Meta-Learning for Robotic Vision Applications

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften

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

genehmigte
Dissertation

von

Ning Gao

aus Heilongjiang

Tag der mündlichen Prüfung: 14.02.2025
1. Referent: Prof. Dr. Gerhard Neumann
2. Referent: Prof. Dr. Vincent Lepetit

Abstract

This thesis presents innovative advancements in robotic vision systems through the use of meta-learning techniques. The research primarily addresses key challenges in improving model generalization, interpretability, and adaptability across various vision tasks. The core contributions include novel algorithms and practical implementations designed to enhance the efficiency and accuracy of these systems.

Key contributions and findings are as follows:

- Meta-Learning for Vision Regression Tasks: The research begins with an in-depth analysis of the factors affecting the performance of meta-learning in vision regression tasks. It introduces Functional Contrastive Learning (FCL), a novel method that significantly improves the generalization capabilities of Conditional Neural Processes (CNPs). FCL enhances task expressivity and reduces prediction errors, demonstrated through rigorous experiments on new datasets.
- GAML: Geometry-Aware Meta-Learner for Cross-Category 6D Pose Estimation: GAML learns object representation in a category-agnostic way across object categories. A neural process-based meta-learning approach is employed to train an encoder to capture texture and geometry of an object in a latent representation, based on very few RGB-D images and ground-truth keypoints. The latent representation is then used by a simultaneously meta-trained decoder to predict the 6D pose of the object in new images. Furthermore, a novel geometry-aware decoder is introduced for the keypoint prediction using a Graph Neural Network (GNN), which explicitly takes geometric We generate a new fully-annotated synthetic datasets from Multiple Categories in Multiple Scenes (MCMS) for evaluation.
- SA6D: Self-Adaptive Few-Shot 6D Pose Estimator: The thesis presents SA6D, a self-adaptive method for 6D pose estimation of novel and

occluded objects using minimal reference images. This approach includes an online self-adaptation module, a region proposal module, and a refinement module, ensuring robust and accurate pose predictions. The method shows superior performance in handling occlusions and minimal reference scenarios, making it suitable for practical robotic applications.

- Interpretable Object Abstraction via Clustering-Based Slot Initialization: To improve object abstraction and interpretability, the thesis proposes a clustering-based slot initialization technique. This method employs clustering algorithms like k-means and mean-shift to enhance initial object representations, leading to improved object discovery and novel view synthesis. The technique's effectiveness is validated through extensive experiments, showing significant advancements over traditional methods.
- Meta-Learning Regrasping Strategies for Physical-Agnostic Objects: The final research focus is on meta-learning strategies for regrasping objects with diverse physical properties. By learning task embeddings that capture the physical characteristics of objects, the proposed methods enable robots to dynamically adapt their grasping strategies. The research includes the development of heterogeneous datasets and extensive simulation and real-world experiments, highlighting the methods' practical applicability.

The thesis concludes by summarizing the significant contributions and findings. It emphasizes the potential impact of the proposed methods on advancing robotic vision systems and suggests directions for future research, including the integration of these techniques into more complex robotic systems and exploring additional applications in computer vision and robotics.

Overall, this thesis provides substantial advancements in the field of robotic vision, offering innovative solutions to enhance adaptability, accuracy, and interpretability through meta-learning techniques.

Zusammenfassung

Diese Doktorarbeit präsentiert innovative Fortschritte in robotischen Visionssystemen durch den Einsatz von Meta-Learning-Techniken. Die Forschung befasst sich hauptsächlich mit den zentralen Herausforderungen der Verbesserung der Modellverallgemeinerung, Interpretierbarkeit und Anpassungsfähigkeit über verschiedene Vision-Aufgaben hinweg. Die Kernbeiträge umfassen neuartige Algorithmen und praktische Implementierungen, die darauf abzielen, die Effizienz und Genauigkeit dieser Systeme zu verbessern.

Wichtige Beiträge und Ergebnisse sind wie folgt:

- Meta-Learning für Vision-Regression-Aufgaben: Die Forschung beginnt mit einer eingehenden Analyse der Faktoren, die die Leistung von Meta-Learning bei Vision-Regression-Aufgaben beeinflussen. Es wird das Functional Contrastive Learning (FCL) eingeführt, eine neuartige Methode, die die Verallgemeinerungsfähigkeit von Conditional Neural Processes (CNPs) erheblich verbessert. FCL erhöht die Ausdruckskraft der Aufgaben und reduziert die Vorhersagefehler, was durch rigorose Experimente an neuen Datensätzen demonstriert wird.
- GAML: Geometry-Aware Meta-Learner für kategorieübergreifende 6D-Posenschätzung: GAML erlernt die Objektrepräsentation in einer kategorieunabhängigen Weise über verschiedene Objektkategorien hinweg. Ein auf neuralen Prozessen basierender Meta-Learning-Ansatz wird verwendet, um einen Encoder zu trainieren, der Textur und Geometrie eines Objekts in einer latenten Repräsentation erfasst, basierend auf sehr wenigen RGB-D-Bildern und Ground-Truth-Schlüsselpunkten. Die latente Repräsentation wird dann von einem gleichzeitig metatrainierten Decoder genutzt, um die 6D-Position des Objekts in neuen Bildern vorherzusagen. AuSSerdem wird ein neuartiger, geometriebewusster Decoder eingeführt, der für die Schlüsselpunktschätzung ein Graph Neural Network (GNN) verwendet, das explizit geometrische Informationen berücksichtigt. Zur Evaluierung erstellen wir ein neues,

vollständig annotiertes synthetisches Datenset aus mehreren Kategorien in mehreren Szenen (MCMS).

- SA6D: Selbstanpassender Few-Shot 6D Pose Estimator: Die Arbeit stellt SA6D vor, eine selbstanpassende Methode zur 6D-Pose-Schätzung neuartiger und verdeckter Objekte unter Verwendung minimaler Referenzbilder. Dieser Ansatz umfasst ein Online-Selbstanpassungsmodul, ein Region-Vorschlagsmodul und ein Verfeinerungsmodul, die robuste und genaue Pose-Vorhersagen gewährleisten. Die Methode zeigt eine überlegene Leistung im Umgang mit Verdeckungen und minimalen Referenzszenarien und eignet sich somit für praktische robotische Anwendungen.
- Interpretierbare Objektabstraktion durch Cluster-basierte Slot Initialisierung: Um die Objektabstraktion und Interpretierbarkeit zu verbessern, schlägt die Arbeit eine Cluster-basierte Slot-Initialisierungstechnik vor. Diese Methode verwendet Clustering-Algorithmen wie k-means und mean-shift, um die anfänglichen Objektrepräsentationen zu verbessern, was zu einer besseren Objekterkennung und neuartigen Sicht-Synthese führt. Die Wirksamkeit der Technik wird durch umfangreiche Experimente validiert, die signifikante Fortschritte gegenüber herkömmlichen Methoden zeigen.
- Meta-Learning-Strategien für das Umgreifen physikalisch-agnostischer Objekte: Der abschlieSSende Forschungsschwerpunkt liegt auf Meta-Learning-Strategien für das Umgreifen von Objekten mit unterschiedlichen physikalischen Eigenschaften. Durch das Lernen von Aufgaben-Embeddings, die die physikalischen Eigenschaften von Objekten erfassen, ermöglichen die vorgeschlagenen Methoden den Robotern, ihre Greifstrategien dynamisch anzupassen. Die Forschung umfasst die Entwicklung heterogener Datensätze sowie umfangreiche Simulations- und Realwelt-Experimente, die die praktische Anwendbarkeit der Methoden hervorheben.

Die Arbeit schlieSSt mit einer Zusammenfassung der bedeutenden Beiträge und Erkenntnisse. Sie betont das potenzielle Impact der vorgeschlagenen Methoden auf die Weiterentwicklung robotischer Visionssysteme und schlägt Richtungen für zukünftige Forschungen vor, einschlieSSlich der Integration dieser Techniken in komplexere robotische Systeme und der Erforschung weiterer Anwendungen in der Computer Vision und Robotik.

Insgesamt bietet diese Doktorarbeit erhebliche Fortschritte im Bereich der robotischen Vision, indem sie innovative Lösungen zur Verbesserung der Anpassungsfähigkeit, Genauigkeit und Interpretierbarkeit durch Meta-Learning-Techniken bietet.

Contents

Abstract i

Zusammenfassung iii

List of Figures xi

List of Tables xix

1 Introduction 1
1.1 Meta-Learning as Human Cognition 1
1.2 Object and Scene Representation 3
1.3 Robotic Application 6

2 Background 9
2.1 Latent Variable Networks 9
2.2 Meta-Learning 10
2.3 Few-Shot Learning 11
2.4 Notation 12

3 What Matters for Meta-Learning Vision Regression Tasks . . . 15
3.1 Introduction 15
3.2 Related Work 19
3.3 Study Design 21
3.3.1 Problem Setting 21
3.3.2 Datasets 22
3.3.3 Data Augmentation, Domain Randomization, Task Augmentation and Meta Regularization 23
3.3.4 Functional Contrastive Learning (FCL) 24
3.3.5 Objective Functions and Evaluation Metrics 25
3.4 Experiments 26
3.4.1 Training Details 27

3.4.2 Results and Discussions 27
3.4.3 Limitations 38
3.5 Conclusion 40

4 GAML: Geometry-Aware Meta-Learner for Cross-Category 6D Pose Estimation 41
4.1 Introduction 41
4.2 Related Work 44
4.3 Preliminary - Conditional Neural Processes 46
4.4 Approach 47
4.4.1 Overview 47
4.4.2 Feature Extraction 48
4.4.3 Meta-Learner for Keypoint Detection 49
4.4.4 Geometry-Aware Keypoint Decoder 50
4.5 Experiments 52
4.5.1 Datasets 52
4.5.2 Evaluation Metrics 53
4.5.3 Implementation and Training Details 53
4.5.4 Evaluation Results 54
4.5.5 Ablation Study 59
4.5.6 Limitations 63
4.6 Conclusion 63

5 SA6D: Self-Adaptive Few-Shot 6D Pose Estimator for Novel and Occluded Objects 65
5.1 Introduction 65
5.2 Related Work 68
5.3 Preliminaries 70
5.4 Method 70
5.4.1 Online Self-Adaptation Module 71
5.4.2 Region Proposal Module 73
5.4.3 Refinement Module 74
5.5 Experiments 75
5.5.1 Datasets and Metrics 76
5.5.2 Results and Discussion 77
5.5.3 Limitations 87
5.6 Conclusion 93

6 Enhancing Interpretable Object Abstraction via Clustering-based Slot Initialization . . . 95
6.1 Introduction . . . 95
6.2 Guiding Slot Initialization using Clustering . . . 97
6.2.1 Image-Dependent Slot Initialization . . . 97
6.2.2 Permutation-Invariant Slot Initialization . . . 97
6.2.3 Automatic Tuning of the Number of Slots using Mean-Shift . . . 98
6.3 Related Work . . . 99
6.4 Experiments . . . 101
6.4.1 Baselines . . . 101
6.4.2 Datasets and Metrics . . . 101
6.4.3 Implementation details . . . 102
6.4.4 Object Discovery . . . 109
6.4.5 Novel View Synthesis . . . 114
6.5 Conclusion . . . 115

7 Meta-Learning Regrasping Strategies for Physical-Agnostic Objects . . . 121
7.1 Introduction . . . 121
7.2 Related Work . . . 123
7.3 Methodology . . . 126
7.3.1 Datasets with Heterogeneous Physical Properties . . . 126
7.3.2 Simulator . . . 127
7.3.3 Meta-Learn Physical Properties as Task Embeddings . 127
7.4 Experiments . . . 130
7.4.1 Evaluation Metrics . . . 130
7.4.2 Statistics of the Collected Data . . . 131
7.4.3 Baselines . . . 132
7.4.4 Experimental Results . . . 132
7.4.5 Limitations . . . 136
7.5 Conclusion . . . 136

8 Conclusion . . . 137

Bibliography . . . 141

List of Figures

3.1 Meta-learning vision regression tasks are designed to i) identify the queried object from context and predict its position for target images (Distractor), ii) identify the object's canonical pose from context and predict the 1D rotation relative to the canonical pose for target images (ShapeNet1D), iii) predict the 2D rotation w.r.t. the canonical pose with random background (ShapeNet2D). Predictions are performed on unseen objects. 16
3.2 (a) CNP Prediction error (pixel) vs context number for the Distractor task using Max aggregation and Max + FCL (Max_{FCL}). Results are evaluated on novel objects from both intra-category (IC) and unseen cross-category (CC) levels. (b) CNP (CA) Prediction error vs context number for ShapeNet2D using DA + TA. (c) We compare a classical object detection method and CNP (Max) using same dataset for training on Distractor. The classical model is further fine-tuned on each new task. The results are shown in dependence of the number of images used for fine-tuning or as context set.(d) Prediction error between the fine-tuned model and CNP (CA) on ShapeNet1D. 28
3.3 Visualization of latent variables on (a) max aggregation (b) max aggregation + functional contrastive learning (Max_{FCL}). 36
3.4 Examples of Distractor on novel categories (sofa and watercraft) where green dots are ground-truth and blue dots are predicted positions. 37
3.5 Examples of ShapeNet1D on novel categories (piano, bed, bus). . 38
3.6 Examples of ShapeNet2D on novel categories (piano, bed, bus). Predictions are converted to (azimuth, elevation) angles. 39

4.1 Illustration of the difference between traditional instance-level 6D pose estimation methods and our approach. Unlike other methods, our proposed approach generalizes to novel objects given a few context observations. The projected ground-truth keypoints are visualized as blue points in the context images. The predicted segmentation and keypoints are visualized in the target images. . . 42
4.5 Samples from MCMS dataset. 52
4.7 **Qualitative comparison on trained LineMOD objects.** Triangles and circles are the projections of ground-truth and predicted keypoints respectively. It can be observed that keypoint predictions of our method are more accurate. 55
4.8 **Qualitative comparison on new LineMOD objects.** Compared with FFB6D, the pose estimation on new objects of our GAML model is more accurate. 56
4.9 **Qualitative results on Toy-MCMS.** Our model can handle large intra-category variations. The car category is illustrated as an example. 56
4.12 Limitations of the proposed method. 63

