconstruction (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021), semantic segmentation (Di et al., 2022)) during training for enhancing the latent feature extraction procedure. The exploration of better feature extraction network structures to enhance the generalizability of these methods was also proposed Zheng, C. Wang, et al., 2023.

## 2.2 Challenges

Visual 6D object pose estimation faces several challenges due to the complexity of real-world environments and object characteristics. These challenges can significantly impact the accuracy and

robustness of pose estimation algorithms. In this section, we discuss some of the key challenges and their implications in detail.

**Background clutter.** Background clutter refers to the presence of irrelevant or distracting elements in the scene that can interfere with object detection and pose estimation. This challenge is particularly problematic in environments with complex backgrounds, where the object of interest may blend into the surroundings or where the feature extraction procedure could be disrupted by the clutter. This interference makes it difficult for the algorithm to distinguish between the object and the background.

**Illumination changes.** Illumination changes, such as variations in lighting conditions, can significantly affect the appearance of objects in the scene. This can lead to variations in the object's appearance, making it challenging for visual pose estimation and tracking algorithms to accurately determine the object's pose across different lighting conditions.

**Textureless objects.** Textureless objects, which lack distinct surface features or patterns, are challenging for pose estimation algorithms because they provide limited visual cues for determining the object's pose. Without sufficient texture information, algorithms may struggle to accurately estimate the object's pose, especially in cluttered or low-texture environments.

**Symmetric objects.** Symmetric objects, which exhibit symmetry along one or more axes, pose a challenge for pose estimation algorithms because multiple poses may be consistent with the observed image data. This ambiguity makes it difficult for algorithms to uniquely determine the object's pose based solely on visual information, especially when the object is partially occluded or viewed from a challenging angle.

**Occlusions.** Occlusions occur when part of the object is obstructed from view by another object or occluder in the scene. Occlusions can make it challenging for pose estimation algorithms to accurately determine the object's pose, as the visible parts of the object may not provide sufficient information to identify the target object or to infer its full pose.

**Noise robustness.** Noise in the input data, such as sensor noise, compression artifacts, or the outliers of detected object point clouds, can negatively impact the accuracy of pose estimation algorithms by introducing spurious information that may be misinterpreted as valid object features. Robustness to noise is crucial for ensuring reliable pose estimation results in real-world applications.

**Similar-looking distractors.** Similar-looking distractors refer to objects or elements in the scene that visually resemble the target object, potentially leading to confusion for pose estimation and tracking algorithms. These distractors can cause algorithms to mistakenly identify them as the target object, leading to errors in pose estimation.

**Intra-category variation.** Intra-category variation refers to the variability in object appearance and geometry within the same object category. Objects within the same category can exhibit significant variations in size, shape, texture, and other visual properties, making it challenging for pose estimation algorithms to generalize across all instances of a category. Addressing intra-category variation requires category-level algorithms to learn robust representations that can capture the underlying similarities and differences between different instances of the same object category.

In summary, visual 6D object pose estimation and tracking face many challenges in real-world situations, especially when several prevalent challenges occur together. My work focuses on the prevalent challenges in the field, such as variation within object categories, dealing with noise, handling symmetric and textureless objects, and coping with occlusions. By addressing these

challenges, my work aims to make object pose estimation and tracking algorithms more accurate and reliable, pushing the field forward.

## 2.3 Related work

In the review of related work, I focus on methods that are directly relevant to my research. Specifically, I review methods that address the challenges in this field that are pertinent to my work. This selective review allows for a focused discussion on approaches that have implications for the advancement of my research objectives.

### 2.3.1 Instance-level: toward real-world challenges

Researchers have explored strategies to overcome challenges in robust visual 6D object pose estimation and tracking for several decades. Traditional methods primarily use handcrafted features and fixed matching procedures for pose estimation (D. G. Lowe, 1999). For instance, Rothganger, Lazebnik, Schmid, and Jean Ponce (2006) use local geometric image descriptors and multi-view spatial constraints for pose estimation, demonstrating a certain level of robustness to background clutter. S. Hinterstoisser et al. (2011) uses line segments as feature descriptors followed by feature-based template matching, effectively handles texture-less objects under cluttered backgrounds. Papazov and Burschka (2010) adopts local keypoints as features with RANSAC (Fischler and Bolles, 1981), aimed to enhance the algorithms's robustness to occlusions. (Stefan Hinterstoisser, Vincent Lepetit, Ilic, Holzer, et al., 2013) propose to use color gradient and surface normal as the feature descriptor for template-matching. While effective in certain situations, these methods often need extensive hyper-parameter tuning when applied to new scenarios or objects, limiting their generalization ability.

To overcome these challenges, Wohlhart and Vincent Lepetit (2015) introduced a deep-learning approach that uses Convolutional Neural Networks (CNNs) to extract descriptors capturing object identity and 3D pose. These descriptors are then utilized in template matching to determine the object pose, improving generalizability, scalability, and robustness, particularly for objects with poor textures. Similarly, Kehl, Milletari, et al. (2016) also applied deep learning to extract feature descriptors, focusing on features extracted from local image patches for template matching. PoseCNN (Xiang et al., 2018) and SSD-6D (Kehl, Manhardt, et al., 2017) leveraged deep learning for end-to-end object pose estimation. They consider the object as a whole and estimate its pose directly using deep neural networks. Additionally, they each propose strategies to address the challenge of rotation ambiguity: SSD-6D leverages a decomposed model pose space, while PoseCNN introduces separate training loss terms for symmetric and non-symmetric objects. BB8 (Rad and Vincent Lepetit, 2017) improved the robustness to occlusion by estimating the 2D projections of an object's 3D bounding box corners in the image. These sparse keypoints are then used for correspondence matching to determine the object's pose. BB8 also addresses rotation ambiguity by mapping output rotations into a specified range during training. AAE (Sundermeyer, Z. Marton, et al., 2019) addressed challenges such as textureless and symmetric objects using a deep learning-based augmented auto-encoder framework. It trained an auto-encoder to reconstruct objects despite illumination changes, background clutter, and occlusions in the input image. Object pose was then recovered using a template-matching procedure based on latent features extracted by the auto-encoder. MPAAE (Sundermeyer, Durner, et al., 2020) extended AAE by structuring multi-path decoders with a shared encoder for object feature extraction, enabling the generalizability to previously unseen objects with a given object 3D model. Despite their advancements, they can struggle in situations with higher levels of occlusion. Post-refinement techniques like Iterative Closest Point (ICP) (Arun, T. S. Huang, and Blostein, 1987) are commonly employed to alleviate this issue, but they often come at the cost of slower inference speeds.

To enhance the performance when objects are partially occluded, researchers have devel-

oped methods that use dense correspondence matching. These approaches begin by estimating dense correspondences and then recovering the object's pose by matching these correspondences, often employing algorithms like RANSAC and PnP. For instance, PVNet (Peng et al., 2019) regresses pixel-wise unit vectors pointing from each pixel to predefined keypoints, then calculates the 6D pose using RANSAC and an uncertainty-driven PnP algorithm. Pix2Pose (Park, Patten, and Vincze, 2019) utilizes pixel-wise 2D-3D correspondences, where an auto-encoder estimates each pixel's corresponding 3D coordinates on the object's canonical frame. Densefusion (C. Wang, D. Xu, et al., 2019) enhances occlusion robustness by incorporating depth information and fusing RGB and depth features for pixel-wise pose estimation. DPOD (Zakharov, Shugurov, and Ilic, 2019) also estimates 2D-3D pixel-wise correspondences and includes a deep learning-based pose refinement structure for improved results. DPOD (Zakharov, Shugurov, and Ilic, 2019) also estimates 2D-3D pixel-wise correspondences and includes a deep learning-based pose refinement scheme for improved results. PVN3D (Y. He, W. Sun, et al., 2020) proposed a voting network to extract point-wise 3D keypoints and then use least square fitting for object pose estimation. FFD6D (Y. He, H. Huang, et al., 2021) then extended PVN3D and enhanced the feature extraction with a bidirectional fusion strategy for fusing the information from the depth and RGB image. GDR-Net (G. Wang et al., 2021) focused on enhancing the correspondence feature extraction using geometry as a guide, and trained the network with Path-PnP, enabling end-to-end regression of object pose. These methods have demonstrated improved robustness to occlusion. However, they have limitations in handling objects with insufficient textures.

To further enhance the pose estimation for textureless objects, CosyPose (Labbé et al., 2020a) leveraged multi-view consistency for scene reconstruction and multiple refinement processes for improved pose accuracy. It demonstrates impressive performance on static objects under occlusion. However, the time-consuming refinement procedures limited its applicability for real-time applications. SurfEmb (Haugaard and Buch, 2021) proposed to learn dense and continuous 2D-3D correspondence distributions with a contrastive loss. It uses a key-query structure

consisting of two models to map object coordinates and color images into embedding distributions, resulting in improved robustness for textureless objects. Trans6D (Z. Zhang et al., 2023) and CRT-6D (Castro and Kim, 2023) investigated Transformer (Dosovitskiy et al., 2021) in 6D object pose estimation scheme for better perceptiveness of object's local and global feature, further enhanced the performance.

Meanwhile, methods that incorporate temporal information have been developed to improve the robustness of handling dynamic real-world objects. Traditional tracking methods (Choi and Christensen, 2013; Schmidt, Newcombe, and Fox, 2014; D. J. Tan and Ilic, 2014; D. J. Tan, Tombari, et al., 2015) are effective in tracking objects under partial occlusion and background clutter. However, these methods often struggle with generalizability and scalability to new scenarios. Recent advancements, particularly those based on deep learning, aim to track objects by utilizing motion information from two consecutive time steps. For instance, Garon and Lalonde (2017) explored the use of deep learning for direct object tracking. Their approach involves using a predicted image, generated by rendering the object model at the predicted pose's view, along with the observed image as input. The neural network then estimates the residual pose, resulting in improved performance. Similarly, DeepIM (Yi Li et al., 2020), $se(3)$-TrackNet (B. Wen, Mitash, et al., 2020) and Marougkas et al. (2020) used the predicted and observed images as input for object tracking under occlusion. $se(3)$-TrackNet introduced a novel neural network architecture to disentangle feature encoding and utilizes a 3D orientation representation using Lie algebra. Marougkas et al. (2020) proposed a multi-attentional convolutional architecture for real-time 6D object pose tracking, integrating multiple soft spatial attention modules into a CNN architecture to address challenges like background clutter and occlusion. However, these methods have limitations, as utilizing only two consecutive time steps may not fully leverage object motion. MaskUKF (Piga et al., 2021) enhances motion tracking robustness by leveraging Unscented Kalman Filter (Wan and Van Der Merwe, 2000). PoseRBPF (Deng, Mousavian, et al., 2019) proposed a novel object pose tracking structure by integrating Rao-Blackwellized particle filters with Augmented Auto-Encoders.

Despite its effectiveness in robustly tracking textureless and symmetric objects, PoseRBPF struggles with tracking objects under occlusion due to its reliance on direct object reconstruction using a single image. The development of learning-based 6D pose tracking schemes has been limited due to the lack of rich video datasets.

One of the contributions of this thesis, TP-AE (Zheng, Leonardis, et al., 2022), is the first method to address a challenging real-life scenario where three main challenges coexist: textureless and symmetric objects under occlusion, with real-time performance.

### 2.3.2 Category-level: Toward generalizability and higher precision

Instance-level methods are limited to handling a small set of specific objects, requiring the availability of 3D models for each object. To expand the algorithm's capability to work with a broader range of objects, researchers have explored estimating object poses at the category level. One early approach (Sahin and Kim, 2019) proposed a part-based random forest method that used skeleton representations to handle shape variations within object categories. While effective, this method is limited in handling complex shapes. H. Wang et al. (2019) explored deep learning for category-level object pose estimation using a reconstruction and matching-based approach. They first estimated the visible surface of the target object in a *Normalized Object Coordinate Space (NOCS)* and then recovered the object pose by iteratively matching the NOCS map with the observed object. Due to the simplicity and effectiveness of this approach, other works have followed this structure and incorporated new strategies to improve the accuracy and generalizability of category-level object pose estimation. For example, SPD (Tian, Ang Jr, and G. H. Lee, 2020) focused on enhancing the reconstruction procedure and suggested using category-level shape prior, which is a representative shape for objects in a category (often the mean shape), to improve the estimation of the NOCS maps. SGPA (K. Chen and Dou, 2021) proposed a method to dynamically adapt the shape prior to the observed object, resulting in significant performance improvement.

ANCSH (X. Li, H. Wang, Yi, L. J. Guibas, et al., 2020) extended the NOCS representation to the parts of articulated objects for category-level articulated object pose estimation. T. Lee, B.-U. Lee, et al. (2021) suggested adding a metric-scale object reconstruction branch in addition to the normalized scale. UDA-COPE (T. Lee, B. Lee, et al., 2021) utilized domain adaptation techniques, specifically the teacher-student self-supervised learning scheme, within the reconstruction and matching-based framework. This approach aims to reduce the reliance on manually labeled training data and enhance generalizability. Self-DPDN (J. Lin, Wei, Ding, et al., 2022) further enhances this by incorporating a deep prior deformation network within the self-supervised learning scheme. TTA-COPE (T. Lee, Tremblay, et al., 2023) improves generalizability by integrating test-time adaptation techniques. Despite their effectiveness, the time-consuming iterative correspondence matching procedure prevented them from being used in real-time applications.

To address real-time application needs, FS-Net (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021) explored using local geometric features for category-level object estimation. They employed a 3D Graph Convolution (3D-GC) based auto-encoder with an auxiliary shape reconstruction branch to extract latent geometric features of the target object and directly regress the pose based on these features. This end-to-end 3D-GC-based structure showed enhanced sensitivity to geometric structures, resulting in improved pose estimation accuracy and real-time performance. Subsequently, SAR-Net (H. Lin, Z. Liu, C. Cheang, et al., 2022) extended FS-Net by incorporating a symmetric-aware shape reconstruction procedure, improving rotation and size estimation. GPV-Pose (Di et al., 2022) further enhanced the 3D-GC-based structure by adopting confidence-aware estimations and integrating loss terms that constrain the pose and geometric consistency between different estimations, resulting in further performance improvements. Following these advancements, SSP-Pose (R. Zhang, Di, Manhardt, et al., 2022) enhanced GPV-Pose by leveraging shape prior with an additional shape deformation module. RBP-Pose (R. Zhang, Di, Lou, et al., 2022) then integrated residual bounding box projection to further improve object size estimation. Despite their improved performance and real-time speed, these 3D graph convolution-based methods

have limitations. They only use local geometric features, which restricts their effectiveness to simple-shaped objects and makes them sensitive to input noise.

In my thesis, HS-Pose (Zheng, C. Wang, et al., 2023) aims to improve the generalizability and robustness of category-level pose estimation for complex-shaped objects and noise, making them more suitable for practical applications. Another approach I explore is category-level object pose refinement, which enhances precision by estimating pose errors. CATRE (X. Liu et al., 2022) was the first to address this, using category-level shape prior and a framework to estimate pose errors between a transformed shape prior and visual input. My contribution, GeoReF (Zheng, Tse, et al., 2024), extends CATRE by addressing geometric variation between observed objects and transformed shape priors, further improving pose estimation accuracy for practical use.

## 2.4 Summary

The field of visual 6D object pose estimation and tracking has evolved over decades, progressing from traditional handcrafted methods to modern deep-learning approaches. Algorithms in this field can be categorized based on their goals (instance-level and category-level), methods (pose estimation, tracking, and refinement), and processing techniques (template-matching, correspondence-matching, and regression-based methods). These methods focused on various challenges, including background clutter, illumination changes, textureless and symmetric objects, occlusions, noise, similar-looking distractors, and intra-category variations. Current efforts are focused on enhancing algorithm robustness in real-world scenarios and generalizability to unseen objects. My thesis work aims to advance these goals by developing accurate, reliable, and generalizable algorithms for real-life applications.

# Chapter Three

# Tackling Real-World Challenges: Symmetry, Textureless, and Occlusion

*This chapter presents work that was published at the **IEEE International Conference on Robotics and Automation (ICRA) 2022** in Philadelphia (PA), USA (Zheng, Leonardis, et al., 2022). The title of the paper is **TP-AE: Temporally primed 6D object pose tracking with auto-encoders**. Tze Ho Elden Tse assisted with the experiment setting and proofreading, and Nora Horanyi and Hua Chen helped with proofreading.*

Fast and accurate tracking of an object's motion is one of the key functionalities of a robotic system for achieving reliable interaction with the environment. This paper focuses on the instance-level six-dimensional (6D) pose tracking problem with a *symmetric* and *textureless* object under *occlusion*. We propose a **T**emporally **P**rimed 6D pose tracking framework with **A**uto-**E**ncoders (TP-AE) to tackle the pose tracking problem. The framework consists of a prediction step and a temporally primed pose estimation step. The prediction step aims to quickly and efficiently generate a guess on the object's real-time pose based on historical information about the target object's motion. Once the prior prediction is obtained, the temporally primed pose estimation step embeds the prior pose into the RGB-D input, and leverages auto-encoders to reconstruct the target object with higher quality under occlusion, thus improving the framework's performance.

Extensive experiments show that the proposed 6D pose tracking method can accurately estimate the 6D pose of a symmetric and textureless object under occlusion, and significantly outperforms the state-of-the-art on T-LESS dataset while running in real-time at 26 FPS.

## 3.1 Introduction

Thanks to the rapid development of reliable mechanical structures, highly efficient actuators, and powerful algorithms, robotic systems have been deployed into various real-world applications such as mobile manipulation (Kaelbling and Lozano-Pérez, 2012), legged systems (Tremblay et al., 2018), robotic manipulation (Deng, Xiang, et al., 2020), and so on. With the increasing need for interacting with the environment, accurate detection and tracking of a target object become a core functionality for modern robotic systems.

This paper focuses on the instance-level six-dimensional (6D) pose tracking problem. Under various robotic application scenarios, the target objects to be manipulated or interacted with are possibly symmetric and textureless. Furthermore, the target object may be occluded by the environment or other objects. In these situations, estimating the target object's pose becomes much more challenging. Unlike the conventional pose estimation problems based on single RGB(-D) data, pose tracking leverages historical information about the target object's movement to assist in obtaining the desired pose. Incorporating the historical information and considering the pose tracking problem allows for dealing with challenging scenarios involving *symmetric* and *textureless* objects under *occlusion*.

To solve the pose tracking problem, we develop a ***T**emporally **P**rimed 6D object pose tracking framework with **A**uto-**E**ncoders (TP-AE)*. The proposed framework first predicts the 6D pose of the target object from a historical pose sequence. Then it uses the prediction to assist the visual-based pose estimation given the real-time RGB-D measurement. For the prediction step,

we propose to use temporal pose information to encode the raw RGB-D image stream information. Once the prediction is generated, the temporally primed pose estimation step adjusts the predicted pose via a reconstructed pose reference generated by auto-encoders. By resorting to the auto-encoder-based reconstruction, the proposed refinement scheme effectively handles *symmetric* and *textureless* objects under *occlusion*. In Figure 3.1, we demonstrate the performance of the TP-AE framework in a challenging scenario, showcasing its superiority over state-of-the-art approaches such as PoseRBPF (Deng, Mousavian, et al., 2019) and CosyPosee (Labbé et al., 2020a) for symmetric and textureless objects under occlusion.

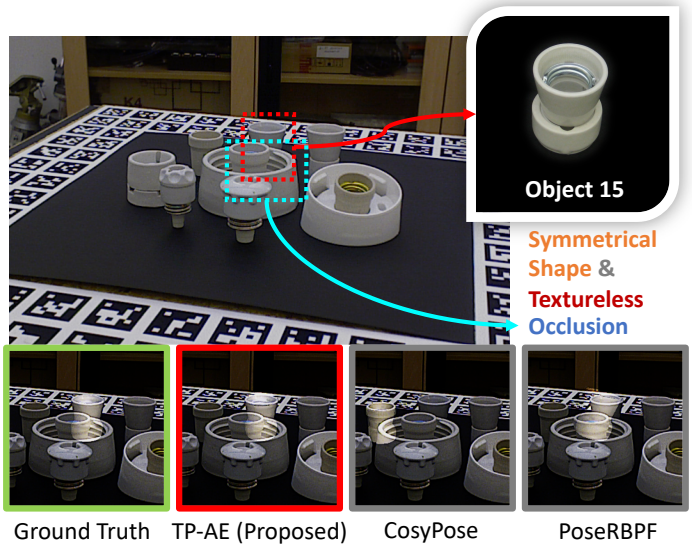



Figure 3.1: Illustration of the challenging scenario and the performance of the proposed TP-AE framework.

### 3.1.1 Related works

The work in this chapter is related to previous work on instance-level 6D object pose estimation and tracking.

**6D object pose estimation.** 6D object pose estimation problem that aims to estimate an object's 6D pose from a single image without temporal information has been extensively studied in the literature over several decades. Classical methods (D. G. Lowe, 1999; Rothganger, Lazebnik, Schmid, and J. Ponce, 2006) achieved high precision while requiring prohibitive hyper-parameter tuning when applied to new scenarios. Recently, deep learning-based methods have shown better generalization ability in challenging scenarios involving *symmetric* and *textureless* objects under *occlusion*. For example, (Manhardt, Arroyo, et al., 2018; Pitteri et al., 2019) aim to address rotational ambiguity. (Oberweger, Rad, and Vincent Lepetit, 2018; Y. He, W. Sun, et al., 2020; Zakharov, Shugurov, and Ilic, 2019) focus on occlusion. (Stefan Hinterstoisser, Vincent Lepetit, Ilic, Holzer, et al., 2013; Rios-Cabrera and Tuytelaars, 2013) consider textureless objects. More recently, various approaches have been proposed to deal with mixed challenges. For instance, (Rad and Vincent Lepetit, 2017; Hodan, Barath, and Jiri Matas, 2020) handle symmetric objects under occlusion. Estimating poses for textureless objects under occlusion has been investigated in (G. Gao et al., 2021; Y. He, H. Huang, et al., 2021; C. Wang, D. Xu, et al., 2019; Shi et al., 2021; Labbé et al., 2020a). Symmetric and textureless objects have been considered in (Sundermeyer, Z.-C. Marton, et al., 2018; Sundermeyer, Durner, et al., 2020; Zhigang Li and Ji, 2020). However, none of them can address symmetry, textureless, and occlusion simultaneously.

**6D object pose tracking.** As a natural extension of classical pose estimation problem, pose tracking problem try to incorporate temporal information to achieve higher pose estimation accuracy, which offers opportunities to simultaneously address all three aforementioned challenges. Traditional pose tracking methods (Choi and Christensen, 2013; Schmidt, Newcombe, and Fox, 2014) rely on hand-crafted likelihood functions, which is hard to generalize to new scenarios. Due to the lack of rich video datasets, the development of learning-based 6D pose tracking schemes remains limited. Along this direction, pioneering works such as (D. J. Tan, Tombari, et al., 2015; Garon and Lalonde, 2017; B. Wen, Mitash, et al., 2020; Marougkas et al., 2020; Yi Li et al., 2020)

try to utilize the relationship between current image with the last image to aid estimating objects' pose. However, focusing on only two consecutive steps is restrictive in fully characterizing objects' motion. To this end, (Deng, Mousavian, et al., 2019; Piga et al., 2021) use Rao-Blackwellized Particle Filter (Murphy and Russell, 2001) and Unscented Kalman Filter (Wan and Van Der Merwe, 2000), respectively, to encode motion information that is subsequently used for object tracking. Nonetheless, the performance of (Deng, Mousavian, et al., 2019) under occlusion remains unsatisfying due to the lack of utilization of temporal information in the object reconstruction phase.

Despite these recent advances on pose estimation and pose tracking, how to construct a reliable and efficient pose tracking framework for *symmetric* and *textureless* objects under *occlusion* remains a challenging problem.

### 3.1.2 Contributions

The main contributions of this paper are as follows. First, we propose, to the authors' best knowledge, the first neural-network-based prior pose generation scheme. This scheme exploits the target object's pose history of any length to better predict the object's future pose. By working with the pose information encoded in the complete movement of the object, the proposed scheme is computationally friendly and generates more accurate predictions in cases where the object moves with non-constant velocity. Second, we develop a novel temporally-primed pose estimation scheme consisting of a pose-image fusion and auto-encoder-based pose estimation, which improves the learning performance of the residual pose. The pose-image fusion scheme helps with reconstructing the intact appearance and point cloud from only partially observed measurements. Combined with the latent codes and features available from the auto-encoders, the overall temporally-primed pose estimation scheme successfully handles *symmetric* and *textureless* objects under *occlusion*. Third, the overall framework achieves not only a state-of-the-art performance in standard 6D object pose estimation dataset benchmarks (especially for the T-LESS dataset that contains numerous sce-

narios with symmetric and textureless objects under occlusion AR: 82.3 vs. 73), but also real-time speed (26 FPS).

### 3.1.3 Framework overview

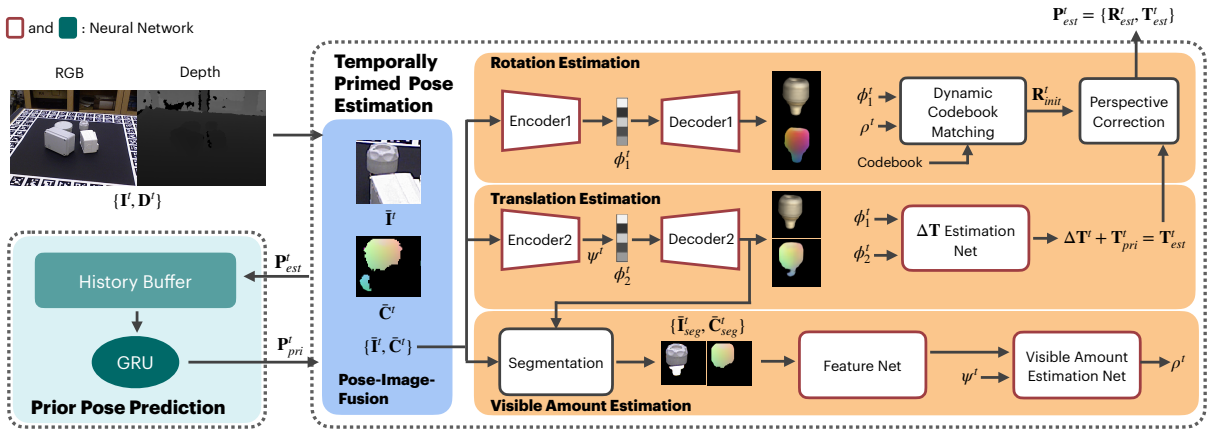



Figure 3.2: Overall structure of the proposed TP-AE framework.

Fig. 3.2 shows an overview of the proposed framework. Generally speaking, the proposed framework contains a prior pose prediction unit and a temporally primed pose estimation unit. At each time step, the prior pose prediction unit first generates a predicted 6D pose based on the historical pose estimations of the target object. Then, such a predicted 6D pose together with the real-time measured RGB-D data is fed into the temporally primed pose estimation unit to further adjust the predicted pose to obtain the final result.

To achieve fast prediction and account for the lack of large video datasets, the proposed framework takes the historically estimated 6D pose sequence as input to a GRU-based neural network to generate the prediction. Once the prediction is obtained, a pose-image-fusion module merges the predicted pose and the input RGB-D image to generate a RGB-Cloud pair. Then, the RGB-Cloud pair is fed into three branches to estimate the object's rotation, translation, and visible amount, respectively.

## 3.2 Prior Pose Prediction

As one of the major differences from standard pose estimation, pose tracking strategies have access to historical information about the movement of the target object, which has a strong implication on determining the object's current pose. The first important question to be answered is how to extract such key information encoded in the historical data.

An intuitive approach to incorporate temporal data is to train a neural network that directly maps the historical image streams to the current pose. Such an approach requires substantial data, which is not practically feasible due to the lack of available datasets. Alternatively, we use historical pose trajectory to represent the object motion without extra data collection effort. Since the pose trajectories of different objects can be shared, one can use the pose trajectories of any object across different datasets to train a prior pose prediction network and then apply it to predict the pose of an unseen object. Moreover, using historical pose sequence for pose prediction runs faster than using an image stream, which is essential for object tracking tasks.

Our prior pose prediction unit consists of a buffer and a prior pose prediction net. The buffer stores the previously estimated poses of the target object. Let $l$ be the size of the buffer and let $\mathbf{P}^t_{est} = \{\mathbf{R}^t, \mathbf{T}^t\}$ with $\mathbf{R}^t \in SO(3)$ and $\mathbf{T}^t \in \mathbb{R}^3$ be the estimated pose of the target object at time $t$, the sequence of estimated poses is denoted by $\{\mathbf{P}^i_{est}\}^{t-1}_{i=t-l-1}$. Given this sequence, the prior pose prediction net uses a GRU network Cho et al., 2014 to regress the prior pose $\mathbf{P}^t_{pri}$.

By using the pose trajectories for prior pose prediction, we can train our prediction net more robustly using additional random data augmentation on the pose trajectories.

**Remark 1** *During inference, the initial prior pose is generated by the existing single-image 6D pose estimation approach, as there is no historical poses for prediction.*

### 3.2.1 Network architecture and loss function

We use a single GRU layer connected to two dense layers to predict the prior pose. During implementation, a 6D parameterization of the rotation space $SO(3)$ proposed by (Y. Zhou et al., 2019) is adopted to respect the continuity of the rotation space. Consequently, the input to the neural network is a vector in $\mathbb{R}^{l\times 9}$ with $l$ being the size of the buffer. The output is a vector in $\mathbb{R}^9$ parameterizing the 6D pose of the target object, including a 3D translation vector and a 6D parameterization of the orientation state.

The total loss contains translation loss and rotation loss. We adopt the $\ell_2$ norm as the translation loss function and calculate the rotation loss following (Y. Zhou et al., 2019). Then, the pose prediction loss is:

$$\mathcal{L}oss_{\text{pri}} = \mathcal{L}oss_{\text{pri},\mathbf{R}} + \beta\mathcal{L}oss_{\text{pri},\mathbf{T}} \tag{3.1}$$

where $\beta$ is a hyperparameter weighting the importance of the translation loss relative to the rotation loss.

## 3.3 Temporally Primed Pose Estimation

This module visually estimates an object's pose assisted by the prior pose. It first fuses the prior pose with the real-time RGB-D data to generate an RGB-Cloud pair, then feeds the pair to three modules to estimate the object's rotation, translation, and visibility. To robustify the pose estimation network against occlusion, we leverage object reconstruction in both the rotation and the translation estimation module.

Reconstruction networks like auto-encoders can extract the low-dimensional representation of objects, which are commonly called latent codes. Roughly speaking, auto-encoder-based strategies first extract the latent code of an object by supervising the reconstruction procedure, in which

the decoder needs to recover the object using the latent code provided by the encoder. Then, the latent code is used for pose estimation to boost the inference speed without needing a decoder. For such method, acquiring high-quality latent codes under occlusion is crucial for pose estimation accuracy.

In this sequel, we first discuss how the prediction and input RGB-D data can be expressed in a unified way to support object reconstruction, then develop the proposed auto-encoder-based strategy that particularly addresses the *symmetry*, *textureless* and *occlusion* challenges.

### 3.3.1 Pose-image-fusion

Given the prior pose prediction $\mathbf{P}^t_{\text{pri}}$ and the real-time RGB-D image $(\mathbf{I}^t, \mathbf{D}^t)$ at a generic time $t$, with $\mathbf{I}^t$ being the RGB data and $\mathbf{D}^t$ being the depth data, we first need to provide a unified way of representing the information encoded therein. For this purpose, we propose to use a cropped RGB-Cloud image pair $(\bar{\mathbf{I}}^t, \bar{\mathbf{C}}^t)$ which carries visually sensible prior information for the reconstruction network. The procedure of obtaining the cropped RGB-Cloud image pair from the given inputs $(\mathbf{P}^t_{\text{pri}}, \mathbf{I}^t, \mathbf{D}^t)$ consists of three main steps.

First, we backproject the depth image using the camera's intrinsic parameters to recover its point cloud image $\mathbf{C}^t$, in which every pixel stores the recovered 3D coordinate of the corresponding pixel in the depth image. Working with the point cloud image helps the network leverage the object's geometric structure.

Then, we extract the potentially useful area by cropping the input RGB image and the recovered point cloud data with an enlarged object's bounding box (Bbox) centering at the predicted translation $\mathbf{T}^t_{pri}$. To improve the robustness of the overall scheme, a Bbox scaling factor $\delta$ depending on the object's diameter $d$ is introduced as follows:

$$\delta = \max\{2\sqrt{3}\epsilon_T/d, s_{\min}\}, \tag{3.2}$$

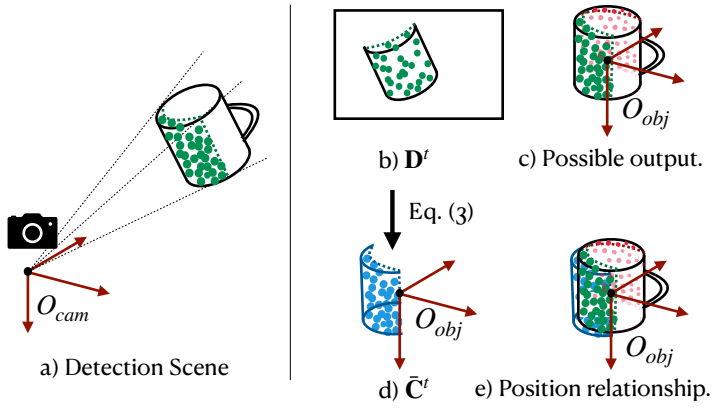



Figure 3.3: Visual depiction of the motivation behind prior pose embedding.

where $\epsilon_T$ is the maximum allowable displacement between the ground truth (GT) translation and the prediction and $s_{min}$ is a pre-specified minimum scaling factor. This cropped RGB image used as $\bar{\mathbf{I}}^t$.

The point cloud image crops, $\tilde{\mathbf{C}}^t$, is then transformed into the frame defined by the predicted pose $\mathbf{P}^t_{\text{pri}}$ and then normalized with a scaling factor associated with the Bbox:

$$\bar{\mathbf{C}}^t = \mathbf{R}^T_{pri}(\tilde{C}^t - \mathbf{T}^t_{pri})/|0.5\delta|. \tag{3.3}$$

In essence, the above operations transform the input RGB-D image to the RGB-Cloud pair $(\bar{\mathbf{I}}^t, \bar{\mathbf{C}}^t)$. By doing so, the object's position in the RGB-Cloud pair indicates the distance from $\mathbf{T}^t_{pri}$ to GT position. The point-wise fusion in (3.3) makes the transformed point cloud robust to occlusion, as the prior pose can be inferred even when part of the point cloud is invisible.

Figure 3.3 provides a visual depiction of the motivation behind incorporating prior information into the point cloud. As illustrated, the recovered point cloud in $\mathbf{D}^t$ can be mapped to either the green (GT) or red surface (if no prior information is provided) in c). By leveraging the prior pose, the transformed point cloud in $\bar{\mathbf{C}}^t$ (d) aligns more closely with the GT point cloud distribution (e), facilitating pose recovery and smoothing the tracking process.

The RGB-Cloud pair, comprising appearance and transformed geometric information, is

well-suited for object reconstruction networks to robustly recover textureless objects under occlusion.

### 3.3.2 Auto-encoder and matching-based rotation estimation

Once the RGB-Cloud pair is generated, we exploit the auto-encoder-based strategy to address the main challenges in estimating the object's pose. We first generate the latent code of the object to achieve robustness to occlusion, then get the rotation by a dynamic codebook matching method. This method naturally handles textureless and symmetric objects, as the code matching compares the intact object's feature among different rotations rather than local features.

#### Latent code extraction and codebook generation

We train an auto-encoder to extract the latent code that encodes the object's rotation while invariant to translation. To do so, we generate the target RGB-Cloud pair in which the object is in the center while maintaining its orientation (See the output of Decoder1 in Figure 3.2) by fusing the GT pose with the GT image. The GT image shows the intact object against a black background with constant lighting at the GT pose. After the training, a codebook is generated by collecting the latent codes of the object in different rotations.

#### Dynamic latent code matching

Typically, a code matching procedure compares the cosine similarity between the input latent code with the codebook and then obtains the rotation using the highest similarity score (Sundermeyer, Z.-C. Marton, et al., 2018; Zhigang Li and Ji, 2020). However, this might fail when the latent code is of poor quality (*e.g.* for almost invisible objects). Therefore, we enhance the occlusion

robustness of the matching phase by inducing the historical information into this phase and using the object's visible amount to balance currently observed information and historical information. Specifically, during inference, we generate two cosine similarity score lists by comparing the codebook with the latent code of the input RGB-Cloud pair and that of the historical RGB-Cloud pair. The historical RGB-Cloud pair is obtained by fusing the prior pose with a synthetically generated RGB-D image in which the target object is viewed from the prior pose perspective. Denoting the first score list as $\mathbf{S}^t_{obs}$ and the second score list as $\mathbf{S}^t_{his}$, the final score is constructed as follows

$$\mathbf{S}^t = \begin{cases} \mathbf{S}^t_{obs}, & \rho^t > \sigma \\ \frac{\rho^t}{\sigma} \cdot \mathbf{S}^t_{obs} + \frac{\sigma - \rho^t}{\sigma} \cdot \mathbf{S}^t_{his}, & \rho^t \leq \sigma \end{cases}, \tag{3.4}$$

where $\sigma$ is a threshold to indicate whether the dynamic adjustment is used. We then get an initial rotation estimation $\mathbf{R}^t_{init}$ using the highest score from the final score list $\mathbf{S}^t$. To account for the rotation ambiguity caused by translation as mentioned in (Kundu, Yin Li, and Rehg, 2018; Sundermeyer, Z. Marton, et al., 2019), we correct $\mathbf{R}^t_{init}$ by first finding a rotation transformation $\Delta(\mathbf{R}^t)$ that aligns the direction of the estimated translation $\mathbf{T}^t_{est}$ to camera's z-axis, then getting the final rotation estimation by $\mathbf{R}^t_{est} = \Delta(\mathbf{R}^t)^{-1}\hat{\mathbf{R}}^t_{init}$.

### 3.3.3 Auto-encoder based translation estimation

The estimation of translation can also leverage the latent code of an auto-encoder to achieve robustness to occlusion. Specifically, we train an auto-encoder (Auto-Encoder 2 in the proposed framework Fig. 3.2) to reconstruct the intact object while preserving its location in the RGB-Cloud pair. The reconstruction target is generated by fusing the prior pose with the GT image. Motivated by the observation that the object's location in the RGB-Cloud pair provides information about the translation difference between the prior prediction $\mathbf{T}^t_{pri}$ and the ground truth $\mathbf{T}^t_{GT}$, we concatenate the latent code of rotation auto-encoder and translation auto-encoder to an estimation network that generates the desired adjustment $\Delta\mathbf{T}^t$. Then, the resulting translation estimation is simply given

by:

$$\mathbf{T}_{est}^{t} = \Delta\mathbf{T}^{t} + \mathbf{T}_{pri}^{t}. \tag{3.5}$$

### 3.3.4 Visible amount estimation

As it is important whether the target object is still being tracked correctly, this module estimates the object's visibility. We define the visible amount $\rho^t$ as the ratio of visible and total object pixels. As shown in Figure 3.2, we use the feature of the intact object (the output of Encoder2) and the occluded object (the output of Feature Net) as the input of the Visible Amount Estimation Net to regress $\rho^t$. During testing, $\rho$ is used to i) trigger re-initialization of tracking when it is lower than a pre-defined level and ii) enhance rotation estimation under severe occlusion (See Eq.(3.4)).

### 3.3.5 Network architecture and loss function

The network architecture of the used auto-encoders is the same as AAE Sundermeyer, Z.-C. Marton, et al., 2018, except that the channel number of input and output images are set to 6. The structure of the Feature Net, the Visible Amount Estimation Net, and the $\Delta\mathbf{T}$ Estimation Net is shown in Fig. 3.4. The loss function for training the temporally primed pose estimation unit includes three terms: reconstruction loss $\mathcal{L}oss_{rec}$, translation loss $\mathcal{L}oss_{\Delta\mathbf{T}}$, and the visible amount loss $\mathcal{L}oss_{\rho}$.

The object reconstruction loss is the region weighted sum of the pixel-wise losses between the reconstructed and the target RGB-Cloud pair. We divide the pixels into three regions, the matched object region, the matched background region, and the mismatch region. Denoting the set of pixels belonging to the object in the target crops and the reconstructed crops as $V$ and $\hat{V}$, respectively, the object matching region is $V \cap \hat{V}$, the mismatch region is $(V - \hat{V}) \cup (\hat{V} - V)$, and all other pixels belong to the background region. By denoting the $\ell_2$ loss of the $i^{th}$ pixel as

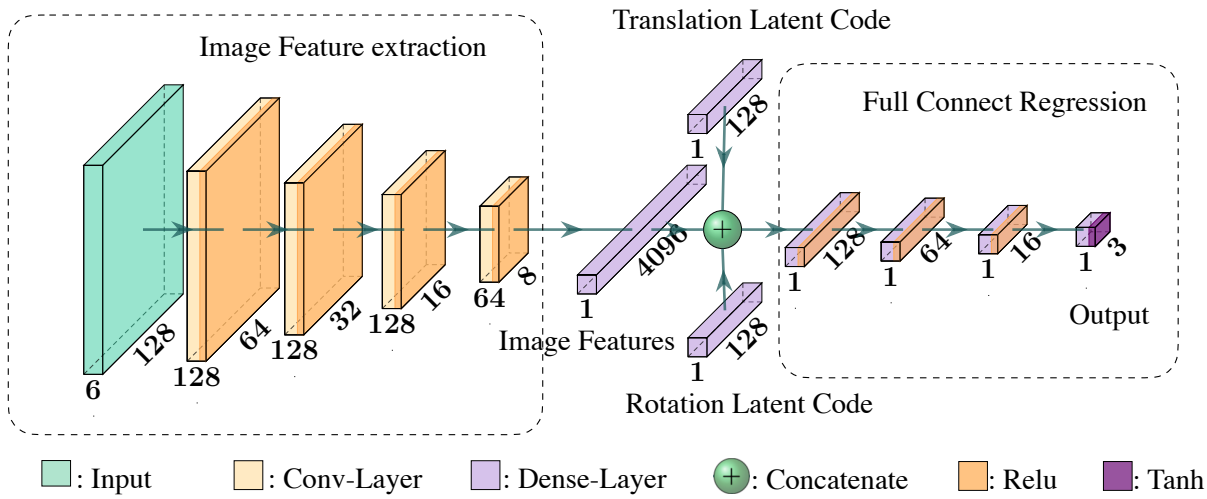



Figure 3.4: The architecture of Feature Net, the Visible Amount Estimation Net, and the $\Delta\mathbf{T}$ Estimation Net.

$\mathcal{L}oss_{px,i}$, the object reconstruction loss is:

$$\mathcal{L}oss_{rec} = \Sigma_{i\in\mathbf{I}}(\gamma \cdot \mathcal{L}oss_{px,i}) \tag{3.6}$$

where $\mathbf{I}$ is the input crops and $\gamma \in \{\gamma_1, \gamma_2, \gamma_3\}$ is the region-based weight, in which $\gamma_1, \gamma_2, \gamma_3$ is used for the mismatched region, the matched object region, and matched background region, respectively. By setting $\gamma_1 > \gamma_2 > \gamma_3$, the auto-encoder is guided to focus on aligning the silhouette.

We use $\ell_2$ loss for translation loss $\mathcal{L}oss_{\Delta\mathbf{T}}$ and visible amount loss $\mathcal{L}oss_{\rho}$. The total estimation loss $\mathcal{L}oss_{est}$ is:

$$\mathcal{L}oss_{est} = \lambda_1\mathcal{L}oss_{rec} + \lambda_2\mathcal{L}oss_{\rho} + \lambda_3\mathcal{L}oss_{\Delta\mathbf{T}}. \tag{3.7}$$

Empirically, the parameters are chosen as $\beta = 0.1$, $\lambda_1 = \lambda_3 = 1$, $\lambda_2 = 0.5$, $\gamma_1 = 3$, $\gamma_2 = 2$ and $\gamma_3 = 1$ to achieve balance between all losses.

## 3.4 Experiments

We provide the implementation details and the experiment results of the proposed framework in this section.

### 3.4.1 Baseline methods

The result of PoseRBPF is available from (Deng, Mousavian, et al., 2019). The single-image single-object (siso) result of CosyPose is available from its official website (Labbé et al., 2020b). We use the RGB version of CosyPose since its average recall ($AR_{vsd}$) performance is lower when applying the ICP refinement according to the BOP challenge results (Hodan, Michel, et al., 2018). The rest of the results are from corresponding papers.

### 3.4.2 Datasets

We use T-LESS (Hodaň, Haluza, et al., 2017) and YCB-V (Xiang et al., 2018) to evaluate our framework since other existing tracking datasets are either limited in size (Fäulhammer et al., 2015), not accurately labeled (Lai et al., 2011), or unsuitable for our problem setting, such as (Garon and Lalonde, 2017; Wu et al., 2017).

**1) T-LESS** is widely used for pose estimation and best fits our problem setting. It contains 10K test images and 39K training images, in which all 30 industrial objects are symmetric and textureless. The testing scenarios include various occlusion levels, from non-occlusion to full occlusion. We use VSD metric (Hodaň, Jiří Matas, and Obdržálek, 2016) for evaluation. The recall accuracy $AR_vsd$ is reported at $err_{vsd} < 0.3$ with a tolerance $\tau = 20mm$ and $> 10\%$ object visibility.

**2) YCB-V** contains 92 RGB-D videos (12 for testing) with 130K real images and 80K synthetic images. It provides 21 daily objects with various shapes and texture levels. We use the ADD-S (Stefan Hinterstoisser, Vincent Lepetit, Ilic, Holzer, et al., 2013) as the metric, where a pose is regarded as correct if the average distance of the model points to the nearest estimated points is less than $10\%$ of the model diameter. Following PoseCNN (Xiang et al., 2018), we report the area under the accuracy-threshold curve (AUC) for pose evaluation.

### 3.4.3 Implementation details

We conduct all the experiments using one NVIDIA RTX 2080Ti GPU and an Intel i9-CPU@3.30GHZ. During training, Adam optimizer is adopted with 15000 training steps and a batch size of 64. Similar to AAE, we use domain randomization methods such as {*Gaussian blur, add, light, multiply, contract, Gaussian noise, Coarse Dropout*} for training images. Image backgrounds are augmented by the images from (S. Hinterstoisser et al., 2011). Occlusion is simulated by randomly rendering two 3D models of other objects around the GT position of the target object. We set $\epsilon_T$ to 28.8mm and $s_{min}$ to 1.3 for pose-image fusion, and the RGB-Cloud pair is scaled to $128 \times 128$ before being fed into the encoders. Pose trajectories for training the prior pose prediction net are augmented with rotation shift, translation shift, noise addition, random drop, and permutation. The buffer size $l$ is 10. During inference, we initialize the prior pose using CosyPose (1-view version). Same as PoseRBPF, we take the one with the highest confidence score from the pose hypotheses as the initial pose. Re-initialization is triggered when: 1) $\rho < 0.2$, and 2) the rotation tracking fails with the same threshold as PoseRBPF (Deng, Mousavian, et al., 2019), which is 0.6 for latent code comparison. The viewpoint number of the codebook is the same as PoseRBPF (184464). We set $\sigma$ of dynamic codebook matching to 0.5 for the T-LESS dataset. Since the pose label of YCB-V is not very accurate, $\sigma$ is set to 1 as compensation.

Table 3.1: Ablation study results (AR: Average Recall).

| Item | Comparison | Train Data Type | $AR_{vsd}$ |
|---|---|---|---|
| Full net | Train data | Mixed | **82.3** |
| | Train data | Synthetic | 77.2 |
| [AS-1] | RGB (AAE +Retina + ICP) | Synthetic | 57.1 |
| | PIEM: RGB-D without Eq. 3.3 | Synthetic | 60.2 |
| | PIEM: RGB-D with Eq. 3.3 | Synthetic | 77.2 |
| [AS-2] | LSTM | Mixed | 81.4 |
| | GRU | Mixed | **82.3** |
| [AS-3] | Without perspective correction | Mixed | 80.6 |
| [AS-5] | No dynamic codebook matching | Mixed | 81.6 |
| [AS-6] | Refine CosyPose Labbé et al., 2020a (siso, acc: 63.8%) | Mixed | 74.3 |

### 3.4.4 Ablation study

We conducted an intensive ablation study using the T-LESS dataset to validate our framework design. Full evaluation results are shown in Table 3.1.

**[AS-1] Pose-image-fusion.** We evaluate the pose-image fusion module by comparing the performance of the auto-encoders trained with i) RGB-only synthetic (syn.) images, ii) RGB-D synthetic images without (w/o) Eq. (3.3), and iii) RGB-D synthetic images with Eq. (3.3). For the first one, we referenced the result of AAE. The increased pose estimation accuracy in Table 3.1 and the improved image reconstruction quality under occlusion shown in Figure 3.5 both confirm the effectiveness and motivation for pose-image fusion.

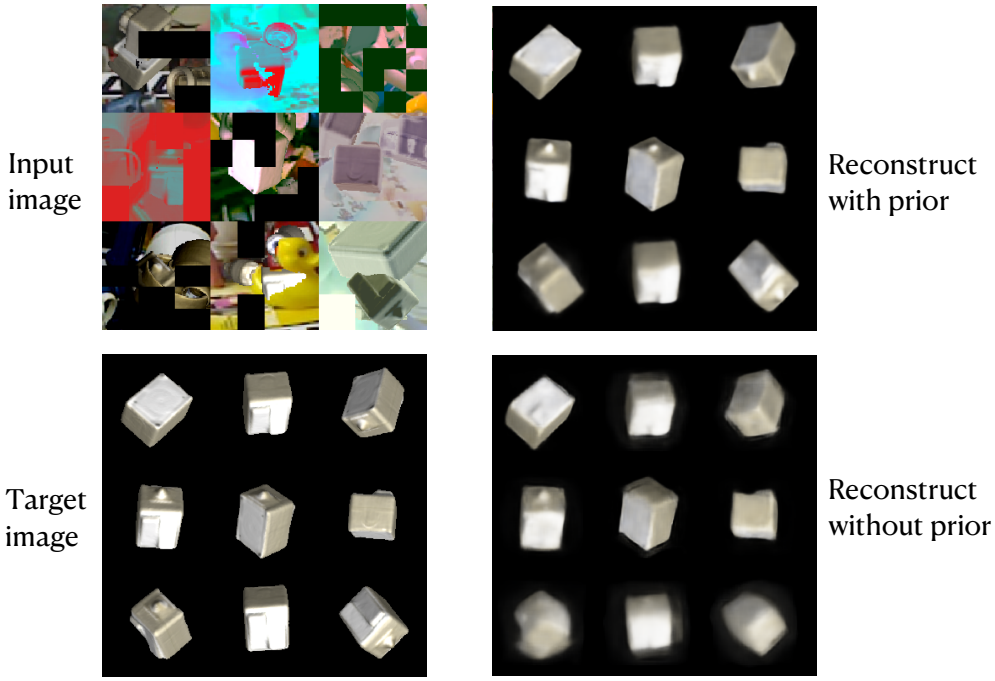



Figure 3.5: Reconstructed images from Decoder1 when without and with the Eq. (3.3) in the pose-image-fusion module.

**[AS-2] Prior pose prediction.** To investigate the suitable network for pose prediction, we compared the LSTM (Hochreiter and Schmidhuber, 1997) and GRU network. As the motion pattern is similar between the training set and test set of both T-LESS and YCB-V, training directly on the training set does not reflect the effectiveness of the proposed module. We thus use the pose trajectories extracted from other datasets [1] and test the trained model on YCB-V and T-LESS. Note that this will make the task harder as the model needs to overcome the domain gap between the training and testing data. Table 3.1 shows that both GRU and LSTM can deal well with the domain gap, and the GRU performs better than LSTM by $0.9\%$.

[1]The pose trajectories are extracted from the test set of the following datasets: OPT (Wu et al., 2017), YCB-M (Grenzdörffer, Günther, and Hertzberg, 2020), TUD-L (Hodan, Michel, et al., 2018), and HO-3D (Hampali et al., 2020). Note that the size of the test set is much smaller than the training set, we thus use several datasets.

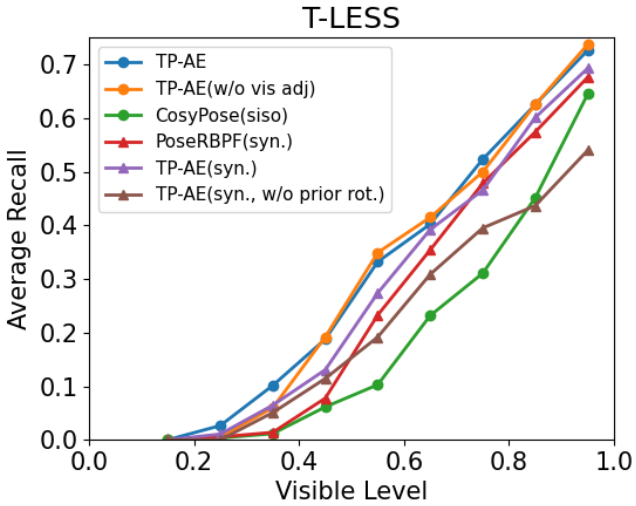



Figure 3.6: Pose accuracy under a range of occlusion levels.

**[AS-3] Perspective correction.** The mean recall dropped by $1.7\%$ when no perspective correction was carried out.

**[AS-4] Pose accuracy distribution under occlusion.** Figure 3.6 shows the pose estimation accuracy under a varying level of occlusion. We compared our method with PoseRBPF and CosyPose on the T-LESS test set of the BOP challenge. As shown in the figure, our proposed method outperforms them across all occlusion levels. Moreover, the improvements are more significant when the visibility is under $0.5$.

**[AS-5] No dynamic codebook matching.** Table 3.1 shows that the performance drops by $0.5\%$ when only using simple codebook matching (use $S^t_{obs}$ directly). The influence of dynamic codebook matching on different occlusion levels is shown in Figure 3.6, which indicates that dynamic matching is effective on all the occlusion levels when the model is trained only on synthetic data, but in case of mixed data is more effective when the visible amount is lower than 0.5.

Table 3.2: Performance Comparison on the T-LESS test set (Primesense)

| Type | Method | $AR_{vsd}$ | Speed |
|---|---|---|---|
| RGBD | **TP-AE** (mixed data) | **82.3** | **26** |
| | **TP-AE** (synthetic data only) | 77.2 | **26** |
| | AAE (Sundermeyer, Z.-C. Marton, et al., 2018) + (ICP) | 57.1 | 2 |
| | PoseRBPF (Deng, Mousavian, et al., 2019) | 72.9 | 10 |
| RGB | AAE (Sundermeyer, Z.-C. Marton, et al., 2018) | 18.4 | 5.9 |
| | PoseRBPF (Deng, Mousavian, et al., 2019) | 41.5 | **11.5** |
| | CosyPose (siso) | **63.8** | - |
| | Pix2Pose (Park, Patten, and Vincze, 2019) | 29.5 | 0.6 |
| | PFRL (Shao et al., 2020) + AAE | 51.53 | 4.2 |
| D | StablePose (Shi et al., 2021) | 73 | 2.5 |
| Rotation Tracking | **TP-AE** (mixed data) | **93.4** | 26 |
| | **TP-AE** (synthetic data only) | 92.7 | 26 |
| | AAE(GT BBox) | 72.8 | - |
| | PoseRBPF(GT BBox) | 85.3 | - |
| GT Re-initialization | **TP-AE** (mixed data) | **84** | 26 |
| | **TP-AE** (synthetic data only) | 79.8 | 26 |

**[AS-6] Refinement** In addition to the object tracking task, we were also interested in how our proposed method could be used to refine the pose estimation method by taking the estimated pose of other approaches as the prior pose. We report a $10.5\%$ increase on CosyPose when using our approach as pose refinement without any iteration.

### 3.4.5 Comparison with state-of-the-art methods

**Results on the T-LESS dataset.** Table 3.2 presents the evaluation results on the T-LESS dataset. We included results from training on synthetic (syn.) data for a fair comparison with AAE and PoseRBPF. For rotation tracking comparison, we used the GT to provide 2D bounding boxes for AAE and PoseRBPF. We also reported results when using the GT poses for initialization. Our framework demonstrates better performances among its competitors on the T-LESS dataset. Qualitative results are shown in Figure 3.7, where the target objects (#2, #5, #9, #17, #24) and their pose are highlighted with a white overlay on the input image. The results are displayed from left to right as GT, TP-AE, CosyPose, and PoseRBPF, respectively.

**Results on the YCB-V Dataset.** We compare our results with state-of-the-art methods in Table 3.3. We only use $20\%$ of the images of the YCB-V training set for training and got comparable results with other approaches.

## 3.5 Conclusion

In this paper, we proposed a novel TP-AE object pose tracking framework that can robustly handle symmetric and textureless objects under occlusion. Our method outperforms the state-of-the-art, while also running in real-time (26 FPS). We successfully demonstrated that embedding tempo-

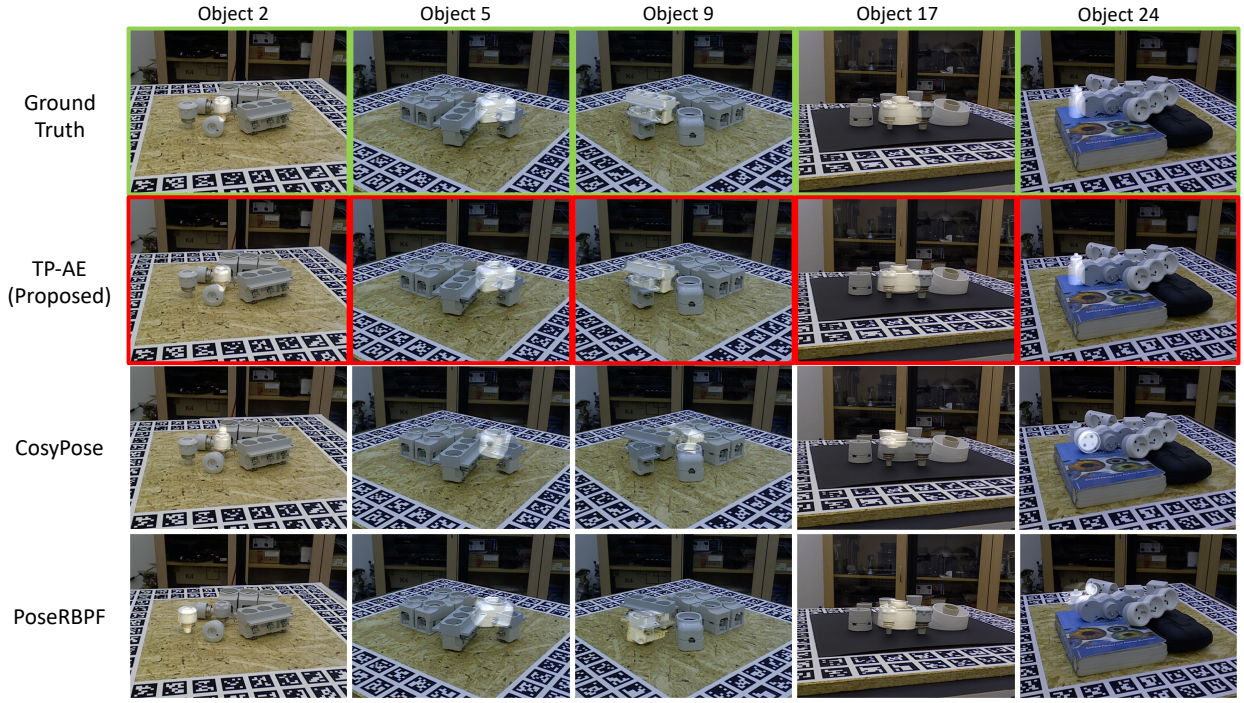



Figure 3.7: Qualitative results on the T-LESS dataset.

ral information in our proposed framework can increase the pose estimation accuracy by a large margin. We also demonstrated the generalizability of our prediction network and its robustness under disturbances. In addition, we reported a thorough analysis on the effectiveness of perspective correction. As a future work, the proposed method could achieve an improved performance when combined with a refinement process.

Table 3.3: AUC Performance on the YCB-V dataset.

| Type | RGB-D Method | ADD-S | Speed |
|---|---|---|---|
| Tracking | **TP-AE** (mixed) | 93.8 | 26 |
| | **TP-AE** (syn) | 92.5 | 26 |
| | PoseRBPF (200 particles) | 93.3 | 5 |
| | DeepIM (Yi Li et al., 2020) + PoseCNN (Xiang et al., 2018) (4 it) | 94.0 | 6 |
| | MaskUKF (Piga et al., 2021) | 94.2 | 52.6 |
| | $se(3)$-TrackNet (B. Wen, Mitash, et al., 2020) | **95.52** | **90.0** |
| Estimation | PoseCNN (ICP) (Xiang et al., 2018) | 93.0 | $< 0.1$ |
| | PVN3D (w/o refinement) (Y. He, W. Sun, et al., 2020) | 95.5 | 5 |
| | Densefusion (w/o refinement) (C. Wang, D. Xu, et al., 2019) | 91.2 | 20 |
| | G2L-Net (w/o refinement) (W. Chen, Jia, H. J. Chang, Duan, and Leonardis, 2020) | 92.4 | 21 |
| | FFB6D (Y. He, H. Huang, et al., 2021) | **96.6** | 13.3 |

# Chapter Four

# Enhancing Generalizability to Category-level Objects

*This chapter presents work that was published at the **The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2023** in Vancouver, Canada (Zheng, C. Wang, et al., 2023). The paper title is **HS-Pose: Hybrid Scope Feature Extraction for Category-Level Object Pose Estimation**. Chen Wang assisted with figure drawing and proofreading, Yinghan Sun helped with tables and figures, and Esha Dasgupta and Hua Chen helped with proofreading.*

In this paper, we focus on the problem of category-level object pose estimation, which is challenging due to the large intra-category shape variation. 3D graph convolution (3D-GC) based methods have been widely used to extract local geometric features, but they have limitations for complex shaped objects and are sensitive to noise. Moreover, the scale and translation invariant properties of 3D-GC restrict the perception of an object's size and translation information. In this paper, we propose a simple network structure, the *HS-layer*, which extends 3D-GC to extract hybrid scope latent features from point cloud data for category-level object pose estimation tasks. The proposed HS-layer: 1) is able to perceive local-global geometric structure and global information, 2) is robust to noise, and 3) can encode size and translation information. Our experiments show that the simple replacement of the 3D-GC layer with the proposed HS-layer on the baseline

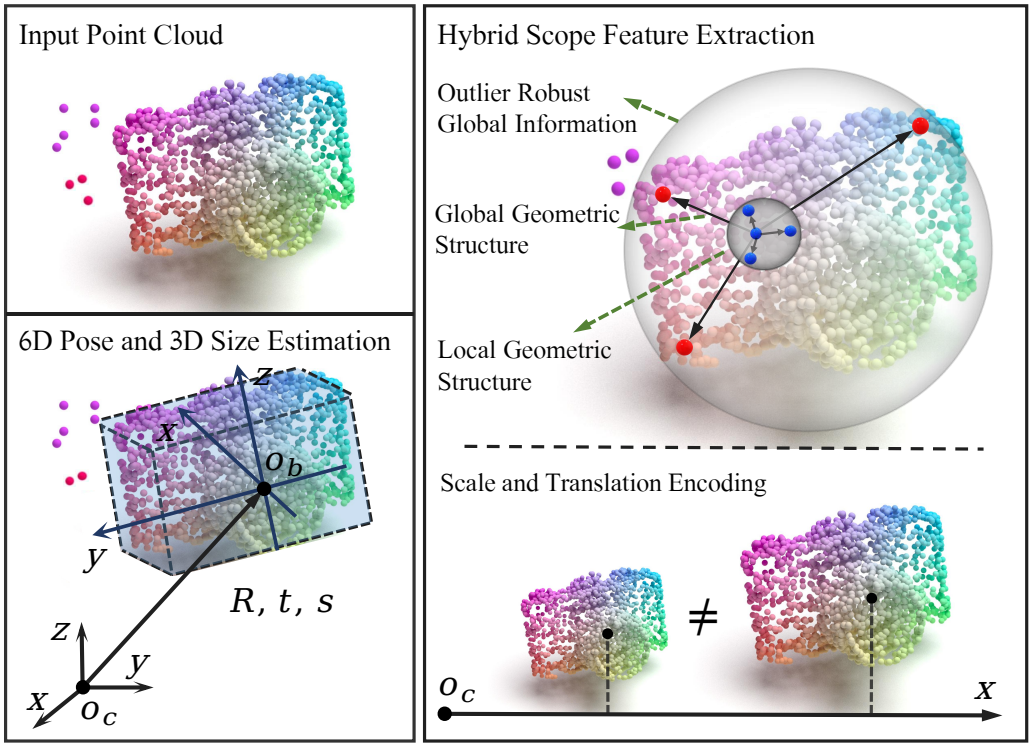



Figure 4.1: Illustration of the hybrid scope feature extraction of the HS-layer.

method (GPV-Pose) achieves a significant improvement, with the performance increased by **14.5%** on $5°2$cm metric and **10.3%** on $\text{IoU}_{75}$. Our method outperforms the state-of-the-art methods by a large margin (**8.3%** on $5°2$cm, **6.9%** on $\text{IoU}_{75}$) on REAL275 dataset and runs in real-time (50 FPS)[1].

## 4.1 Introduction

Accurate and efficient estimation of an object's pose and size is crucial for many real-world applications (B. Wen and Bekris, 2021), including robotic manipulation (Kothari et al., 2017), augmented reality (Y. Su, Rambach, et al., 2019), and autonomous driving, among others. In these applications, it is essential that pose estimation algorithms can handle the diverse range of objects

[1]Code is available: https://github.com/Lynne-Zheng-Linfang/HS-Pose

encountered in daily life. While many existing works (Jiang et al., 2022; Hansheng Chen et al., 2022; Mo et al., 2022; Y. Xu et al., 2022) have demonstrated impressive performance in estimating an object's pose, they typically focus on only a limited set of objects with known shapes and textures, aided by CAD models. In contrast, category-level object pose estimation algorithms (Sahin and Kim, 2019; Weng et al., 2021; C. Wang, Martín-Martín, et al., 2020; J. Lin, Wei, Zhihao Li, et al., 2021; Di et al., 2022) address all objects within a given category and enable pose estimation of unseen objects during inference without the target objects' CAD models, which is more suitable for daily-life applications. However, developing such algorithms is more challenging due to the shape and texture diversity within each category.