5.1 We present a generalizable and category-agnostic few-shot 6D object pose estimator using a small number of posed RGB-D images as reference. Compared to existing methods, our approach provides robust and accurate predictions on novel objects against occlusions without requiring retraining or any object information. 66
5.2 **Overview.** SA6D includes three modules: i) The *online self-adaptation module* discovers and segments the target object (*milk cow*) from a cluttered scene giving a few posed RGB-D images as reference. Subsequently, the canonical object point cloud model from the reference images and the local model from the test image are constructed based on the segments. ii) The *region proposal module* outputs a robust region of interest (ROI) of the target object against occlusion by incorporating visual and geometric features. A coarse 6D pose is then estimated by comparing the cropped test and reference images using Gen6D (Liu et al., 2022) and iii) further fine-tuned using ICP (Rusinkiewicz and Levoy, 2001). . . 68

5.3 **Online self-adaptation module**. A pretrained segmentor φ is first applied on reference images to predict segmentations. Meanwhile, an adaptive segmentor $\varphi*$ is initialized from φ. With the ground-truth translation of the target object in the reference images $\mathrm{T}_{\mathrm{ref}}$, the object center can be reprojected to the image. For each reference image, one segment is chosen as a positive sample if it includes the reprojected object center while the remaining segments are considered as negative samples. Subsequently, an object-level representation of each segment is computed by averaging the pixel-wise dense features from $\varphi*$. A contrastive loss is then applied over the positive and negative object representations and updates $\varphi*$ iteratively. After adaptation, $\varphi*$ generates the target object representation $r*$ by averaging over all positive representations from reference images. Given a test image, we obtain the representation of each candidate segment in the same way and compute the cosine similarity between each candidate and $r*$, where the most similar candidate is chosen as the segment of the target object. Meanwhile, the canonical and local object models are computed based on the segments and depth images. . . 71
5.4 **Qualitative results.** The green bounding box denotes the ground-truth pose and blue denotes the prediction. In SA6D, blue denotes prediction before refinement while red is the final prediction. . . . 72
5.5 Analysis of the number of (a) reference and (b) online iterations. (c) An example of proposed ROI from SA6D (red) and Gen6D (blue), the red cross denotes the target object. 78
5.6 Examples of using RPM-Net for point cloud registration instead of ICP. The yellow point cloud denotes the reconstructed object point cloud model and the blue one denotes the prediction after transformation using the predicted pose from RPM-Net. Better overlapping between two point clouds indicates better performance. RPM-Net cannot generalize on unseen objects and is prone to get stuck in local optima. 80
5.7 Discussion. (a) A false positive sample is selected given the reprojected center of the target object (*milk cow*) is occluded by another object (*yellow rabbit*). Nevertheless, (b) SA6D provides robust prediction with explainable confidence scores. 81
5.8 Online-Adaptation results on challenging scenes against severe occlusion and truncation. Three candidates with the highest confidence scores are visualized in order. 82

5.9 Robust prediction of target segmentation on LineMOD. Three candidates with the highest scores are visualized in order. 83
5.10 Robust prediction of target segmentation on LineMOD-OCC. Three candidates with the highest scores are visualized in order. . 84
5.11 Robust prediction of target segmentation on HomebrewedDB. Three candidates with the highest scores are visualized in order. . 85
5.12 Failure cases. Using ICP in the refinement module leads to a worse prediction than the initial prediction. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP. 86
5.13 Prediction on LineMOD dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP. 88
5.14 Prediction on LineMOD-OCC dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP. 89
5.15 Prediction on HomebrewedDB dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP. 90
5.16 Prediction on FewSOL dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP. 91
5.17 Prediction on Wild6D dataset with 20 reference images. The red bounding box is the ground-truth pose and the green bounding box denotes the prediction. 92

6.1 The network architecture. Instead randomizing slot initialization from a common distribution widely used in prior work, we initialize slot representations conditioned on the input features. A clustering algorithm and a mapping layer are adopted. 96
6.2 The framework architecture for slot initialization for slot attention. The top row is the original architecture. 103
6.3 The framework architecture for IODINE based extensions. The original starts directly at iteration 1 with slots drawn out of the standard gaussian distribution with $(\mu, \sigma) = (0, 1)$. 104
6.4 The Direct mapping approach. Slots are identical to the cluster centers chosen by the clusterization algorithm. 104
6.5 The Small MLPs mapping approach. It extends the direct mapping approach by a nonlinear network between cluster centers and slots. 104
6.6 The Large MLPs approach. 105
6.7 The permutation invariant Pseudoweights mapping. 106
6.8 The permutation invariant mapping between 6 cluster centers and 3 slots. For this example all slots and cluster centers are of dimension D=1, to keep it simple. The pseudoweights tensor has high values in black squares and low values in white squares. If the blue and yellow slot change their position, the slots won't change their initialization. 107
6.9 Qualitative results on MDS dataset. 109
6.10 The slot-wise predicted masks and reconstructed scenes on MDS dataset. 110
6.11 The original Slot Attention model struggles with overlapped objects. 110
6.12 The slot-wise predicted masks and reconstructed scenes on CLEVR6 dataset. 111
6.13 Qualitative results on Chairs dataset. 111
6.14 The slot-wise predicted masks and reconstructed scenes on Chairs dataset. 112
6.15 Qualitative comparison of generalization on CLEVR10 while the models are trained with CLEVR6. 116
6.16 Another qualitative comparison of generalization on CLEVR10. . 117
6.17 Qualitative results on the object discovery task. Notably, the *mean-shift* (MS) versions can recover detailed appearance over all datasets with even better quality than original input for IODINE-based models in MDS dataset. 118

6.18 Qualitative results of slot-wise reconstructed scenes (left) and masks (right). *Mean-shift* models disentangle the objects better than others and recover more details. 118
6.19 Qualitative results on increasing objects. The models are trained on CLEVR6 but evaluated on CLEVR10 with larger number of objects. 119
6.20 Failure cases on Chairs dataset where *k-means* and *Psuedoweights* (PW) cannot disentangle the objects and use each individual slot for specific areas. 119
6.21 Qualitative results on novel view synthesis. Our models can represent the chairs with more details than the original uORF. 120

7.1 (a) Existing datasets such as Jacquard dataset (Depierre et al., 2018) exhibit a variety of textures and geometries. In contrast, (b) our research centers on heterogeneous physical properties (mass distribution and friction coefficients) across the object. For instance, the red part denotes higher mass density and friction while yellow denotes lower mass density. (c) We further evaluate our method in a sim-to-real scenario. 122
7.2 We generate two types of datasets. Each shape is considered as one category with numerous instances incorporating varying combinations of mass distribution, friction coefficient, and size. Different colors represent distinct physical properties for visualization. . . . 126
7.3 The dataset is split into context and target sets for each object. The context set includes the depth images x w.r.t. the grasp candidates, the distance z between the grasp candidates and the gripper, and the binary grasp labels indicating if the grasp succeeds. In contrast, the target set lacks labels during the inference phase. The data is split randomly between context and target sets for each training iteration. 128
7.4 The structure of ConDex. 129
7.5 pipeline 129
7.6 **Statistics of *Letters* dataset.** Each curve represents one category comprising numerous instances. The collected data exhibit a normal distribution centered around a positive rate of approximately 50%. 131
7.7 Results are evaluated on *Letters* dataset over 50 objects from intra- and cross-categories, each object is grasped 30 times. The dashed line denotes the performance with random grasping. 133

7.8 Evaluation on *Bottles* dataset from intra- and cross-categories (i.e., object *2* and *6*). The dashed line denotes the performance with random grasping. 133
7.9 **Error rate vs. context number on cross-category.** 450 object instances are evaluated from two previously unseen categories from *Letters* dataset. Results are presented with a maximum of 20 context points during evaluation, while a maximum of 15 context points is provided during training. The dashed line denotes the performance of DexNet. 134
7.10 Experiments on a real robot involves evaluating the success of manipulation based on the criteria that a successful grasp results in the bottle being successfully placed inside the designated box. The bottle is filled with different quantities of material to acquire diverse mass distributions. 135
7.11 **Sim-to-real evaluation.** The results stem from experiments conducted on a real robot, with the models being entirely trained using the *Bottles* dataset obtained through rollouts within the Mujoco simulation. 135

List of Tables

3.1 Prediction error (pixel) on euclidean distance in the 2D image plane for Distractor. Different aggregation methods and augmentations are employed. The first row shows results for intra-category (IC) evaluations, the second row for cross-category (CC). 27
3.2 Pascal1D pose estimation error. MSE and standard deviations are calculated with 5 random seeds. 27
3.3 ShapeNet1D pose estimation error(°). Results are calculated with 5 random seeds except for MAML. The first row presents results for IC and the second row for CC. 28
3.4 Comparison of different augmentation techniques on ShapeNet2D. Results are calculated with 3 random seeds using CNP (CA) as baseline. 30
3.5 Performance on ShapeNet1D using small (S), medium (M) and large (L) training dataset sizes for CNP with cross-attention (CA) and Max aggregation. The first row presents results for intra-category (IC) and the second row for cross-category (CC) evaluation. MSE and standard deviations are calculated with 5 random seeds. 30
3.6 Comparison of aggregation methods on ShapeNet2D using DR+DA+TA. Results are calculated with 3 random seeds. 32
3.7 Comparison of different data augmentation techniques on ShapeNet2D using CNP (CA) + DR as baseline. 33
3.8 Analysis of FCL + CNP on different choices of positive pairs using: i) the same context set with different augmentations (Same Ctx), ii) different context sets from the same task (Diff Ctx), iii) context and target sets (Ctx & Target). Prediction error (pixel) is calculated with 3 random seeds. 34
3.9 Results of the evaluation on ShapeNet1D using different temperature values in FCL. 34
3.10 Results of the evaluation on ShapeNet2D using different temperature values in FCL. 35

3.11 Analysis of latent task representation on Distractor between Max and Max_{FCL} using various clustering metrics. 35
3.12 Performance of MMAML (Vuorio et al., 2019) on ShapeNet1D. . 37

3 Evaluation results on LineMOD dataset. 54
4.2 Single category - car evaluation on Toy-MCMS dataset 55
4.3 Multi-category evaluation on PBR-MCMS dataset 57
4.4 Multi-category evaluation on Toy-MCMS dataset 58
4.5 Comparison between GAML and fine-tuned FFB6D on PBR-MCMS using ADD metric. 59
4.6 Multi-category evaluation on Occlusion-MCMS dataset 61
4.7 GAML network architecture. 62
4.8 ADD Results on PBR-MCMS using different number k of neighbors in GNN decoder. 62
4.9 ADD Results of CNP and ANP on Toy dataset. 63

5.1 Evaluation of ADD-0.1d on LineMOD, LineMOD-OCC, and HomeBrewedDB datasets against category-agnostic baselines. . . 75
5.2 Evaluation on FewSOL (Padalunkal et al., 2022) dataset over 336 objects using 8 reference images. 75
5.3 Evaluation on Wild6D (Fu and Wang, 2022) dataset against category-level baselines. 75
5.4 Evaluation on Gen6D with different object diameters as prior knowledge. Results are averaged over objects for each dataset. . . 79
5.5 Evaluation on LineMOD using LineMOD-OCC as reference. . . . 80

6.1 Quantitative results on the object discovery task. 112
6.2 Evaluation with different number of iterations (5 iterations are used for training). In particular, our models achieve significant improvement already at the first iteration. 113
6.3 Results of novel view synthesis on Chairs-diverse. 114

1 Introduction

1.1 Meta-Learning as Human Cognition

The rapid advancement of machine learning techniques has led to significant breakthroughs in various domains, from natural language processing to computer vision. However, traditional machine learning models are typically designed to learn a specific task from scratch, requiring vast amounts of data and computational resources. This approach often fails to generalize well to new, unseen tasks, posing a significant limitation in real-world applications where data may be expensive to obtain.

Humans are able to rapidly learn the fundamentals of new tasks within minutes of experience based on prior knowledge. For instance, humans can classify novel objects by capturing the distinguishable properties (e.g., textures, shapes and scales) from only a few examples where the discriminative ability is associated with prior knowledge. In many real-world applications, it is preferable to acquire relative small but related datasets rather than single but large dataset in order to learn versatile and transferable skills. For instance, the robot learning to push the drawer is also expected to be able to pull the drawer back. Furthermore, the robot should also adapt quickly to new but similar tasks such as pushing the door without extensive data. Meta-learning is proposed to learn relevant knowledge from various tasks and generalize to unseen tasks with only a few samples. Of the various meta-learning algorithms, MAML-based models (Finn et al., 2017; Nichol et al., 2018; Finn et al., 2019; Yoon et al., 2018) and Neural Processes (NPs) (Garnelo et al., 2018a; Garnelo et al., 2018b; Kim et al., 2019; Gordon et al., 2020) are two variants which are receiving increasing attention in the recent years. Both algorithms try to learn good prior knowledge from related tasks without expanding the learned parameters or sacrificing efficiency at inference. While these methods have shown promising results in many domains, such as few-shot classification (Nichol et al., 2018; Sung et al., 2018; Snell et al., 2017; Ren et al., 2018; Vinyals et al., 2016b) and

hyperparameter optimization (Wei et al., 2021; Volpp et al., 2020; Franceschi et al., 2018), an extensive study on meta-learning vision regression tasks has not yet been conducted. This is in particular true for NPs which have mostly been investigated on tasks with low-dimensional input such as function regression or pixel-wise completion (Wang and Van Hoof, 2020; Norcliffe et al., 2021; Louizos et al., 2019a; Lee et al., 2020).

We make two major contributions to the largely unexplored area of meta-learning on high-dimensional input tasks. On the algorithmic level, inspired by SimCLR (Chen et al., 2020c), we propose an improvement to NPs by employing contrastive learning at the functional space (FCL) and still train the model in an end-to-end fashion. On the experimental side, we propose two application datasets, object discovery and pose estimation, which are based on high-dimensional inputs and require the meta-learning models to learn and reason at an image level. These two applications serve us as a testbed where we investigate different algorithmic choices as well as data and task augmentation techniques for meta-learning. For both applications, we evaluate the performance on novel objects at both **intra-category** (IC) and **cross-category** (CC) levels. The results on Distractor show that our proposed algorithmic improvements significantly increase the performance, indicating our methods can enhance the task expressivity. The results on pose estimation demonstrate that meta-learning can successfully be applied to predict poses of unknown objects, which has a huge potential in robotic grasping and virtual/augmented reality (VR/AR).