In recent years, category-level object pose estimation research (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021; R. Zhang, Di, Lou, et al., 2022; R. Zhang, Di, Manhardt, et al., 2022) has advanced rapidly by adopting state-of-the-art deep learning methods. (H. Wang et al., 2019; D. Chen et al., 2020) gain the ability to generalize by mapping the input shape to normalized or metric-scale canonical spaces and then recovering the objects' poses via correspondence matching. Better handling of intra-category shape variation is also achieved by leveraging shape priors (Tian, Ang Jr, and G. H. Lee, 2020; K. Chen and Dou, 2021; R. Zhang, Di, Manhardt, et al., 2022), symmetry priors (H. Lin, Z. Liu, C. Cheang, et al., 2022), or domain adaptation (T. Lee, B. Lee, et al., 2021; J. Lin, Wei, Ding, et al., 2022). Additionally, (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021) enhances the perceptiveness of local geometry, and (Di et al., 2022; R. Zhang, Di, Lou, et al., 2022) exploit geometric consistency terms to improve the performance further.

Despite the remarkable progress of existing methods, there is still room for improvement in the performance of the category-level object pose estimation. Reconstruction and matching-based methods (H. Wang et al., 2019; Tian, Ang Jr, and G. H. Lee, 2020; T. Lee, B. Lee, et al., 2021) are usually limited in speed due to the time-consuming correspondence-matching procedure. Recently, various methods (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021; H. Lin, Z. Liu, C. Cheang, et al., 2022; Di et al., 2022; R. Zhang, Di, Manhardt, et al., 2022; R. Zhang, Di, Lou, et al., 2022)

built on 3D graph convolution (3D-GC) (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020) have achieved impressive performance and run in real-time. They show outstanding local geometric sensitivity and the ability to generalize to unseen objects. However, only looking at small local regions impedes their ability to leverage the global geometric relationships that are essential for handling complex geometric shapes and makes them vulnerable to outliers. In addition, the scale and translation invariant nature of 3D-GC restrict the perception of object size and translation information.

To overcome the limitations of 3D-GC in category-level object pose estimation, we propose the hybrid scope latent feature extraction layer (HS-layer), which can perceive both local and global geometric relationships and has a better awareness of translation and scale. Moreover, the proposed HS-layer is highly robust to outliers. Figure 4.1 shows an illustration of the hybrid scope feature extraction of the HS-layer. To demonstrate the effectiveness of the HS-layer, we replace the 3D-GC layers in GPV-Pose (Di et al., 2022) to construct a new category-level object pose estimation framework, HS-pose. This framework significantly outperforms the state-of-the-art method and runs in real-time. Our approach extends the perception of 3D-GC to incorporate other essential information by using two parallel paths for information extraction. The first path encodes size and translation information (STE), which is missing in 3D-GC due to its invariance property. The second path extracts outlier-robust geometric features using the receptive field with the feature distance metric (RF-F) and the outlier-robust feature extraction layer (ORL).

The main contribution of this paper is as follows:

- We propose a network architecture, the hybrid scope latent feature extraction layer (HS-layer), that can simultaneously perceive local and global geometric structure, encode translation and scale information, and extract outlier-robust feature information. Our proposed HS-layer balances all these critical aspects necessary for category-level pose estimation.
- We use the HS-layer to develop a category-level pose estimation framework, HS-Pose, based

on GPV-Pose. The HS-Pose, when compared to its parent framework, has an advantage in handling complex geometric shapes, capturing object size and translation while being robust to noise.

- We conduct extensive experiments and show that the proposed method can handle complex shapes and outperforms the state-of-the-art methods by a large margin while running in real-time (50FPS).

## 4.2 Related Works

The work in this chapter is related to previous work on instance-level and category-level 6D object pose estimation.

**Instance-level object pose estimation.** Instance-level object pose estimation estimates the pose of known objects with the 3D CAD model provided. Existing methods usually achieve the pose using end-to-end regression (Kehl, Manhardt, et al., 2017; Labbé et al., 2020a; Yi Li et al., 2020), template matching (V. N. Nguyen et al., 2022; D. Cai, Heikkilä, and Rahtu, 2022; Shugurov et al., 2022), or 2D-3D correspondence-matching (Tremblay et al., 2018; J. Sun et al., 2022; Hansheng Chen et al., 2022; Haugaard and Buch, 2021; Merrill et al., 2022; Y. He, W. Sun, et al., 2020). End-to-end regression-based methods estimate object pose directly from the visual observations and have a high inference speed. Template matching methods recover the object pose by comparing the visual observation and usually exhibit robustness to textureless objects. (Stefan Hinterstoisser, Vincent Lepetit, Rajkumar, et al., 2016; Vidal, C.-Y. Lin, and Martí, 2018) use the 3D models as templates, which achieve high accuracy but suffer from low matching speed. In recent years, latent feature-based template matching methods (Sundermeyer, Z. Marton, et al., 2019; Sundermeyer, Durner, et al., 2020; Zheng, Leonardis, et al., 2022; Deng, Mousavian, et al.,

2019) have achieved real-time performance and have gained popularity. 2D-3D correspondence matching-based methods (Y. Su, Saleh, et al., 2022; Zakharov, Shugurov, and Ilic, 2019) first estimate the 2D-3D correspondences and then retrieve the objects' pose by PnP methods. They show outstanding results for textured objects. The correspondences can be sparse bounding box corners (Rad and Vincent Lepetit, 2017; Tekin, Sinha, and Pascal Fua, 2018), or distinguishable points on the object's surface (Park, Patten, and Vincze, 2019; Zhigang Li, G. Wang, and Ji, 2019; Peng et al., 2019). While the aforementioned methods have shown impressive capabilities in estimating object pose, their applicability is limited to a few objects and usually needs the corresponding CAD models.

**Category-level object pose estimation.** Category-level methods estimate the pose of unseen objects within specific categories (J. Lin, Wei, Zhihao Li, et al., 2021; Irshad et al., 2022; Manhardt, Nickel, et al., 2020; Sahin and Kim, 2019). NOCS (H. Wang et al., 2019) suggests mapping the input shape to a normalized canonical space (NOCS) and retrieving the pose by point matching. (Irshad et al., 2022; D. Chen et al., 2020; T. Lee, B. Lee, et al., 2021) enhance NOCS using a shape prior (Tian, Ang Jr, and G. H. Lee, 2020), mapping the shape to a metric scale space (D. Chen et al., 2020), or domain adaptation (T. Lee, B. Lee, et al., 2021). (K. Chen and Dou, 2021; J. Lin, Wei, Ding, et al., 2022) leverage structural similarity between the shape prior and the observed object. TransNet (H. Zhang et al., 2022) extends the targets to transparent objects. However, they show limited speed and are unsuitable for real-time applications. CATRE (X. Liu et al., 2022) explored real-time pose refinement for pose estimation. FS-Net (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021) explored local geometric relationships using 3D-GC (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020), which shows robustness to rotation estimation and runs in real-time. (Di et al., 2022; H. Lin, Z. Liu, C. Cheang, et al., 2022; R. Zhang, Di, Manhardt, et al., 2022; R. Zhang, Di, Lou, et al., 2022) inherit the utilization of 3D-GC and enhance the pose estimation performance in different ways. SAR-Net (H. Lin, Z. Liu, C. Cheang, et al., 2022) proposes shape alignment

and symmetry-aware shape reconstruction. GPV-Pose (Di et al., 2022) presents geometric-pose consistency terms and point-wise bounding box (Bbox) voting. (R. Zhang, Di, Manhardt, et al., 2022; R. Zhang, Di, Lou, et al., 2022) further enhance (Di et al., 2022) by shape deformation (R. Zhang, Di, Manhardt, et al., 2022) and residual Bbox voting (R. Zhang, Di, Lou, et al., 2022). Nonetheless, they only look at local geometric relationships and are limited in handling more complex shapes.

## 4.3 Methodology

This paper considers the category-level pose estimation problem of estimating the 6D pose and 3D size of an arbitrary instance in the same category based on visual observation. In particular, our approach estimates the 3D rotation $\mathbf{R} \in SO(3)$, the 3D translation $\boldsymbol{t} \in \mathbb{R}^3$, and the size $\boldsymbol{s} \in \mathbb{R}^3$ of object instances based on a depth image, the objects' categories, and segmentation masks. The segmentation mask and category information can be generated by object detectors (*e.g.* MaskRCNN (K. He et al., 2017)). We use point cloud data $\mathcal{P} \in \mathbb{R}^{N\times 3}$ as the direct input of our network, which is achieved by back-projecting the segmented depth data and downsampling.

Due to the fact that geometric features are essential for determining an object's pose across different shapes, the 3D graph convolution (3D-GC) (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020) is widely adopted in recent category-level object pose estimation methods (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021; Di et al., 2022; R. Zhang, Di, Manhardt, et al., 2022; R. Zhang, Di, Lou, et al., 2022; H. Lin, Z. Liu, C. Cheang, et al., 2022). In particular, GPV-Pose (Di et al., 2022) uses a 3D-GCN encoder, formed by 3D-GC layers, together with geometric consistency terms for category-level object pose estimation and achieves state-of-the-art performance. However, 3D-GC cannot perceive global geometric features, limiting its capability to handle complex geometric shapes and being sensitive to noise. Also, it is invariant to scale and translation,

which contradicts category-level pose estimation tasks (*i.e.*, size and translation estimation).

In this paper, we propose the hybrid scope geometric feature extraction layer (HS-layer) which is based on 3D-GC and keeps its local geometric sensitivity while extending it to have the following characteristics: 1) perception of global geometric structural relationships, 2) robustness to noise, and 3) encoding of size and translation information, particularly for category-level object pose estimation tasks.

### 4.3.1 Background of 3D-GC

The core unit of 3D-GC is a deformable kernel that generalizes the convolution kernel used in 2D image processing to deal with unstructured point cloud data. In particular, a 3D-GC kernel $K^S$ is defined as:

$$K^S = \{(\boldsymbol{k}_C, \boldsymbol{w}_C), (\boldsymbol{k}_1, \boldsymbol{w}_1), \ldots, (\boldsymbol{k}_S, \boldsymbol{w}_S)\}, \tag{4.1}$$

where $S$ is the total number of support vectors, $\boldsymbol{k}_C = [0, 0, 0]^T$ is the central kernel point, $\{\boldsymbol{k}_s \in \mathbb{R}^3\}_{s=1}^{S}$ are the support kernel vectors and $\boldsymbol{w}$ is the weight associated with each kernel vector. The 3D-GC kernel performs a convolution on the receptive field $R^M(\boldsymbol{p}_i)$, which is the point along with its neighbors and their associated features $\boldsymbol{f}$:

$$R^M(\boldsymbol{p}_i) = \{(\boldsymbol{p}_i, \boldsymbol{f}_i), (\boldsymbol{p}_m, \boldsymbol{f}_m) | \boldsymbol{p}_m \in \mathcal{N}^M(\boldsymbol{p}_i)\}. \tag{4.2}$$

Here $\mathcal{N}^M(\boldsymbol{p}_i)$ is the set of the $M$ nearest neighbor points of $\boldsymbol{p}_i$. In particular, in (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020) the receptive field with point distance metric (RF-P) is used for finding which of the nearest neighbors is within the point distance metric:

$$\text{dist}_p(\boldsymbol{p}_i, \boldsymbol{p}_j) = \|\boldsymbol{p}_i - \boldsymbol{p}_j\|. \tag{4.3}$$

For more details, the readers can refer to the original work (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020). It should be noted that 3D-GC has size and translation invariance by design.

Although this invariance may be benefit tasks like segmentation and classification, it harms the pose estimation task as the size and translation are the targets to estimate.

### 4.3.2 Overall framework

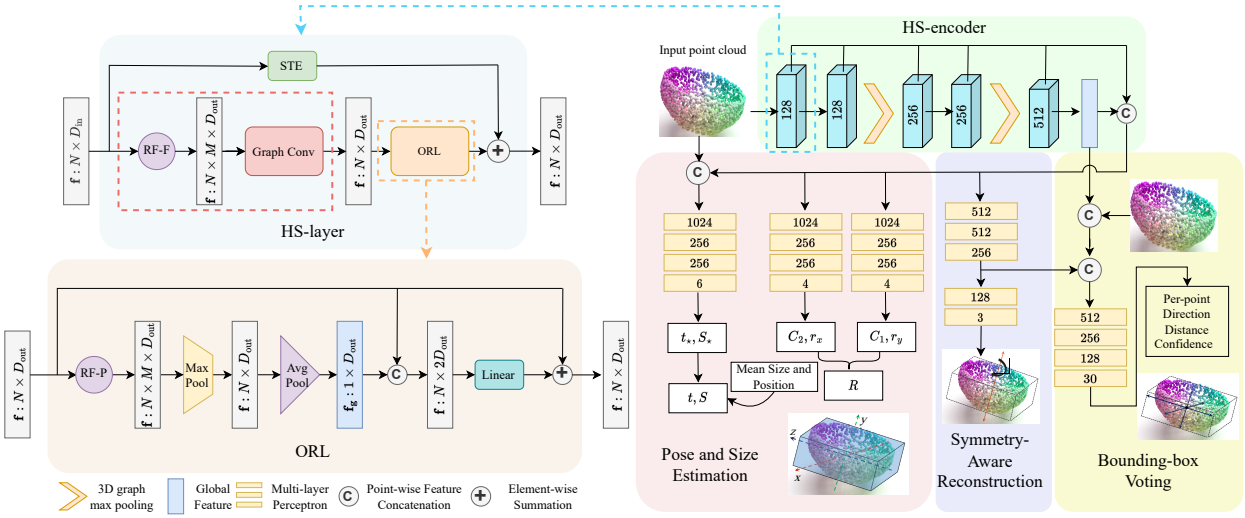



Figure 4.2: Overview of the proposed HS-Pose.

The overview of the framework, HS-Pose, is shown in Figure 4.2. We use the proposed HS-layer to form an encoder (HS-encoder) to extract the hybrid scope latent features from the input point cloud data. Then, the extracted latent features are fed into the three downstream branches for object pose estimation, symmetry-aware point cloud reconstruction, and bounding box voting, respectively. To demonstrate the effectiveness of the proposed HS-layer, which can be inserted into any category-level object pose estimation method, we construct our hybrid scope pose estimation network (HS-Pose) based on the state-of-the-art 3D-GC based GPV-Pose with minimal modification. Specifically, we only replace the 3D-GC layers of the 3D-GCN encoder of GPV-Pose with the HS-layer and keep all the other settings the same as the original GPV-Pose, which include network layers, network connection structure, and the downstream branches. Therefore, the extracted features from the encoder along with the input point cloud are fed into three modules for object

pose regression, symmetric-based point cloud reconstruction, and bounding box voting. During inference, only the encoder and the pose regression module are used.

Inside the HS-layer, we extract the hybrid scope latent features of the input using two parallel paths. The first path performs scale and translation encoding (STE), which provides essential information for size and translation estimation. The second path extracts outlier-robust geometric features by leveraging local and global geometric relationships, as well as global information in two phases. In the first phase, we form the receptive fields of points based on their feature distances (RF-F), then feed them to a graph convolution (GC) layer to extract high-level geometric features. The output of the GC layer is taken as the second phase's input and passes through an outlier-robust feature extraction layer (ORL), where each point feature is adjusted by an outlier robust global information. The final output of the HS-layer is the element-wise summation of the features of both paths.

### 4.3.3 Scale and translation encoding (STE)

As mentioned earlier, even though 3D-GC provides geometric features crucial in rotation estimation, it loses the essential translation and scale information necessary for pose estimation. To address this problem, existing 3D-GC-based methods try to use another network for translation and size estimation (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021) or concatenate the point cloud data with the extracted features for downstream estimation tasks with the assistance of other modules (*i.e.*, bounding box voting) (Di et al., 2022; R. Zhang, Di, Manhardt, et al., 2022; R. Zhang, Di, Lou, et al., 2022). While these methods are effective and all achieve improvements from the baseline, we emphasize the scale and translation information is beneficial during the latent feature extraction phase.

As shown in Figure 4.2, our suggestion is to connect in parallel a linear layer (see STE in

HS-layer in the figure) to the geometric extraction path and then perform element-wise summation for their output features:

$$\boldsymbol{f}_n^{\text{out}} = \boldsymbol{g}(\boldsymbol{f}_n) + \mathbf{h}(\boldsymbol{f}_n), \tag{4.4}$$

where $\mathbf{h}$ and $\boldsymbol{g}$ apply linear transformation and geometric feature extraction on the features of the points, respectively, and $\boldsymbol{f}_n$ is the $n$-th point's feature. In particular, we use the points' positions for size and translation encoding in the first layer since there are no features in the original point cloud. Our ablation study in Table 4.1 shows that this design choice keeps the advantage of geometric feature extraction, and boosts the performance of translation and scale estimation.

### 4.3.4 Receptive field with feature distance (RF-F)

As introduced in Sec. 4.3.1, 3D-GC learns awareness of local geometric features by forming receptive fields with point Euclidean distance metric (RF-P) and then using the deformable kernel-based graph convolution to extract geometric features for the receptive fields. However, RF-P restricts the perception to small local regions. Even though the perceived regions can be enlarged when cooperating with 3D graph pooling, it can not perceive the global geometric relationships essential for complex geometric structures. This limitation is also exhibited in the performance of category-level object pose estimation tasks (Di et al., 2022), where the methods show impressive capability in handling simple geometric shapes (*e.g.* bowl) while encountering difficulty with more complex shapes (*e.g.* mug and camera). However, this limitation has not been well addressed. To this end, we extend the 3D-GC and propose a simple manner to leverage global geometric structural relationships.

We suggest forming the receptive field with the feature distance metric (RF-F). Specifically, we find $\boldsymbol{p}_i$'s neighbors using the feature distance metric:

$$\text{dist}_f(\boldsymbol{p}_i, \boldsymbol{p}_m) = \|\boldsymbol{f}_i - \boldsymbol{f}_m\| . \tag{4.5}$$

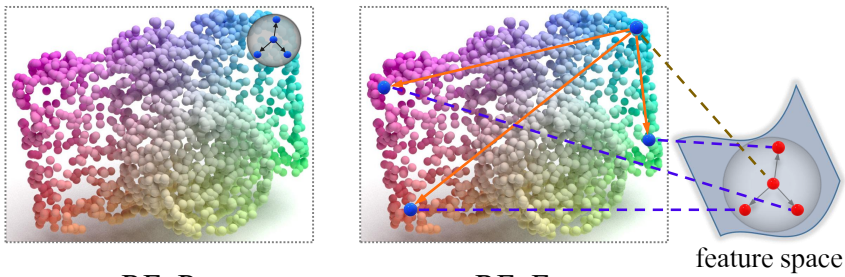



Figure 4.3: The illustration and comparison of the receptive field between RF-P and RF-F.

In other words, with the feature distance metric, the distance between two points is the Euclidean distance between their associated features. We denote the corresponding receptive fields as $R_f^M(\boldsymbol{p}_i)$. This receptive field has the advantage that it is not restricted to local regions; distant points with similar features can also be included.

Figure 4.3 shows the difference between RF-P and RF-F. RF-F can capture a larger receptive field and, therefore, can capture geometric relationships in a larger area, while the RF-P always formed with local regions.

For initialization, in the first layer, we use RF-P and set all the features $\boldsymbol{f}$ to 1. The RF-F is used in the following layers for extracting higher-level geometric relationships.

### 4.3.5 Outlier robust feature extraction layer (ORL)

3D-GC's sensitivity to noise influences the category-level methods (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021; Di et al., 2022; R. Zhang, Di, Lou, et al., 2022; R. Zhang, Di, Manhardt, et al., 2022) that are based on it. To address this problem, we introduce an outlier robust feature extraction layer (ORL) on top of the 3D-GC layer, which enhances the method's robustness to noise. The ORL is constructed as follows. Denote the input to this layer as $\{(\boldsymbol{p}_1, \boldsymbol{f}_1), \ldots, (\boldsymbol{p}_N, \boldsymbol{f}_N)\}$, where $\boldsymbol{f}_n \in \mathbb{R}^D$ is the feature of point $\boldsymbol{p}_n$. In Figure 4.4, $X$ represents the input point cloud of a

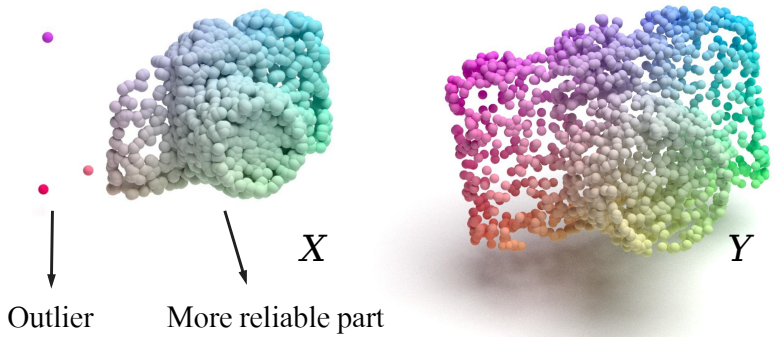



Figure 4.4: The design intuition of the outlier robust feature extraction layer (ORL).

camera, including outliers, while $Y$ depicts the complete shape. Outliers can be distracting, and their features ($\boldsymbol{f}$) should not be relied upon. Having a perception of global information, especially the more reliable part, helps the network gain resistance to noise. To focus on the global information of the more reliable part, we need a mechanism to alleviate the deviation caused by the outliers. Using the global average or maximum pooling directly is limited in addressing this, as all points are taken equally in the pooling procedure.

To lower outliers' influence, we propose using the local region as a guide to extract the global feature. As shown in Figure 4.2 (see ORL), we first use RF-P to find the $M$ nearest neighbors of each point $\mathcal{N}_p^M(\boldsymbol{p}_n)$. Then, we extract the channel-wise max features of $\mathcal{N}_p^M(\boldsymbol{p}_n)$ using a maximum pooling layer. It should be noted that the points in the reliable parts are more likely to be presented in other points' receptive fields and thus contribute more to the results of the max pooling. The output of the max pooling layer is then passed to a global average pooling layer to get the global feature $\boldsymbol{f}^{\text{global}}$. We then generate an adjusting feature using the $\boldsymbol{f}^{\text{global}}$ and the original input per-point feature $\boldsymbol{f}_n$ by first concatenating them and then feeding them to a linear layer. The final output of ORL is the result of the summation of the adjusting feature and the input features $\boldsymbol{f}_n$ of this layer.

## 4.4 Experiments

**Implementation details.** To rigorously verify the effectiveness of the proposed HS-layer and ensure a fair comparison with the baseline GPV-Pose, we construct the HS-Pose by replacing GPV-Pose's 3D-GC layer with the HS-layer while keeping the overall network structure and network parameters identical to the GPV-Pose, as shown in Figure 4.2. For a fair comparison, we choose 10 neighbors for the RF-F, consistent with the RF-P in GPV-Pose. The neighbor number of ORL is the same as the RF-F. No other parameters need to be set for the HS-layer as they only depend on the input and output. We also keep the settings, data augmentation strategy, loss terms, and their parameters, the same as those in GPV-Pose's official code[2]. Following GPV-Pose, the off-the-shelf object detector MaskRCNN (K. He et al., 2017) is employed to generate instance segmentation masks, and 1028 points are randomly sampled as the input to the network. The code is developed using `PyTorch`. We run all experiments on a computer equipped with an Intel(R) Core(TM) i9-10900K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. All categories are trained together with a batch size of 32, and the training epochs are set to 150 and 300 for REAL275 and CAMERA25 datasets, respectively. The Ranger optimizer (L. Liu et al., 2019; Yong et al., 2020; M. R. Zhang et al., 2019) is used with the learning rate starting at $1e^{-4}$ and then decreasing based on a cosine schedule for the last $28\%$ training phase.

**Baseline methods.** We use GPV-Pose (Di et al., 2022) as the baseline for the ablation study. Since GPV-Pose did not provide the performance of $10°$2cm, 2cm, and $5°$, we generate them using their official code[2]. To ensure a fair comparison of their relative speeds, we report GPV-Pose's speed on our machine using the same evaluation code as ours. The results of the other methods are taken directly from the corresponding papers.

[2] https://github.com/lolrudy/GPV_Pose

**Datasets.** We evaluate our method on REAL275 (H. Wang et al., 2019) and CAMERA25 (H. Wang et al., 2019), the two most popular benchmark datasets for category-level object pose estimation. REAL275 is a real-world dataset that provides 7k RGB-D images in 13 scenes. It contains 6 categories of objects (can, laptop, mug, bowl, camera, and bottle), and every category contains 6 instances. The training data comprises 4.3k images from 7 scenes, with 3 objects from each category shown in different scenes. The testing data includes 2.7k images from 6 scenes and 3 objects from each category. CAMERA25 is a synthetic RGB-D dataset that contains the same categories as REAL275. It provides 1085 objects for training and 184 for testing. The training set contains 275K images, and the testing set contains 25K.

**Evaluation metrics.** Following (R. Zhang, Di, Lou, et al., 2022; Di et al., 2022), we use the mean average precision (mAP) of the *3D Intersection over Union (IoU)* with thresholds of $25\%$, $50\%$, and $75\%$ to evaluate the object's size and pose together. We evaluate the rotation and translation estimation performance using the metrics of $5^\circ$, $10^\circ$, 2cm and 5cm, which means an estimation is considered correct if its corresponding error is lower than the threshold. The pose estimation performance is also evaluated using the combination of rotation and translation thresholds: $5^\circ$2cm, $5^\circ$5cm, $10^\circ$2cm, and $10^\circ$5cm.

### 4.4.1 Ablation study

To validate the proposed architecture, we conduct intensive ablation studies using the REAL275 (H. Wang et al., 2019) dataset. We incrementally add the proposed strategies (STE, RF-F, and ORL) on the baseline (GPV-Pose) to study their influences, with full results shown in Table 4.1.

We chose GPV-Pose as our baseline for two key reasons. Firstly, for its effectiveness demonstration, GPV-Pose is a leading method in category-level object pose estimation based on 3D-GC, making it ideal for showing how each component of our HS-layer affects pose estima-

Table 4.1: Ablation studies on REAL275.

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined.

| Row | Method | $IoU_{25}$ | $IoU_{50}$ | $IoU_{75}$ | $5°2cm$ | $5°5cm$ | $10°2cm$ | $10°5cm$ | 2cm | $5°$ | Speed(FPS) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A0 | GPV-Pose (baseline) | 84.2 | **83.0** | 64.4 | 32.0 | 42.9 | 55.0 | 73.3 | 69.7 | 44.7 | **69** |
| B0 | A0 + STE | 84.2 | 82.2 | 73.1 | 36.4 | 45.1 | 62.2 | 76.7 | 75.6 | 47.4 | <u>66</u> |
| B1 | A0 + RF-F | 84.2 | <u>82.8</u> | 67.7 | 38.9 | 52.3 | 62.1 | 81.8 | 71.7 | 56.1 | 65 |
| B2 | A0 + STE + RF-F | 84.1 | 82.0 | 72.0 | 42.7 | 53.7 | 63.4 | 79.2 | 75.7 | 57.0 | 64 |
| C0 | A0 + STE + RF-F + Average Pool | 84.1 | 81.7 | 73.4 | 43.7 | 54.8 | 65.7 | 81.6 | 75.7 | <u>58.5</u> | 62 |
| C1 | A0 + STE + RF-F + Max Pool | 84.2 | 81.7 | <u>74.8</u> | 44.3 | 54.5 | 66.9 | 81.8 | 77.3 | 58.1 | 62 |
| **D0** | A0 + STE + RF-F + ORL (**Full**) | 84.2 | 82.1 | 74.7 | **46.5** | <u>55.2</u> | <u>68.6</u> | <u>82.7</u> | **78.2** | 58.2 | 50 |
| E0 | D0: Neighbor number: $10 \rightarrow 20$ | **84.3** | <u>82.8</u> | **75.3** | <u>46.2</u> | **56.1** | **68.9** | **84.1** | <u>77.8</u> | **59.1** | 38 |

tion performance. Secondly, for comparison with other strategies, methods like SSP-Pose and RBP-Pose are built on GPV-Pose. SSP-Pose uses prior shape information and a shape deformation module, while RBP-Pose enhances GPV-Pose with a residual bounding box projection (SPRV) module and a shape deformation module. Comparing with these methods demonstrates how our STE and RF-F strategies perform relative to the state-of-the-art. Specifically, we show STE's effectiveness by comparing with SSP-Pose, and we highlight the impact of the RF-F approach by comparing with RBP-Pose.

**[AS-1] Scale and translation encoding (STE).** To demonstrate the effectiveness of STE and highlight the significance of scale and translation awareness when extracting latent features, we parallelly connected a single linear layer to each 3D-GC layer in the encoder of the GPV-Pose. The results in Table 4.1, specifically the [B0] row, indicate that the inclusion of STE has a significant positive impact on scale and translation estimation (**8.7%** improvement on $IoU_{75}$ and **5.9%** improvement on 2cm) while also slightly improving rotation estimation ($2.7\%$ improvement on $5°$). As shown in Table 4.5, such a simple addition even outperforms the SSP-Pose in several strict metrics ($IoU_{75}$, $5°2$cm, and $5°5$cm) and shows a notable improvement of $6.8\%$ on the $IoU_{75}$ metric,

despite that the SSP-Pose extends the GPV-Pose using a much more complex shape deformation module. The experiment results demonstrate the effectiveness of STE.

**[AS-2] Receptive field with feature distance (RF-F).** To show the usefulness of the proposed RF-F strategy and to demonstrate the importance of the global geometric relationships, we apply RF-F on GPV-Pose. From the results in Table 4.1 ([B1]), we see that RF-F has a substantial impact on rotation estimation and brings a performance leap by **11.4%** on $5^\circ$ metric. In addition, it improves the performance on $\text{IoU}_{75}$ and 2cm by $3.3\%$ and $2.0\%$, respectively, thanks to the fact that having a sense of the global geometric relationships is helpful in finding the object's center and shape boundary. When comparing the experimental results with the state-of-the-art methods in Table 4.5, our simple RF-F strategy achieves comparable performance with the state-of-the-art methods and outperforms them on the stricter metrics (*e.g.* $5^\circ$2cm and $5^\circ$5cm).

**[AS-3] The combination of RF-F and STE.** To exhibit the benefit of leveraging global geometric relationships and size-translation awareness, we conduct an experiment that combines RF-F and STE. As shown in [B2], the cooperation of RF-F and STE enhances each other and contributes to a better performance than their individual results. When compared with the baseline method, GPV-Pose, the combination of RF-F and STE improves $5^\circ$5cm by **10.8%**, $5^\circ$ by **12.3%** and $\text{IoU}_{75}$ by **7.6%**.

**[AS-4] Outlier robust feature extraction layer (ORL).** To demonstrate the effectiveness of the ORL, we add the ORL on top of [AS-3]. The results shown in the [D0] row of Table 4.1 demonstrate that using global features to adjust per-point feature extraction is helpful for both pose and size estimation with an improvement of $5.2\%$ ($10^\circ$2cm) and $2.7\%$ ($\text{IoU}_{75}$), respectively. To check the effectiveness of the outlier robust global feature, we further conduct two experiments by replacing the outlier robust global feature with two popular global pooling methods: average

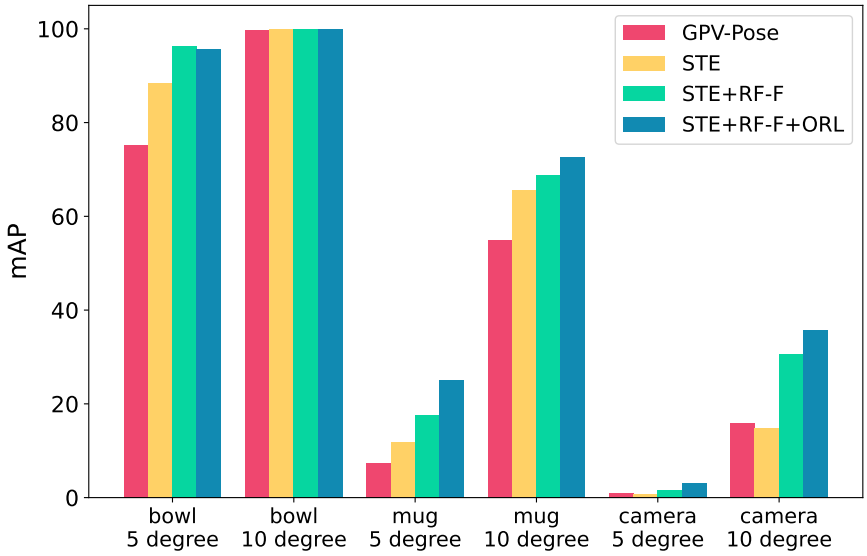



Figure 4.5: The rotation estimation performance of the proposed three strategies and GPV-Pose on categories with different geometric complexity.

pooling [C0] and max pooling [C1]. The results of [D0], [C0], and [C1] all show the contribution of global information to pose estimation. The comparison between [D0] and [C0, C1] shows that the outlier robust global feature plays a positive role and enhances the overall performance.

**[AS-5] Capability of handling complex shapes.** To exhibit the proposed method's capability in handling complex geometric shapes, we compare the rotation estimation results of the three proposed strategies (STE, RF-F, and ORL) and GPV-Pose on categories with different shape complexity in Figure 4.5. The figure illustrates the rotation estimation mAP ($5^\circ$ and $10^\circ$) for objects with different geometric complexities, with the bottle being the simplest and the camera the most complex shape. Our method significantly improves the rotation estimation for the simple shape (bowl) to nearly $100\%$ and notably enhances the rotation mAP for more complex objects (mug and camera). This demonstrates the method's ability to handle objects across different shape-complexities. The figure also demonstrates the effectiveness of leveraging global geometric relationships (STE+RF-F *vs*. STE) and shows the usefulness of outlier robust global information guided feature extraction in ORL (STE+RF-F+ORL *vs*. STE+RF-F).

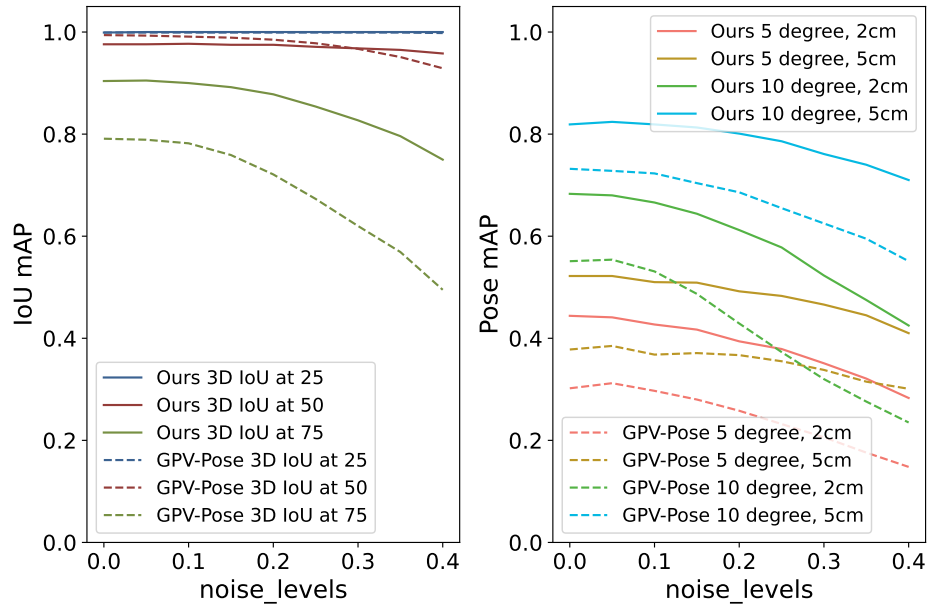



Figure 4.6: The comparison of outlier resistance between GPV-Pose and the proposed method.

**[AS-6] Noise resistance.** To demonstrate the outlier robustness of our method, we tested it alongside GPV-Pose under different outlier ratios (ranging from $0.0\%$ to $40.0\%$). Outliers are defined as points that do not belong to the target object, and the outlier ratio represents the ratio of outlier points to the total number of points in the input point cloud. We utilized the REAL275 dataset for testing and generated noisy input data by sampling points from both the background and the object region based on the specified outlier ratio. Our method consistently outperforms GPV-Pose by a significant margin across various outlier ratios, demonstrating superior stability as the outlier ratio increases. For a fair comparison, both our method and GPV-Pose were tested on the same noisy dataset. Figure 4.6 provides a visual representation of these results.

Table 4.2: Performance of our method when changing the neighbor number of RF-F.

The neighbor number of the ORL is fixed to 10 in this experiment. Overall best results are in bold, and the second-best results are underlined.

| Neighbor Number | 3 | 5 | 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|---|
| 5°2cm | 39.8 | 41.5 | **46.5** | <u>46.1</u> | 44.3 | 41.6 |
| 5°5cm | 49.2 | 51.4 | <u>55.2</u> | **56.7** | 54.7 | 54.4 |
| $IoU_{75}$ | 72.9 | 72.8 | **74.7** | <u>73.4</u> | **74.7** | 71.9 |
| Speed (FPS) | **64** | 60 | 50 | 41 | 34 | 30 |

### 4.4.2 Influence of neighbor numbers

To understand how neighbor numbers affect performance, we tested our method with different neighbor numbers (from 3 to 40) in the RF-F and ORL components[3]. We conducted experiments in three groups: 1) changing RF-F's neighbor number with fixed ORL's neighbor number, 2) changing ORL's neighbor number with fixed RF-F's neighbor number, and 3) changing both neighbor numbers simultaneously. Results indicate that optimal performance occurs within a specific neighbor number range. Using the same neighbor numbers in ORL and RF-F improves performance, with the best precision achieved when both neighbor numbers are 20 or 30. We also show results for 20 neighbors, which outperform those with 10 neighbors, in row [E0] of Table 4.1. Notably, for a fair comparison with GPV-Pose and to focus on the HS-layer's design, we use the results with 10 neighbors (as GPV-Pose) in all tables and figures unless specified otherwise.

**Change in RF-F's neighbor number.** Table 4.2 displays our method's performance with different RF-F neighbor numbers, with fixed ORL neighbor number at 10. Increasing RF-F's neighbor number leads to longer processing times, with a certain range (around 10-20 neighbors) producing

[3]Due to GPU memory limitations, we adjusted the batch size for each neighbor number (16 for 3-20 neighbors, 8 for 30-40 neighbors).

better precision. Speed decreases from 64 FPS to 30 FPS when increasing neighbors from 3 to 40. Performance on $5°2$cm peaks at 10 neighbors (46.5%), then declines to 41.6% at 40 neighbors. Generally, using 10 neighbors for RF-F achieves the best balance of performance and speed, as too few or too many neighbors can degrade precision. Insufficient or excessive neighbor numbers can adversely affect precision, as too few neighbors may not fully capture global geometric features, while too many may obscure geometric structural information in the receptive field.

Table 4.3: Performance of our method when changing the neighbor number of ORL.

The neighbor number of RF-F is fixed to 10 in this experiment. Overall best results are in bold, and the second-best results are underlined.

| Neighbor Number | 3 | 5 | 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|---|
| $5°2$cm | 43.3 | <u>43.6</u> | **46.5** | 42.7 | 43.1 | 39.4 |
| $5°5$cm | 53.1 | 53.0 | <u>55.2</u> | **55.3** | 54.4 | 53.7 |
| $\text{IoU}_{75}$ | 74.6 | 72.8 | **74.7** | 72.7 | 73.9 | 71.1 |
| Speed (FPS) | **52** | 51 | 50 | 48 | 48 | 49 |

**Change in ORL's neighbor number.** Table 4.3 shows the proposed method's performance with different ORL neighbor numbers, with fixed RF-F neighbor number at 10. Speed remains relatively stable, dropping by only 4 FPS from 3 to 40 neighbors. Precision benefits from an appropriate range of neighbor numbers, with 10 neighbors performing better than other values. Compared to RF-F, ORL's neighbor number has a lesser impact on speed, likely due to finding neighbors in a lower dimensional space.

**Simultaneous change in neighbor numbers.** Table 4.4 presents the method's performance when both ORL and RF-F neighbor numbers change simultaneously. Optimal performance occurs around 10-30 neighbors, where using the same number of neighbors in both components improves precision. Increasing neighbor numbers in both components helps balance finding reliable points

Table 4.4: Our performance when changing the neighbor number of the ORL and RF-F together. The neighbor numbers of ORL and RF-F are the same in this experiment. Overall best results are in bold, and the second-best results are underlined.

| Neighbor Number | 3 | 5 | 10 | 20 | 30 | 40 |
|---|---|---|---|---|---|---|
| $5°2$cm | 39.0 | 41.7 | **46.5** | 46.2 | <u>46.4</u> | 39.6 |
| $5°5$cm | 47.6 | 52.8 | 55.2 | <u>56.1</u> | **56.6** | 55.8 |
| $IoU_{75}$ | 73.1 | 72.7 | 74.7 | **75.3** | <u>75.2</u> | 70.3 |
| Speed (FPS) | 64 | 59 | 50 | 38 | 30 | 26 |

and rejecting outliers, leading to better performance. However, too many neighbors can still degrade performance by including too much noise.

### 4.4.3 Comparison with state-of-the-art methods

**Results on the REAL275 dataset.** We compare HS-Pose with state-of-the-art methods in Table 4.5, focusing on depth-only methods for pose estimation to ensure a fair comparison. As shown in the table, our method outperforms the state-of-the-art in all metrics except for IoU$50$, where it also shows comparable performance. HS-Pose achieves real-time performance and outperforms the second-ranked method on strict metrics by a large margin, with improvements of **8.3%** on $5°2$cm, **7.1%** on $5°5$cm, and **6.9%** on IoU$75$. Figure 4.7 presents the average precision of each category under different thresholds, comparing HS-Pose with GPV-Pose.

Furthermore, we compare our depth-only method with RGB-D-based approaches in Table 4.6. HS-Pose achieves competitive results with RGB-D-based approaches, ranking first in 5 out of 9 metrics and second in the rest. We significantly outperform them in several pose estimation metrics, such as **56.1%** (HS-Pose) *vs*. $50.7\%$ on $5°5$cm metric, and **84.1%** (HS-Pose) *vs*. $78.4\%$ on $10°5$cm metric. Many of these approaches were trained with synthetic data or used mixed training

Table 4.5: Comparison with the state-of-the-art methods (depth only) on REAL275 dataset.

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined.

| Method | $IoU_{25}$ | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 10°10cm | Speed(FPS) |
|---|---|---|---|---|---|---|---|---|---|
| SAR-Net (H. Lin, Z. Liu, C. Cheang, et al., 2022) | - | 79.3 | 62.4 | 31.6 | 42.3 | 50.3 | 68.3 | - | 10 |
| FS-Net [4] (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021) | 84.0 | 81.1 | 63.5 | 19.9 | 33.9 | - | 69.1 | 71.0 | 20 |
| UDA-COPE (T. Lee, B. Lee, et al., 2021) | - | 79.6 | 57.8 | 21.2 | 29.1 | 48.7 | 65.9 | - | - |
| SSP-Pose (R. Zhang, Di, Manhardt, et al., 2022) | 84.0 | 82.3 | 66.3 | 34.7 | 44.6 | - | 77.8 | 79.7 | 25 |
| RBP-Pose (R. Zhang, Di, Lou, et al., 2022) | - | - | 67.8 | 38.2 | 48.1 | 63.1 | 79.2 | - | 25 |
| GPV-Pose Di et al., 2022 | 84.1 | **83.0** | 64.4 | 32.0 | 42.9 | 55.0 | 73.3 | 74.6 | **69** |
| **Ours** | **84.2** | 82.1 | **74.7** | **46.5** | **55.2** | **68.6** | **82.7** | **83.7** | 50 |

Table 4.6: Comparison with the state-of-the-art methods (RGB-D) on the REAL275 dataset.

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined. *Type* lists the input data type for pose estimation. *Syn.* denotes whether the synthetic data is used during training.

| Method | Type | Syn. | $IoU_{25}$ | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 10°10cm | Speed(FPS) |
|---|---|---|---|---|---|---|---|---|---|---|---|
| NOCS (H. Wang et al., 2019) | RGB-D | ✓ | **84.9** | 80.5 | 30.1 | - | 9.5 | 13.8 | 26.7 | 26.7 | 5 |
| CASS (D. Chen et al., 2020) | RGB-D | ✓ | 84.2 | 77.7 | 15.3 | 19.5 | 23.5 | 50.8 | 58.0 | 58.3 | - |
| SPD (Tian, Ang Jr, and G. H. Lee, 2020) | RGB-D | ✓ | 83.4 | 77.3 | 53.2 | 19.3 | 21.4 | 43.2 | 54.1 | - | 4 |
| DualPoseNet (J. Lin, Wei, Zhihao Li, et al., 2021) | RGB-D | ✓ | - | 79.8 | 62.2 | 29.3 | 35.9 | 50.0 | 66.8 | - | 2 |
| SGPA (K. Chen and Dou, 2021) | RGB-D | ✓ | - | 80.1 | 61.9 | 35.9 | 39.6 | 61.3 | 70.7 | - | - |
| CR-Net (J. Wang, K. Chen, and Dou, 2021) | RGB-D | ✓ | - | 79.3 | 55.9 | 27.8 | 34.3 | 47.2 | 60.8 | - | - |
| Self-DPDN (J. Lin, Wei, Ding, et al., 2022) | RGB-D | ✓ | - | **83.4** | **76.0** | 46.0 | 50.7 | **70.4** | 78.4 | - | - |
| **Ours (10 neighbors)** | D | | 84.2 | 82.1 | 74.7 | **46.5** | 55.2 | 68.6 | 82.7 | 83.7 | 50 |
| **Ours (20 neighbors)** | D | | 84.3 | 82.8 | 75.3 | 46.2 | **56.1** | 68.9 | **84.1** | **85.2** | 38 |

with CAMERA25 and REAL275, resulting in larger training sets with over 1K objects. Additionally, they often exhibit limited inference speed. In contrast, HS-Pose is trained using REAL275 with only 1.6k images and 18 objects, achieving real-time performance.

**Results on the CAMERA25 Dataset.** The performance comparison of the proposed method and the state-of-the-art is shown in Table 4.7. Our method ranks top and second on all the metrics

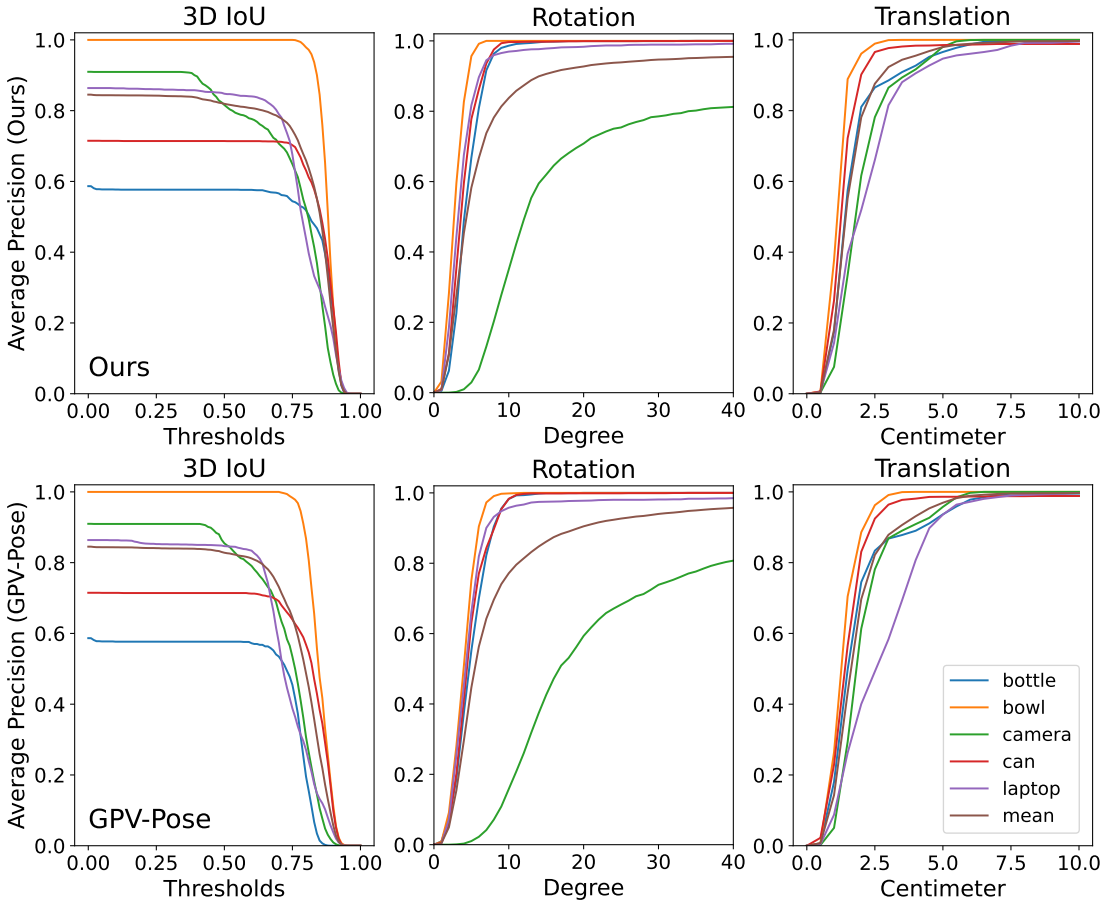



Figure 4.7: Per-category comparison between our method and GPV-Pose.

without prior information. Of the four scores ranked second, three are close to the tops with negligible differences ($0.1\%$ on $10°5$cm and $\text{IoU}_{75}$ metrics, and $0.2\%$ on $5°2$cm metric). It is also worth noting that CAMERA25 is a synthetic dataset that contains no noise, so one main contribution of the proposed method, noise robustness, is not reflected in this dataset. However, this contribution can be identified by comparing the proposed and the state-of-the-art methods' performance on the CAMERA25 and the REAL275 dataset. The REAL275 dataset contains the same object categories as the CAMERA25 but is real-world collected and contains complex noise. It can be observed that the performance drop of our method is much less than other methods when encountering real-world noises in the REAL275. This demonstrates that our method is more noise-robust compared with other methods.

[4]We use the result provided by GPV-Net, which is higher than the reported result in the FS-Net paper.

Table 4.7: Comparison with state-of-the-art methods (depth-only) on the CAMERA25 dataset.

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined. *Prior* denotes whether the method uses shape priors.

| Method | Prior | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm |
|---|---|---|---|---|---|---|---|
| SAR-Net (H. Lin, Z. Liu, C. Cheang, et al., 2022) | ✓ | 86.8 | 79.0 | 66.7 | 70.9 | 75.3 | 80.3 |
| SSP-Pose (R. Zhang, Di, Manhardt, et al., 2022) | ✓ | - | 86.8 | 64.7 | 75.5 | - | 87.4 |
| RBP-Pose (R. Zhang, Di, Lou, et al., 2022) | ✓ | 93.1 | <u>89.0</u> | **73.5** | <u>79.6</u> | **82.1** | **89.5** |
| GPV-Pose (Di et al., 2022) | | **93.4** | 88.3 | 72.1 | 79.1 | - | 89.0 |
| **Ours** | | <u>93.3</u> | **89.4** | <u>73.3</u> | **80.5** | <u>80.4</u> | <u>89.4</u> |

We also compare the proposed method with the state-of-the-art RGB-D-based methods on the CAMERA25 dataset and show the results in Table 4.8. We achieved top and second scores on 5 out of 6 metrics (4 tops and 1 second) without the need for RGB data or shape prior.

Table 4.8: Comparison with state-of-the-art methods (RGB-D) on the CAMERA25 dataset.

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined. *Type* lists the input data type for pose estimation. *Syn.* denotes whether the synthetic data is used during training.

| Method | Type | Prior | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm |
|---|---|---|---|---|---|---|---|---|
| SPD(Tian, Ang Jr, and G. H. Lee, 2020) | RGB-D | ✓ | 93.2 | 83.1 | 54.3 | 59.0 | 73.3 | 81.5 |
| CR-Net (J. Wang, K. Chen, and Dou, 2021) | RGB-D | ✓ | **93.8** | 88.0 | 72.0 | 76.4 | <u>81.0</u> | 87.7 |
| SGPA (K. Chen and Dou, 2021) | RGB-D | ✓ | 93.2 | 88.1 | 70.7 | 74.5 | **82.7** | 88.4 |
| NOCS (H. Wang et al., 2019) | RGB-D | | 83.9 | 69.5 | 32.3 | 40.9 | 48.2 | 64.6 |
| DualPoseNet (J. Lin, Wei, Zhihao Li, et al., 2021) | RGB-D | | 92.4 | 86.4 | 64.7 | 70.7 | 77.2 | 84.7 |
| **Ours (10 neighbors)** | D | | 93.3 | **89.4** | <u>73.3</u> | <u>80.5</u> | 80.4 | <u>89.4</u> |
| **Ours (20 neighbors)** | D | | <u>93.4</u> | <u>89.3</u> | **74.0** | **82.0** | 80.3 | **90.2** |

**Qualitative results.** Figure 4.8 presents a qualitative comparison between GPV-Pose and our method. Ground truth results are shown with white lines, GPV-Pose's results in blue, and ours in green. For symmetric objects like the bowl, bottle, and can, correct estimated rotations align with the symmetry axis. As can be seen from the figure, our method achieves a better size and pose estimation (*e.g.* the first three columns), shows robustness to occlusion (*e.g.* the laptop in the last column), and handles complex shapes better (*e.g.* the cameras and mugs in each column).

### 4.4.4 Per-Category results on REAL275 and CAMERA25

The per-category results trained on the REAL275 and CAMERA25 datasets are shown in Table 4.9 and Table 4.10, respectively.

Table 4.9: Per-category results of our method on REAL275 dataset.

| Category | $IoU_{25}$ | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 10°10cm | 5° | 10° | 2cm | 5cm | 10cm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bottle | 57.7 | 57.7 | 54.8 | 43.0 | 53.1 | 80.0 | 95.4 | 98.5 | 66.9 | 99.2 | 81.0 | 96.5 | 99.5 |
| bowl | 100.0 | 100.0 | 100.0 | 92.1 | 95.6 | 96.5 | 100.0 | 100.0 | 95.6 | 100.0 | 96.5 | 100.0 | 100.0 |
| camera | 90.9 | 82.3 | 65.2 | 2.3 | 3.1 | 28.3 | 35.7 | 35.8 | 3.1 | 35.8 | 60.9 | 98.1 | 100.0 |
| can | 71.4 | 71.4 | 70.5 | 68.6 | 75.0 | 90.0 | 98.5 | 98.5 | 77.8 | 99.7 | 90.0 | 98.6 | 98.8 |
| laptop | 86.1 | 84.9 | 67.1 | 49.3 | 79.5 | 52.4 | 94.0 | 96.8 | 80.8 | 96.9 | 52.4 | 94.6 | 99.5 |
| mug | 99.2 | 96.4 | 90.8 | 23.8 | 25.0 | 64.6 | 72.6 | 72.6 | 25.0 | 72.6 | 88.5 | 100.0 | 100.0 |
| average | 84.2 | 82.1 | 74.7 | 46.5 | 55.2 | 68.6 | 82.7 | 83.7 | 58.2 | 84.0 | 78.2 | 98.0 | 99.6 |

### 4.4.5 Inference speed

Our method achieves an inference speed of 50 frames per second (FPS) on a computer equipped with an Intel(R) Core(TM) i9-10900K CPU, 32 GB RAM, and an NVIDIA GeForce RTX 3090 GPU. It's important to note that inference speeds can vary depending on the hardware configuration

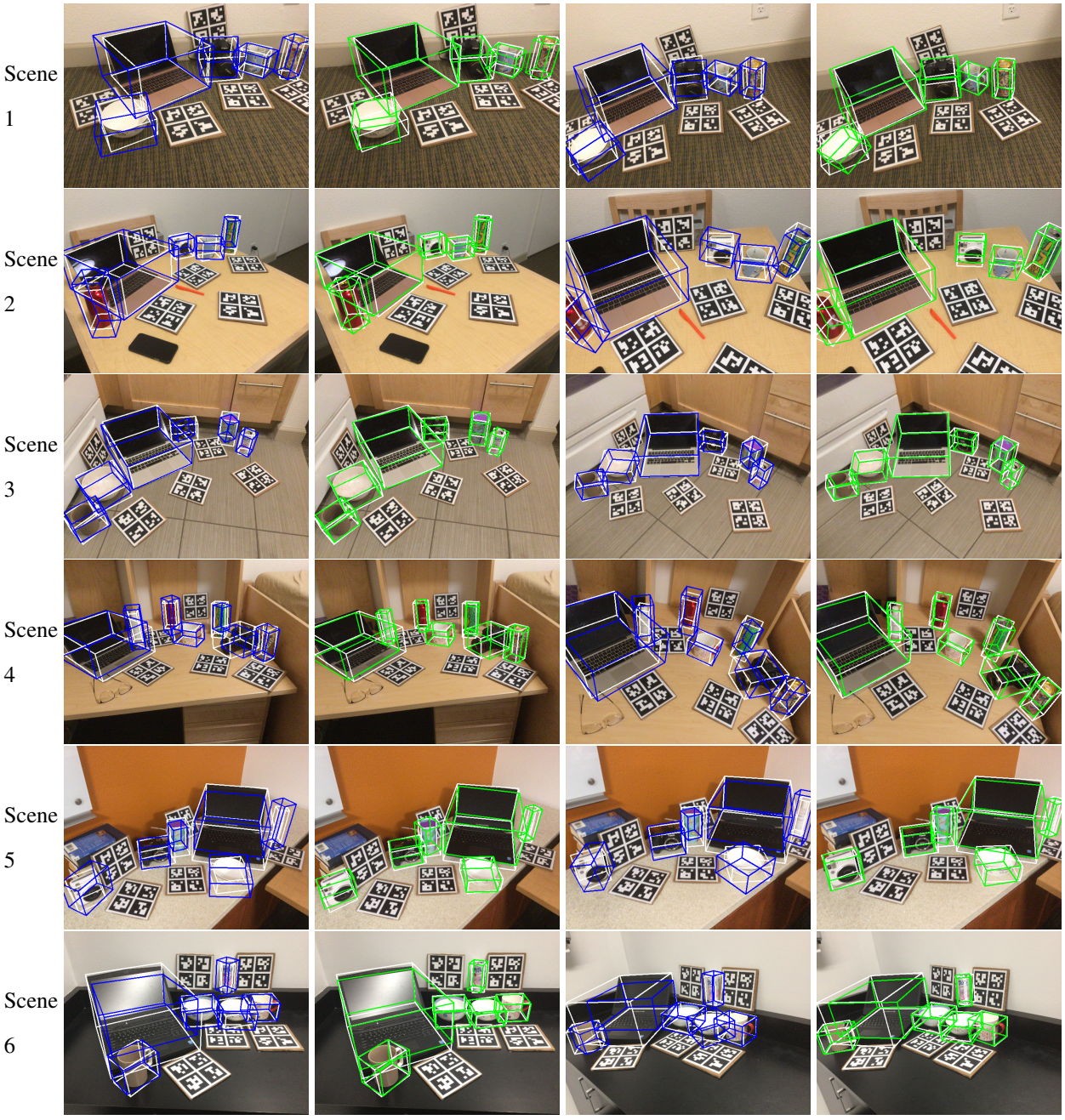



Figure 4.8: Qualitative results of our method (green line) and the GPV-Pose (blue line) on the REAL275 dataset.

Table 4.10: Per-category results of our method on CAMERA25 dataset.

| category | $IoU_{25}$ | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 10°10cm | 5° | 10° | 2cm | 5cm | 10cm |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| bottle | 93.9 | 93.8 | 90.9 | 80.1 | 96.7 | 80.7 | 97.8 | 99.4 | 98.5 | 99.8 | 80.7 | 97.9 | 99.5 |
| bowl | 96.9 | 96.8 | 96.8 | 98.4 | 98.6 | 99.4 | 99.8 | 99.8 | 98.7 | 99.8 | 99.4 | 99.8 | 99.9 |
| camera | 94.8 | 85.4 | 74.3 | 51.2 | 55.1 | 65.0 | 70.6 | 70.9 | 55.5 | 71.4 | 86.9 | 99.0 | 99.6 |
| can | 92.5 | 92.4 | 92.2 | 99.0 | 99.4 | 99.0 | 99.5 | 99.5 | 99.9 | 100.0 | 99.0 | 99.5 | 99.6 |
| laptop | 98.4 | 97.4 | 90.6 | 75.6 | 85.2 | 81.1 | 92.7 | 97.0 | 89.0 | 97.1 | 83.3 | 95.6 | 99.9 |
| mug | 94.1 | 93.8 | 91.9 | 35.4 | 47.9 | 57.4 | 76.2 | 76.2 | 49.1 | 76.9 | 75.9 | 99.5 | 99.6 |
| average | 95.1 | 93.3 | 89.4 | 73.3 | 80.5 | 80.4 | 89.4 | 90.5 | 81.8 | 90.8 | 87.5 | 98.6 | 99.7 |

of the machine. Therefore, we use the speed results to demonstrate the real-time performance of our method without emphasizing direct speed comparisons with other methods.

To provide a fair speed comparison with the baseline method, GPV-Pose, we report the speed of GPV-Pose on our machine using the same evaluation code as ours. GPV-Pose achieved a speed of 69 FPS on our system, which is notably faster than the speed reported in the original paper (20 FPS). This speed difference can be attributed to two main factors: machine differences and evaluation code optimization.

**Machine differences.** The original paper for GPV-Pose reports speed tests on a single TITAN X GPU, while we tested GPV-Pose on a single RTX 3090 GPU with an Intel(R) Core(TM) i9-10900K CPU, 32 GB RAM. This difference in hardware configuration resulted in a speed of 33 FPS on our machine.

**Evaluation code optimization.** Our evaluation code is a refactored version of GPV-Pose's code. We change some for-loop operations to batch operations and remove unnecessary calculations (*e.g.* the bounding box voting and symmetric point cloud reconstruction) during inference. These

optimizations significantly improved the speed from 33 FPS to 69 FPS. Importantly, all the changes have passed unit tests to to ensure they produced the same results as the original code.

## 4.5 Conclusion

In this paper, we proposed a hybrid scope latent feature extraction layer, the HS-layer, and used it to construct a category-level object pose estimation framework HS-Pose. Based on the advantages of the HS-layer, HS-Pose can handle complex shapes, capture an object's size and translation, and is robust to noise. The capability of the overall framework is demonstrated in the experiments. The comparisons with the existing methods show that our HS-Pose achieves state-of-the-art performance. In future work, we plan to apply our proposed HS-layer to other problems where unstructured data needs to be processed, and the combination between the local and the global information becomes critical.

# Chapter Five

# Achieving High-Precision in Category-level Applications

*This chapter presents work that has been accepted by* ***The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR) 2024*** *in Seattle (WA), USA (Zheng, Tse, et al., 2024). Tze Ho Elden Tse and Chen Wang assisted with figure drawing and draft writing, Yinghan Sun helps with tables and figures, and Hua Chen helps with the draft writing.*

Object pose refinement is essential for robust object pose estimation. Previous work has made significant progress toward instance-level object pose refinement. Yet, category-level pose refinement is a more challenging problem due to the large shape variations of different objects within a category and the discrepancy between the target object and the shape prior. To address these challenges, we introduce a novel architecture for category-level object pose refinement. Our approach integrates an HS-layer and learnable affine transformations, which aims to enhance the extraction and alignment of geometric information. Additionally, we introduce a cross-cloud transformation mechanism that efficiently merges diverse data sources. Finally, we push the limits of our model by incorporating the shape prior information for translation and size error prediction. We conducted extensive experiments to demonstrate the effectiveness of the proposed framework. Through extensive quantitative experiments, we demonstrate significant improvement over

the baseline method by a large margin across all metrics, *i.e.* $\text{IoU}_{75}$ and $10°2\text{cm}$ improvement by **8.2%** and **10.5%**, respectively. [1]

## 5.1 Introduction

Understanding an object's pose is crucial for a wide range of real-world applications, including robotic manipulation (Kappler et al., 2018; Morgan et al., 2021; K. Zhang et al., 2023), augmented reality (Marder-Eppstein, 2016; Nee et al., 2012), and autonomous driving (Y. Su, Rambach, et al., 2019; Kothari et al., 2017). Significant progress has been made for object pose estimation (Hanzhi Chen et al., 2023; Hodan, Barath, and Jiri Matas, 2020; Hai et al., 2023; Yang and Pavone, 2023) and pose refinement (Yi Li et al., 2018; Iwase et al., 2021; B. Wen, Mitash, et al., 2020; Castro and Kim, 2023; Z. Zhang et al., 2023) using the object's CAD model. Despite the promising performance, the reliance on accurate instance-level CAD models limits their generalizability to everyday objects. Category-level methods have therefore been proposed to overcome this limitation (Zheng, C. Wang, et al., 2023; J. Lin, Wei, Ding, et al., 2022; X. Liu et al., 2022; K. Chen, James, et al., 2023). The objective of this line of work focuses on estimating object poses within a category given category-level shape priors. As a result, they face unique challenges as there exist diverse shape variations in each object category. We illustrate these shape variations in Fig. 5.1, where *SP-m* is the mean shape of a category, and *SP-1* and *SP-2* are randomly sampled shapes from the categories.

Recently, there have been remarkable advancements in category-level object pose estimation (Zheng, C. Wang, et al., 2023; Di et al., 2022; R. Zhang, Di, Lou, et al., 2022; W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021; J. Lin, Wei, Ding, et al., 2022), primarily due to effective utilization of geometric information through 3D graph convolution (Z.-H. Lin, S.-Y. Huang, and

[1] Project page: https://lynne-zheng-linfang.github.io/georef.github.io

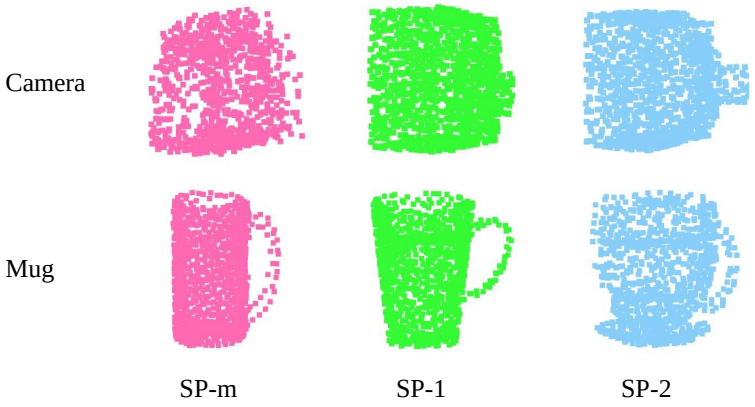



Figure 5.1: Examples of the shape variation.