Recently, there has been effort in using meta-learning for object pose estimation, e.g., on the Pascal1D dataset (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021; Ni et al., 2021). However, Pascal1D contains only a limited number of different objects and the output space is limited to 1D rotation in azimuth axis. Furthermore, the mean-square-error (MSE) loss function, which is employed by previous work, is inappropriate for training, since it does not consider the similarity of coterminal angles. In order to further understand the performance on pose estimation, we generate new datasets with increasing task diversities, e.g., random background, cross-categorical object variations and 2D rotation. Since the background is generated from real-world images instead of being left blank as in prior work, our datasets significantly increase the task difficulty and narrow the gap between toy and real world applications. The results demonstrate that meta-learning can successfully be applied to predict poses of unknown objects, which has a huge potential in robot grasping, virtual reality (VR) and augmented reality (AR).

Prior work (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021; Ni et al., 2021) on Pascal1D also demonstrates that meta-learning algorithms suffer from overfitting, especially with limited training data. Our work analyzes the effect of different techniques commonly adopted in recent meta-learning methods (i.e., data augmentation, task augmentation, regularization and domain randomization) on aforementioned datasets. We empirically find that the meta-learning algorithms employed in our work ultimately lead to overfitting regardless of dataset size for both applications. Moreover, our work shows that the results in prior work (Rajendran et al., 2020; Yin et al., 2020), where MAML typically performs best for such tasks, are misleading. In particular, we find Conditional Neural Processes (CNPs) (Garnelo et al., 2018a) are more flexible and efficient than MAML in the investigated pose regression tasks. Additionally, we find that MAML (Finn et al., 2017) suffers from underfitting especially on large-scale datasets and depends heavily on hyperparameter tuning.

1.2 Object and Scene Representation

Object and scene representation is a fundamental area of research in computer vision and artificial intelligence, focusing on how visual information is captured, processed, and understood by machines. Effective representation of objects and scenes is vital for various applications, including autonomous driving, robotic manipulation, augmented reality, and image recognition. Accurate representation allows systems to perform tasks such as identifying and localizing objects, understanding spatial relationships, and recognizing complex scenes with high precision. This capability is essential for developing intelligent systems that can operate in dynamic and unstructured environments.

Object and scene representation pose several challenges:

- Variability in Appearance: Objects can appear in numerous forms, sizes, and poses, and they can be viewed from different angles under varying lighting conditions. Capturing this variability requires robust and adaptable representation models.
- Occlusion and Clutter: In real-world scenarios, objects are often partially occluded by other objects or present in cluttered environments.

Effective representation must account for such occlusions and still accurately identify and localize objects.

- Hierarchical Structure: Scenes often contain multiple objects with complex hierarchical relationships. Understanding these relationships is critical for accurate scene interpretation.
- Real-Time Processing: Many applications, such as autonomous driving and robotic manipulation, require real-time processing of visual information. Efficient representation models must balance accuracy with computational efficiency.

Recent advances in deep learning have significantly improved object and scene representation capabilities. Methods such as Faster R-CNN (Ren et al., 2015), YOLO (Redmon et al., 2016), and SSD (Liu et al., 2016) have revolutionized object detection and recognition by achieving remarkable accuracy and speed. These models utilize deep learning techniques to identify and localize objects within images with high precision. A newer approach involves using slot-based representations to encode objects in a scene. Beyond 2D images, representing objects and scenes in 3D provides a more comprehensive understanding of the environment. Techniques such as 3D point clouds (Qi et al., 2017a) and voxel-based representations (Wu et al., 2015) enable the capture of depth information, allowing for more accurate modeling of spatial relationships and object geometries. Moreover, advancements like Gen6D (Liu et al., 2022) offer generalizable 6D object pose estimation, providing robust and accurate 3D object representations in cluttered and dynamic environments.

Accurately estimating the 6D pose of novel objects is critical for robotic grasping, especially for the tabletop setup. Prior work has investigated instance-level 6D pose estimation (Wang et al., 2019a; Peng et al., 2019; He et al., 2020; He et al., 2021), where the objects are predefined. Although achieving satisfying performance, these methods are prone to overfit to specific objects and suffer from poor generalization. Due to the high variety of objects with different colors and shapes in the real-world, it is impractical to retrain the model every time new objects come in, which is time-consuming and data inefficient.

Recently, several approaches (Wang et al., 2019b; Chen et al., 2020b; Wang et al., 2020a; Chen et al., 2021b; Chen et al., 2020d; Chen and Dou, 2021; Fu and Wang, 2022; Zhang et al., 2022) have been proposed for category-level 6D pose estimation instead of specific objects where they map different

instances of each category into a unified representational space based on RGB or RGB-D features. However, the assumption of a unified categorical space potentially leads to a decrease in performance in case of strong object variations. However, conditioning on specific object categories limits the generalization to objects from novel categories with strong object variations. Meanwhile, some approaches (Li et al., 2022; Liu et al., 2022; Gao et al., 2022b; Park et al., 2020; He et al., 2022b; Shugurov et al., 2022) investigate generalizable 6D pose estimation as a few-shot learning problem, i.e., predicting the 6D pose of novel and category-agnostic objects given a few labeled reference images with the known pose of the novel object to define the object canonical coordinates. Although achieving promising results, these methods so far only work well on non-occluded and object-centric images, i.e., without the interference of other objects. This limits the generalization to real-world scenarios with multiple objects in cluttered and occluded scenes. Furthermore, additional object information is required such as object diameter (Liu et al., 2022), mesh model (Shugurov et al., 2022; Li et al., 2022), object 2D bounding box (He et al., 2022b) or ground-truth mask (Park et al., 2020; Lin et al., 2021b), which is not always available for novel object categories. Our method aims to enable a fully generalizable few-shot 6D object pose estimation (FSPE) model.

In summary, we identify the primary challenges that are not adequately addressed by the current state-of-the-art methods (Liu et al., 2022; Park et al., 2020; He et al., 2022b; Shugurov et al., 2022) as follows:

- The category-agnostic 6D pose estimation in cluttered scenes with heavy occlusions is performing poorly.
- The object-centric reference images from cluttered scenes are cropped by ground-truth segmentation or bounding box of the target object, which limits the generalization in real-world scenarios.
- The prediction should not require any form of prior information from the unseen object such as diameter, mask or object mesh model.
- The requirement of extensive reference images covering all different view-points is not practical.

Object-centric representations using slots have shown good performance in object detection (Li et al., 2021a; Locatello et al., 2020), segmentation (Kabra et al., 2021; Greff et al., 2019) and tracking (Wu et al., 2021; Kipf et al., 2022; Li et al., 2020b) tasks. Slot Attention mechanisms (Locatello et al., 2020) dynamically allocate and update slots to represent different objects,

enabling efficient and interpretable scene decomposition. Slots are a set of latent variables. The core idea behind slot representation is to allocate distinct "slots" to represent different objects or parts of a scene. Each slot is designed to capture specific features or attributes of an object, enabling the model to disentangle and isolate various components of the visual input. This modular approach contrasts with traditional dense representations, where all information is entangled in a single high-dimensional vector. The common approach is to frame disentangled and structured slot representations of the compositional scene with some iterative refinement mechanisms in a self-supervised manner, e.g., using softmax-based attention (Locatello et al., 2020) or variational inference (Greff et al., 2019). The idea is to improve the sample efficiency and generalization of capturing the structured environment to unseen compositions or objects. However, Effectively scaling slot-based models to handle high-resolution images and complex scenes with numerous objects remains an open research problem. Ensuring stable and efficient training of slot-based models can be challenging, particularly when dealing with diverse and complex datasets. Moreover, the number of slots needs to be specified beforehand on each dataset, which limits the generalization and flexibility. In addition, a random slot initialization from a common distribution is widely used in prior works, which lacks consideration between the slots and the perceptual input. Consequently, the quality of the following iterative slot refinement is also affected by the sub-optimal initialization.

1.3 Robotic Application

One of the most crucial advantages of meta-learning is its ability to enable robots to adapt rapidly to new and unseen tasks. Traditional machine learning models require large amounts of labeled data and extensive training to perform effectively. However, robots often operate in dynamic and unpredictable environments where they encounter new objects and tasks frequently. Meta-learning equips robots with the capability to generalize from previous experiences, allowing them to learn new tasks with few examples. This is particularly beneficial in applications such as robotic manipulation, where the variability of objects and tasks can be vast.Meta-learning improves the generalization capabilities of robotic systems, allowing them to perform well across a wide range of scenarios. By training on a diverse set of tasks, meta-learning algorithms develop a meta-knowledge that can be applied to new situations.

This is crucial for real-world robotic applications where the environments are often cluttered, objects can be occluded, and the physical properties of objects can vary significantly. Meta-learning enables robots to handle these variations more effectively, resulting in more robust performance.

Grasp detection is one of the fundamental problems in robotic manipulation, which has led to great progress in recent years thanks to the advancements of deep learning techniques. Grasp detection is normally formed as finding the stable grasp position w.r.t. the geometry of the object, the configuration of the end-effector, and the specific manipulation tasks (Newbury et al., 2022). Extensive studies have investigated various end-effector such as parallel jaw (Mahler et al., 2016; Mahler et al., 2017a), suction gripper (Eppner et al., 2016; Mahler et al., 2017b) and multi-finger gripper (Mayer et al., 2022) using RGB-D (Jiang et al., 2011; Lenz et al., 2013) or depth (Morrison et al., 2019) images from synthetic (Schaub and Schöttl, 2020) or real-scene (Song et al., 2019; Fang et al., 2020; Levine et al., 2016) datasets, with the purpose of increasing the generalization on unseen objects (Kalashnikov et al., 2018), closing the sim-to-real gap (Kleeberger et al., 2020; Quillen et al., 2018), and increasing the robustness against occlusion (Breyer et al., 2021). Recently, Huang et al. (2022), Farias et al. (2022), and Rho et al. (2021) further improve the performance on grasping deformable objects.

Although these methods have achieved promising results on various open datasets, it is noteworthy that these datasets predominantly adhere to a homogeneous assumption of the physical properties. The 3D objects (Chang et al., 2015; Kasper et al., 2012; Depierre et al., 2018) in simulation and 3D printed objects (Morrison et al., 2020) are typically treated as entire entities, neglecting the consideration of diverse material properties and friction coefficients for individual components, while most real-world datasets (Çalli et al., 2015; Singh et al., 2014; Lenz et al., 2013; Choi et al., 2008) frequently exhibit a variety of textures and geometries but tend to feature uniform distribution of mass. Crucially, neither the training nor evaluation stages explicitly incorporate physical properties. Thus, most vision-based grasp detection algorithms rely solely on geometries and textures. This limitation becomes evident in practical scenarios, exemplified by an instance where methods relying solely on object geometry and texture fail to effectively lift an object due to the oversight of variations in part density or friction coefficients.

Meta-learning plays a pivotal role in improving grasp detection, object manipulation, and task planning. For instance, grasp detection involves identifying

stable grasp positions for various objects, which can be highly variable in shape, size, and material properties. Meta-learning allows robots to generalize grasp strategies from previously encountered objects to new ones, improving their ability to handle diverse and novel objects. Additionally, integrating meta-learning with existing robotic frameworks, such as DexNet-2.0, enhances the robot's ability to infer physical properties from visual input, leading to more effective and reliable manipulation. In contrast to prior works, we explicitly construct various objects with distinct mass distributions and friction coefficients and employ a shared model to acquire knowledge of these properties purely from depth images, emphasizing the significance of discerning physical attributes solely from visual input. We frame this challenge as a few-shot learning problem, i.e., wherein the physical properties of each object must be gleaned from contextual information derived from a limited number of grasp trials. In essence, our method aims to emulate human learning principles by:

- Accumulatively acquiring knowledge of physical properties through previous experiences.
- Facilitating learning during both online and offline inference processes.
- Seamlessly integrating into existing grasp pipelines without compromising performance.
- Enhancing real-world performance while leveraging knowledge gained from simulation.

Conditional Neural Processes (CNP) (Garnelo et al., 2018a) has shown advances in few-shot classification and regression tasks (Gao et al., 2022b), characterized by rapid adaptation and inference capabilities. In our work, we integrate CNP into DexNet-2.0 (Mahler et al., 2017a) with minimal alterations to the original grasp frameworks and encode the object's physical properties as latent representations derived from contextual information. To address the limited availability of appropriate datasets, we create two synthetic datasets characterized by distinguishable physical properties in comparison to existing ones. Subsequently, we conduct performance evaluations using the Pybullet and Mujoco simulators, including novel objects from both intra-category (IC) and cross-category (CC). Furthermore, we extend our evaluation to real-world scenarios, where the model is exclusively trained in Mujoco, facilitating the investigation of the sim-to-real gap.

2 Background

2.1 Latent Variable Networks

Latent variable models (LVMs) have become a fundamental aspect of deep learning, providing powerful tools for uncovering hidden structures within data. These models (Kingma and Welling, 2014; Sohn et al., 2015; Higgins et al., 2017) assume that observed data is influenced by underlying, unobserved variables known as latent variables. By incorporating these latent variables, LVMs enable the modeling of complex, high-dimensional data, capturing intricate patterns and dependencies that are not directly observable.

Latent variables are inferred from observed data and represent hidden factors that affect the observations. In the context of deep learning, these latent variables are often used to learn representations that capture the essential features of the data, facilitating tasks such as dimensionality reduction, data generation, and clustering. Several latent variable models have been developed and adapted within deep learning frameworks, each serving different purposes and applications:

- Variational Autoencoders (VAEs): VAEs (Kingma and Welling, 2014) are a type of generative model that combines neural networks with probabilistic modeling. They encode data into a latent space and then decode it back to the original space, allowing for the generation of new data points.
- Generative Adversarial Networks (GANs): GANs (Goodfellow et al., 2014) consist of two neural networks, a generator and a discriminator, that compete against each other. The generator creates data samples, while the discriminator evaluates them. This adversarial process helps the generator learn to produce realistic data.

- Neural Processes (NPs): NPs (Garnelo et al., 2018a) are a family of models that generalize Gaussian Processes using neural networks to handle high-dimensional data. They model distributions over functions and are used for tasks like few-shot learning and uncertainty estimation.

Latent variable models in deep learning are applied across various domains, including generating text, translating languages, and performing sentiment analysis using VAEs and GANs. In computer vision, VAEs can generate and reconstruct images, enhance image resolution, and perform style transfer. In healthcare, LVMs is also capable of analyzing medical images, predicting disease progression, and personalizing treatment plans.

2.2 Meta-Learning

Meta-learning, often described as "learning to learn" is an advanced machine learning approach where models are trained to adapt quickly to new tasks with minimal data. This concept is particularly valuable in deep learning, where traditional models typically require large amounts of data and extensive training to perform effectively on new tasks. Meta-learning aims to create more flexible and efficient models capable of generalizing across a wide variety of tasks.