Y.-C. F. Wang, 2020). In applications that require high precision, it is common to employ an object pose refinement procedure in conjunction with pose estimation. This involves an initial pose estimation algorithm determining the object pose, followed by a refinement step to further enhance the accuracy of the initially estimated pose by predicting and correcting its error. However, while instance-level object pose refinement has been extensively studied, category-level pose refinement remained unexplored until the introduction of CATRE (X. Liu et al., 2022). By leveraging initial object pose and size estimations, CATRE achieves category-level pose refinement by iteratively aligning the observed target object point cloud with the category-level shape prior. This pipeline is shown to be effective by improving the accuracy of the initial pose and size estimations.

While CATRE has proven to be effective in many scenarios, it is limited by the reliance on the PointNet (Qi, H. Su, et al., 2017) encoder, which is primarily designed for classification and segmentation tasks. This design choice limits its ability to capture essential and fine-grained geometric relationships for accurate pose estimation and refinement. This ability is particularly important in category-level pose refinement as there exists diverse shape variations between inputs. Consequently, CATRE obtains suboptimal performance by direct application of 3D graph convolution. In addition, as CATRE treats the point cloud and the shape prior features separately until a later stage of the network, they potentially miss out on the benefits of integrating these features earlier. Moreover, their approach does not incorporate shape prior information into the translation

and scale estimation module, which presents another area for potential improvement.

In this paper, we introduce a novel architecture for category-level object pose refinement which aims to address the limitations mentioned above. To better extract both local and global geometric information, we incorporate an HS layer into our feature extraction process. We apply learnable affine transformations to the features to address the geometric discrepancies between the observed point cloud and the shape prior. This enables the network to align these features more effectively. In addition, we propose a cross-cloud transformation mechanism that is specifically designed to enhance the merging of information between the observed point clouds and the shape prior. This mechanism enables more efficient integration of information between the two sources. Finally, we push the limit of our model by incorporating shape prior information to more accurately predict errors in translation and size estimation.

Our extensive experimental results on two category-level object pose datasets demonstrate that our proposed model to be effective in addressing the problem of shape variations in category-level object pose refinement, and consequently outperforms the state-of-the-art significantly. To the best of our knowledge, our proposed method is the first to successfully address the shape variation issue which is common in category-level pose refinement. Specifically, to enable graph convolution to be effective in capturing geometric relationships between different shapes, we propose an adaptive affine transformation matrix that aligns the observed point clouds and the shape prior. Additionally, the proposed cross-cloud transformation mechanism effectively fuses features from different input point clouds and brings further performance improvements.

Our contributions are as follows:

- We introduce a novel architecture to specifically address the shape variations issue in category-level object pose refinement. Our proposed method results in consistent performance gain and exhibits better generalization ability.

- We propose a unique cross-cloud transformation mechanism which efficiently merges diverse information from observed point clouds and shape priors.
- We conduct extensive experiments on two category-level object pose datasets to validate our proposed method. On the REAL275 dataset, our method significantly outperforms SPD by 39.1% increase in the $5°5$cm metric. Additionally, we achieve 10.5% improvement in the $10°2$cm metric over the state-of-the-art method, CATRE.

## 5.2 Related Work

The work in this chapter is related to previous work on 6D object pose estimation and refinement at the instance level and category level.

**Instance-level object pose estimation and refinement.** Instance-level approaches estimate the pose of the target object given known 3D CAD models. They can be briefly divided into correspondence-matching methods and template-matching methods. Correspondence matching methods (Merrill et al., 2022; Rad and Vincent Lepetit, 2017; W. Chen, Jia, H. J. Chang, Duan, and Leonardis, 2020; Y. He, W. Sun, et al., 2020; Tremblay et al., 2018; Y. Su, Saleh, et al., 2022; Zakharov, Shugurov, and Ilic, 2019; J. Sun et al., 2022; Hansheng Chen et al., 2022; Haugaard and Buch, 2021; Tekin, Sinha, and Pascal Fua, 2018; J. Zhou et al., 2023) matches the outstanding features of the observed object images with its model. Template matching methods (Sundermeyer, Z. Marton, et al., 2019; Zheng, Leonardis, et al., 2022; Shugurov et al., 2022; D. Cai, Heikkilä, and Rahtu, 2022; Sundermeyer, Durner, et al., 2020; Stefan Hinterstoisser, Vincent Lepetit, Rajkumar, et al., 2016; Vidal, C.-Y. Lin, and Martí, 2018; V. N. Nguyen et al., 2022) compares the images or extracted features with the pre-generated templates. As the initial pose estimates can be noisy to various factors such as occlusions, object pose refinement (Yi Li et al., 2018; Labbé et al., 2020a; B. Wen, Mitash,

et al., 2020) is shown to be useful in improving the performance of instance-level methods. Even though they achieved impressive over the target object, the reliance on object CAD models limited their generalizability for handling everyday objects. In this paper, we consider a more challenging problem setting where only the category-level shape prior is provided.

**Category-level object pose estimation and refinement.** Both tasks mainly focus on addressing the shape variation between the objects. The pioneering work NOCS (H. Wang et al., 2019) tackles the shape discrepancy by recovering the normalized visible shape of the target object and achieving the pose by point cloud matching. A series of methods extend this structure by leveraging different information such as domain adaptation (T. Lee, B. Lee, et al., 2021), different reconstruction space (D. Chen et al., 2020), shape prior (Tian, Ang, and G. H. Lee, 2020; T. Lee, B. Lee, et al., 2021; D. Chen et al., 2020; Irshad et al., 2022), and structural similarities (K. Chen and Dou, 2021; J. Lin, Wei, Ding, et al., 2022). However, this line of work is often limited in speed due to the iterative point matching. Another series of work starts with FS-Net (W. Chen, Jia, H. J. Chang, Duan, Shen, et al., 2021), which adopts 3D graph convolution (3D-GC) (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020) to obtain geometric sensitivity. Due to its effectiveness and real-time performance, graph convolution is widely adopted in recent methods with an enhancement in directions including loss function (Di et al., 2022), bounding box voting (R. Zhang, Di, Lou, et al., 2022), and shape deformation (R. Zhang, Di, Manhardt, et al., 2022). HS-Pose (Zheng, C. Wang, et al., 2023) extends the geometric feature extraction from local to global, which enhances the capability to handle objects with complex shapes. The research on category-level refinement began recently with the proposal of CATRE (X. Liu et al., 2022). It introduced an effective pipeline that leverages shape priors and a focalization strategy for pose refinement and effectively improves the initial pose estimations. In this paper, we extend the CATRE and tackle the geometric variation issue within the framework of category-level pose refinement.

## 5.3 Methodology

### 5.3.1 Problem formulation

In this paper, we tackle the problem of category-level object pose refinement. Given the initial pose and size estimation ($\mathbf{R}_0$, $\mathbf{t}_0$, $\mathbf{s}_0$), along with the observed point cloud $\mathcal{O} \in \mathbb{R}^{N^O \times 3}$ and the shape prior $\mathcal{P} \in \mathbb{R}^{N^P \times 3}$, we aim to predict the estimation error ($\Delta\mathbf{R}$, $\Delta\mathbf{t}$, $\Delta\mathbf{s}$) between the initial estimations and the ground truths. The pose refinement algorithm $\phi$ can be described as:

$$(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s}) = \phi(\mathbf{R}_0, \mathbf{t}_0, \mathbf{s}_0, \mathcal{O}, \mathcal{P}). \tag{5.1}$$

This pose refinement algorithm $\phi$ can be applied iteratively to improve the refinement performance.

### 5.3.2 Preliminaries

Our proposed category-level object pose refinement framework builds upon previous work, CATRE (X. Liu et al., 2022) which we briefly review in the following.

CATRE is the first framework that considers the problem of category-level pose refinement. It predicts the error between the ground truth and the estimated poses by aligning the input point clouds and the categorical shape priors. Specifically, the network architecture of CATRE consists of four components: a) point cloud focalization, b) shared encoder, c) rotation prediction, and d) translation and size prediction. In point clouds focalization, the observed point clouds $\mathcal{O}$ and the shape prior $\mathcal{P}$ are first aligned with the initial pose and size estimation $[\mathbf{R}_0, \mathbf{t}_0, \mathbf{s}_0]$:

$$\begin{aligned} \hat{\mathcal{O}} &= \{\hat{o}_i | \hat{o}_i = o_i - \mathbf{t}_0, o_i \in \mathcal{O}\}, \\ \hat{\mathcal{P}} &= \{\hat{p}_i | \hat{p}_i = \text{diag}(\mathbf{s}_0)\mathbf{R}_0 p_i, p_i \in \mathcal{P}\}, \end{aligned} \tag{5.2}$$

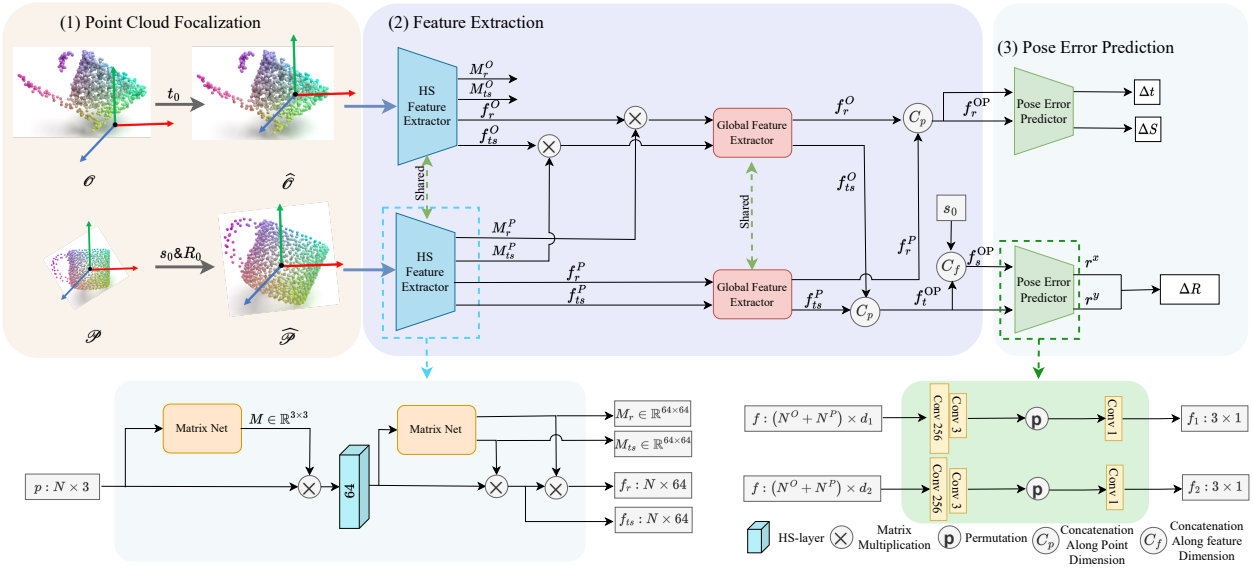



Figure 5.2: Overall structure of the proposed method.

where diag$(\cdot)$ converts a vector to a diagonal matrix. The focalized observed point cloud $\hat{\mathcal{O}}$ and the focalized shape prior points $\hat{\mathcal{P}}$ contain full information required to predict the estimation error $(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s})$. First, a PointNet-based shared encoder is used to extract features from the two focalized point clouds independently. Then, both the extracted features are used for $\Delta\mathbf{R}$ estimation, while the global feature of the focalized observed point cloud along with the $\mathbf{s}_0$ are used for $\Delta\mathbf{t}$ and $\Delta\mathbf{s}$. The initial estimates are updated using the predicted error $(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s})$. Finally, the updated estimates are used to predict the error again in which this process is iterative and the estimations are refined progressively and continuously.

### 5.3.3 Overall structure of GeoReF

The overall framework of our proposed object pose refinement approach is shown in Fig. 5.2. This framework comprises three principal components: 1) point cloud focalization, 2) feature extraction, and 3) pose error prediction. We follow CATRE and use the same point cloud focalization module. We apply focalization and extract features from both the observed point cloud and the shape prior by our Feature Extraction component. Then, we predict the estimation errors

$(\Delta\mathbf{R}, \Delta\mathbf{t}, \Delta\mathbf{s})$ using the extracted features in the pose error prediction component.

### 5.3.4 Graph convolution with learnable affine transformation (LAT)

Geometric structural information is effective in estimating an object's pose for category-level object pose estimations. However, as shown in the ablation study [AS-1], directly applying the 3D graph convolutions (*e.g.*, HS-Encoder (Zheng, C. Wang, et al., 2023) and 3DGCN-Encoder (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020)) to category-level object pose refinement tasks results in poor performance, as we show in the ablation study. This is due to the differences in task nature between the pose estimation and the pose refinement. In the pose estimation task, the network only needs to extract geometric structural information from a single point cloud. On the contrary, in the pose refinement framework, the network not only needs to extract the geometric structural information from the two input point clouds but also needs to establish the geometric correspondences between different object shapes. However, establishing geometric correspondence becomes challenging due to the issue of shape variation issue in category pose refinement.

To address the aforementioned problem, we propose to use learnable affine transformations (LATs). By employing LATs, the network can dynamically adjust the input point cloud and the point features which enables better establishment of geometric correspondences between the two different input shapes. Specifically, we apply three LATs (as shown in the bottom left of Fig. 5.2, where the Matrix Net outputs the learnable affine transformations): The first LAT $M \in \mathbb{R}^3$ is applied to the input point cloud in the Euclidean space. The second LAT $M_{\text{ts}}$ is applied to the extracted translation and size features $f_{\text{ts}}$. The third LAT $M_r$ is applied to the extracted rotation feature $f_r$. With this approach, our method can better utilize the valuable geometric features in pose refinement.

### 5.3.5 Cross-cloud transformation (CCT) for information mixing

In pose refinement, effectively blending information from the focalized observed object and the shape prior is crucial for enabling the network to align them accurately. However, in CATRE, the data from the observed point cloud and the shape prior are processed independently until the late rotation prediction stage, where they are merely concatenated, limiting the effectiveness of the alignment. To address this problem, we introduce a novel cross-cloud transformation mechanism that effectively mixes the geometric information from the shape of prior features into the features of the observed point cloud. In particular, we use the feature transformation matrices $M_r^P$, $M_{\text{ts}}^P$ of shape prior to transforming the features of the observed point cloud:

$$f_r^O = M_r^P f_r^O, \tag{5.3}$$

$$f_{\text{ts}}^O = M_{\text{ts}}^P f_{\text{ts}}^O. \tag{5.4}$$

### 5.3.6 Integrating shape prior in pose estimation

The information contained in the shape prior is crucial for the network to align the observed point cloud and the shape prior. For the rotation error prediction, the information contained in the shape prior is the essential information. For the translation and size prediction, this information can also be utilized by the network to adjust the learned geometric features accordingly. Therefore, unlike CATRE, which relied solely on features extracted from the observed point cloud to predict $(\Delta\mathbf{t}, \Delta\mathbf{s})$, our approach also incorporates the information from shape prior to predict them. In particular, we not only mix the information using the previous CCT mechanism, but also concatenate the features from the shape prior and observed point cloud to obtain mixed features in a similar way as the rotation estimation. We utilize $f_{\text{t}}^{\text{OP}}$ and $f_{\text{s}}^{\text{OP}}$ which contain both information from shape prior and observed point cloud like $f_r^{\text{OP}}$ to predict $(\Delta\mathbf{t}, \delta)$.

We use two pose error predictors of the same network architecture to predict the rotation

error and the translation and size error, respectively. Note that the weights of these two pose error predictors are not shared. The network structure of the pose error predictor is shown in Fig. 5.2. The pose predictor takes in two features, passes them through two same paths separately, and obtains two vectors in $\mathbb{R}^3$. In the translation and size branch, the pose error predictor takes in $f_{\mathrm{t}}^{\mathrm{OP}}$ and $f_{\mathrm{s}}^{\mathrm{OP}}$ and passes them through the two paths, and the output two vectors are regarded as $\Delta\mathbf{t}$ and $\Delta\mathbf{s}$, respectively. For the rotation error prediction, the mixed rotation features $f_r^{\mathrm{OP}}$ are copied and passed through the two paths in the pose error predictor. The two output vectors are regarded as $r_x$ and $r_y$, where $r_x$ and $r_y$ are the first and second axes of the rotation error matrix $\Delta\mathbf{R}$. The third column $r_z$ of $\Delta\mathbf{R}$ can be found by:

$$r_z = r_x \times r_y. \tag{5.5}$$

### 5.3.7 Detailed network architectures

The network structure of the HS Feature Extractor and the Pose Error Predictor is shown in Fig. 2 of the main paper. The structure of the Pose Error Predictor for $\Delta\mathbf{R}$ estimation and the $\Delta\mathbf{t}, \Delta\mathbf{s}$ estimation are identical, we follow the CATRE (X. Liu et al., 2022) and use 3 Convolution-1D layers with permutation before the final layer to generate the pose errors. For the Matrix Net, we follow PointNet (Qi, H. Su, et al., 2017) first use 3 Convolution-1D layers with $[64, 128, 1024]$ output dimensions and a kernel size of 1 to extract the dense point features, then the features going through a maximum pooling layer and 3 liner layers with [512, 256, $\boldsymbol{f}_{\mathrm{LAT}}$] to generate the matrix. For the first Matrix Net that generates the adaptive affine transformation (LAT) for the input point cloud, $\boldsymbol{f}_{\mathrm{LAT}}$ is 9. For the second Matrix Net, $\boldsymbol{f}_{\mathrm{LAT}}$ is 8192, as it outputs two LATs with the matrix size of $\mathbb{R}^{64\times64}$. In the final structure of the GeoReF, we use two HS-layers to replace the first two Convolution-1D layers in the second Matrix Net, which in our experiments, show slightly better results than without HS-layers (See Table 5.1 [B0, G0] for the performance comparison). The structure of the Global Feature Extractor is shown in Fig. 5.3, we use 1 layer of HS-layer and 2

Convolution-1D layers with the output size of $[128, 512, 1024]$ to extract dense point features, and then apply maximum pooling to get the global feature. Finally, the global feature is concatenated with the input features for the outputs.

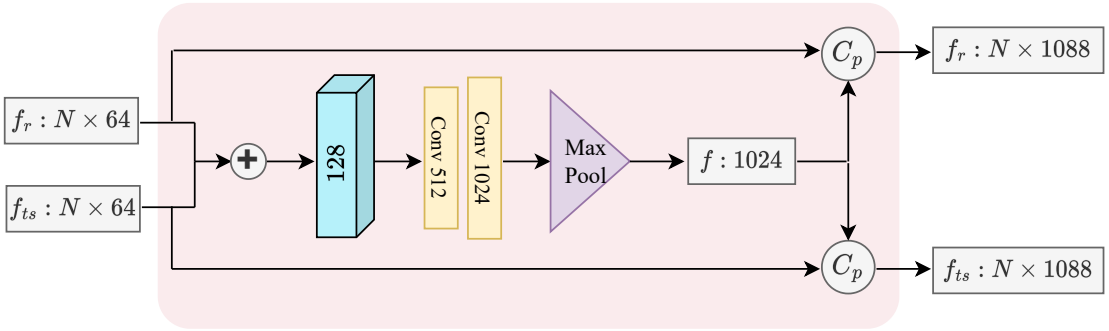



Figure 5.3: Structure of the global extractor

## 5.4 Experiments

**Implementation details.** We implement and experiment with our method using an RTX 4090 GPU with a batch size of 12 and 150 training epochs. We follow CATRE (X. Liu et al., 2022) and adopt its loss functions and the basic data augmentation strategies including random dropping points, adding Gaussian noise, random pose perturbations, etc. We set the number of points for both the observed points and shape prior to be 512. We train the network using Ranger optimizer (L. Liu et al., 2019; Yong et al., 2020; M. R. Zhang et al., 2019) with a base learning rate of $10^{-4}$ and anneal the learning rate from the $72\%$ of the total epoch based on cosine schedule.

**Baselines.** We use CATRE as the baseline for our ablation study as it is the state-of-the-art category-level object refinement method. As CATRE did not provide the results of $\text{IoU}_{50}$, we obtained them using their official pre-trained model and kept the rest of the reported metric scores consistent with the corresponding paper. For fair comparisons, we use the same initial estima-

tions as CATRE, which is the pose estimation results of SPD (Tian, Ang, and G. H. Lee, 2020). The result of replacing PointNet with the 3DGC-Encoder in the ablation study is provided by CATRE. We also apply our method to other state-of-the-art category-level object pose estimation approaches (Zheng, C. Wang, et al., 2023; Di et al., 2022; J. Lin, Wei, Ding, et al., 2022) to demonstrate the effectiveness of our proposed refinement method. For the pose refinement on HS-Pose (Zheng, C. Wang, et al., 2023) and RBP-Pose (R. Zhang, Di, Lou, et al., 2022), we compute the initial estimations using their official pre-trained models. The results of other methods are taken directly from their paper.

**Datasets.** As we focus on the problem of shape variation between input object point clouds, we choose two popular category-level object pose estimation benchmarks to verify our approach, *i.e.*, REAL275 (H. Wang et al., 2019) and CAMERA25 (H. Wang et al., 2019). They both contain 6 object categories with multiple levels of shape complexities, *i.e.*, bowl, can, bottle, laptop, mug, and camera. REAL275 contains 36 objects in 13 real-world scenes with 7k RGB-D images in total. Among them, 16 objects in 7 scenes are used for training, resulting in 4.3k images in training. CAMERA25 is a large synthetic RGB-D dataset. It provides 1085 objects and 275k RGB-D images for training, and 184 objects and 25k images for testing.

**Evaluation metrics.** Following (Zheng, C. Wang, et al., 2023; X. Liu et al., 2022), we evaluate our method using: 1) The mean average precision (mAP) of the *3D Intersection over Union (IoU)* at different thresholds ($50\%$ and $75\%$) to evaluate the pose and size estimation together [2]. 2) The pose metric at $n^\circ m$cm defines a pose as correct if the rotation error is below $n^\circ$ and the translation error is below $m$ cm. Here, we use $5^\circ$, $10^\circ$, 2cm, and 5cm as the thresholds.

[2]Note that there was a small mistake with the IoU computation from the original benchmark evaluation code (H. Wang et al., 2019), we follow (X. Liu et al., 2022) to recalculate the IoU metrics for the SOTA methods.

### 5.4.1 Ablation study

To verify the proposed architecture, we conducted comprehensive ablation studies on the REAL275 dataset using the initial pose estimations from SPD (Tian, Ang, and G. H. Lee, 2020). We present a quantitative comparison of our method with various key components disabled to motivate our design choices in Table 5.1.

Table 5.1: Ablation studies on the REAL275 dataset.

Higher score indicates better performance. In the 'Row' column, the code in bold means the strategies taken in the final structure. In the 'Method' column, the notation '$X$:$Y$' denotes module $Y$ from structure $X$, '$X$+$Y$' means add module $Y$ to $X$, and '$X \rightarrow Y$' indicates replacing $X$ with $Y$.

| Row | Method | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 2cm | 5° |
|---|---|---|---|---|---|---|---|---|---|
| A0 | CATRETian, Ang, and G. H. Lee, 2020 (baseline) | 77.0 | 43.6 | 45.8 | 54.4 | 61.4 | 73.1 | 75.1 | 58.0 |
| **B0** | **Ours**: E0 + Cross-Cloud Transformation | **79.2** 2.2↑ | **51.8** 8.2↑ | **54.4** 8.6↑ | **60.3** 5.9↑ | **71.9** 10.5↑ | **79.4** 6.3↑ | **81.9** 6.8↑ | **64.3** 6.3↑ |
| C0 | A0: PointNet → HS-Encoder | 71.0 | 30.1 | 41.9 | 45.9 | 60.6 | 70.3 | 71.9 | 48.7 |
| C1 | A0: PointNet → 3DGCN-Encoder | - | 28.4 | 36.0 | 43.4 | - | - | 68.0 | 47.7 |
| **D0** | A0 + prior in ST branch | 77.1 | 45.8 | 48.0 | 54.6 | 63.8 | 72.5 | 77.9 | 59.2 |
| **E0** | D0: PointNet → HS-layer+LAT | 79.4 | 51.0 | 52.4 | 58.6 | 69.4 | 77.7 | 80.4 | 62.4 |
| E1 | B0: No LAT on input points | 76.1 | 39.3 | 46.6 | 53.0 | 65.4 | 74.8 | 78.0 | 58.2 |
| E2 | B0: No LAT on features | 78.5 | 48.8 | 47.4 | 53.0 | 67.4 | 75.0 | 80.4 | 57.4 |
| E3 | B0: No LAT on the rotation feature | 79.8 | 50.6 | 50.4 | 56.2 | 68.6 | 76.3 | 80.2 | 60.8 |
| F0 | E0+ Global Concatenation Fusion | 77.7 | 48.4 | 47.8 | 54.5 | 67.1 | 75.2 | 80.1 | 59.4 |
| G0 | B0: No HS-layer in Matrix Net | 77.8 | 50.2 | 54.1 | 60.1 | 70.5 | 78.0 | 81.2 | 63.6 |

**[AS-1] Using geometric features directly.** To illustrate the limitations of existing geometric-based encoder under the problem of shape variations, we replace the encoder of CATRE with two robust geometric-based point cloud convolutional structures, namely 3DGCN-Encoder (Z.-H. Lin, S.-Y. Huang, and Y.-C. F. Wang, 2020) and HS-Encoder (Zheng, C. Wang, et al., 2023). 3DGCN is a widely adopted graph convolution in existing category-level object pose estimation algorithms, while HS-Encoder is a recent architecture that achieves state-of-the-art performance in

category-level object pose estimation. However, as shown in Table 5.1 [C0, C1], even though both HS-Encoder and 3DGCN-Encoder are powerful in finding an object's pose from individual inputs, they failed to manage the pose refinement scenarios when there exist shape variations between the target object and the shape prior. We observed both encoders result in a performance drop when compared to the original CATRE with $\text{IoU}_{75}$ of 37.3% (HS-Encoder) *vs*. 43.7% (CATRE), $5°5$cm of 43.4% (3DGCN-Encoder) *vs*. 53.3% (CATRE).

**[AS-2] Use prior features in translation and scale estimation.** To validate that the information of shape prior is also important in scale and translation error prediction, we add the features of the prior shape to the scale and translation branch by using the same network architecture as the rotation branch. As shown in Table 5.1 [D0], incorporating shape prior information in translation and size estimation enhanced the overall performance by $2.2\%$ improvement on $5°2$cm metric and $2.4\%$ on $10°2$cm metric.

**[AS-3] Use learnable affine transformation (LAT) for geometric features.** To demonstrate the effectiveness of LAT in addressing the shape variation issue, we conduct ablation studies on applying the proposed LAT to the input point cloud and the extracted geometric features in the feature space. The result is shown in Table 5.1 [E0]. Compared to using geometric features directly (Table 5.1 [C0]), LATs brings a significant boost on all the metrics, with $\text{IoU}_{75}$ improved by **20.9%**, $5°2$cm improved by **10.5%**, and $5°5$cm improved by **12.7%**. The resulted network also significantly outperforms the PointNet-based encoder (Table 5.1[D0]) on all the metrics with $\text{IoU}_{75}$ and $10°5$cm improved by **5.2%**, $5°2$cm improved by **6.4%**, and $10°2$cm improved by **5.6%**. These results verified the effectiveness of LAT on geometric features.

**[AS-4] The influence of each learnable affine transformation (LAT).** To further demonstrate the influences of each LAT, we conducted three experiments by gradually disabling LAT from the

framework: 1) without the LAT on the input point cloud, 2) without applying LATs on features, and 3) without independent LATs on the rotation features, where a single LAT is used for the rotation, translation, and scale features. The results are shown in Table 5.1 [E1-E3]. Compared to the directly using geometric features (Table 5.1 [C0]), we show that LAT can significantly enhance the network by around $10\%$ improvements on $\text{IoU}_{75}$ and $7\%$ on $5°5\text{cm}$ metric in Table 5.1 [E1-E3]. We verify that the combination of them consistently results in better performance.

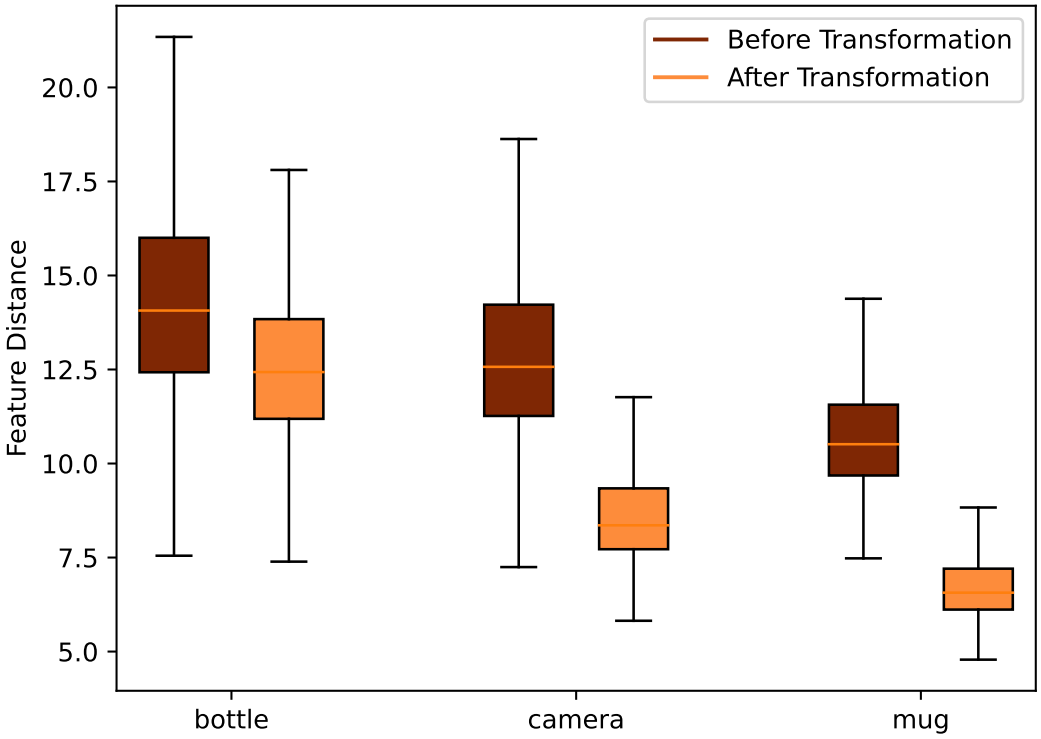



Figure 5.4: Feature distances between the shape prior and the input point cloud before and after applying the cross-cloud transformation.

**[AS-5] Cross-cloud transformation (CCT) based feature fusion** To demonstrate that it is important to have a good feature fusion strategy in the feature extraction phase, we conducted experiments on two different feature fusion strategies. One of them is the widely adopted feature fusion strategy, where the global feature of one point cloud is concatenated with the features of another point cloud and then goes through convolutional layers for feature fusion. Another one is the pro-

posed CCT-based feature fusion. As shown in Table 5.1 [F0], applying global concatenation-based fusion does not enhance the overall performance, and even results in worse performance in all the metrics with the $5°2$cm significantly decreased by $4.6\%$, and $5°5$cm decreased by $3.1\%$ (compared to Table 5.1 [D0]). On the contrary, as shown in Table 5.1 [B0], our simple CCT-based feature fusion shows its effectiveness by improving all the pose metrics by around $2.0\%$. To visualize the effect of CCT, we show a statistics plot of feature distances before and after CCT on objects with different shape complexities of the CAMERA25 test set. In this experiment, the initial pose of the shape prior is aligned with the ground truth pose to guarantee that the observed variations in feature distance are solely attributable to differences in shape. As shown in Fig 5.4, the feature distance between the shape prior and the input target shrinks significantly after applying CCT.