Meta-learning operates on two levels:

- Meta-Level Learning: This involves learning the best strategies for quick adaptation across multiple tasks. The meta-learner acquires knowledge from a diverse set of tasks to optimize its learning process.
- Task-Level Learning: At this level, the model applies the strategies learned at the meta-level to adapt quickly to a specific new task with only a few examples.

Two prominent approaches in meta-learning within deep learning are Model-Agnostic Meta-Learning (MAML) (Finn et al., 2017) and Neural Processes (NPs) (Garnelo et al., 2018a; Garnelo et al., 2018b; Kim et al., 2019; Gordon et al., 2020).Model-Agnostic Meta-Learning (MAML) seeks to find an optimal set of initial parameters that can be quickly fine-tuned for new tasks. The method involves a two-stage optimization process:

- Outer Loop: The meta-learner updates the initial parameters by evaluating performance across a variety of tasks.
- Inner Loop: For each specific task, the model rapidly fine-tunes these parameters using a small dataset, allowing for quick adaptation.

Neural Processes (NPs): Garnelo et al. (2018a) proposed Neural Processes as a combination of neural networks and Gaussian Processes. NPs model distributions over functions, conditioned on a set of context points. Variants like Convolutional Conditional Neural Processes (CCNPs) (Gordon et al., 2020) and Attentive Neural Processes (ANPs) (Kim et al., 2019) enhance this approach by integrating convolutional kernels and attention mechanisms to improve the model's ability to handle uncertainty and variability in data.

Meta-learning has significant implications in various domains such as computer vision (Chen et al., 2020c; Raghu et al., 2019), natural language processing (NLP) (Bansal et al., 2020; Dou et al., 2019), and robotics (Wang and He, 2019; Clavera et al., 2018).

2.3 Few-Shot Learning

Few-shot learning is a subfield of machine learning that focuses on the challenge of training models to perform well with limited data. Traditional machine learning models typically require large amounts of labeled data to achieve high performance. However, in many real-world scenarios, such as medical diagnosis, robotics, and natural language processing, collecting and labeling large datasets can be impractical, expensive, or time-consuming.

Few-shot learning aims to overcome these limitations by enabling models to generalize from a small number of examples. This is achieved by learning a meta-knowledge or a meta-model that captures common patterns and representations across various tasks. The goal is to make the model capable of rapid adaptation with minimal data for the new task. Metric learning approaches, such as Siamese Networks (Koch et al., 2015), Prototypical Networks (Snell et al., 2017), and Matching Networks (Vinyals et al., 2016a) learn a similarity measure between data points. These models embed the data into a feature space where similar examples are closer together, enabling classification by comparing distances to prototype representations. Few-shot learning has also been significantly advanced through the application of transfer learning and

meta-learning techniques, enabling models to leverage knowledge from related tasks to improve performance on new, unseen tasks with limited data. These methods have been successfully applied in various domains, including computer vision, natural language processing, and reinforcement learning, demonstrating their broad applicability and potential.

It is good to note that, both meta-learning and few-shot learning aim to enable models to quickly adapt to new tasks with minimal data. They are designed to overcome the limitations of traditional machine learning models that require large amounts of labeled data. Meta-learning is a broader concept that encompasses various techniques aimed at improving the learning process itself while few-shot learning is a specific problem setting within meta-learning that focuses on the ability to learn new tasks from a small number of examples. It is a subset of meta-learning that emphasizes rapid learning with minimal data. Furthermore, meta-learning can be applied to a variety of architectures depending on the meta-learning algorithm. For instance, MAML (Finn et al., 2017) can be used with any model that uses gradient-based optimization, while other meta-learning approaches might use specialized architectures like Neural Processes (Garnelo et al., 2018a; Garnelo et al., 2018b; Gordon et al., 2020; Kim et al., 2019). In contrast, few-shot learning often employs specific architectures designed for effective learning from few examples, such as Siamese Networks (Koch et al., 2015), Prototypical Networks (Snell et al., 2017), and Matching Networks (Vinyals et al., 2016b), which are tailored to measure similarities between examples. In this thesis, meta-learning is often combined with few-shot learning.

2.4 Notation

We now briefly describe both MAML and NPs in a unified way. We assume that all tasks are sampled from the same distribution $p(\mathcal{T})$, each task $\mathcal{T}_i$ includes a context set $\mathcal{D}_C^i = \{(x_{C,1}, y_{C,1}), ..., (x_{C,K}, y_{C,K})\}_i$ and a target set $\mathcal{D}_T^i = \{(x_{T,1}, y_{T,1}), ..., (x_{T,M}, y_{T,M})\}_i$ where K and M are the number of samples in each set which could be different for each task. The entire training dataset is denoted as $\mathcal{D} = \{\mathcal{D}_C^i, \mathcal{D}_T^i\}_{i=1}^N$ where N is the number of tasks sampled for training. During inference, the model is tested on a new task $\mathcal{T}^* \sim p(\mathcal{T})$ given a small context set, from which it has to infer a new function $f^* : (\mathcal{D}_C^*, x_T^*) \rightarrow \hat{y}_T^*$. In meta-learning, there are two types of learned

parameters, the first is the meta-parameters θ, which are learned during a meta-training phase using $\mathcal{D}$. The second is task-specific parameters ϕ^* which are updated based on samples from a new task $\mathcal{D}_C^*$ conditioned on the learned meta-parameters θ. Predictions can be constructed as $\hat{y}_T^* = f_{\theta,\phi^*}(x_T^*)$, where f is the meta-model parameterized by θ and ϕ^*. Both MAML and NPs form task-specific ϕ^* as $p(\phi^*|D_c^*, \theta)$ which conditioned on meta-parameters θ and context set sampled from task.

MAML considers both θ and ϕ^* as weights of neural networks and ϕ is updated by gradient optimization on new task, while CNP considers only θ as neural weights. Different from MAML, which updates ϕ^* by gradient optimization on the new task samples, CNP takes ϕ^* as task representation and predicts it from the context set as $\phi^* = \bigoplus_{i=1}^{K} h_\theta(x_{C,i}^*, y_{C,i}^*)$. Here $\bigoplus$ is a permutation invariant operator since task representation should not be affected by context orders , h is an encoder parameterized by θ. Subsequently, a decoder g_θ will take ϕ^* as an additional input and output $\hat{y}_T^* = g_\theta(x_T^*, \phi^*)$. Note that meta-parameters θ are fixed after meta-training phase, therefore CNPs don't not require any fine-tuning as MAML.

3 What Matters for Meta-Learning Vision Regression Tasks

3.1 Introduction

Humans are able to rapidly learn the fundamentals of new tasks within minutes of experience based on prior knowledge. For instance, humans can classify novel objects by capturing the distinguishable properties (e.g., textures, shapes and scales) from only a few examples. In many real-world applications, it is preferable to acquire relative small but related datasets rather than single but large dataset in order to learn versatile and transferable skills. For instance, the robot learning to push the drawer is also expected to be able to pull the drawer back. Furthermore, the robot should also adapt quickly to new but similar tasks such as pushing the door without extensive data. Meta-learning is proposed to learn relevant knowledge from various tasks and generalize to unseen tasks with only a few samples. Of the various meta-learning algorithms, MAML-based models (Finn et al., 2017; Nichol et al., 2018; Finn et al., 2019; Yoon et al., 2018) and Neural Processes (NPs) (Garnelo et al., 2018a; Garnelo et al., 2018b; Kim et al., 2019; Gordon et al., 2020) are two variants which are receiving increasing attention in the recent years. Both algorithms try to learn good prior knowledge from related tasks without expanding the learned parameters or sacrificing efficiency at inference. While these methods have shown promising results in many domains, such as few-shot classification (Nichol et al., 2018; Sung et al., 2018; Snell et al., 2017; Ren et al., 2018; Vinyals et al., 2016b) and hyperparameter optimization (Wei et al., 2021; Volpp et al., 2020; Franceschi et al., 2018), an extensive study on meta-learning vision regression tasks has not yet been conducted. This is in particular true for NPs which have mostly been investigated on tasks with low-dimensional input such as function regression or pixel-wise completion (Wang and Van Hoof, 2020; Norcliffe et al., 2021; Louizos et al., 2019a; Lee et al., 2020).

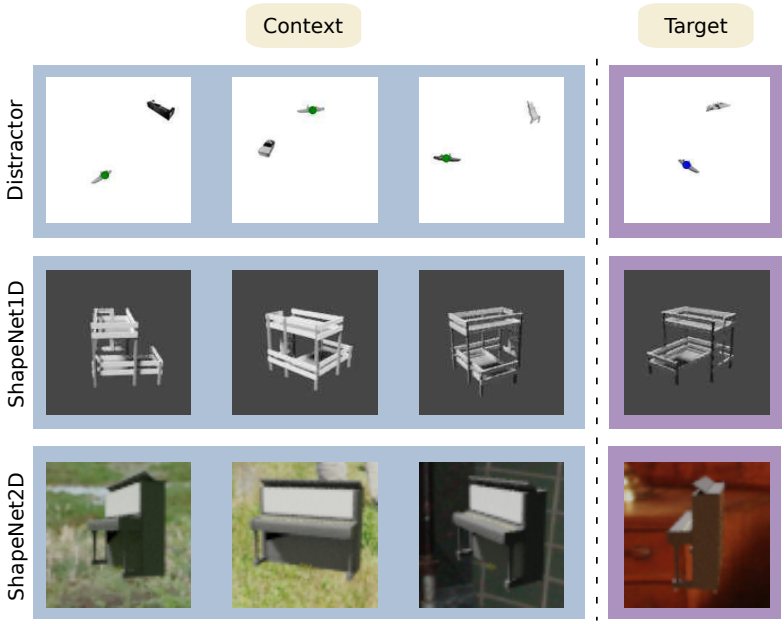


Figure 3.1: Meta-learning vision regression tasks are designed to i) identify the queried object from context and predict its position for target images (Distractor), ii) identify the object's canonical pose from context and predict the 1D rotation relative to the canonical pose for target images (ShapeNet1D), iii) predict the 2D rotation w.r.t. the canonical pose with random background (ShapeNet2D). Predictions are performed on unseen objects.

In this paper, we make two major contributions to the largely unexplored area of meta-learning on high-dimensional input tasks. On the algorithmic level, inspired by SimCLR (Chen et al., 2020c), we propose an improvement to NPs by employing contrastive learning at the functional space (FCL) and still train the model in an end-to-end fashion. On the experimental side, we propose two application datasets, object discovery and pose estimation, which are based on high-dimensional inputs and require the meta-learning models to learn and reason at an image level. Afterwards, we investigate different algorithmic choices as well as data and task augmentation techniques for meta-learning on these two applications. In order to evaluate their performance on high-dimensional inputs, we consider two applications which require the model to learn and reason at an image level, namely object discovery and pose estimation.

For the first application we create a regression task called “Distractor" (see Fig. 3.1), where each image contains two objects, the queried object and a distractor object, placed at random positions. The goal of this task is to identify the queried object and predict its position in the image plane. Unlike previous tasks such as image completion, where each pixel is considered as an independent input, our task requires the model to learn a high-level representation from the entire image. The second application (i.e., pose estimation) is inspired by prior work (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021; Ni et al., 2021) on the Pascal1D dataset. As this dataset shows limited object variations and features only 1D rotation around the azimuth axis, we generate two new datasets with increasing task diversity, e.g., by introducing random background, cross-categorical object variations and 2D rotation. Since the background is generated from real-world images instead of blank as in prior work, our datasets significantly increase the task difficulty and allow us to perform a thorough investigation of the performance for the considered meta-learning approaches. Examples of our datasets are shown in Fig. 3.1 where i) ShapeNet1D contains 1D rotations as in Pascal1D, however with larger object variations and ii) ShapeNet2D features 2D rotation and random background.

For both applications, we evaluate the performance on novel objects at both **intra-category** (IC) and **cross-category** (CC) levels. The results on Distractor show that our proposed algorithmic improvements significantly increase the performance, indicating our methods can enhance the task expressivity. The results on pose estimation demonstrate that meta-learning can successfully be applied to predict poses of unknown objects, which has a huge potential in robotic grasping and virtual/augmented reality (VR/AR). In this work, we mainly investigate Conditional Neural Processes (CNP) (Garnelo et al., 2018a) which is a deterministic variant of NPs, since we empirically find using stochastic distribution over latent space doesn’t benefit the performance on our tasks. Inspired by SimCLR (Chen et al., 2020c), we propose functional contrastive learning (FCL) on top of CNP using context and target set, which trains in an end-to-end fashion without sacrificing computational efficiency.

Recently, there has been effort in using meta-learning for object pose estimation, e.g., on the Pascal1D dataset (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021; Ni et al., 2021). However, Pascal1D contains only a limited number of different objects and the output space is limited to 1D rotation in azimuth axis. Furthermore, the mean-square-error (MSE) loss function, which is employed by previous work, is inappropriate for training, since it does not consider the similarity of coterminal angles. In order to further understand the

performance on pose estimation, we generate new datasets with increasing task diversities, e.g., random background, cross-categorical object variations and 2D rotation. Since the background is generated from real-world images instead of being left blank as in prior work, our datasets significantly increase the task difficulty and narrow the gap between toy and real world applications. The results demonstrate that meta-learning can successfully be applied to predict poses of unknown objects, which has a huge potential in robot grasping, virtual reality (VR) and augmented reality (AR). Examples of our dataset are shown in Fig. 3.1 where i) ShapeNet1D contains 1D rotation like in Pascal1D, however with larger object variations and ii) ShapeNet2D features 2D rotation and random background. Inference is conducted on novel objects from both intra- and cross-category levels.

Prior work (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021; Ni et al., 2021) on Pascal1D also demonstrates that meta-learning algorithms suffer from overfitting, especially with limited training data. Our work analyzes the effect of different techniques commonly adopted in recent meta-learning methods (i.e., data augmentation, task augmentation, regularization and domain randomization) on aforementioned datasets. We empirically find that the meta-learning algorithms employed in our work ultimately lead to overfitting regardless of dataset size for both applications. Moreover, our work shows that the results in prior work (Rajendran et al., 2020; Yin et al., 2020), where MAML typically performs best for such tasks, are misleading. In particular, we find Conditional Neural Processes (CNPs) (Garnelo et al., 2018a) are more flexible and efficient than MAML in the investigated pose regression tasks. Additionally, we find that MAML (Finn et al., 2017) suffers from underfitting especially on large-scale datasets and depends heavily on hyperparameter tuning. Moreover, we find that using cross-attention (CA) can alleviate the problem of overfitting on pose estimation but not on the Distractor task, which indicates that cross-attention is more effective on object-centric inputs. More results are shown in Section 3.4. without sacrificing computational time using fast attention module from Performer (Choromanski et al., 2021).

The primary contributions of this work can be summarized as follows: (1) We investigate meta-learning algorithms on vision regression tasks and demonstrate their ability to tackle structured problems. (2) We propose functional contrastive learning on the task representation of CNPs and thereby improve its expressivity. (3) We quantitatively analyze various deep learning techniques to alleviate meta overfitting. Our results rectify misleading conceptions from prior work, e.g., that MAML performs best for such tasks. We also present

insights and practical recommendations on designing and implementing meta-learning algorithms on vision regression tasks, and hope our work facilitates future work on meta-learning vision regression tasks.

3.2 Related Work

Meta-Learning. In meta-learning, also known as *learning to learn*, a learning agent gains meta knowledge from previous learning episodes or different domains and then uses this acquired knowledge to improve the learning on future tasks (Hospedales et al., 2021). MAML is an optimization-based meta-learning method and represents the meta knowledge as the model parameters, where learning good initial parameters can enable quick adaptation to new tasks with only few update steps on a small number of samples (Finn et al., 2017). Different from MAML, Neural Processes (NPs) constitute a class of neural latent variable models and interpret meta-learning as conditional few-shot function regression (Garnelo et al., 2018b). Similar to Gaussian Processes, NPs model distributions over functions conditioned on contexts (Garnelo et al., 2018b; Kim et al., 2019; Gordon et al., 2019). Meta-learning algorithms have been applied successfully in low-dimensional function regression (Garnelo et al., 2018a; Garnelo et al., 2018b; Kim et al., 2019; Wang and Van Hoof, 2020), image completion (Gordon et al., 2020; Louizos et al., 2019a; Lee et al., 2020), few-shot classification (Nichol et al., 2018; Sung et al., 2018; Snell et al., 2017; Ren et al., 2018; Vinyals et al., 2016b; Vuorio et al., 2019), reinforcement learning (Yoon et al., 2018; Rakelly et al., 2019; Yu et al., 2019; Gupta et al., 2018; Gondal et al., 2021), and neural architecture search (NAS) (Liu et al., 2019; Zoph and Le, 2017; Elsken et al., 2020; Lee et al., 2021a). Recent works (Finn et al., 2019; Rajendran et al., 2020; Yin et al., 2020; Yao et al., 2021) go one step further and apply meta-learning to pose estimation using gray-scale images. However, in these studies, the prediction is restricted to 1D rotation and the employed loss function is ill-posed as it does not take the periodicity of rotation into consideration. Moreover, FRCL (Gondal et al., 2021) proposes to improve meta-learning by adding contrastive representation learning from disjoint context sets. A follow-up work (Kallidromitis et al., 2021) further extends this idea to time series data by combining contrastive learning with ConvNP (Gordon et al., 2020). However, in contrast to these two methods which need to learn a representation in a self-supervised way and fine-tune on downstream tasks subsequently, we use functional contrastive learning (FCL)

between context and target sets and train in an end-to-end fashion. In our work, we investigate 2D pose regression on RGB-images with additional random background of real images, where the increasing task difficulty further presents how expressive the latent variable models can achieve. Furthermore, we propose a new paradigm of vision-based regression task called Distractor, which arises more interest in the applications of exploring meta-learning algorithms.

Meta Overfitting. It is well-known that meta-learning algorithms suffer from two notorious types of overfitting: i) **Memorization overfitting** occurs when the model only conditions on the input to predict the output instead of relying on the context set (Yin et al., 2020); ii) **Learner overfitting** happens when the prediction model and meta-learner overfit only to the training tasks and cannot generalize to novel tasks even though the prediction can condition on the context set (Rajendran et al., 2020). Recently, different methods have been proposed to mitigate those overfitting issues, e.g., adding a regularization term on weights to restrict the memorization (Yin et al., 2020). However, tuning a regularization term between underfitting and overfitting is challenging (Li et al., 2020a). Subsequently, a related work (Rajendran et al., 2020) applied task augmentation which helps both memorization and learner overfitting. Meanwhile, Yao et al. (2021) proposed MetaMix and Channel Shuffle to linearly combine features of context and target sets and replace channels with samples from different tasks. Furthermore, Ni et al. (2021) empirically showed that data augmentation can also alleviate meta overfitting. Moreover, they find that employing data augmentation on target set achieves better performance. However, extensive comparisons on how these methods perform individually or combinedly are missing. In this work, we separate these techniques into data augmentation (DA), task augmentation (TA), meta-regularization (MR) and domain randomization (DR), and quantitatively compare them in different combinations on the two aforementioned applications in order to arrive at a better understanding and consistent comparisons.

MAML aims to learn a good prior which could benefits all the sampled tasks and assumes the optimal parameters could be reached within few steps updates.

3.3 Study Design

3.3.1 Problem Setting

Considering regression tasks with high dimensional input using neural process models is not well explored, we consider two types of image-based regression tasks, namely object discovery and pose estimation. First, we propose a non-trivial object discovery task called Distractor, which is only used for evaluating CNP variants. In contrast to existing object detection tasks (Lin et al., 2014; Geiger et al., 2012; Everingham et al., 2010; Russakovsky et al., 2015) that are designed to specify all object instances from an input image, our task aims to i) distinguish the queried object from other distractors and additionally ii) predict its 2D location in the image plane. Therefore, it is essential to learn a distinctive embedding ϕ^* that can represent various queried objects given their associated context images $\{x^*_{C,i}\}_{i=1}^K$ and corresponding positions $\{y^*_{C,i}\}_{i=1}^K$. Note that the distractors are sampled randomly from all categories and in many cases their appearances closely resemble the queried object. Hence, it is expected that aggregating multiple context pairs helps extracting expressive information to disambiguate the tasks and thus improve the performance. Hence, aggregating multiple context pairs is expected to fuse more expressive information in order to disambiguate the tasks and improve the performance.

The second task, pose estimation, is evaluated on three datasets, namely Pascal1D, ShapeNet1D and ShapeNet2D with incremental difficulty, caused e.g., by extending inference to unseen cross-category objects, adding random backgrounds and extending 1D rotations to 2D rotations. Note that in this task, each object has a random canonical pose, which has to be learned from a context set D^*_C where $\{y^*_{C,i}\}_{i=1}^K$ are the ground-truth rotations of context images $\{x^*_{C,i}\}_{i=1}^K$.

We use these tasks for an exhaustive evaluation of meta-learning algorithms: i) We evaluate the performance of CNPs using different aggregation operators, i.e., mean (Garnelo et al., 2018a), max, bayesian aggregation (BA) (Volpp et al., 2021) and cross-attention (CA) (Kim et al., 2019). ii) We evaluate MAML on Pascal1D and ShapeNet1D following Rajendran et al. (2020) and Yin et al. (2020) and compare it with different CNP variants. iii) Furthermore, we investigate meta overfitting with respect to different choices, e.g., augmentations, regularization, aggregation operators and task properties. iv) Moreover, we combine functional contrastive learning (FCL) with CNPs and compare it

with original CNPs. Pascal1D is proposed and conducted by prior work (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021), including 65 objects in total where each object contains 100 images with random 1D rotation in azimuth. However, this dataset has limited task diversity and uses inappropriate objective function in prior work. Therefore, we create ShapeNet1D and ShapeNet2D datasets. ShapeNet1D is similar as Pascal1D but with larger task diversity and modified objective function while ShapeNet2D is generated with 2D rotation and using random background instead of static blank scene.

3.3.2 Datasets

Most current famous object detection tasks, for instance, MSCOCO (Lin et al., 2014), KITTI (Geiger et al., 2012), Pascal VOC (Everingham et al., 2010) and ILSVRC (Russakovsky et al., 2015), encourage methods to detect and classify all objects in the image. We create a object discovery task, in contrast, aims to identify and predict solely the properties of the queried object insteaf of others. More spacifically, each image contains two objects with random position. The model is essential to learn which object is queried from context set and further predict its 2D position in image plane. Note that this task is non-trivial since traditional object detector cannot differentiate the queried object. We generate **Distractor** that contains 12 object categories from ShapeNetCoreV2 (Chang et al., 2015), where each category includes 1000 randomly sampled objects. For each object we create 36 128×128 gray-scale images, containing two objects with random azimuth rotation and 2D position (see Fig. 3.1). The data generation is based on an extended version of a prior open-source pipeline (Gordon et al., 2019). We choose 10 categories for training, where we reserve 20% of the data for intra-category (IC) evaluation. The remaining 2 categories are only used for cross-category (CC) evaluation. The second dataset, **Pascal1D** (Yin et al., 2020), contains 65 objects from 10 categories. We randomly select 50 objects for training and the other 15 objects for testing. 100 128×128 gray-scale images are rendered for each object with a random rotation in azimuth angle normalized between $[0, 10]$. Since the performance is limited due to the size of the dataset, we generate a larger dataset, **ShapeNet1D** which includes 30 categories. 27 of these are used during training and IC evaluation, the other 3 categories are used for CC evaluation. For each training category, we randomly sample 50 objects for training and 10 for IC evaluation while CC evaluation is performed on 20 objects for each unseen category. To further increase the task difficulty, we create **ShapeNet2D** which includes 2D

rotations. We restrict the azimuth angles to the range $[0°, 180°]$ in order to reduce the effect of symmetric ambiguity while elevations are restricted to $[0°, 30°]$. Furthermore, we use RGB images and employ randomly sampled real-world images from SUN2012 (Xiao et al., 2010) as background instead of static background. Note that all tasks are evaluated on both intra- and cross-category level except Pascal1D.

3.3.3 Data Augmentation, Domain Randomization, Task Augmentation and Meta Regularization

Data Augmentation (DA). We use standard image augmentation tecnniques in our work, i.e., *Dropout* and *Affine* for all tasks, and an additional *CropAndPad* for all pose regression tasks. Furthermore, we employ *Contrast*, *Brightness* and *Blur* for ShapeNet2D.

Domain Randomization (DR). For ShapeNet2D, we additionally employ DR (Tobin et al., 2017) by regenerating background images for all training data after every 2k training iterations while the data used for evaluation remain the same.

Task Augmentation (TA). Task augmentation adds randomness to each task in order to encourage the meta-learner to learn non-trivial solutions instead of simply memorizing the training tasks. Following Rajendran et al. (2020), we sample random noise $\epsilon^{(t)}$ from a discrete set for each task and create new tasks by adding the noise to the regression targets: $D_C^{(t)} = \{x_{C,i}^{(t)},\ y_{C,i}^{(t)} + \epsilon^{(t)}\}_{i=1}^{K}$ and $D_T^{(t)} = \{x_{T,i}^{(t)},\ y_{T,i}^{(t)} + \epsilon^{(t)}\}_{i=1}^{M}$. Specifically, we sample 2D position noise from a discrete set $\epsilon \in \{0, 1, 2, ..., 16\}^2$ for Distractor. For Pascal1D, we use the same noise set $\{0., 0.25, 0.5, 0.75\}$ as proposed in Yin et al. (2020) and Rajendran et al. (2020) while $\{0., 0.125, 0.25, ..., 2\}$ for ShapeNet1D. In ShapeNet2D, we first only add random noise in the azimuth angle from the discrete set $\{-10°, -9°, ..., 20°\}$ and in a second step add additional elevation noise from the set $\{-5°, -4°, ..., 10°\}$ for further comparison.

Meta Regularization (MR). Following Yin et al. (Yin et al., 2020), we employ MR on the weights θ of the neural networks. Furthermore, we find that it is crucial to fine-tune the coefficient β which modulates the regularizer and task information stored in the meta-parameters θ. In our experiments, we use $\beta = 1e^{-4}$ for Pascal1D, $1e^{-7}$ for ShapeNet1D and ShapeNet2D.

3.3.4 Functional Contrastive Learning (FCL)

The representations learned by CNP are invariant under permutation of the elements within a given context set. This property is achieved by a permutation invariant aggregation mechanism, e.g., max aggregation. However, another desirable property of the representation is invariance across context sets of the same task. In particular, the representations of different context sets belonging to the same task should be close to each other in the embedding space, while representations of different tasks should be farther apart. To achieve this, we add an additional contrastive loss at the functional space and train the model in an end-to-end fashion. The contrastive cross-entropy loss is defined as follows (Chen et al., 2020c):

$$\mathcal{L}_{\mathrm{FCL}} = -\frac{2}{N}\sum_{t=1}^{N}\log\frac{\exp(\mathrm{sim}(\phi_C^{(t)}\cdot\phi_T^{(t)})/\tau)}{D(\phi_C^{(t)})D(\phi_T^{(t)})}, \tag{3.1}$$

where N denotes the number of tasks per batch. $(\phi_C^{(t)}, \phi_T^{(t)})$ denotes a positive pair of latent representations of a given task obtained from context and target set respectively. More specifically, the pairs are obtained via max aggregation $\phi_C^{(t)} = \max(r_{C,1}^{(t)}, \ldots, r_{C,K}^{(t)})$ and $\phi_T^{(t)} = \max(r_{T,1}^{(t)}, \ldots, r_{T,M}^{(t)})$, where K denotes the number of context pairs per task and M the number of target pairs per task. $\max$ returns the element-wise maximum over the latent variables $r_i = h_\theta(x_i, y_i)$ which are output by the encoder network h_θ for each context pair (x_i, y_i). τ is a temperature parameter, similar to SimCLR (Chen et al., 2020c), we consider τ as a hyperparameter which is crucial for learning good representations. $\mathrm{sim}(\cdot)$ is the cosine similarity and $D(\phi_i^t)$ sums the similarity of all positive and negative pairs for ϕ_i^t:

$$D(\phi_i^t) = \sum_{k=1}^{N}\sum_{j\in\{C,T\}}\mathbb{1}_{[\{k\neq t\}\vee\{j\neq i\}]}\exp\left(\frac{\mathrm{sim}(\phi_i^t\cdot\phi_j^k)}{\tau}\right), \tag{3.2}$$

where $\mathbb{1}_{[\{k\neq t\}\vee\{j\neq i\}]} \in \{0,1\}$ is an indicator evaluating to 1 only if the representations are sampled from different tasks or different sets. The log-value in Eq. (3.1) can be interpreted as the weighted importance of the positive pair. Therefore, this loss function encourages the model to obtain large similarity for positive pairs and small for negative pairs.