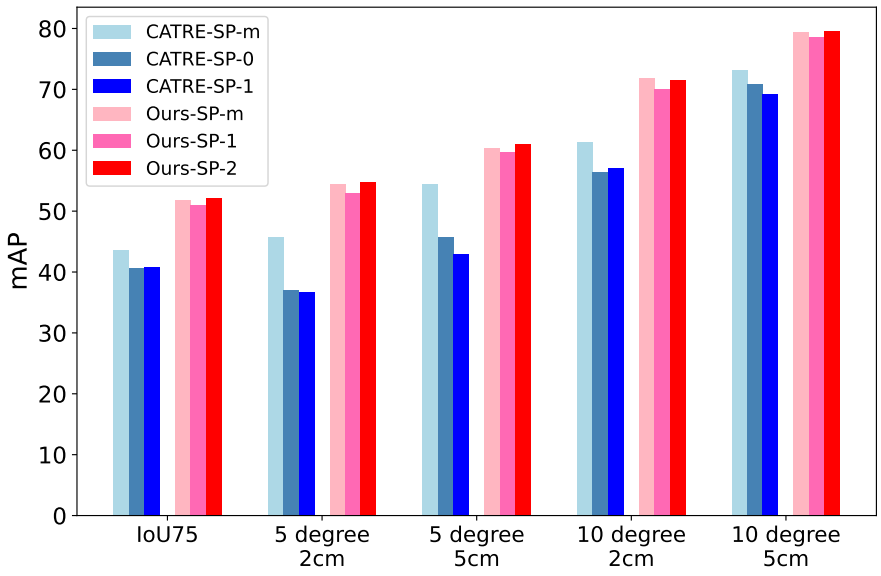
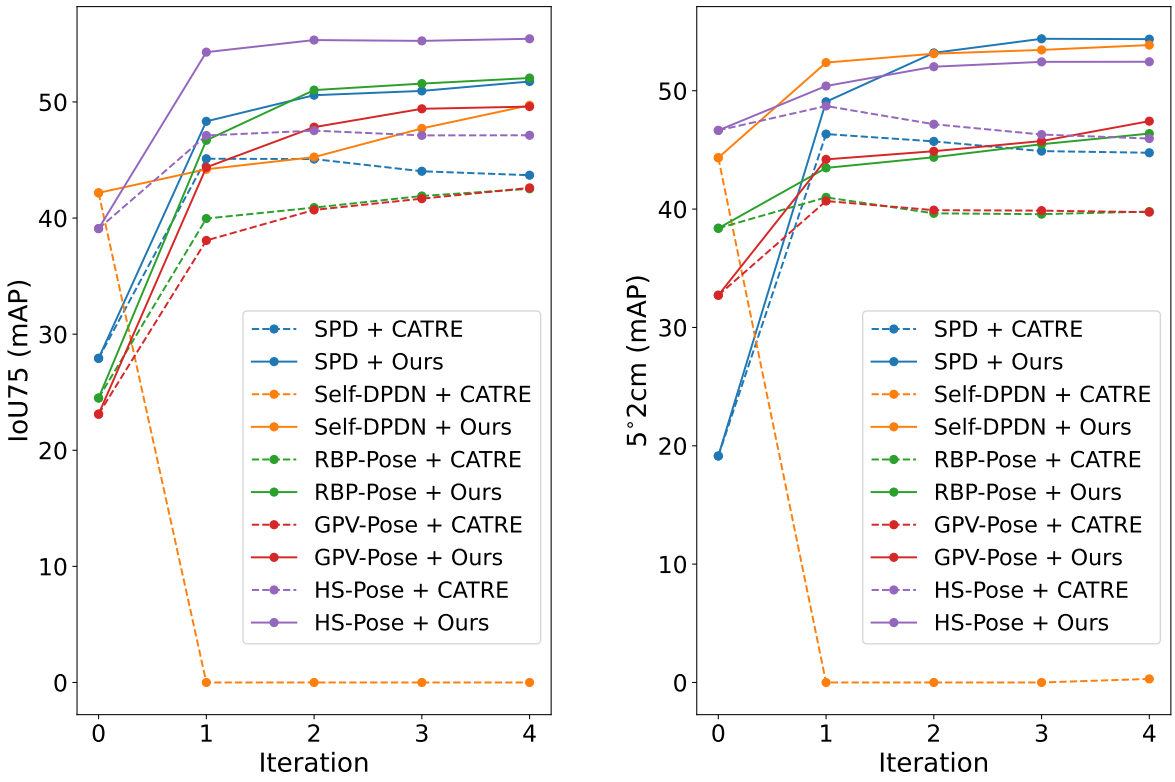



Figure 5.5: The performance of our method and the CATRE under different shape priors.

**[AS-6] Handle shape variations.** To demonstrate that the proposed method can handle the shape variations, we replaced the original shape prior with two randomly sampled models from CAMERA25 and trained on REAL275. Note that the original shape prior represents the mean shape of a category, and the new models are randomly sampled. Therefore, there are larger shape variations with certain target objects. We present the results in Fig. 5.5, where we denote the shape prior

as *SP-m*, the two sampled models as *SP-1* and *SP-2*, respectively. The figure shows that CATRE performs best when using the mean shape of the category, while its performance drops dramatically on the randomly sampled shape priors. In contrast, our method exhibits robustness to shape variations introduced by different shape priors and consistently delivers strong performance.




(a) (b)

Figure 5.6: Comparison between CATRE and our method on different initial estimations across different refining iterations: (a) $\text{IoU}_{75}$ performance comparison. (b) $5°2\text{cm}$ performance comparison.

**[AS-7] Refinement with different initial estimations.** To demonstrate our model's robustness to different initial estimations, we compare it with CATRE using initial estimations from five pose estimation methods with varying performance (Tian, Ang, and G. H. Lee, 2020; J. Lin, Wei, Ding, et al., 2022; R. Zhang, Di, Lou, et al., 2022; Di et al., 2022; Zheng, C. Wang, et al., 2023).

Figure 5.6 presents line graphs illustrating the performance changes of our method and CATRE during iteration. Our method (solid lines) consistently improves initial estimation performance, while CATRE (dashed lines) fails with Self-DPDN's initial estimations (J. Lin, Wei, Ding, et al., 2022). Furthermore, our method shows continuous improvement over iterations, contrasting with CATRE's performance drop on some of the initial estimations after the first iteration (dashed lines in Fig. 5.6). Quantitative results in Table 5.2 further highlight the robustness differences. Our method significantly enhances initial estimations across different metrics. For example, we improved Self-DPND and GPV-Pose (Di et al., 2022) on the strict $5°2$cm metric by 9.6% and 15.4%, respectively. Additionally, we boosted GPV-Pose and HS-Pose (Zheng, C. Wang, et al., 2023) $\text{IoU}_{75}$ by 26.5% and 15.2%, respectively. However, CATRE fails to refine Self-DPDN (J. Lin, Wei, Ding, et al., 2022) and struggles with high-accuracy initial estimations, such as those from HS-Pose, leading to performance drops on $10°5$cm and $10°2$cm. These results highlight our method's robustness and its ability to improve performance across various initial estimations.

Table 5.2: Performance comparison with CATRE on REAL275 using different initial estimations. Each comparison group contains 3 methods: the initial pose estimation method, refinement using CATRE, and refinement using our method, respectively.

| Method | $\text{IoU}_{75}$ | $5°2$cm | $5°5$cm | $10°2$cm | $10°5$cm |
|---|---|---|---|---|---|
| Self-DPDN (J. Lin, Wei, Ding, et al., 2022) | 42.2 | 44.3 | 50.9 | 65.1 | 78.6 |
| Self-DPDN + CATRE | 0.0 42.2↓ | 0.3 44.0↓ | 5.1 45.8↓ | 0.4 64.7↓ | 6.9 71.7↓ |
| Self-DPDN + Ours | 49.7 7.5↑ | 53.9 9.6↑ | 60.1 9.2↑ | 75.0 9.9↑ | 82.8 4.2↑ |
| GPV-Pose (Di et al., 2022) | 23.1* | 32.0 | 42.9 | 55.0 | 73.3 |
| GPV-Pose + CATRE | 42.6 19.5↑ | 39.7 7.7↑ | 54.1 11.2↑ | 57.1 2.1↑ | 78.0 4.7↑ |
| GPV-Pose + Ours | 49.6 26.5↑ | 47.4 15.4↑ | 57.8 14.9↑ | 68.1 13.1↑ | 81.2 7.9↑ |
| HS-Pose (Zheng, C. Wang, et al., 2023) | 39.1* | 46.5 | 55.2 | 68.6 | 82.7 |
| HS-Pose + CATRE | 47.1 8.0↑ | 48.7 2.2↑ | 59.1 3.9↑ | 67.8 0.8↓ | 81.2 1.5↓ |
| HS-Pose + Ours | 54.3 15.2↑ | 51.7 5.2↑ | 59.6 4.4↑ | 74.3 5.7↑ | 83.8 1.1↑ |

**[AS-8] Effect of number of iterations** We find that our proposed method's performance saturates after four iterations. Therefore, we set the iteration number to four for our experiments. The line graph in Figure 5.6 illustrates this saturation effect, showing that our proposed method consistently outperforms the baseline method and saturates after four iterations.

### 5.4.2 Generalizability test on the CAMERA25 dataset

In real-world applications, category-level algorithms often encounter a substantially larger number of objects than what is represented in their training sets. Therefore, category-level approaches are required to generalize across diverse testing scenarios. To simulate this challenge, we selecte the CAMERA25 dataset, which provides more than 25K RGB-D testing images.

We train our model using mini-sets comprising 2%, 4%, and 6% of the CAMERA25 training data, resulting in training sets of approximately 5K, 10K, and 15K images, respectively, from a total of 275K images. To ensure a balanced distribution of different categories in the sampled mini datasets, we control the number of images per object: 5 images per object for the 2% training set, 10 images for the 4% training set, and 15 images for the 6% training set.

In Table 5.3, the baseline method's performance decreased dramatically when using 2% of the training images, with $\text{IoU}_{75}$ dropped by **12.9%** and $5^\circ$2cm dropped by **9.0%** (see [B0]). On the contrary, our method exhibits a much higher performance when using small datasets for training. Specifically, our method can already outperform the fully-trained CATRE with only **2%** of images (see [B1]). The performance using 4% images is further increased: our method outperforms CATRE with $\text{IoU}_{75}$ and $5^\circ$2cm improved by 3.1% and 3.6%, respectively. Furthermore, we observed that our performance became stable when using 4% of the training set, eliminating the need for further testing on larger data sizes. In contrast, CATRE required additional training data for better performance. This highlights the efficiency and effectiveness of our approach compared

to the baseline

Table 5.3: The generalizability test on the CAMERA25 dataset.

Higher score means better performance. Overall best results are in bold, and the second-best results are underlined.

| Row | Method | Train Data Size | $IoU_{75}$ | $5°2cm$ | $5°5cm$ | $10°2cm$ | $10°5cm$ |
|---|---|---|---|---|---|---|---|
| A0 | CATRE | 275k | 76.1 | 75.4 | 80.3 | 83.3 | 89.3 |
| B0 | CATRE | 5k | 63.2 | 66.4 | 72.3 | 79.4 | 87.4 |
| B1 | Ours | 5k | 77.5 | 75.4 | 81.1 | 83.4 | <u>90.0</u> |
| C0 | CATRE | 10k | 66.5 | 69.7 | 75.5 | 81.8 | 89.1 |
| C1 | Ours | 10k | **79.2** | <u>77.9</u> | <u>84.0</u> | **83.8** | **90.5** |
| D0 | CATRE | 15k | 69.7 | 73.2 | 78.8 | 82.6 | 89.4 |
| D1 | Ours | 15k | <u>78.1</u> | **78.0** | **84.1** | <u>83.6</u> | **90.5** |

### 5.4.3 Comparison with state-of-the-arts

**REAL275.** We conduct pose refinement on SPD (Tian, Ang, and G. H. Lee, 2020) using the proposed approach and compare the resulting performance with the state-of-the-art category-level object pose estimation and refinement methods. As shown in Table 5.4, our method significantly improves the performance of SPD on all the metrics, with $5°5cm$ enhanced by **39.1%**, $5°5cm$ improved by **35.3%**, $IoU_{75}$ enhanced by **24.8%**, $10°2cm$ improved by **24.4%**, and $10°5cm$ improve by **25.4%**. In comparison to our baseline, CATRE, our proposed method demonstrates a substantial improvement across various performance metrics. Specifically, we observe a remarkable enhancement of **10.5%** in $10°2cm$, **8.6%** in $5°2cm$, **7.2%** in $IoU_{75}$, and **6.3%** in $10°5cm$. In addition, compared with SOTA pose estimation methods, the pose estimation results of applying our proposed refinement method on SPD significantly outperformed these methods's results by a large margin. Specifically, the estimation results of applying our method on SPD rank top on 4

Table 5.4: Performance Comparison with other methods on REAL275.

| Method | $IoU_{75}$ | $5°2cm$ | $5°5cm$ | $10°2cm$ | $10°5cm$ |
|---|---|---|---|---|---|
| NOCS (H. Wang et al., 2019) | 9.4 | 7.2 | 10.0 | 13.8 | 25.2 |
| DualPoseNet (J. Lin, Wei, Zhihao Li, et al., 2021) | 30.8 | 29.3 | 35.9 | 50.0 | 66.8 |
| CR-Net (J. Wang, K. Chen, and Dou, 2021) | 33.2 | 27.8 | 34.3 | 47.2 | 60.8 |
| SGPA (K. Chen and Dou, 2021) | 37.1 | 35.9 | 39.6 | 61.3 | 70.7 |
| RBP-Pose (R. Zhang, Di, Lou, et al., 2022) | 24.5 | 38.2 | 48.1 | 63.1 | 79.2 |
| GPV-Pose (Di et al., 2022) | 23.1 | 32.0 | 42.9 | 55.0 | 73.3 |
| HS-Pose (Zheng, C. Wang, et al., 2023) | 39.1 | <u>46.5</u> | <u>55.2</u> | <u>68.6</u> | **82.7** |
| SPD* (Tian, Ang, and G. H. Lee, 2020) | 27.0 | 19.1 | 21.2 | 43.5 | 54.0 |
| SPD*+CATRE (X. Liu et al., 2022) | <u>43.6</u> | 45.8 | 54.4 | 61.4 | 73.1 |
| SPD*+**Ours** | **51.8** | **54.4** | **60.3** | **71.9** | <u>79.4</u> |

out of 5 metrics and rank second on the rest metric, and achieved a **7.9%** increase on the $5°2cm$ metric and **5.1%** enhancement on $5°5cm$. It is worth noting that the purpose of this section is not to compare refinement and estimation methods, as they are designed to address different problems. Instead, our objective here is to demonstrate how our proposed refinement approach can improve the performance of existing pose estimation methods. Therefore, even though our refinement on HS-Pose (Zheng, C. Wang, et al., 2023) produced better performance, as mentioned in the ablation study, we choose to refine weaker initial estimations to show the capability of our approach.

**CAMERA25.** Table 5.5 compares the accuracy of our method with the state-of-the-arts. As discussed in Sec. 5.4.2, our performance stabilizes when using 4% of the full train set. Therefore, we present the results obtained with this training size. As shown in Table 5.5, we greatly enhanced the performance of SPD, resulting in a performance that outperformed state-of-the-art pose estimation methods. Specifically, we improved the performance of SPD (Tian, Ang, and G. H. Lee, 2020) on

$IoU_{75}$ by 32.7%, $5°5$cm by 25.2%, and $5°2$cm by 23.8%. We also outperform the baseline CATRE on $IoU_{75}$ by $3.1\%$, $5°5$cm by $3.7\%$, and $5°2$cm by $2.5\%$. Additionally, we show our results trained using 5k images (2%) of the train set, which already outperforms the state-of-the-art methods.

Table 5.5: Comparison with other methods on the CAMERA25 dataset

Higher score means better performance. Overall best results are in bold. SPD* is the implementation results from CATRE, which is similar to the original SPD results.

| Method | $IoU_{75}$ | $5°2$cm | $5°5$cm | $10°2$cm | $10°5$cm |
|---|---|---|---|---|---|
| NOCS (H. Wang et al., 2019) | 37.0 | 32.3 | 40.9 | 48.2 | 64.6 |
| DualPoseNet (J. Lin, Wei, Zhihao Li, et al., 2021) | 71.7 | 64.7 | 70.7 | 77.2 | 84.7 |
| CR-Net (J. Wang, K. Chen, and Dou, 2021) | 75.0 | 72.0 | 76.4 | 81.0 | 87.7 |
| SGPA (K. Chen and Dou, 2021) | 69.1 | 70.7 | 74.5 | 82.7 | 88.4 |
| SAR-Net (H. Lin, Z. Liu, C. Cheang, et al., 2022) | 62.6 | 66.7 | 70.9 | 75.3 | 80.3 |
| SSP-Pose (R. Zhang, Di, Manhardt, et al., 2022) | - | 64.7 | 75.5 | - | 87.4 |
| RBP-Pose (R. Zhang, Di, Lou, et al., 2022) | - | 73.5 | 79.6 | 82.1 | 89.5 |
| GPV-Pose (Di et al., 2022) | - | 72.1 | 79.1 | - | 89.0 |
| HS-Pose (Zheng, C. Wang, et al., 2023) | - | 73.3 | 80.5 | 80.4 | 89.4 |
| SPD* (Tian, Ang, and G. H. Lee, 2020) | 46.9 | 54.1 | 58.8 | 73.9 | 82.1 |
| SPD*+CATRE (X. Liu et al., 2022) | 76.1 | 75.4 | 80.3 | 83.3 | 89.3 |
| SPD*+**Ours** (2%) | <u>77.5</u> | <u>75.4</u> | <u>81.1</u> | <u>83.4</u> | <u>90.0</u> |
| SPD*+**Ours** | **79.2** | **77.9** | **84.0** | **83.8** | **90.5** |

### 5.4.4 Per-category performance on REAL275 and CAMERA25

We present the per-category object pose refinement results on the REAL275 dataset and the CAMERA25 dataset in Table 5.6 and Table 5.7, respectively. We use SPD (Tian, Ang, and G. H. Lee,

2020) as the initial estimation method and report the performance after 4 refinement iterations. We show that our method largely improved the initial performance.

Table 5.6: Per-category results of our method on REAL275 dataset.

| Method | category | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 10°10cm | 5° | 2cm |
|---|---|---|---|---|---|---|---|---|---|---|
| SPD | bottle | 49.9 | 13.1 | 21.6 | 23.2 | 69.4 | 76.0 | 87.1 | 35.9 | 80.7 |
| SPD+**Ours** | bottle | 49.8 | 36.2 | 64.8 | 68.0 | 82.5 | 88.6 | 100.0 | 82.5 | 89.1 |
| SPD | bowl | 100.0 | 77.1 | 50.5 | 54.0 | 75.8 | 80.3 | 80.3 | 54.0 | 94.7 |
| SPD+**Ours** | bowl | 100.0 | 91.9 | 91.2 | 95.6 | 95.4 | 100.0 | 100.0 | 95.7 | 95.4 |
| SPD | camera | 43.4 | 3.4 | 0.0 | 0.0 | 0.2 | 0.2 | 0.2 | 0.0 | 34.8 |
| SPD+**Ours** | camera | 78.4 | 12.4 | 2.1 | 2.1 | 17.9 | 18.8 | 18.9 | 2.2 | 58.3 |
| SPD | can | 70.0 | 29.8 | 37.9 | 42.7 | 80.4 | 91.6 | 91.6 | 45.5 | 87.1 |
| SPD+**Ours** | can | 70.3 | 36.7 | 75.6 | 78.6 | 96.0 | 99.9 | 99.9 | 80.7 | 96.0 |
| SPD | laptop | 82.0 | 35.5 | 4.6 | 7.0 | 24.5 | 65.3 | 65.9 | 7.1 | 29.1 |
| SPD+**Ours** | laptop | 80.8 | 73.9 | 67.6 | 91.8 | 68.9 | 94.4 | 95.6 | 92.5 | 69.3 |
| SPD | mug | 66.5 | 8.7 | 0.3 | 0.3 | 10.3 | 10.4 | 10.4 | 0.3 | 85.2 |
| SPD+**Ours** | mug | 96.2 | 59.5 | 24.8 | 25.9 | 70.7 | 74.8 | 74.8 | 25.9 | 89.9 |

### 5.4.5 Qualitative results

We provide qualitative comparisons of the pose estimation results by SPD, CATRE on SPD, and our method on SPD in Fig. 5.7 and 5.8. It's worth noting that for symmetric objects like the bowl, bottle, and can, correct estimated rotations are considered when the symmetry axis is aligned. As shown in Fig. 5.7, our method on SPD achieves the best size and pose estimations. In particular, by considering the camera in the first column of Fig. 5.7, all of the comparison methods struggle to estimate its orientation. We also provide qualitative examples of the iteration process in Fig. 5.8,

Table 5.7: Per-category results of our method on CAMERA25 dataset.

| Method | category | $IoU_{50}$ | $IoU_{75}$ | 5°2cm | 5°5cm | 10°2cm | 10°5cm | 10°10cm | 5° | 2cm |
|---|---|---|---|---|---|---|---|---|---|---|
| SPD | bottle | 88.9 | 64.5 | 63.8 | 82.8 | 69.2 | 92.4 | 97.3 | 86.8 | 69.8 |
| SPD+**Ours** | bottle | 89.4 | 73.8 | 73.8 | 94.2 | 74.2 | 95.1 | 99.4 | 98.4 | 74.2 |
| SPD | bowl | 95.9 | 80.6 | 83.4 | 83.7 | 95.8 | 96.3 | 96.3 | 83.7 | 99.2 |
| SPD+**Ours** | bowl | 96.0 | 94.7 | 97.9 | 98.2 | 99.5 | 99.8 | 99.8 | 98.2 | 99.6 |
| SPD | camera | 61.9 | 4.7 | 27.3 | 29.3 | 72.9 | 78.6 | 78.6 | 29.5 | 89.8 |
| SPD+**Ours** | camera | 81.6 | 67.7 | 83.1 | 87.2 | 90.8 | 95.2 | 95.2 | 87.2 | 93.9 |
| SPD | can | 90.2 | 87.2 | 98.1 | 98.2 | 99.4 | 99.6 | 99.6 | 98.2 | 99.6 |
| SPD+**Ours** | can | 90.3 | 89.8 | 99.9 | 100.0 | 99.9 | 100.0 | 100.0 | 100.0 | 99.9 |
| SPD | laptop | 93.3 | 17.7 | 35.0 | 41.9 | 61.0 | 80.5 | 84.5 | 43.7 | 65.5 |
| SPD+**Ours** | laptop | 95.3 | 81.3 | 74.0 | 85.5 | 77.4 | 91.8 | 95.8 | 89.1 | 77.9 |
| SPD | mug | 82.7 | 24.1 | 15.5 | 15.5 | 44.1 | 44.1 | 44.1 | 15.9 | 99.6 |
| SPD+**Ours** | mug | 89.8 | 67.7 | 39.0 | 39.0 | 61.0 | 61.0 | 61.0 | 39.4 | 99.9 |

with white lines representing the ground truth. Our method demonstrates faster convergence and more accurate final result than CATRE. Furthermore, we present additional qualitative results of our method test on different REAL275 test scenes in Fig. 5.9 and Fig. 5.10, with the performance differences highlighted by red arrows.

### 5.4.6 Inference speed

On a machine with an Intel 13900k CPU and a Nvidia RTX 4090 GPU, the speed of our proposed method is 67.5 FPS for 1 iteration, and 22.3 FPS when using 4 iterations.

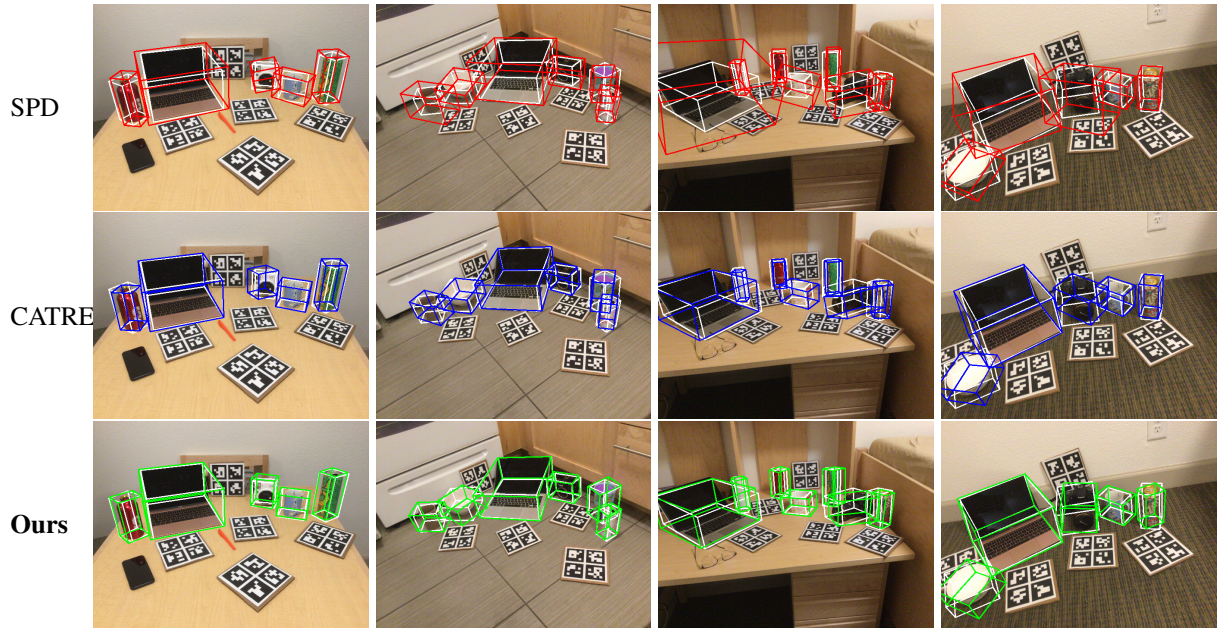



Figure 5.7: Qualitative comparison between the proposed method and CATRE using SPD as the initial estimation.

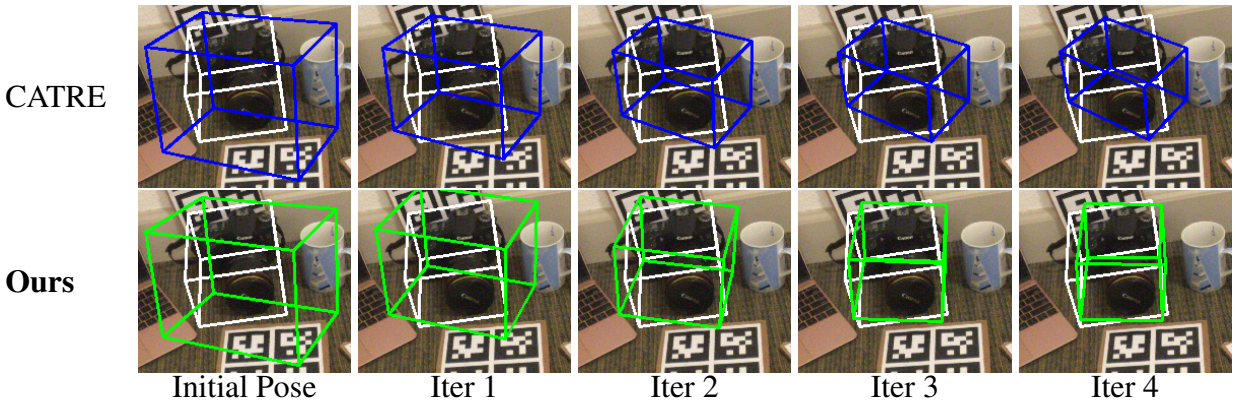



Figure 5.8: Comparison between the proposed method (row #2) and the baseline method (row #1) during a complete refinement iteration, both utilized the SPD as the initial estimation.

## 5.5 Conclusion

In this work, we proposed a novel category-level object pose refinement method which targeted at addressing the challenge of shape variation. We shown that the geometric structural information can be aligned by our adaptive affine transformations. We also demonstrated that the cross-

cloud transformation mechanism can efficiently merge information from distinct point clouds. We further incorporated shape prior information and observed improvements in translation and size predictions. We verified that each of our technical components contributed meaningfully through extensive ablations. We believe our method sets a strong baseline for future study and opens up new possibilities to handling more complex shapes, *i.e*. articulated objects.

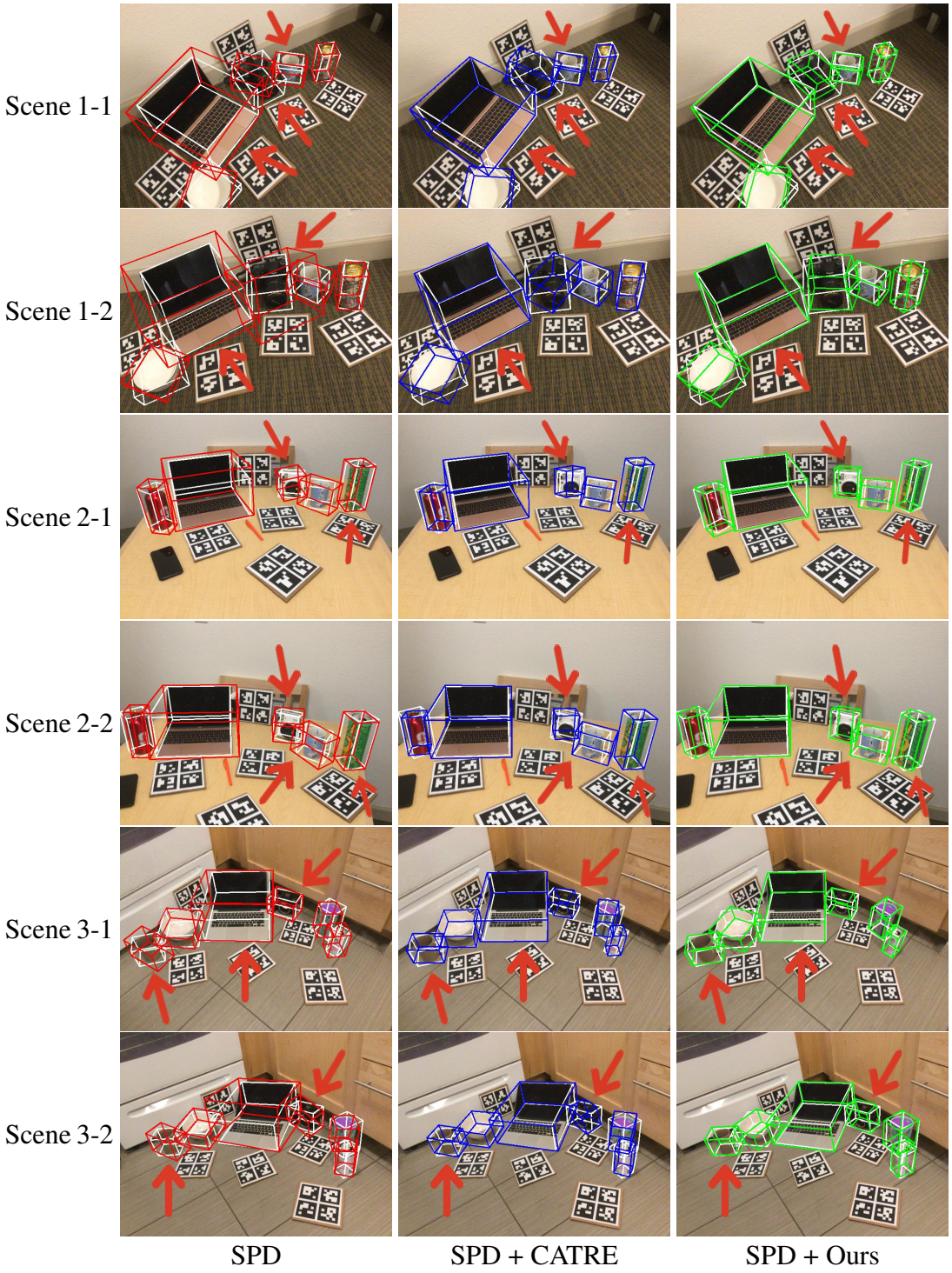



Figure 5.9: More qualitative comparison between the proposed method (column #3) and the baseline method (column #2) use the SPD (column #1) as the initial estimation (scene 1 to scene 3).

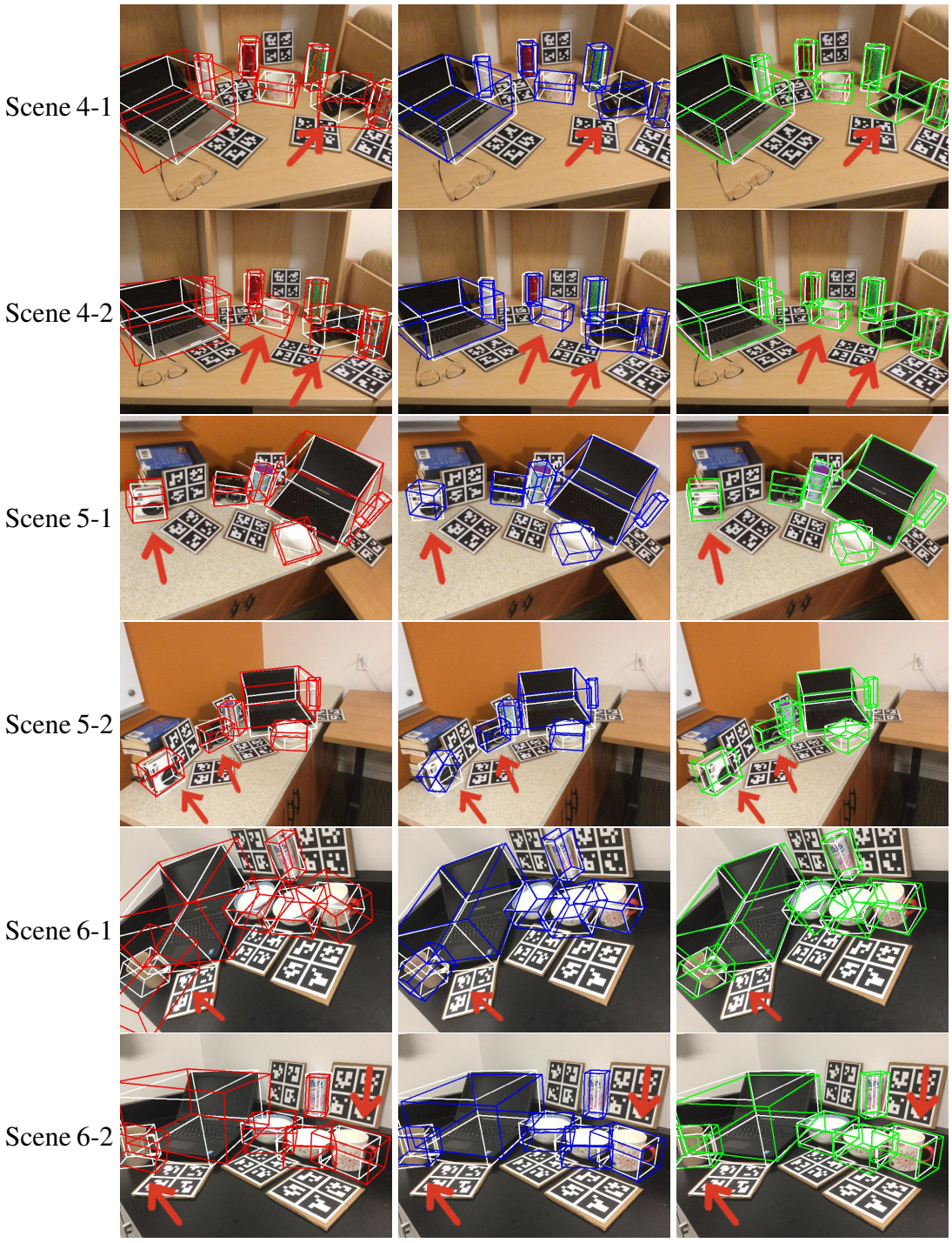



Figure 5.10: More qualitative comparison between the proposed method (column #3) and the baseline method (column #2) use the SPD (column #1) as the initial estimation (scene 4 to scene 6).