3.3.5 Objective Functions and Evaluation Metrics

Pascal1D. Following prior work (Yin et al., 2020; Rajendran et al., 2020; Yao et al., 2021) we conduct experiments using the MSE score between predicted and ground-truth azimuth rotation for both training and evaluation. However, this loss function does not take the ambiguity of coterminal angles into account. Hence, it can hamper the training process, e.g., predicting 359° for a ground-truth angle of 0° incurs a higher loss than predicting 180°. Nevertheless, we follow the same setup to obtain a fair comparison to prior works.
In addition, we take data augmentation into comparison which is also missing in prior works. Afterwards, we make an improvement on the loss function in ShapeNet1D. Moreover, we test two variants of CNP, the first one uses mean aggregation which is same as Yin et al. (2020) and Rajendran et al. (2020), the second one is with cross attention module which is similar as ANP (Kim et al., 2019) but only the deterministic path and without self-attention module. Here we denote the second variant as ANP for simplicity.

We find this task has some inappropriate design for both training and evaluation. First, meta-learning algorithms require large dataset during meta-training in order to have a good generalize ability. This dataset in contrast contains only 50 objects which is far from enough. Because of the limited training data, meta-learning methods tend to simply memorize the training objects instead of learning from context, which hinders generalization on novel objects during test. The results in table Table 3.2 also shows the mean prediction error is around 30 degree after denormalizing the value. Thus, it is unpersuasive to say meta-learning can help predict poses of novel objects even with meta augmentation techniques. Another more critical point we found during implementation is that,

ShapeNet1D. Instead of using MSE score, we use the “cosine-sine-loss" for training and a prediction error defined in terms of the angular degree for evaluation. The loss of a single sample is defined as:

$$\mathcal{L} = |\cos(y) - \cos(y^*)|^2 + |\sin(y) - \sin(y^*)|^2, \tag{3.3}$$

where N is the number of objects per minibatch, M is the number of test images per object, y^* is the ground-truth rotation and y the predicted rotation. The prediction error used for evaluation is defined as follows:

$$\mathcal{E} = \min\{\mathcal{E}_{y^+,y^*}, \mathcal{E}_{y^-,y^*}, \mathcal{E}_{y,y^*}\}, \tag{3.4}$$

where

$$\mathcal{E}_{y^{\pm},y^*} = |y \pm 360 - y^*|, \mathcal{E}_{y,y^*} = |y - y^*|. \tag{3.5}$$

ShapeNet2D. During training, the model need to implicitly encode all relevant information into the latent space apart from background. We represent the 2D rotation as quaternion in both training and evaluation. The loss of a single sample is accordingly defined as follows:

$$\mathcal{L} = \min\left\{\left|q^* - \frac{q}{\|q\|}\right|, \left|q^* - \frac{-q}{\|q\|}\right|\right\}, \tag{3.6}$$

where q^* denotes the ground-truth unit quaternion and q denotes the predicted quaternion. We empirically find that using this objective function achieves a better performance than constraining the scalar part of q to be positive. We hypothesize that enforcing the scalar constraint breaks the continuity of the rotation representation and therefore hampers training.

3.4 Experiments

In this section we present experimental results[1], perform a thorough analysis and provide insights and recommendations. Instead of presenting the results following the task sequence, we structure this section by different algorithmic choices and perform a systematic comparison over all tasks by raising different questions.

The experiments are organized as follows: Section 3.4.1 introduce the training details, Section 3.4.2 introduces approaches for mitigating meta-learning overfitting, presents the results on Distractor, gives an empirical and analytical comparison between MAML and CNP, analyzes the performance of CNP with respect of context size and sample efficiency, and lastly demonstrates ablation studies.

[1] Codes and data are available at https://github.com/boschresearch/what-matters-for-meta-learning

3.4.1 Training Details

For all tasks, we use 500k training iterations for CNPs and 70k for MAML. Furthermore, the best model on the intra- and cross-category dataset is saved during training. This leads to better models than early stopping with manually defined intervals. All experiments are conducted on a single NVIDIA V100-32GB GPU. Distractor and ShapeNet2D need around $3 - 5$ days for training, depending on different choices of augmentations, Pascal1D needs 8 hours and ShapeNet1D around 12 hours.

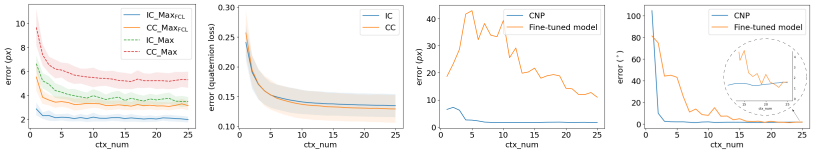
3.4.2 Results and Discussions

Methods	Mean	Max	BA	CA	Max_{FCL}
No Aug	6.02	5.11	4.63	5.13	**3.70**
	6.89	6.17	5.91	6.39	**4.61**
DA	2.67	2.45	2.44	2.65	**2.00**
	4.10	3.75	3.97	4.08	**3.05**
TA	6.29	6.18	6.33	6.32	**5.45**
	7.19	7.04	7.02	7.02	**6.66**
TA+DA	3.20	3.09	2.65	3.05	**2.60**
	6.07	5.14	4.67	4.98	**3.90**

Table 3.1: Prediction error (pixel) on euclidean distance in the 2D image plane for Distractor. Different aggregation methods and augmentations are employed. The first row shows results for intra-category (IC) evaluations, the second row for cross-category (CC).

Methods	MAML	CNP (Mean)	CNP (CA)
No Aug	1.69 (0.22)	5.28 (0.51)	4.66 (0.74)
MR	1.90 (0.27)	2.96 (0.21)	3.33 (0.27)
TA	**1.02 (0.06)**	**1.98 (0.22)**	**1.36 (0.25)**
DA	2.10 (0.09)	3.69 (0.13)	2.90 (0.03)
TA+DA	1.31 (0.14)	2.29 (0.19)	1.77 (0.33)

Table 3.2: Pascal1D pose estimation error. MSE and standard deviations are calculated with 5 random seeds.



(a) Distractor (b) ShapeNet2D (c) Distractor (d) ShapeNet1D

Figure 3.2: (a) CNP Prediction error (pixel) vs context number for the Distractor task using Max aggregation and Max + FCL (Max_{FCL}). Results are evaluated on novel objects from both intra-category (IC) and unseen cross-category (CC) levels. (b) CNP (CA) Prediction error vs context number for ShapeNet2D using DA + TA. (c) We compare a classical object detection method and CNP (Max) using same dataset for training on Distractor. The classical model is further fine-tuned on each new task. The results are shown in dependence of the number of images used for fine-tuning or as context set.(d) Prediction error between the fine-tuned model and CNP (CA) on ShapeNet1D.

Methods	MAML	CNP (Max)	CNP (CA)
No Aug	25.27	14.97 (0.37)	8.19 (0.30)
	21.63	18.09 (0.21)	9.13 (0.18)
MR	13.23	12.71 (0.26)	8.87 (0.36)
	16.55	14.77 (0.35)	8.43 (0.39)
TA	23.01	10.89 (0.27)	7.92 (0.25)
	20.59	14.43 (0.55)	9.18 (0.50)
DA	14.69	8.64 (0.21)	6.24 (0.15)
	16.02	9.87 (0.35)	6.54 (0.19)
TA+DA	17.96	7.66 (0.18)	**5.81 (0.23)**
	18.79	8.66 (0.19)	**6.23 (0.12)**
TA+DA+FCL	–	7.82 (0.08)	6.44 (0.36)
	–	8.84 (0.04)	6.74 (0.20)
TA+DA+MR	13.45	10.54 (0.37)	8.28 (0.17)
	14.44	10.76 (0.30)	8.04 (0.10)

Table 3.3: ShapeNet1D pose estimation error(°). Results are calculated with 5 random seeds except for MAML. The first row presents results for IC and the second row for CC.

MAML or CNPs? We compare MAML and CNPs on two pose estimation datasets, Pascal1D and ShapeNet1D. We obtain similar results as prior work (Yin et al., 2020; Rajendran et al., 2020) on Pascal1D, where MAML performs better than CNPs and the latter shows more severe overfitting (see Table 3.2).

However, Table 3.3 illustrates that both CNP variants outperform MAML with a large margin on ShapeNet1D. It is good to note that, the prediction errors of all methods in Table 3.2 after denormalizing are larger than 30°, indicating the experiments of prior work on Pascal1D simply used too little meta-data to make informative conclusions about the quality of different algorithms. Our interpretation is that MAML tries to learn a good initial prior (global optimum) which needs to be optimized on each specific task (fine-tuned optimum) within few samples and updates. On small datasets, MAML can easily find a global optimum that satisfies all the training tasks. At the same time MAML also overfits less, since the fine-tuning from global to fine-tuned optimum happens during inference time. However, finding a global optimum is getting difficult for large-scale datasets due to the increasing task diversity. Consequently, more samples and updates are necessary to fine-tune the task-specific parameters ϕ, which also explains why MAML is sensitive to hyperparameter tuning (Antoniou et al., 2019). Furthermore, MAML shows much longer training times than CNPs, which limits us to conduct exhaustive comparisons on more complicated tasks such as Distractor or ShapeNet2D. In contrast, CNPs use the local parameterization ϕ as a fixed dimensional output of the encoder, which forces the model to learn an informative low-rank representation from the contexts. Meanwhile, increasing data and task diversity will encourage the model to extract more expressive and mutual-exclusive task representations. In addition, CNP explicitly disentangles the meta-parameters θ and task-specific parameter ϕ during meta-train and reduces the learnable parameters ϕ to a fixed number during meta-test compared to MAML and thus reduces the number of parameters that are adaptable at test time.

DA, DR, TA or MR? From the results of different experiments presented in Table 3.1, Table 3.2, Table 3.3 and Table 3.4, it is obvious that DA improves the performance across all tasks and methods. Table 3.4 also shows the importance of DR on ShapeNet2D which cannot be simply compensated by DA. TA hinders the performance on Distractor but benefits all pose regression tasks. The reason for this is that, for Distractor, TA increases task complexity by shifting the origin of the image plain by the sampled noise, thus creating N^2 copies of the original task, where $\mathrm{N} = 16$ is the number of non-zero elements in the noise set. However, since these task copies live in independent coordinate frames, the increased task diversity is irrelevant to the original task. For pose regression tasks, by contrast, TA augments the canonical poses of the existing data, which coherently benefits the original task as the augmented canonical

Methods	IC ($1e^{-2}$)	CC ($1e^{-2}$)
None	38.33 (0.33)	39.81 (0.31)
DR	18.67 (0.13)	20.05 (0.12)
DR+MR	27.89 (0.61)	28.99 (0.46)
DR+TA$_{\text{azi}}$	16.94 (0.13)	18.42 (0.26)
DR+TA$_{\text{azi+ele}}$	16.62 (0.12)	17.76 (0.35)
DA	19.32 (0.09)	17.98 (0.09)
DR+DA	14.26 (0.09)	13.91 (0.14)
DR+DA+TA$_{\text{azi+ele}}$	14.12 (0.14)	13.59 (0.10)
DR+DA+TA$_{\text{azi+ele}}$ + FCL	**14.01 (0.09)**	**13.32 (0.18)**

Table 3.4: Comparison of different augmentation techniques on ShapeNet2D. Results are calculated with 3 random seeds using CNP (CA) as baseline.

Methods	CA$_{\text{S}}$	CA$_{\text{M}}$	CA$_{\text{L}}$	Max$_{\text{S}}$	Max$_{\text{M}}$	Max$_{\text{L}}$
No Aug	18.60 (0.78)	12.08 (0.44)	8.19 (0.30)	30.44 (0.82)	18.86 (0.34)	14.97 (0.37)
	19.95 (1.08)	12.62 (0.87)	9.13 (0.18)	30.59 (1.14)	21.78 (0.47)	18.09 (0.21)
TA	18.69 (0.87)	10.70 (0.98)	7.92 (0.25)	21.67 (0.66)	13.69 (0.27)	10.89 (0.27)
	19.24 (0.79)	12.05 (0.73)	9.18 (0.50)	23.60 (0.88)	16.76 (0.62)	14.43 (0.55)
TA+DA	7.86 (0.21)	6.32 (0.11)	**5.81** (**0.23**)	11.00 (0.16)	8.23 (0.34)	7.66 (0.18)
	7.49 (0.35)	6.48 (0.41)	**6.23** (**0.12**)	12.98 (0.48)	9.65 (0.40)	8.66 (0.19)

Table 3.5: Performance on ShapeNet1D using small (S), medium (M) and large (L) training dataset sizes for CNP with cross-attention (CA) and Max aggregation. The first row presents results for intra-category (IC) and the second row for cross-category (CC) evaluation. MSE and standard deviations are calculated with 5 random seeds.

poses remain in the coordinate frame of the original task. Therefore, even though TA increases the cross-entropy $\mathcal{H}(Y|X)$ for both cases as demanded in MRM (Rajendran et al., 2020), only the pose regression tasks gain additional benefits. MR results in underfitting as combining MR with augmentations leads to worse performance than using the same augmentations alone for both ShapeNet1D and ShapeNet2D (see Table 3.3 and Table 3.4). MR alleviates overfitting without the use of other augmentations. However, combining MR with sophisticated augmentations can improve the performance in comparison to using solely augmentations (Table 3.3 and Table 3.4). Furthermore, MR requires extensive fine-tuning on the regularization parameter β to modulate between underfitting and overfitting.

Effect of the context set size in CNPs. We compare the prediction error w.r.t. the size of the context set for Distractor (see Fig. 3.2a) and ShapeNet2D (see Fig. 3.2b). Both figures show that increasing the context set size benefits the performance, indicating that both Max and CA aggregations can merge useful information from different context pairs and thereby reduce the task ambiguity. In addition, we find that the model can further improve the performance given the size of context set surpasses the maximum number used for training (15 for both tasks). In particular, there is a small performance gap between intra- and cross-category evaluation for Distractor which is however absent for ShapeNet2D. We believe this indicates that Distractor has more task ambiguity than pose estimation and thus explains why Distractor gains more benefits from FCL than ShapeNet2D (see Table 3.1 and Table 3.4).