# Chapter Six

# Discussion

In this thesis, I have explored the field of visual 6D object pose estimation and tracking, with a focus on developing algorithms to fit realistic applications better. I have presented three approaches to address outstanding problems in this domain, including robustness in challenging scenarios (TP-AE, Chapter 3), generalization to category-level objects (HS-Pose, Chapter 4), and achieving high precision in category-level applications (GeoReF, Chapter 5). While these approaches demonstrate significant advancements, there remain several limitations and avenues for future work that warrant detailed discussion.

## 6.1 Limitations

**Instance-level: TP-AE.** Despite the advancements made by the proposed Temporally Primed Auto-Encoder (TP-AE) in instance-level object pose estimation and tracking, several limitations persist. One notable challenge is the algorithm's performance in long-term tracking scenarios, particularly when the tracked object undergoes complete occlusion and then reappears. While TP-AE has shown improvements in tracking objects across various occlusion levels, accurately re-establishing the object's identity before and after full occlusion remains a daunting task, especially

in environments with multiple identical objects. The algorithm's vulnerability to similar-looking distractors is another persistent limitation. During re-initialization, if a similar-looking object is present, the algorithm may mistakenly recognize it as the target object and track it instead. Furthermore, while TP-AE has demonstrated significant improvements over other methods when trained solely on synthetic data, its performance still falls short compared to training with a combination of synthetic and real data. Bridging the synthetic-real domain gap is crucial to further enhancing TP-AE's performance and applicability in real-world scenarios.

**Category-level: HS-Pose and GeoReF.** The development of category-level 6D object pose estimation is still in its early stages, and thus, the proposed methods, HS-Pose and GeoReF, also face certain limitations. Firstly, accurately obtaining category-level object poses under occlusion remains challenging. While our approaches have significantly improved accuracy and robustness for objects with various shapes, they may struggle when the target object is partially occluded. This highlights the need for enhancements in this area to improve robustness in real-world applications. Secondly, the computational memory consumption of HS-Pose and GeoReF poses a limitation, as the graph convolution involved requires substantial GPU memory. Optimizing these algorithms to reduce their computational demands while maintaining performance could facilitate their deployment on everyday devices. Lastly, a limitation lies in the reliance on object detection methods, as the performance of HS-Pose and GeoReF is closely tied to the object detector's performance. Improvements in object detection could directly benefit the accuracy and efficiency of these methods, indicating a potential area for future research and development.

## 6.2 Future Works

### 6.2.1 Instance-level

To address the limitations of TP-AE in long-term tracking scenarios, specific future directions can be explored. One approach is to enhance the model's memory capabilities by incorporating an additional memory module that stores object information and background information, such as object motion information and where the object disappeared. This memory can help the model maintain the identity of the object before and after occlusion, making it easier to relocate the object in a nearby position and reducing the risk of mistakenly tracking a different object. Additionally, developing more robust re-initialization strategies that use contextual information, such as object appearance and motion cues, can help differentiate between the target object and similar-looking distractors. By combining these approaches, TP-AE may improve its ability to track objects accurately over extended periods, making it more suitable for real-world applications where objects may undergo occlusion and reappearances.

Another important area for future work of TP-AE is to bridge the synthetic-real domain gap. Several strategies can be explored to achieve this. One approach is to adopt domain adaptation methods to align the distributions of synthetic and real images, making synthetic data more representative of real-world scenarios. This can involve using adversarial training or feature-level alignment methods to make the synthetic data indistinguishable from real images. Another strategy is to leverage unsupervised or self-supervised learning methods, which can learn from unannotated real-world data to improve the model's performance on real data. Additionally, active learning approaches can be employed to selectively annotate a small subset of real data that maximally improves the model's performance, reducing the overall annotation burden. By bridging the gap between synthetic and real data, TP-AE can be further enhanced for real-world applications.

### 6.2.2 Category-level

To improve the occlusion robustness for the proposed HS-Pose and GeoReF, one possible future work is to leverage multi-view fusion or integrate temporal information. In Chapter 3, we show how temporal information can be leveraged in instance-level object pose tracking, resulting in enhanced occlusion robustness even when the object is severely occluded. While how to leverage this information in category-level object pose estimation is still an open question, I believe it is an area worth exploring.

To reduce the computational memory consumption of HS-Pose and GeoReF, another future work is to explore a more efficient neural network structure for graph convolution to reduce the memory footprint during computation. Additionally, strategies like model quantization or pruning could be applied to reduce the overall computational demands of the algorithms.

Integrating semantic understanding into pose estimation algorithms can be another future work to improve the robustness of HS-Pose and GeoReF. By incorporating information about object categories and relationships between objects, they can better understand the context of the scene, leading to more accurate and reliable pose estimations, and reducing the reliance on additional object detectors.

Another future work for GeoReF is exploring category-level object pose refinement without a shape-prior. Although shape priors are widely used in the literature, and our method is not strictly tied to specific priors, refining object poses without a shape prior could benefit the algorithm to be widely adapted to different real-world applications. Refining without shape prior remains an open challenge, which I am interested in investigating in future work, showing the potential for more progress in the field.

For HS-Pose, in future work, I also plan to apply the proposed HS-layer to other problems where unstructured data needs to be processed, and the combination between the local and global

information becomes critical.

# Chapter Seven

# Conclusion

In this thesis, I have explored the challenging field of visual 6D object pose estimation and tracking, with a primary goal of developing algorithms that are not only accurate but also practical for real-world applications. Through this journey, three novel approaches have been introduced, each addressing crucial aspects of this field: robustness in challenging scenarios, generalization to category-level objects, and achieving high precision in category-level applications.

One of the key contributions of this work is TP-AE (Chapter 3), a robust object pose tracking framework designed to handle symmetric and textureless objects under occlusion. By integrating object motion information into the auto-encoder-based reconstruction process and leveraging latent feature matching, TP-AE significantly improves pose estimation accuracy. Experimental validation on benchmark datasets demonstrates TP-AE's superiority over state-of-the-art methods while maintaining real-time performance.

Another significant contribution is HS-Pose (Chapter 4), a framework that focuses on enhancing the generalizability to previously unseen objects within a known category. The hybrid scope latent feature extraction layer (HS-layer), introduced in this framework, adeptly captures local and global geometric structural information, encodes size and translation information, and remains robust against noise. HS-Pose demonstrates its proficiency in handling complex shapes,

accurately capturing object size and translation, and exhibiting robustness across diverse outlier levels. Comparative experiments with existing methods highlight the substantial performance gain of HS-Pose, establishing it as the new benchmark in real-time category-level object pose estimation.

Furthermore, the thesis presents GeoReF (Chapter 5), a method aimed at advancing category-level object pose estimation for high-precision applications. GeoReF targets the challenge of shape variation in category-level object pose refinement, achieving accurate pose refinement through the alignment of geometric structural information and efficient merging of information from disparate point clouds. Extensive ablations confirm the meaningful contributions of each proposed component. Experiments showed that GeoReF sets a strong baseline for future study and opens up new possibilities for handling more complex shapes, *i.e*. articulated objects.

In conclusion, the methodologies introduced through TP-AE, HS-Pose, and GeoReF represent significant advancements in visual 6D object pose estimation and tracking. These approaches not only set a new standard for performance, robustness, and adaptability but also hold promise for various real-world applications. Future research can build upon these foundations to further enhance the capabilities and applicability of object pose estimation and tracking systems.

# References


Aldoma, A., F. Tombari, J. Prankl, A. Richtsfeld, L. Di Stefano, and M. Vincze (2013). "Multi-modal cue integration through Hypotheses Verification for RGB-D object recognition and 6DOF pose estimation". In: *2013 IEEE International Conference on Robotics and Automation*, pp. 2104–2111. DOI: 10.1109/ICRA.2013.6630859.

Altschul, Stephen F, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman (1997). "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs". In: *Nucleic acids research* 25.17, pp. 3389–3402.

Andriluka, Mykhaylo, Umar Iqbal, Eldar Insafutdinov, Leonid Pishchulin, Anton Milan, Juergen Gall, and Bernt Schiele (June 2018). "PoseTrack: A Benchmark for Human Pose Estimation and Tracking". In: pp. 5167–5176. DOI: 10.1109/CVPR.2018.00542.

Arun, K. S., T. S. Huang, and S. D. Blostein (1987). "Least-Squares Fitting of Two 3-D Point Sets". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-9.5, pp. 698–700. DOI: 10.1109/TPAMI.1987.4767965.

Azad, Pedram, Tamim Asfour, and Rüdiger Dillmann (2009). "Accurate shape-based 6-DoF pose estimation of single-colored objects". In: *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 2690–2695. DOI: 10.1109/IROS.2009.5354606.

Badrinarayanan, V., A. Kendall, and R. Cipolla (Dec. 2017). "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation". In: *IEEE Transactions on Pat-*

*tern Analysis and Machine Intelligence* 39.12, pp. 2481–2495. ISSN: 1939-3539. DOI: 10.1109/TPAMI.2016.2644615.

Baker, Stephen, Jonathan Hardy, Kenneth E Sanderson, Michael Quail, Ian Goodhead, Robert A Kingsley, Julian Parkhill, Bruce Stocker, and Gordon Dougan (2007). "A novel linear plasmid mediates flagellar variation in Salmonella Typhi". In: *PLoS Pathog* 3.5, e59.

Barsom, E., Maurits Graafland, and Marlies Schijven (Oct. 2016). "Systematic review on the effectiveness of augmented reality applications in medical training". In: *Surgical Endoscopy* 30. DOI: 10.1007/s00464-016-4800-6.

Bay, Herbert, Tinne Tuytelaars, and Luc Van Gool (2006). "SURF: Speeded Up Robust Features". In: *Computer Vision – ECCV 2006*. Ed. by Aleš Leonardis, Horst Bischof, and Axel Pinz. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 404–417. ISBN: 978-3-540-33833-8.

Besl, P.J. and Neil D. McKay (1992). "A method for registration of 3-D shapes". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14.2, pp. 239–256. DOI: 10.1109/34.121791.

Bimbo, Joao, Shan Luo, Kaspar Althoefer, and Hongbin Liu (Jan. 2016). "In-Hand Object Pose Estimation Using Covariance-Based Tactile To Geometry Matching". In: *IEEE Robotics and Automation Letters* 1, pp. 1–1. DOI: 10.1109/LRA.2016.2517244.

Brachmann, Eric, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother (2014). "Learning 6D Object Pose Estimation Using 3D Object Coordinates". In: *Computer Vision – ECCV 2014*. Ed. by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars. Cham: Springer International Publishing, pp. 536–551. ISBN: 978-3-319-10605-2.

Burchfiel, Benjamin and George Konidaris (2019). *Probabilistic Category-Level Pose Estimation via Segmentation and Predicted-Shape Priors*. arXiv: 1905.12079 [cs.CV].

Cai, Dingding, Janne Heikkilä, and Esa Rahtu (June 2022). "OVE6D: Object Viewpoint Encoding for Depth-Based 6D Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6803–6813.

Cai, Hongping, Tomas Werner, and Jiri Matas (July 2013). “Fast Detection of Multiple Textureless 3-D Objects”. In: *Computer Vision Systems*, pp. 103–112. DOI: 10.1007/978-3-642-39402-7_11.

Cai, Yingjie, Kwan-Yee Lin, Chao Zhang, Qiang Wang, Xiaogang Wang, and Hongsheng Li (June 2022). “Learning a Structured Latent Space for Unsupervised Point Cloud Completion”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5543–5553.

Calli, Berk, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron Dollar (July 2015). “The YCB Object and Model Set: Towards Common Benchmarks for Manipulation Research”. In: *Proceedings of IEEE International Conference on Advanced Robotics (ICAR)*.

Cao, Zhe, Yaser Sheikh, and Natasha Banerjee (May 2016). “Real-time scalable 6DOF pose estimation for textureless objects”. In: pp. 2441–2448. DOI: 10.1109/ICRA.2016.7487396.

Castro, Pedro and Tae-Kyun Kim (Jan. 2023). “CRT-6D: Fast 6D Object Pose Estimation With Cascaded Refinement Transformers”. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5746–5755.

Chang, Angel X., Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu (2015). *ShapeNet: An Information-Rich 3D Model Repository*. cite arxiv:1512.03012. URL: http://arxiv.org/abs/1512.03012.

Chen, Dengsheng, Jun Li, Zheng Wang, and Kai Xu (June 2020). “Learning Canonical Shape Space for Category-Level 6D Object Pose and Size Estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Hansheng, Pichao Wang, Fan Wang, Wei Tian, Lu Xiong, and Hao Li (2022). *EPro-PnP: Generalized End-to-End Probabilistic Perspective-n-Points for Monocular Object Pose Estimation*. DOI: 10.48550/ARXIV.2203.13254. URL: https://arxiv.org/abs/2203.13254.

Chen, Hanzhi, Fabian Manhardt, Nassir Navab, and Benjamin Busam (June 2023). "TexPose: Neural Texture Learning for Self-Supervised 6D Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4841–4852.

Chen, Hua, Linfang Zheng, and Wei Zhang (2020). "Optimal Control Inspired Q-Learning for Switched Linear Systems". In: *2020 American Control Conference (ACC)*, pp. 4003–4010. DOI: 10.23919/ACC45564.2020.9147818.

Chen, Kai and Qi Dou (2021). "SGPA: Structure-guided prior adaptation for category-level 6d object pose estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2773–2782.

Chen, Kai, Stephen James, Congying Sui, Yun-Hui Liu, Pieter Abbeel, and Qi Dou (2023). "StereoPose: Category-Level 6D Transparent Object Pose Estimation from Stereo Images via Back-View NOCS". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2855–2861. DOI: 10.1109/ICRA48891.2023.10160780.

Chen, Wei, Xi Jia, Hyung Jin Chang, Jinming Duan, and Aleš Leonardis (June 2020). "G2L-Net: Global to Local Network for Real-Time 6D Pose Estimation With Embedding Vector Features". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Wei, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Aleš Leonardis (June 2021). "FS-Net: Fast Shape-Based Network for Category-Level 6D Object Pose Estimation With Decoupled Rotation Mechanism". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1581–1590.

Chen, Xiaozhi, Huimin Ma, Ji Wan, Bo Li, and Tian Xia (July 2017). "Multi-view 3D Object Detection Network for Autonomous Driving". In: pp. 6526–6534. DOI: 10.1109/CVPR.2017.691.

Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (Oct. 2014). "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the*

*2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, pp. 1724–1734. DOI: 10.3115/v1/D14-1179. URL: https://www.aclweb.org/anthology/D14-1179.

Choi, Changhyun and Henrik Christensen (Oct. 2012). "3D Textureless Object Detection and Tracking: An Edge-based approach". In: DOI: 10.1109/IROS.2012.6386065.

— (Nov. 2013). "RGB-D object tracking: A particle filter approach on GPU". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1084–1091. DOI: 10.1109/IROS.2013.6696485.

Collet, Alvaro, Manuel Martinez, and Siddhartha Srinivasa (Sept. 2011). "The MOPED framework: Object recognition and pose estimation for manipulation". In: *International Journal of Robotics Research (IJRR)* 30, pp. 1284–1306. DOI: 10.1177/0278364911401765.

Correll, Nikolaus, Kostas Bekris, Dmitry Berenson, Oliver Brock, Albert Causo, Kris Hauser, Kei Okada, Alberto Rodríguez, Joseph Romano, and Peter Wurman (Jan. 2016). "Lessons from the Amazon Picking Challenge". In.

Danielczuk, Michael, Matthew Matl, Saurabh Gupta, Andrew Li, Andrew Lee, Jeffrey Mahler, and Ken Goldberg (2019). "Segmenting Unknown 3D Objects from Real Depth Images using Mask R-CNN Trained on Synthetic Data". In: *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*.

Deng, Xinke, Junyi Geng, Timothy Bretl, Yu Xiang, and Dieter Fox (2022). "iCaps: Iterative Category-Level Object Pose and Shape Estimation". In: *IEEE Robotics and Automation Letters* 7.2, pp. 1784–1791. DOI: 10.1109/LRA.2022.3142441.

Deng, Xinke, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox (June 2019). "PoseRBPF: A Rao-Blackwellized Particle Filter for6D Object Pose Estimation". In: *Proceedings of Robotics: Science and Systems*. FreiburgimBreisgau, Germany. DOI: 10.15607/RSS.2019.XV.049.

Deng, Xinke, Yu Xiang, Arsalan Mousavian, Clemens Eppner, Timothy Bretl, and Dieter Fox (2020). "Self-supervised 6D Object Pose Estimation for Robot Manipulation". In: *2020*

*IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3665–3671. DOI: 10.1109/ICRA40945.2020.9196714.

Di, Yan, Ruida Zhang, Zhiqiang Lou, Fabian Manhardt, Xiangyang Ji, Nassir Navab, and Federico Tombari (June 2022). "GPV-Pose: Category-Level Object Pose Estimation via Geometry-Guided Point-Wise Voting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6781–6791.

Doosti, Bardia, Shujon Naha, Majid Mirbagheri, and David Crandall (2020). "HOPE-Net: A Graph-based Model for Hand-Object Pose Estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby (2021). *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. arXiv: 2010.11929 [cs.CV].

Doumanoglou, Andreas, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim (June 2016). "Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd". In: pp. 3583–3592. DOI: 10.1109/CVPR.2016.390.

Fäulhammer, T., A. Aldoma, M. Zillich, and M. Vincze (2015). "Temporal integration of feature correspondences for enhanced recognition in cluttered and dynamic environments". In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3003–3009. DOI: 10.1109/ICRA.2015.7139611.

Fischler, M. and R. Bolles (1981). "Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography". In: *Communications of the ACM* 24.6, pp. 381–395. URL: /brokenurl#%20http://publication.wilsonwong.me/load.php?id=233282275.

Gao, Ge, Mikko Lauri, Xiaolin Hu, Jianwei Zhang, and Simone Frintrop (2021). "CloudAAE: Learning 6D Object Pose Regression with On-line Data Synthesis on Point Clouds". In: *CoRR* abs/2103.01977. arXiv: 2103.01977. URL: https://arxiv.org/abs/2103.01977.

Gao, Wei and Russ Tedrake (2021). “kPAM 2.0: Feedback Control for Category-Level Robotic Manipulation”. In: *CoRR* abs/2102.06279. arXiv: 2102.06279. URL: https://arxiv.org/abs/2102.06279.

Garon, Mathieu and Jean-François Lalonde (Mar. 2017). “Deep 6-DOF Tracking”. In: *IEEE Transactions on Visualization and Computer Graphics* PP. DOI: 10.1109/TVCG.2017.2734599.

Geiger, Andreas (June 2012). “Are we ready for autonomous driving? The KITTI vision benchmark suite”. In: pp. 3354–3361.

Grenzdörffer, Till, Martin Günther, and Joachim Hertzberg (May 2020). “YCB-M: A Multi-Camera RGB-D Dataset for Object Recognition and 6DoF Pose Estimation”. In: *International Conference on Robotics and Automation (ICRA)*.

Güler, R. A., N. Neverova, and I. Kokkinos (June 2018). “DensePose: Dense Human Pose Estimation in the Wild”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7297–7306. DOI: 10.1109/CVPR.2018.00762.

Gulrajani, Ishaan, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville (2017). “Improved Training of Wasserstein GANs”. In: *CoRR* abs/1704.00028. arXiv: 1704.00028. URL: http://arxiv.org/abs/1704.00028.

Gupta, Kartik, Lars Petersson, and Richard Hartley (2019). *CullNet: Calibrated and Pose Aware Confidence Scores for Object Pose Estimation*. arXiv: 1909.13476 [cs.CV].

Gupta, Saurabh, Pablo Arbelaez, Ross Girshick, and Jitendra Malik (June 2015). “Aligning 3D models to RGB-D images of cluttered scenes”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4731–4740. DOI: 10.1109/CVPR.2015.7299105.

Hai, Yang, Rui Song, Jiaojiao Li, Mathieu Salzmann, and Yinlin Hu (June 2023). “Rigidity-Aware Detection for 6D Object Pose Estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8927–8936.

Hampali, Shreyas, Mahdi Rad, Markus Oberweger, and Vincent Lepetit (2020). “HOnnotate: A method for 3D Annotation of Hand and Object Poses”. In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haralick, R.M., H. Joo, C. Lee, X. Zhuang, V.G. Vaidya, and M.B. Kim (1989). "Pose estimation from corresponding point data". In: *IEEE Transactions on Systems, Man, and Cybernetics* 19.6, pp. 1426–1446. DOI: 10.1109/21.44063.

Harris, Chris and Carl Stennett (1990). "RAPID-a video rate object tracker." In: *BMVC*, pp. 1–6.

Haugaard, Rasmus Laurvig and Anders Glent Buch (2021). "SurfEmb: Dense and Continuous Correspondence Distributions for Object Pose Estimation with Learnt Surface Embeddings". In: *CoRR* abs/2111.13489. arXiv: 2111.13489. URL: https://arxiv.org/abs/2111.13489.

He, Kaiming, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick (2017). "Mask R-CNN". In: *CoRR* abs/1703.06870. arXiv: 1703.06870. URL: http://arxiv.org/abs/1703.06870.

He, Yisheng, Haoqiang Fan, Haibin Huang, Qifeng Chen, and Jian Sun (2022). *Towards Self-Supervised Category-Level Object Pose and Size Estimation*. arXiv: 2203.02884 [cs.CV].

He, Yisheng, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun (June 2021). "FFB6D: A Full Flow Bidirectional Fusion Network for 6D Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3003–3013.

He, Yisheng, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun (June 2020). "PVN3D: A Deep Point-Wise 3D Keypoints Voting Network for 6DoF Pose Estimation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hinterstoisser, S., S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit (Nov. 2011). "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes". In: *International Conference on Computer Vision (ICCV)*, pp. 858–865. DOI: 10.1109/ICCV.2011.6126326.

Hinterstoisser, Stefan, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, Vincent Lepetit, and bullet Lepetit (May 2012). "Gradient Response Maps for Real-Time Detection of Texture-Less Objects". In: *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI)* 1. DOI: 10.1109/TPAMI.2011.206.

Hinterstoisser, Stefan, Vincent Lepetit, Slobodan Ilic, Pascal Fua, and Nassir Navab (Sept. 2010). "Dominant Orientation Templates for Real-Time Detection of Texture-Less Objects". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2257–2264. DOI: 10.1109/CVPR.2010.5539908.

Hinterstoisser, Stefan, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab (2013). "Model Based Training, Detection and Pose Estimation of Texture-Less 3D Objects in Heavily Cluttered Scenes". In: *Asian Conference on Computer Vision (ACCV)*, pp. 548–562. ISBN: 978-3-642-37331-2.

Hinterstoisser, Stefan, Vincent Lepetit, Naresh Rajkumar, and Kurt Konolige (2016). "Going Further with Point Pair Features". In: *Computer Vision – ECCV 2016*. Springer International Publishing, pp. 834–848. DOI: 10.1007/978-3-319-46487-9_51. URL: https://doi.org/10.1007%2F978-3-319-46487-9_51.

Hoang, Dinh-Cuong, Todor Stoyanov, and Achim J Lilienthal (2019). "Object-rpe: Dense 3d reconstruction and pose estimation with convolutional neural networks for warehouse robots". In: *2019 European Conference on Mobile Robots (ECMR)*. IEEE, pp. 1–6.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780.

Hodan, Tomas, Daniel Barath, and Jiri Matas (June 2020). "EPOS: Estimating 6D Pose of Objects With Symmetries". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hodan, Tomas, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiří Matas, and Carsten Rother (2018). "BOP: Benchmark for 6D Object Pose Estimation: 15th European Conference, Munich, Germany, September 8-14, Proceedings, Part X". In: pp. 19–35. ISBN: 978-3-030-01248-9. DOI: 10.1007/978-3-030-01249-6_2.

Hodan, Tomas, Xenophon Zabulis, Manolis Lourakis, Stepan Obdrzalek, and Jiri Matas (Sept. 2015). "Detection and fine 3D pose estimation of texture-less objects in RGB-D images". In: pp. 4421–4428. DOI: 10.1109/IROS.2015.7354005.

Hodaň, Tomáš, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis (2017). "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Hodaň, Tomáš, Jiří Matas, and Štěpán Obdržálek (2016). "On Evaluation of 6D Object Pose Estimation". In: *ECCV 2016 Workshops*, pp. 606–619. ISBN: 978-3-319-49409-8.

Hoque, Sabera, Shuxiang Xu, Ananda Maiti, Yuchen Wei, and Md Yasir Arafat (2023). "Deep learning for 6D pose estimation of objects—A case study for autonomous driving". In: *Expert Systems with Applications* 223, p. 119838.

Horanyi, Nora, Linfang Zheng, Eunji Chong, Aleš Leonardis, and Hyung Jin Chang (June 2023). "Where Are They Looking in the 3D Space?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 2678–2687.

Hu, Yinlin, Pascal Fua, Wei Wang, and Mathieu Salzmann (2020). "Single-stage 6d object pose estimation". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2930–2939.

Hu, Yinlin, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann (June 2019). "Segmentation-Driven 6D Object Pose Estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huttenlocher, D. P., G. A. Klanderman, and W. J. Rucklidge (Sept. 1993). "Comparing images using the Hausdorff distance". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15.9, pp. 850–863. ISSN: 1939-3539. DOI: 10.1109/34.232073.

Irshad, Muhammad Zubair, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon (2022). "ShAPO: Implicit Representations for Multi Object Shape Appearance and Pose Optimization". In: URL: https://arxiv.org/abs/2207.13691.

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

Jiang, Xiaoke, Donghai Li, Hao Chen, Ye Zheng, Rui Zhao, and Liwei Wu (2022). "Uni6D: A Unified CNN Framework without Projection Breakdown for 6D Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11174–11184.

Jocher, Glenn, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, TaoXie, Jiacong Fang, imyhxy, Kalen Michael, Lorna, Abhiram V, Diego Montes, Jebastin Nadar, Laughing, tkianai, yxNONG, Piotr Skalski, Zhiqiang Wang, Adam Hogan, Cristi Fati, Lorenzo Mammana, AlexWang1900, Deep Patel, Ding Yiwei, Felix You, Jan Hajek, Laurentiu Diaconu, and Mai Thanh Minh (Feb. 2022). *ultralytics/yolov5: v6.1 - TensorRT, TensorFlow Edge TPU and OpenVINO Export and Inference*. Version v6.1. DOI: 10.5281/zenodo.6222936. URL: https://doi.org/10.5281/zenodo.6222936.

Kaelbling, Leslie Pack and Tomás Lozano-Pérez (2012). "Unifying perception, estimation and action for mobile manipulation via belief space planning". In: *2012 IEEE International Conference on Robotics and Automation*, pp. 2952–2959. DOI: 10.1109/ICRA.2012.6225237.

Kappler, Daniel, Franziska Meier, Jan Issac, Jim Mainprice, Cristina Garcia Cifuentes, Manuel Wüthrich, Vincent Berenz, Stefan Schaal, Nathan Ratliff, and Jeannette Bohg (2018). "Real-Time Perception Meets Reactive Motion Generation". In: *IEEE Robotics and Automation Letters* 3.3, pp. 1864–1871. DOI: 10.1109/LRA.2018.2795645.

Kart, Ugur, Alan Lukezic, Matej Kristan, Joni-Kristian Kamarainen, and Jiri Matas (June 2019). "Object Tracking by Reconstruction With View-Specific Discriminative Correlation Filters". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Katz, Dov, Moslem Kazemi, J. Bagnell, and Anthony Stentz (May 2013). "Interactive segmentation, tracking, and kinematic modeling of unknown 3D articulated objects". In: pp. 5003–5010. ISBN: 978-1-4673-5641-1. DOI: 10.1109/ICRA.2013.6631292.

Kehl, Wadim, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab (Oct. 2017). "SSD-6D: Making RGB-Based 3D Detection and 6D Pose Estimation Great Again". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 1530–1538. DOI: 10.1109/ICCV.2017.169.

Kehl, Wadim, Fausto Milletari, Federico Tombari, Slobodan Ilic, and Nassir Navab (2016). "Deep learning of local rgb-d patches for 3d object detection and 6d pose estimation". In: *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, pp. 205–220.

Kosaka, A. and G. Nakazawa (1995). "Vision-based motion tracking of frigid objects using prediction of uncertainties". In: *Proceedings of 1995 IEEE International Conference on Robotics and Automation*. Vol. 3, 2637–2644 vol.3. DOI: 10.1109/ROBOT.1995.525655.

Kothari, Nikunj, Misha Gupta, Leena Vachhani, and Hemendra Arya (Jan. 2017). "Pose estimation for an autonomous vehicle using monocular vision". In: pp. 424–431. DOI: 10.1109/INDIANCC.2017.7846512.

Krull, Alexander, Eric Brachmann, Frank Michel, Michael Ying Yang, Stefan Gumhold, and Carsten Rother (Dec. 2015). "Learning Analysis-by-Synthesis for 6D Pose Estimation in RGB-D Images". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Kundu, Abhijit, Yin Li, and James Rehg (June 2018). "3D-RCNN: Instance-Level 3D Object Reconstruction via Render-and-Compare". In: pp. 3559–3568. DOI: 10.1109/CVPR.2018.00375.

Labbé, Yann, Justin Carpentier, Mathieu Aubry, and Josef Sivic (Aug. 2020a). "CosyPose: Consistent multi-view multi-object 6D pose estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Labbé, Yann, Justin Carpentier, Mathieu Aubry, and Josef Sivic (2020b). *Ylabbe/Cosypose: Code FOR "CosyPose: Consistent MULTI-VIEW multi-object 6D Pose Estimation", ECCV 2020. Pre-trained model ids: refiner-bop-tless-synt+real–881314 and detector-bop-tless-synt+real–452847*. URL: https://github.com/ylabbe/cosypose.

Lai, K., L. Bo, X. Ren, and D. Fox (2011). "A large-scale hierarchical multi-view RGB-D object dataset". In: *2011 IEEE International Conference on Robotics and Automation*, pp. 1817–1824. DOI: 10.1109/ICRA.2011.5980382.

Lee, Taeyeop, Byeong-Uk Lee, Myungchul Kim, and In So Kweon (2021). "Category-Level Metric Scale Object Shape and Pose Estimation". In: *IEEE Robotics and Automation Letters* 6.4, pp. 8575–8582. DOI: 10.1109/LRA.2021.3110538.

Lee, Taeyeop, Byeong-Uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In So Kweon, and Kuk-Jin Yoon (2021). "UDA-COPE: Unsupervised Domain Adaptation for Category-level Object Pose Estimation". In: *CoRR* abs/2111.12580. arXiv: 2111.12580. URL: https://arxiv.org/abs/2111.12580.

Lee, Taeyeop, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon (June 2023). "TTA-COPE: Test-Time Adaptation for Category-Level Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21285–21295.

Lepetit, Vincent, Francesc Moreno-Noguer, and Pascal Fua (Feb. 2009). "EPnP: An accurate O(n) solution to the PnP problem". In: *International Journal of Computer Vision (IJCV)* 81. DOI: 10.1007/s11263-008-0152-6.

Li, Chi, Jin Bai, and Gregory D. Hager (Sept. 2018). "A Unified Framework for Multi-View Multi-Class Object Pose Estimation". In: *The European Conference on Computer Vision (ECCV)*.

Li, Fu, Shishir Reddy Vutukur, Hao Yu, Ivan Shugurov, Benjamin Busam, Shaowu Yang, and Slobodan Ilic (Oct. 2023). "NeRF-Pose: A First-Reconstruct-Then-Regress Approach for Weakly-Supervised 6D Object Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pp. 2123–2133.