CNPs vs pretrained models. It is a common practice in vision task to pretrain a model on a large-scale dataset (e.g., ImageNet (Deng et al., 2009)) in order to obtain good prior features and reduce training time. To conduct a fair comparison of this approach to our model regarding data efficiency, we first pretrain a classical object detection model jointly over all tasks using the same training data as for CNPs. After training has finished, we fine-tune the pretrained model further on each specific new task using different numbers of images. Results are shown in Fig. 3.2c for Distractor and Fig. 3.2d for ShapeNet1D, where the horizontal axis denotes the number of images used for fine-tuning or as contexts for CNPs, respectively. Both figures show that CNPs outperform the pretrained model especially for small numbers of contexts. In the Distractor task, CNP (Max) outperforms the fine-tuned model with a large margin after 25 context images are given. Note that CNPs are capable of transferring to various tasks simultaneously. In contrast, the pretrained model requires separate tuning on each given task, which results in a decreased performance on prior learning tasks.

Which aggregation methods should I use? Cross-attention (CA) performs better than mean aggregation on Pascal1D (see Table 3.2) and Max on ShapeNet1D (see Table 3.3), while it achieves a similar performance to Max aggregation and BA on ShapeNet2D (see Table 3.6). In contrast, mean aggregation used in the original CNP performs the worst on both Pascal1D and ShapeNet2D. Our interpretation is that Mean assigns the same importance to each context while the other aggregation operators can allocate different weights. In Max, the posterior entropy of the represented function is non-increasing given more con-

Methods	IC ($1e^{-2}$)	CC ($1e^{-2}$)
CNP+Mean	15.04 (0.08)	15.45 (0.13)
CNP+Max	14.20 (0.06)	**13.56** (0.28)
CNP+BA	14.16 (0.08)	**13.56** (0.18)
CNP+CA	**14.12** (0.14)	13.59 (0.10)

Table 3.6: Comparison of aggregation methods on ShapeNet2D using DR+DA+TA. Results are calculated with 3 random seeds.

texts (Naderiparizi et al., 2020). Max assigns a weight of one to a context and zero to all others for each dimension of the representation while BA assigns the weights predicted by another neural network. Meanwhile, CA assigns importance by comparing the similarity between context inputs $\{x_C^i\}_{i=1}^K$ and target input x_T at the feature space. BA encodes the importance directly in the latent space with additional estimated uncertainty and applies the Bayes rule to update the posterior. Therefore it helps to identify the task as long as more informative context is added

Furthermore, we find that CA achieves competitive results on all pose estimation tasks but performs slightly worse than BA and Max on Distractor (see Table 3.1) though still better than mean aggregation. This indicates that CA helps in learning representations for object-centric images. Distractor, however, contains objects with random locations, requiring the model to disregard positional information. Methods like CA, which compare similarity between contexts and target over feature space, face inherent difficulties on Distractor. This is due to the fact that CNNs, owing to their translational equivariant nature, are prone to encode some positional information into the extracted image features. Consequently, CA, which compares the similarity directly on this feature space, inevitably forces the model to focus on positional similarity, which leads to a suboptimal allocation of importance.

How much meta-data is essential? We split the training data of ShapeNet1D into subsets of three different sizes, with 10 objects per category for the small dataset (S), 30 objects per category for the medium dataset (M) and 50 for the large dataset (L). Afterwards, we test the performance of CNP with Max aggregation and CA on each of them. The results in Table 3.5 show that Max overfits on the small dataset by simply memorizing all training tasks while CA

works much better. Moreover, CA trained on small dataset achieves a comparable performance with Max on large dataset after using TA and DA, and even outperforms Max on the cross-category level. Thus, we conclude that using CA in combination with augmentation techniques can drastically alleviate the overfitting problem and therefore requires less meta-data on object-centric vision tasks than Max. In contrast, MAML performs much worse on ShapeNet1D (L) (see Table 3.3) than CNPs and thus hardly profits from an increased dataset.

Data augmentation. Table 3.7 shows the effect of each individual data augmentation technique on ShapeNet2D. The first row contains results obtained with all techniques applied jointly. In the other rows, one of the techniques is removed respectively. We find that removing *Affine* leads to the worst performance which indicates that object-centric pose regression tasks are more sensitive to scale. On the other hand, omitting *CropAndPad* even leads to an performance increase.

Methods	Val	Test
All	0.1417	0.1410
w/o CropAndPad	0.1412	0.1368
w/o Affine	**0.1623**	**0.1743**
w/o Dropout	0.1452	0.1445
w/o Contrast	0.1482	0.1406
w/o Brightness	0.1454	0.1380
w/o Blur	0.1426	0.1422

Table 3.7: Comparison of different data augmentation techniques on ShapeNet2D using CNP (CA) + DR as baseline.

Does FCL improve CNPs? Table 3.1 shows the evaluation on Distractor using different aggregation methods where Max_{FCL} denotes Max aggregation with FCL. Modulating task representation by functional contrastive learning (FCL) alleviates meta overfitting across all augmentation levels and thus achieves a significant improvement in performance. Fig. 3.2a further compares the performance of Max and Max_{FCL} for different context set sizes, showing that our methods can differentiate the queried object and distractors well, even for very small context sets. Furthermore, we investigate the influence of FCL on

Methods	IC	CC
Same Ctx	2.30 (0.04)	3.46 (0.06)
Diff Ctx	2.16 (0.05)	3.25 (0.05)
Ctx & Target	**2.00** (**0.02**)	**3.05** (**0.08**)

Table 3.8: Analysis of FCL + CNP on different choices of positive pairs using: i) the same context set with different augmentations (Same Ctx), ii) different context sets from the same task (Diff Ctx), iii) context and target sets (Ctx & Target). Prediction error (pixel) is calculated with 3 random seeds.

the predicted task representations over all 12 categories using different clustering metrics, where the results show that FCL leads to a more dispersed latent distribution compared to the original CNPs, which can improve generalization capability to unseen tasks.

FCL on different sets. We compare FCL on three choices of positive pairs: i) We use the same context set but with different data augmentations. ii) We use different context sets sampled from the same task. iii) We use context and target sets from the same task. We test the performance on Distractor using Max aggregation and DA. For each choice, we run three experiments with different seeds and present the average performance in Table 3.8. Compared to Table 3.1, all three choices consistently outperform CNP (Max) while using FCL on context and target sets achieves the best performance.

Functional Contrastive Learning on CNPs

τ	1.0	0.5	0.2	0.07	0.007
IC	8.5550	8.9810	8.8551	**7.8196**	8.1409
CC	10.4660	10.5135	10.5604	**8.8420**	9.3846

Table 3.9: Results of the evaluation on ShapeNet1D using different temperature values in FCL.

A grid search on hyperparameter τ is very expensive especially on vision tasks. Therefore, we search only on a discrete set $\{0.007, 0.7, 0.2, 0.5, 1.0\}$ and find that $\tau = 0.07$ shows the best performance on ShapeNet1D and $\tau = 0.007$ on ShapeNet2D. The results are shown in Table 3.9 and Table 3.10.

τ	1.0	0.5	0.2	0.07	0.007
IC	0.1564	0.174	0.1962,	0.1441	**0.1401**
CC	0.1594	0.1758	0.2089	0.1390	**0.1332**

Table 3.10: Results of the evaluation on ShapeNet2D using different temperature values in FCL.

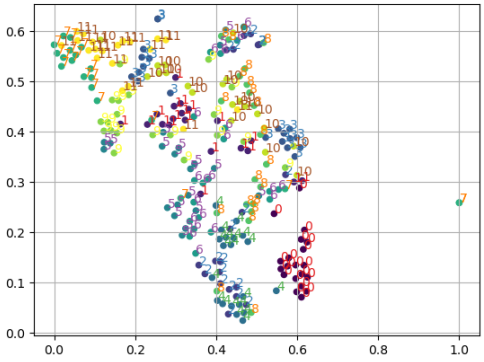
Methods	Max	Max_{FCL}
ARI ↑	0.21	0.20
MI ↑	1.13	1.03
SS ↑	0.31	0.15
CHI ↑	118.73	18.90
DBI ↓	1.00	1.65

Table 3.11: Analysis of latent task representation on Distractor between Max and Max_{FCL} using various clustering metrics.

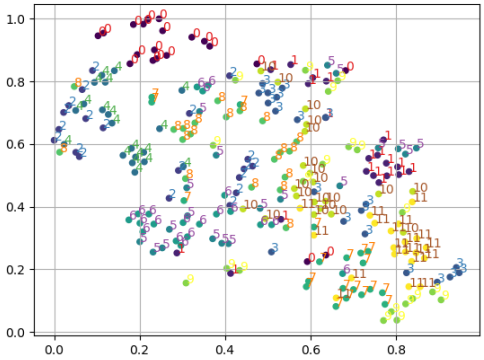
For the Distractor task, we visualize the task representation obtained for novel objects in Fig. 3.3 where each color or number denotes one category and each point denotes the representation of each novel object. $\{10, 11\}$ are the novel categories *{sofa, watercraft}*. Note that each object is considered as a single task and all tasks are learned in a category-agnostic manner. This figure indicates that Max_{FCL} can better shrink the distance between similar objects and repel the different ones implicitly. For instance, without a contrastive loss there is one outlier in Fig. 3.3a that is far away in representation space from the other objects. In particular, some samples are not well clustered based on categories, which is due to the high object variations within the same category.

Furthermore, we investigate the influence of FCL on the predicted task representations over all 12 categories using five clustering metrics, namely Adjusted Rand Index (ARI), Mutual Information (MI), Silhouette Score (SS), Calinski-Harabasz Index (CHI) and Davies-Bouldin Index (DBI). Results are shown in Table 3.11. FCL leads to a more dispersed latent distribution compared to the original CNP, which reduces the vacancy in the latent space and thus improve the generalization ability to unseen tasks.

Additional Results. We have evaluated MMAML (Vuorio et al., 2019), a conditional variant of MAML, on ShapeNet1D based on reviewer's recom-



(a) Max



(b) Max_{FCL}

Figure 3.3: Visualization of latent variables on (a) max aggregation (b) max aggregation + functional contrastive learning (Max_{FCL}).

mendation in Table 3.12. The results is worse than MAML, indicating that the designed task-aware modulation in MMAML doesn't benefit our tasks.

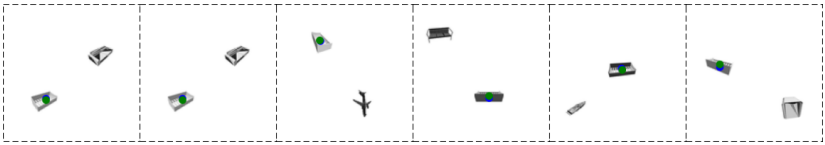
Task Augmentation The angular orientation of Pascal1D is normalized to $[0, 10]$ whereas ShapeNet1D uses radians with range $[0, 2\pi]$. For ShapeNet2D,

MMAML	No Aug	DA	TA	DA+TA
IC	19.6900	26.3624	19.0705	27.4973
CC	20.6123	26.4090	19.4285	27.3120

Table 3.12: Performance of MMAML (Vuorio et al., 2019) on ShapeNet1D.

the azimuth angles are restricted to the range $[0°, 180°]$ in order to reduce the effect of symmetric ambiguity while elevations are restricted to $[0°, 30°]$. we add random noise to both azimuth and elevation angles and then convert the rotation to quaternions for training.

Data Augmentation *Affine* scales images between $80\% - 120\%$ of their size along x and y axis and translate the images between $-10\% - 10\%$ relative to the image height and width, and fills random value for the newly created pixels. *Dropout* either drops random 1%-10% of all pixels or random image patches with $2\% - 25\%$ of the original image size. *CropAndPad* pads each side of the images less than 5% of the image size using random value or the closest edge value. For ShapeNet2D, we furthermore add *GammaContrast* with a range $[0.5, 2.0]$, *AddToBrightness* with a range $[-30, 30]$ and *AverageBlur* using a window of $k \times k$ neighbouring pixels where $k \in [1, 3]$. We use the open-source package (Jung et al., 2020) for all data augmentations.



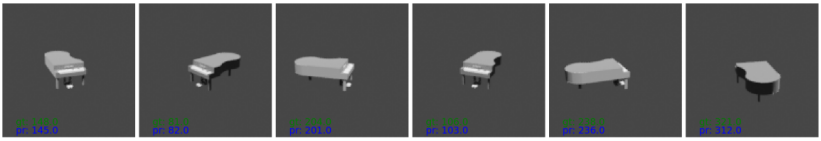
(a) Distractor sofa



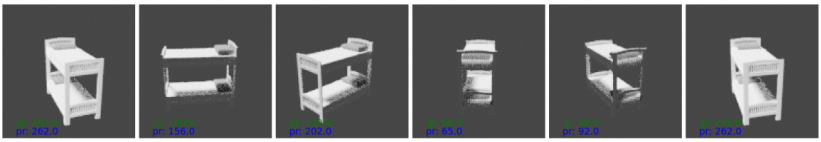
(b) Distractor watercraft

Figure 3.4: Examples of Distractor on novel categories (sofa and watercraft) where green dots are ground-truth and blue dots are predicted positions.

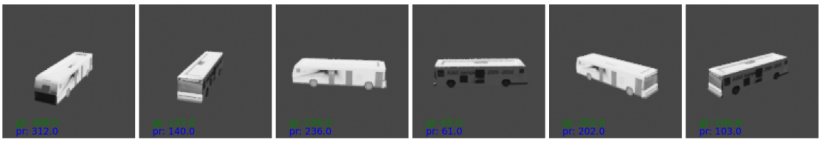
Meta Regularization Yin et al.(Yin et al., 2020) employ regularization on weights, the loss function is defined as:



(a) ShapeNet1D piano



(b) ShapeNet1D bed



(c) ShapeNet1D bus

Figure 3.5: Examples of ShapeNet1D on novel categories (piano, bed, bus).

$$\mathcal{L} = \mathcal{L}_O + \beta D_{\mathrm{KL}}(q(\theta; \theta_\mu, \theta_\sigma)||r(\theta)) \tag{3.7}$$

where $\mathcal{L}_O$ denotes the original loss function defined individually in Distractor and pose estimation. meta-parameters θ denote the parameters which are not used to adapt to the task training data. Function $r(\theta)$ is a variational approximation to the marginal which is set to $\mathcal{N}(\theta; 0, I)$ in Yin et al.(Yin et al., 2020). We follow the same setup in our experiments.

Examples of Inference Results

We visualize examples of evaluation on novel cateogories in Fig. 3.4 for Distractor, Fig. 3.5 for ShapeNet1D and Fig. 3.6 for ShapeNet2D.