Li, Xiaolong, He Wang, Li Yi, Leonidas Guibas, A Lynn Abbott, and Shuran Song (2020). "Category-Level Articulated Object Pose Estimation". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Li, Xiaolong, He Wang, Li Yi, Leonidas J. Guibas, A. Lynn Abbott, and Shuran Song (June 2020). "Category-Level Articulated Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Yi, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox (Sept. 2018). "DeepIM: Deep Iterative Matching for 6D Pose Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

— (Mar. 2020). "DeepIM: Deep Iterative Matching for 6D Pose Estimation". In: *International Journal of Computer Vision* 128. DOI: 10.1007/s11263-019-01250-9.

Li, Zhigang and Xiangyang Ji (2020). "Pose-guided Auto-Encoder and Feature-Based Refinement for 6-DoF Object Pose Regression". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 8397–8403. DOI: 10.1109/ICRA40945.2020.9196953.

Li, Zhigang, Gu Wang, and Xiangyang Ji (Oct. 2019). "CDPN: Coordinates-Based Disentangled Pose Network for Real-Time RGB-Based 6-DoF Object Pose Estimation". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Liang, Guoyuan, Fan Chen, Yu Liang, Yachun Feng, Can Wang, and Xinyu Wu (2021). "A manufacturing-oriented intelligent vision system based on deep neural network for object recognition and 6d pose estimation". In: *Frontiers in neurorobotics* 14, p. 616775.

Lin, Haitao, Zichang Liu, Chi-Hou Cheang, Lingwei Zhang, Yanwei Fu, and X. Xue (2021). "DONet: Learning Category-Level 6D Object Pose and Size Estimation from Depth Observation". In: *ArXiv* abs/2106.14193.

Lin, Haitao, Zichang Liu, Chilam Cheang, Yanwei Fu, Guodong Guo, and Xiangyang Xue (June 2022). "SAR-Net: Shape Alignment and Recovery Network for Category-Level 6D Object Pose and Size Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6707–6717.

Lin, Jiehong, Hongyang Li, Ke Chen, Jiangbo Lu, and Kui Jia (2021). "Sparse Steerable Convolutions: An Efficient Learning of SE (3)-Equivariant Features for Estimation and Tracking of Object Poses in 3D Space". In: *Advances in Neural Information Processing Systems* 34.

Lin, Jiehong, Zewei Wei, Changxing Ding, and Kui Jia (2022). *Category-Level 6D Object Pose and Size Estimation using Self-Supervised Deep Prior Deformation Networks*. DOI: 10.48550/ARXIV.2207.05444. URL: https://arxiv.org/abs/2207.05444.

Lin, Jiehong, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li (2021). "DualPoseNet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3560–3569.

Lin, Tsung-Yi, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar (Oct. 2017). "Focal Loss for Dense Object Detection". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Lin, Zhi-Hao, Sheng-Yu Huang, and Yu-Chiang Frank Wang (June 2020). "Convolution in the Cloud: Learning Deformable Kernels in 3D Graph Convolution Networks for Point Cloud Analysis". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1797–1806. DOI: 10.1109/CVPR42600.2020.00187.

Liu, Jian, Wei Sun, Chongpei Liu, Xing Zhang, and Qiang Fu (2023). "Robotic continuous grasping system by shape transformer-guided multi-object category-level 6D pose estimation". In: *IEEE Transactions on Industrial Informatics*.

Liu, Liyuan, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han (2019). *On the Variance of the Adaptive Learning Rate and Beyond*. DOI: 10.48550/ARXIV.1908.03265. URL: https://arxiv.org/abs/1908.03265.

Liu, M., O. Tuzel, A. Veeraraghavan, and R. Chellappa (June 2010). "Fast directional chamfer matching". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1696–1703. DOI: 10.1109/CVPR.2010.5539837.

Liu, Wei, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg (2016). "SSD: Single Shot MultiBox Detector". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 21–37. ISBN: 978-3-319-46448-0.

Liu, Xingyu, Gu Wang, Yi Li, and Xiangyang Ji (2022). "CATRE: Iterative point clouds alignment for category-level object pose refinement". In: *European Conference on Computer Vision (ECCV)*. Springer, pp. 499–516.

Lowe, D. G. (Sept. 1999). "Object recognition from local scale-invariant features". In: *Proceedings of the Seventh IEEE International Conference on Computer Vision*. Vol. 2, 1150–1157 vol.2. DOI: 10.1109/ICCV.1999.790410.

Lowe, David G (1987). "Three-dimensional object recognition from single two-dimensional images". In: *Artificial intelligence* 31.3, pp. 355–395.

— (1992). "Robust model-based motion tracking through the integration of search and estimation". In: *International Journal of Computer Vision* 8.2, pp. 113–122.

— (May 1991). "Fitting Parameterized Three-Dimensional Models to Images". In: 13.5, pp. 441–450. ISSN: 0162-8828. DOI: 10.1109/34.134043. URL: https://doi.org/10.1109/34.134043.

Lukežič, Alan, Ugur Kart, Jani Kapyla, Ahmed Durmush, Joni-Kristian Kamarainen, Jiri Matas, and Matej Kristan (Oct. 2019). "CDTB: A Color and Depth Visual Object Tracking Dataset and Benchmark". In: pp. 10012–10021. DOI: 10.1109/ICCV.2019.01011.

Manhardt, Fabian, Diego Arroyo, Christian Rupprecht, Benjamin Busam, Nassir Navab, and Federico Tombari (Dec. 2018). "Explaining the Ambiguity of Object Detection and 6D Pose from Visual Data". In: *IEEE International Conference on Computer Vision (ICCV)*.

Manhardt, Fabian, Wadim Kehl, Nassir Navab, and Federico Tombari (2018). *Deep Model-Based 6D Pose Refinement in RGB*. DOI: 10.48550/ARXIV.1810.03065. URL: https://arxiv.org/abs/1810.03065.

Manhardt, Fabian, Manuel Nickel, Sven Meier, Luca Minciullo, and Nassir Navab (2020). "CPS: Class-level 6D Pose and Shape Estimation From Monocular Images". In: *CoRR* abs/2003.05848. arXiv: 2003.05848. URL: https://arxiv.org/abs/2003.05848.

Manuelli, Lucas, Wei Gao, Peter R. Florence, and Russ Tedrake (2019). "kPAM: KeyPoint Affordances for Category-Level Robotic Manipulation". In: *CoRR* abs/1903.06684. arXiv: 1903.06684. URL: http://arxiv.org/abs/1903.06684.

Marchand, Eric, Hideaki Uchiyama, and Fabien Spindler (Jan. 2016). "Pose Estimation for Augmented Reality: A Hands-On Survey". In: *IEEE Transactions on Visualization and Computer Graphics* 22. DOI: 10.1109/TVCG.2015.2513408.

Marder-Eppstein, Eitan (July 2016). "Project Tango". In: pp. 25–25. DOI: 10.1145/2933540.2933550.

Marougkas, Isidoros, Petros Koutras, Nikolaos Kardaris, Georgios Retsinas, Georgia Chalvatzaki, and Petros Maragos (2020). "How to track your dragon: A Multi-Attentional Framework for real-time RGB-D 6-DOF Object Pose Tracking". In: *CoRR* abs/2004.10335. arXiv: 2004.10335. URL: https://arxiv.org/abs/2004.10335.

Martin-Martin, Roberto and Oliver Brock (Sept. 2014). "Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors". In: DOI: 10.1109/IROS.2014.6942902.

Martin-Martin, Roberto, Sebastian Höfer, and Oliver Brock (May 2016). "An Integrated Approach to Visual Perception of Articulated Objects". In: DOI: 10.1109/ICRA.2016.7487714.

Merrill, Nathaniel, Yuliang Guo, Xingxing Zuo, Xinyu Huang, Stefan Leutenegger, Xi Peng, Liu Ren, and Guoquan Huang (June 2022). "Symmetry and Uncertainty-Aware Object SLAM for 6DoF Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14901–14910.

Michel, Frank, Alexander Kirillov, Eric Brachmann, Alexander Krull, Stefan Gumhold, Bogdan Savchynskyy, and Carsten Rother (July 2017). "Global Hypothesis Generation for 6D Ob-

ject Pose Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Michel, Frank, Alexander Krull, Eric Brachmann, Michael Ying Yang, Stefan Gumhold, and Carsten Rother (Sept. 2015). "Pose Estimation of Kinematic Chain Instances via Object Coordinate Regression". In: DOI: 10.5244/C.29.181.

Mildenhall, Ben, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng (2020). "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". In: *ECCV*.

Mo, Ningkai, Wanshui Gan, Naoto Yokoya, and Shifeng Chen (June 2022). "ES6D: A Computation Efficient and Symmetry-Aware 6D Pose Regression Framework". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6718–6727.

Moakher, Maher (Apr. 2002). "Means and Averaging in the Group of Rotations". In: *SIAM Journal on Matrix Analysis and Applications* 24. DOI: 10.1137/S0895479801383877.

Morgan, Andrew S., Bowen Wen, Junchi Liang, Abdeslam Boularias, Aaron M. Dollar, and Kostas Bekris (July 2021). "Vision-driven Compliant Manipulation for Reliable; High-Precision Assembly Tasks". In: *Proceedings of Robotics: Science and Systems*. Virtual. DOI: 10.15607/RSS.2021.XVII.070.

Müller, Thomas, Alex Evans, Christoph Schied, and Alexander Keller (Jan. 2022). "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". In: *arXiv:2201.05989*.

Murphy, Kevin and Stuart Russell (2001). "Rao-Blackwellised Particle Filtering for Dynamic Bayesian Networks". In: *Sequential Monte Carlo Methods in Practice*. New York, NY: Springer New York, pp. 499–515. ISBN: 978-1-4757-3437-9. DOI: 10.1007/978-1-4757-3437-9_24. URL: https://doi.org/10.1007/978-1-4757-3437-9_24.

Nee, Andrew, S K Ong, George Chryssolouris, and Dimitris Mourtzis (Dec. 2012). "Augmented reality applications in design and manufacturing". In: *CIRP Annals - Manufacturing Technology* 61, pp. 657–679. DOI: 10.1016/j.cirp.2012.05.010.

Nguyen, Van Nguyen, Yinlin Hu, Yang Xiao, Mathieu Salzmann, and Vincent Lepetit (June 2022). "Templates for 3D Object Pose Estimation Revisited: Generalization to New Objects and Robustness to Occlusions". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6771–6780.

Niemeyer, Michael and Andreas Geiger (2021). "GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields". In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Oberweger, Markus, Mahdi Rad, and Vincent Lepetit (Sept. 2018). "Making Deep Heatmaps Robust to Partial Occlusions for 3D Object Pose Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Ozkul, Emrah and Sarp Kumlu (Dec. 2019). "Augmented Reality Applications in Tourism". In: *International Journal of Contemporary Tourism Research*, pp. 107–122. DOI: 10.30625/ijctr.625192.

Papazov, Chavdar and Darius Burschka (2010). "An efficient ransac for 3d object recognition in noisy and occluded scenes". In: *Asian conference on computer vision*. Springer, pp. 135–148.

Papon, Jeremie and Markus Schoeler (Dec. 2015). "Semantic Pose Using Deep Networks Trained on Synthetic RGB-D". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Park, Kiru, Timothy Patten, and Markus Vincze (Oct. 2019). "Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation". In: *IEEE International Conference on Computer Vision (ICCV)*.

Pauwels, Karl and Danica Kragic (Sept. 2015). "SimTrack: A simulation-based framework for scalable real-time object pose detection and tracking". In: pp. 1300–1307. DOI: 10.1109/IROS.2015.7353536.

Pavlakos, G., X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis (May 2017). "6-DoF object pose from semantic keypoints". In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2011–2018. DOI: 10.1109/ICRA.2017.7989233.

Peng, Sida, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao (June 2019). "PVNet: Pixel-Wise Voting Network for 6DoF Pose Estimation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Piga, Nicola A., Fabrizio Bottarel, Claudio Fantacci, Giulia Vezzani, Ugo Pattacini, and Lorenzo Natale (2021). "MaskUKF: An Instance Segmentation Aided Unscented Kalman Filter for 6D Object Pose and Velocity Tracking". In: *Frontiers in Robotics and AI* 8, p. 38. ISSN: 2296-9144. DOI: 10.3389/frobt.2021.594583. URL: https://www.frontiersin.org/article/10.3389/frobt.2021.594583.

Pitteri, G., M. Ramamonjisoa, S. Ilic, and V. Lepetit (2019). "On Object Symmetries and 6D Pose Estimation from Images". In: *International Conference on 3D Vision (3DV)*, pp. 614–622.

Qi, Charles R., Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas (June 2018). "Frustum PointNets for 3D Object Detection From RGB-D Data". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qi, Charles R., Hao Su, Kaichun Mo, and Leonidas J. Guibas (July 2017). "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qian, Kun, Yanhui Duan, Chaomin Luo, Yongqiang Zhao, and Xingshuo Jing (2023). "Pixel-Level Domain Adaptation for Real-to-Sim Object Pose Estimation". In: *IEEE Transactions on Cognitive and Developmental Systems* 15.3, pp. 1618–1627. DOI: 10.1109/TCDS.2023.3237502.

Rad, Mahdi and Vincent Lepetit (Oct. 2017). "BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth". In: *IEEE International Conference on Computer Vision (ICCV)*.

Redmon, Joseph and Ali Farhadi (2018). *YOLOv3: An Incremental Improvement*. cite arxiv:1804.02767Comme Tech Report. URL: http://arxiv.org/abs/1804.02767.

Rennie, Colin, Rahul Shome, Kostas Bekris, and Alberto De Souza (Sept. 2015). "A Dataset for Improved RGBD-Based Object Detection and Pose Estimation for Warehouse Pick-and-Place". In: *IEEE Robotics and Automation Letters* 1. DOI: 10.1109/LRA.2016.2532924.

Rios-Cabrera, Reyes and Tinne Tuytelaars (Dec. 2013). "Discriminatively Trained Templates for 3D Object Detection: A Real Time Scalable Approach". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 2048–2055. DOI: 10.1109/ICCV.2013.256.

Roberts, Lawrence G (1963). "Machine perception of three-dimensional solids". PhD thesis. Massachusetts Institute of Technology.

Rothganger, Fred, Svetlana Lazebnik, Cordelia Schmid, and J. Ponce (Mar. 2006). "3D Object Modeling and Recognition Using Local Affine-Invariant Image Descriptors and Multi-View Spatial Constraints". In: *International Journal of Computer Vision* 66, pp. 231–259. DOI: 10.1007/s11263-005-3674-1.

Rothganger, Fred, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce (2006). "3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints". In: *International journal of computer vision* 66, pp. 231–259.

Roy, A., Xi Zhang, N. Wolleb, C. P. Quintero, and M. Jägersand (2015). "Tracking benchmark and evaluation for manipulation tasks". In: *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 2448–2453. DOI: 10.1109/ICRA.2015.7139526.

Sahin, Caner, Guillermo Garcia-Hernando, Juil Sock, and Tae-Kyun Kim (2020). "A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators". In: *Image and Vision Computing*, p. 103898. ISSN: 0262-8856. DOI: https://doi.org/10.1016/j.imavis.2020.103898. URL: http://www.sciencedirect.com/science/article/pii/S0262885620300305.

Sahin, Caner and Tae-Kyun Kim (2019). "Category-Level 6D Object Pose Recovery in Depth Images". In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Cham: Springer International Publishing, pp. 665–681. ISBN: 978-3-030-11009-3.

Sattler, Torsten, Qunjie Zhou, Marc Pollefeys, and Laura Leal-Taixe (June 2019). "Understanding the Limitations of CNN-Based Absolute Camera Pose Regression". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schmidt, Tanner, Richard Newcombe, and Dieter Fox (July 2014). "DART: Dense Articulated Real-Time Tracking". In: *Proceedings of Robotics: Science and Systems*. DOI: 10.15607/RSS.2014.X.030.

Schwarz, Katja, Yiyi Liao, Michael Niemeyer, and Andreas Geiger (2020). "GRAF: Generative Radiance Fields for 3D-Aware Image Synthesis". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Segal, Aleksandr, Dirk Haehnel, and Sebastian Thrun (2009). "Generalized-icp." In: *Robotics: science and systems (RSS)*. Vol. 2. 4. Seattle, WA, p. 435.

Shao, Jianzhun, Yuhang Jiang, Gu Wang, Zhigang Li, and Xiangyang Ji (June 2020). "PFRL: Pose-Free Reinforcement Learning for 6D Pose Estimation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shi, Yifei, Junwen Huang, Xin Xu, Yifan Zhang, and Kai Xu (June 2021). "StablePose: Learning 6D Object Poses From Geometrically Stable Patches". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15222–15231.

Shotton, J., B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon (June 2013). "Scene Coordinate Regression Forests for Camera Relocalization in RGB-D Images". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937. DOI: 10.1109/CVPR.2013.377.

Shugurov, Ivan, Fu Li, Benjamin Busam, and Slobodan Ilic (June 2022). "OSOP: A Multi-Stage One Shot Object Pose Estimation Framework". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6835–6844.

Simonyan, Karen and Andrew Zisserman (Sept. 2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *arXiv 1409.1556*.

Song, Chen, Jiaru Song, and Qixing Huang (2020). *HybridPose: 6D Object Pose Estimation under Hybrid Representations*. arXiv: 2001.01869 [cs.CV].

Song, Shuran and Jianxiong Xiao (June 2016). "Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Su, Hao, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas (Dec. 2015). "Render for CNN: Viewpoint Estimation in Images Using CNNs Trained With Rendered 3D Model Views". In: *IEEE International Conference on Computer Vision (ICCV)*.

Su, Yongzhi, Jason Rambach, Nareg Minaskan, Paul Lesur, Alain Pagani, and Didier Stricker (2019). "Deep Multi-state Object Pose Estimation for Augmented Reality Assembly". In: *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, pp. 222–227. DOI: 10.1109/ISMAR-Adjunct.2019.00-42.

Su, Yongzhi, Mahdi Saleh, Torben Fetzer, Jason Rambach, Nassir Navab, Benjamin Busam, Didier Stricker, and Federico Tombari (2022). *ZebraPose: Coarse to Fine Surface Encoding for 6DoF Object Pose Estimation*. DOI: 10.48550/ARXIV.2203.09418. URL: https://arxiv.org/abs/2203.09418.

Sun, Jiaming, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou (2022). "OnePose: One-Shot Object Pose Estimation without CAD Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6825–6834.

Sun, Yinghan, Linfang Zheng, Hua Chen, and Wei Zhang (2023). *Multi-Resolution Planar Region Extraction for Uneven Terrains*. arXiv: 2311.12562 [cs.CV].

Sundermeyer, Martin, Maximilian Durner, En Yen Puang, Zoltan-Csaba Marton, Narunas Vaskevicius, Kai O. Arras, and Rudolph Triebel (June 2020). "Multi-Path Learning for Object Pose Estimation Across Domains". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Sundermeyer, Martin, Zoltan Marton, Maximilian Durner, and Rudolph Triebel (Oct. 2019). "Augmented Autoencoders: Implicit 3D Orientation Learning for 6D Object Detection". In: *International Journal of Computer Vision (IJCV)* 128. DOI: 10.1007/s11263-019-01243-8.

Sundermeyer, Martin, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel (Sept. 2018). "Implicit 3D Orientation Learning for 6D Object Detection from RGB Images". In: *The European Conference on Computer Vision (ECCV)*.

Suwajanakorn, Supasorn, Noah Snavely, Jonathan J Tompson, and Mohammad Norouzi (2018). "Discovery of Latent 3D Keypoints via End-to-end Geometric Reasoning". In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., pp. 2059–2070. URL: http://papers.nips.cc/paper/7476-discovery-of-latent-3d-keypoints-via-end-to-end-geometric-reasoning.pdf.

Tan, David Joseph and Slobodan Ilic (June 2014). "Multi-Forest Tracker: A Chameleon in Tracking". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: 10.1109/CVPR.2014.157.

Tan, David Joseph, Federico Tombari, Slobodan Ilic, and Nassir Navab (Dec. 2015). "A Versatile Learning-Based 3D Temporal Tracker: Scalable, Robust, Online". In: *IEEE International Conference on Computer Vision (ICCV)*, pp. 693–701. DOI: 10.1109/ICCV.2015.86.

Tejani, Alykhan, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim (Sept. 2014). "Latent-Class Hough Forests for 3D Object Detection and Pose Estimation". In: pp. 462–477. DOI: 10.1007/978-3-319-10599-4_30.

Tekin, Bugra, Federica Bogo, and Marc Pollefeys (June 2019). "H+O: Unified Egocentric Recognition of 3D Hand-Object Poses and Interactions". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4506–4515. DOI: 10.1109/CVPR.2019.00464.

Tekin, Bugra, Sudipta Sinha, and Pascal Fua (June 2018). "Real-Time Seamless Single Shot 6D Object Pose Prediction". In: pp. 292–301. DOI: 10.1109/CVPR.2018.00038.

Tian, Meng, Marcelo H Ang, and Gim Hee Lee (2020). "Shape prior deformation for categorical 6d object pose and size estimation". In: *European Conference on Computer Vision (ECCV)*. Springer, pp. 530–546.

Tian, Meng, Marcelo H Ang Jr, and Gim Hee Lee (Aug. 2020). "Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation". In: *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tolstikhin, Ilya, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Martin Keysers, Jakob Uszkoreit, Mario Lučić, and Alexey Dosovitskiy (2021). "MLP-Mixer: An All-MLP Architecture for Vision". In: *NeurIPS 2021 (poster)*. URL: https://arxiv.org/abs/2105.01601.

Tremblay, Jonathan, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield (Sept. 2018). *Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects*. DOI: 10.48550/ARXIV.1809.10790. URL: https://arxiv.org/abs/1809.10790.

Tulsiani, Shubham and Jitendra Malik (June 2015). "Viewpoints and Keypoints". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Umeyama, S. (Apr. 1991). "Least-squares estimation of transformation parameters between two point patterns". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.4, pp. 376–380. ISSN: 1939-3539. DOI: 10.1109/34.88573.

Vacchetti, L., V. Lepetit, and P. Fua (2004). "Stable real-time 3D tracking using online and offline information". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.10, pp. 1385–1391. DOI: 10.1109/TPAMI.2004.92.

Valentin, J., M. Nießner, J. Shotton, A. Fitzgibbon, S. Izadi, and P. Torr (June 2015). "Exploiting uncertainty in regression forests for accurate camera relocalization". In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4400–4408. DOI: 10.1109/CVPR.2015.7299069.

Vidal, Joel, Chyi-Yeu Lin, and Robert Martí (2018). *6D Pose Estimation using an Improved Method based on Point Pair Features*. DOI: 10.48550/ARXIV.1802.08516. URL: https://arxiv.org/abs/1802.08516.

Vidanage, KH, Harinda Fernando, Lakmini Abeywardhana, et al. (2023). "Human, Object and Pose Detection for Theft Prevention through Surveillance System". In: *International Research Journal of Innovations in Engineering and Technology* 7.11, p. 160.

Vu, Van-Duc, Dinh-Dai Hoang, Phan Xuan Tan, Van-Thiep Nguyen, Thu-Uyen Nguyen, Ngoc-Anh Hoang, Khanh-Toan Phan, Duc-Thanh Tran, Duy-Quang Vu, Phuc-Quan Ngo, et al. (2024). "Occlusion-Robust Pallet Pose Estimation for Warehouse Automation". In: *IEEE Access*.

Wan, E.A. and R. Van Der Merwe (2000). "The unscented Kalman filter for nonlinear estimation". In: *Proceedings of the IEEE 2000 Adaptive Systems for Signal Processing, Communications, and Control Symposium (Cat. No.00EX373)*, pp. 153–158. DOI: 10.1109/ASSPCC.2000.882463.

Wang, Chen, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu (2020). "6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints". In: *International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 10059–10066.

Wang, Chen, Danfei Xu, Yuke Zhu, Roberto Martin-Martin, Cewu Lu, Li Fei-Fei, and Silvio Savarese (June 2019). "DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Gu, Fabian Manhardt, Federico Tombari, and Xiangyang Ji (June 2021). "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 (CVPR)*, pp. 16611–16621.

Wang, He, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas (June 2019). "Normalized object coordinate space for category-level 6d object pose and

size estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2642–2651.

Wang, Jiaze, Kai Chen, and Qi Dou (2021). "Category-level 6D object pose estimation via cascaded relation and recurrent reconstruction networks". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 4807–4814.

Wang, Pengyuan, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam (June 2022). "PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation With Photometrically Challenging Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21222–21231.

Wang, Yue, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon (2019). "Dynamic Graph CNN for Learning on Point Clouds". In: *ACM Transactions on Graphics (TOG)*.

Wen, Bowen and Kostas Bekris (2021). "BundleTrack: 6D Pose Tracking for Novel Objects without Instance or Category-Level 3D Models". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8067–8074. DOI: 10.1109/IROS51168.2021.9635991.

Wen, Bowen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal (June 2022). "You Only Demonstrate Once: Category-Level Manipulation from Single Visual Demonstration". In: *Proceedings of Robotics: Science and Systems*. New York City, NY, USA. DOI: 10.15607/RSS.2022.XVIII.044.

Wen, Bowen, Chaitanya Mitash, Baozhang Ren, and Kostas Bekris (July 2020). "se(3)-TrackNet: Data-driven 6D Pose Tracking by Calibrating Image Residuals in Synthetic Domains". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Wen, Xin, Zhizhong Han, Yan-Pei Cao, Pengfei Wan, Wen Zheng, and Yu-Shen Liu (2021). "Cycle4Completion: Unpaired Point Cloud Completion using Cycle Transformation with Miss-

ing Region Coding". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Weng, Yijia, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas (2021). "CAPTRA: Category-level pose tracking for rigid and articulated objects from point clouds". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13209–13218.

Wetterstrand, K A (2016). *DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP)*. URL: www.genome.gov/sequencingcosts.

Wohlhart, Paul and Vincent Lepetit (June 2015). "Learning descriptors for object recognition and 3D pose estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3109–3118. DOI: 10.1109/CVPR.2015.7298930.

Wu, Po-Chen, Yueh-Ying Lee, Hung-Yu Tseng, Hsuan-I Ho, Ming-Hsuan Yang, and Shao-Yi Chien (2017). "A Benchmark Dataset for 6DoF Object Pose Tracking". In: *IEEE International Symposium on Mixed and Augmented Reality (ISMAR Adjunct)*.

Wuthrich, Manuel, Peter Pastor, Mrinal Kalakrishnan, Jeannette Bohg, and Stefan Schaal (Nov. 2013). "Probabilistic Object Tracking using a Range Camera". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3195–3202. DOI: 10.1109/IROS.2013.6696810.

Xiang, Yu, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox (June 2018). "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes". In: *Robotics: Science and Systems (RSS)*. DOI: 10.15607/RSS.2018.XIV.019.

Xiao, J., J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba (June 2010). "SUN database: Large-scale scene recognition from abbey to zoo". In: *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3485–3492. DOI: 10.1109/CVPR.2010.5539970.

Xie, Chulin, Chuxin Wang, Bo Zhang, Hao Yang, Dong Chen, and Fang Wen (June 2021). "Style-Based Point Generator With Adversarial Rendering for Point Cloud Completion". In: *Pro-*

*ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4619–4628.

Xie, Haozhe, Hongxun Yao, Shangchen Zhou, Jiageng Mao, Shengping Zhang, and Wenxiu Sun (2020). "GRNet: Gridding Residual Network for Dense Point Cloud Completion". In: *CoRR* abs/2006.03761. arXiv: 2006.03761. URL: https://arxiv.org/abs/2006.03761.

Xu, Danfei, Dragomir Anguelov, and Ashesh Jain (June 2018). "PointFusion: Deep Sensor Fusion for 3D Bounding Box Estimation". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, Yan, Kwan-Yee Lin, Guofeng Zhang, Xiaogang Wang, and Hongsheng Li (June 2022). "RNNPose: Recurrent 6-DoF Object Pose Refinement With Robust Correspondence Field Estimation and Pose Optimization". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14880–14890.

Yang, Heng and Marco Pavone (June 2023). "Object Pose Estimation With Statistical Guarantees: Conformal Keypoint Detection and Geometric Uncertainty Propagation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8947–8958.

Yong, Hongwei, Jianqiang Huang, Xiansheng Hua, and Lei Zhang (2020). *Gradient Centralization: A New Optimization Technique for Deep Neural Networks*. DOI: 10.48550/ARXIV.2004.01461. URL: https://arxiv.org/abs/2004.01461.

Yuan, Wentao, Tejas Khot, David Held, Christoph Mertz, and Martial Hebert (2018). "PCN: Point Completion Network". In: *CoRR* abs/1808.00671. arXiv: 1808.00671. URL: http://arxiv.org/abs/1808.00671.

Zakharov, Sergey, Ivan Shugurov, and Slobodan Ilic (Oct. 2019). "DPOD: 6D Pose Object Detector and Refiner". In: *The IEEE International Conference on Computer Vision (ICCV)*.

Zhang, Huijie, Anthony Opipari, Xiaotong Chen, Jiyue Zhu, Zeren Yu, and Odest Chadwicke Jenkins (2022). *TransNet: Category-Level Transparent Object Pose Estimation*. DOI: 10.48550/ARXIV.2208.10002. URL: https://arxiv.org/abs/2208.10002.

Zhang, Junzhe, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy (June 2021). "Unsupervised 3D Shape Completion Through GAN Inversion". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1768–1777.

Zhang, Kun, Chen Wang, Hua Chen, Jia Pan, Michael Yu Wang, and Wei Zhang (2023). "Vision-based Six-Dimensional Peg-in-Hole for Practical Connector Insertion". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1771–1777. DOI: 10.1109/ICRA48891.2023.10161116.

Zhang, Michael R., James Lucas, Geoffrey Hinton, and Jimmy Ba (2019). *Lookahead Optimizer: k steps forward, 1 step back*. DOI: 10.48550/ARXIV.1907.08610. URL: https://arxiv.org/abs/1907.08610.

Zhang, Ruida, Yan Di, Zhiqiang Lou, Fabian Manhardt, Federico Tombari, and Xiangyang Ji (2022). *RBP-Pose: Residual Bounding Box Projection for Category-Level Pose Estimation*. DOI: 10.48550/ARXIV.2208.00237. URL: https://arxiv.org/abs/2208.00237.

Zhang, Ruida, Yan Di, Fabian Manhardt, Federico Tombari, and Xiangyang Ji (2022). *SSP-Pose: Symmetry-Aware Shape Prior Deformation for Direct Category-Level Object Pose Estimation*. DOI: 10.48550/ARXIV.2208.06661. URL: https://arxiv.org/abs/2208.06661.

Zhang, Shaobo, Wanqing Zhao, Ziyu Guan, Xianlin Peng, and Jinye Peng (June 2021). "Keypoint-Graph-Driven Learning Framework for Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1065–1073.

Zhang, Wenxiao, Qingan Yan, and Chunxia Xiao (2020). "Detail Preserved Point Cloud Completion via Separated Feature Aggregation". In: *CoRR* abs/2007.02374. arXiv: 2007.02374. URL: https://arxiv.org/abs/2007.02374.

Zhang, Zhongqun, Wei Chen, Linfang Zheng, Aleš Leonardis, and Hyung Jin Chang (2023). "Trans6D: Transformer-Based 6D Object Pose Estimation and Refinement". In: *Computer*

*Vision – ECCV 2022 Workshops*. Ed. by Leonid Karlinsky, Tomer Michaeli, and Ko Nishino. Cham: Springer Nature Switzerland, pp. 112–128. ISBN: 978-3-031-25085-9.

Zheng, Linfang, Aleš Leonardis, Tze Ho Elden Tse, Nora Horanyi, Hua Chen, Wei Zhang, and Hyung Jin Chang (2022). "TP-AE: Temporally Primed 6D Object Pose Tracking with Auto-Encoders". In: *2022 IEEE International Conference on Robotics and Automation (ICRA)*.

Zheng, Linfang, Tze Ho Elden Tse, Chen Wang, Yinghan Sun, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang (June 2024). "GeoReF: Geometric Alignment Across Shape Variation for Category-level Object Pose Refinement". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10693–10703.

Zheng, Linfang, Chen Wang, Yinghan Sun, Esha Dasgupta, Hua Chen, Aleš Leonardis, Wei Zhang, and Hyung Jin Chang (June 2023). "HS-Pose: Hybrid Scope Feature Extraction for Category-Level Object Pose Estimation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17163–17173.

— (n.d.). "Supplementary Material of HS-Pose: Hybrid Scope Feature Extraction for Category-level Object Pose Estimation". In: ().

Zhou, Jun, Kai Chen, Linlin Xu, Qi Dou, and Jing Qin (Oct. 2023). "Deep Fusion Transformer Network with Weighted Vector-Wise Keypoints Voting for Robust 6D Object Pose Estimation". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13967–13977.

Zhou, Yi, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li (June 2019). "On the Continuity of Rotation Representations in Neural Networks". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5738–5746. DOI: 10.1109/CVPR.2019.00589.

Zhu, Menglong, Konstantinos Derpanis, Yinfei Yang, Samarth Brahmbhatt, Mabel Zhang, Cody Phillips, Matthieu Lecce, and Kostas Daniilidis (May 2014). "Single image 3D object detection and pose estimation for grasping". In: pp. 3936–3943. DOI: 10.1109/ICRA.2014.6907430.