3.4.3 Limitations

Scalability Since meta-learning is a type of data-driven algorithm, the generalization depends on the diversity of training tasks. Thus, it is essential



Figure 3.6: Examples of ShapeNet2D on novel categories (piano, bed, bus). Predictions are converted to (azimuth, elevation) angles.

to increase prior knowledge with more sophisticated augmentations. Augmentations used in our work are restricted within existing knowledge, using generative model to enrich training data could be one solution.

Representation The task representation in CNP is learned implicitly using different aggregation methods. However, this aggregation can be considered as a bottleneck and thus loses important information from context. We add contrustive loss after aggregation in order to disentangle different tasks in a category-agnostic way. Future work can investigate further on latent representation using more effective aggregation or compositional representation.

Furthermore, concerning the class of NPs, we restricted ourselves to the deterministic CNPs in the experiments since we empirically found that using stochastic variants such as NPs (Garnelo et al., 2018b) does not benefit our tasks. We leave an in-depth exploration of stochastic NPs on vision tasks to future work.

3.5 Conclusion

In this paper, we investigate MAML and CNPs on several image-level regression tasks and analyze the importance of different choices in mitigating meta overfitting. Furthermore, we provide insights and practical recommendations of different algorithmic choices for CNPs with respect to various task settings. In addition, we combine CNPs with functional contrastive learning in task space and train in an end-to-end manner, which significantly improves the task expressivity of CNPs. We believe that our work can lay the basis for future work on designing and implementing meta-learning algorithms in image-based regression tasks.

4 GAML: Geometry-Aware Meta-Learner for Cross-Category 6D Pose Estimation

4.1 Introduction

Estimating the 6D pose of an object is of practical interest for many real-world applications such as robotic grasping, autonomous driving and augmented reality (AR). Prior work has investigated instance-level 6D pose estimation (Wang et al., 2019a; Peng et al., 2019; He et al., 2020; He et al., 2021), where the objects are predefined. Although achieving satisfying performance, these methods are prone to overfit to specific objects and thus suffer from poor generalization. Due to the high variety of objects with different colors and shapes in the real-world, it is impractical to retrain the model every time new objects come in, which is time-consuming and data inefficient. Recently, this issue has raised increasing attention in the community and several approaches (Wang et al., 2019b; Chen et al., 2020b; Wang et al., 2020a; Chen et al., 2021b; Chen et al., 2020d; Chen and Dou, 2021) have been proposed for category-level 6D pose estimation. NOCS (Wang et al., 2019b) and CASS (Chen et al., 2020b), for example, map different instances of each category into a unified representational space based on RGB or RGB-D features. However, the assumption of a unified space potentially leads to a decrease in performance in case of strong object variations. FS-Net (Chen et al., 2021b) proposes an orientation-aware autoencoder with 3D graph convolutions for latent feature extraction where translation and scale are estimated using a tiny PointNet (Qi et al., 2017a). Furthermore, Chen et al.(Chen et al., 2020d) provide an alternative based on “analysis-by-synthesis" to train a pose-aware image generator, implicitly representing the appearance, shape and pose of the entire object categories. However, these methods require a pretrained object

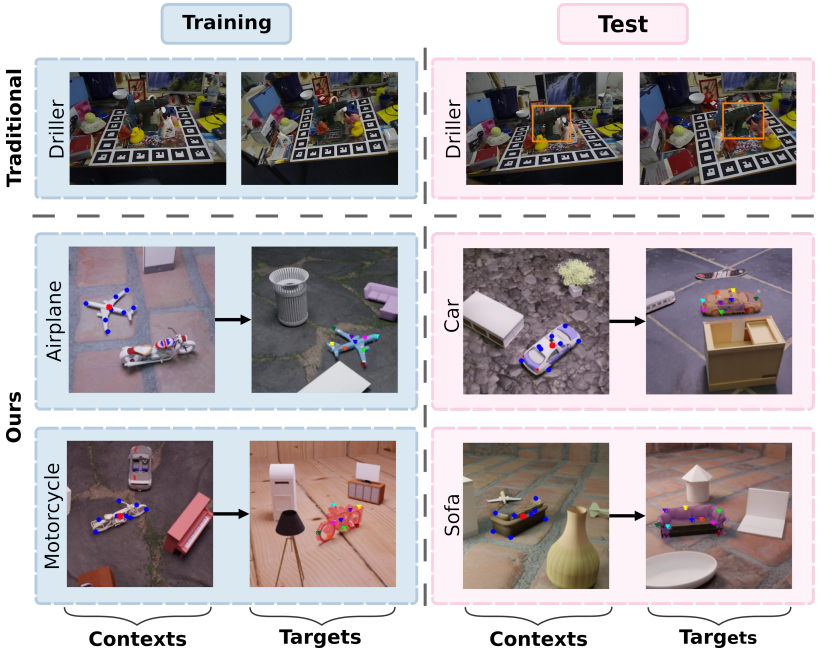


Figure 4.1: Illustration of the difference between traditional instance-level 6D pose estimation methods and our approach. Unlike other methods, our proposed approach generalizes to novel objects given a few context observations. The projected ground-truth keypoints are visualized as blue points in the context images. The predicted segmentation and keypoints are visualized in the target images.

detector on each specific category which limits their generalization ability across categories.

In this paper, we present a new meta-learning based approach to increase the generalization capability of 6D pose estimation. To our knowledge, this is the first work that allows generalization across object categories. The main idea of our method lies in meta-learning object-centric representations in a category-agnostic way. Meta-learning aims to adapt rapidly to new tasks based only on a few examples. More specifically, we employ Conditional Neural Processes (CNPs) (Garnelo et al., 2018a) to learn a latent representation of objects, capturing the generic appearance and geometry. Inference on new objects then merely needs a few labeled examples as input to extract

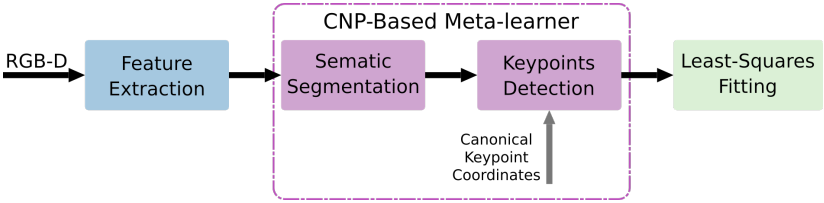


Figure 4.2: Schematic pipeline of our approach.

a respective representation. In particular, fine-tuning on new objects is not necessary. A comparison between traditional instance-level approaches and ours is illustrated in Fig. 4.1.

For feature extraction, we use FFB6D (He et al., 2021), which learns representative features through a fusion network based on RGB-D images. However, instead of directly using the extracted features for downstream applications, i.e., segmentation and keypoint offsets prediction, we add CNP on top of the fusion network to further meta-learn a latent representation for each object. CNP takes in the representative features from a set of context images of an object, together with their ground-truth labels, and yields a latent representation. The subsequent predictions for new target images are conditioned on this latent representation.

To further leverage the object geometry and improve the keypoint prediction, we propose a novel GNN-based decoder which takes predefined canonical keypoints in the object's reference frame as an additional input and encodes local spatial constraints via message passing among the keypoints. Note that the additional input to the GNN does not require any further annotations on top of those existing datasets used by prior keypoint-based methods. The proposed pipeline is illustrated in Fig. 4.2.

Due to the lack of available data for cross-category level 6D pose estimation, we generate our own synthetic dataset for **M**ultiple **C**ategories in **M**ultiple **S**cenes (**MCMS**) using objects from ShapeNet (Chang et al., 2015) and extending the open-source rendering pipeline (Denninger et al., 2019) with online occlusion and truncation checks. This provides us with the flexibility to generate datasets with limited and considerable occlusion respectively.

In summary, the main contributions of this work are as follows:

- We introduce a novel meta-learning framework for 6D pose estimation with strong generalization ability on unseen objects within and across object categories.
- We propose a GNN-based keypoint prediction module that leverages geometric information from canonical keypoint coordinates and captures local spatial constraints among keypoints via message passing.
- We provide fully-annotated synthetic datasets with abundant diversity, which facilitate future research on intra- and cross-category level 6D pose estimation.

4.2 Related Work

6D Pose Estimation. For instance-level 6D pose estimation, methods can be categorized into three classes: correspondence-based, template-based and voting-based methods (Du et al., 2019). Correspondence-based methods aim to find 2D-3D correspondences (Zakharov et al., 2019; Song et al., 2020; Pham et al., 2020) or 3D-3D correspondences (Fischer et al., 2021). Template-based methods, on the other hand, match the inputs to templates, which can be either explicit pose-aware images (Hinterstoisser et al., 2012; Hinterstoisser et al., 2013) or templates learned implicitly by neural networks (Sundermeyer et al., 2018). Voting-based approaches (Peng et al., 2019; He et al., 2020; He et al., 2021) generate voting candidates from feature representations, after which the RANSAC algorithm (Fischler and Bolles, 1981) or a clustering mechanism such as MeanShift (Kobayashi and Otsu, 2010) is applied for selecting the best candidates. Our feature extractor, FFB6D (He et al., 2021), falls into this latter category. FFB6D proposes a bidirectional fusion module to combine appearance and geometry information for feature learning. The extracted features are then used to predict per-point semantic labels and keypoint offsets, after which MeanShift is used to vote for 3D keypoints. Finally, the keypoints are used to predict the final 6D pose by Least-Squares Fitting (Arun et al., 1987).

Recently, category-level 6D object pose estimation has gained increasing attention (Wang et al., 2019b; Chen et al., 2020b; Wang et al., 2020a; Chen et al., 2021b; Chen et al., 2020d). Wang et al.(Wang et al., 2019b) share a canonical representation for all possible object instances within a category

using Normalized Object Coordinate Space (NOCS). However, inferring the object pose by predicting only the NOCS representation is not easy given large intra-category variations (Fan et al., 2021a). To tackle this problem, Tian et al. (2020) accounts for intra-category shape variations by explicitly modeling the deformation from shape prior to object model while CASS (Chen et al., 2020b) generates 3D point clouds in the canonical space using a variational autoencoder (VAE). FS-Net (Chen et al., 2021b) proposes a shape-based model using 3D graph convolutions and a decoupled rotation mechanism to further reduce the sensitivity of RGB features to the color variations. However, these methods model the feature space explicitly on a category-level and therefore have a limited generalization ability across categories. By contrast, our method learns 6D pose estimation in a category-agnostic manner and can handle new objects from unseen categories.

Meta-Learning. Meta-learning, also known as learning to learn, aims to acquire meta knowledge that can help the model to quickly adapt to new tasks with very few samples. In general, meta-learning can be categorized into metric-based (Vinyals et al., 2016b; Snell et al., 2017; Sung et al., 2018), optimization-based (Finn et al., 2017; Nichol et al., 2018; Finn et al., 2019) and model-based (Santoro et al., 2016; Garnelo et al., 2018a; Garnelo et al., 2018b; Kim et al., 2019) methods. Many meta-learning approaches have been applied to computer vision applications, e.g., few-shot image classification (Gidaris and Komodakis, 2018; Zhang et al., 2020; Tseng et al., 2020; Liu et al., 2021b), vision regression (Gao et al., 2022b), object detection (Perez-Rua et al., 2020; Fan et al., 2020; Fan et al., 2021b; Zhang et al., 2021a; Chen et al., 2021a), robotic grasping (Gao et al., 2022a), semantic segmentation (Siam et al., 2019; Li et al., 2020c; Pambala et al., 2021; Zhang et al., 2021b) and 3D reconstruction (Wallace and Hariharan, 2019; Michalkiewicz et al., 2020). Our work is based on Neural Processes (NPs) (Garnelo et al., 2018b; Kim et al., 2019; Louizos et al., 2019b; Gordon et al., 2020; Lee et al., 2020), which fall into the category of model-based meta-learning approaches. NPs have shown promising performance on simple tasks like function regression and image completion. However, their application to 6D pose estimation has not yet been explored properly. We introduce CNP (Garnelo et al., 2018a) to this problem in order to tackle the issue of poor generalization ability of existing methods on both intra- and cross-category level.

Graph Neural Networks. Graph neural networks (GNNs) have been widely applied on vision applications, such as image classification (Lee et al., 2018; Kampffmeyer et al., 2019; Long et al., 2021), semantic segmentation (Qi et al., 2017b; Landrieu and Simonovsky, 2018; Wang et al., 2019c; Liang et al., 2019), and object detection (Hu et al., 2018; Shi and Rajkumar, 2020; Wang et al., 2021c). Recently, many works start using GNNs on human pose estimation (Wang et al., 2020b; Bin et al., 2020; Yang et al., 2021b). Yang et al.(Yang et al., 2021b) derive the pose dynamics from historical pose tracklets through a GNN which accounts for both spatio-temporal and visual information while PGCN (Bin et al., 2020) builds a directed graph over the keypoints of the human body to explicitly model their correlations. DEKR (Geng et al., 2021) adopts a pixel-wise spatial transformer to concentrate on information from pixels in the keypoint regions and dedicated adaptive convolutions to further disentangle the representation. Our approach is based on a similar idea as PGCN, where we take the keypoints in the canonical object coordinates as an additional input in order to leverage the spatial constraints between keypoints. We show that this drastically increases the performance on unseen objects and robustness on occluded scenes.

4.3 Preliminary - Conditional Neural Processes

Conditional Neural Processes (CNPs) (Garnelo et al., 2018a) can be interpreted as conditional models that perform inference for some target inputs x_t conditioned on observations, called “contexts". These contexts consist of inputs x_c and corresponding labels y_c originating from one specific task. Note that in our case, each distinct object is considered as a task.

The basic form of CNP comprises three core components: encoder, aggregator and decoder. The encoder takes a set of M_c context pairs from a given task $C = \{(x_c^i, y_c^i)\}_{i=1}^{M_c}$ and extracts embeddings from each context pair respectively, $r_i = h_\theta(x_c^i, y_c^i),\ \forall (x_c^i, y_c^i) \in C$, where h is a neural network parameterized by θ. Afterwards, the aggregator a summarizes these embeddings using a permutation invariant operator $\otimes$ and yields the global latent variable as task representation: $z = a(r_1, r_2, ..., r_{M_c}) = r_1 \otimes r_2 \otimes ... \otimes r_{M_c}$. Since the size of context set M_c varies and the task representation has to be independent of the order of contexts, a permutation invariant mechanism is essential. Max aggregation is used in our model as we empirically find it outperforms mean aggregation,

which is used in the original CNP. Finally, the decoder performs predictions for a set of target inputs $T = \{x_t^i\}_{i=1}^{M_t}$ conditioned on the corresponding task representation z extracted and aggregated before: $\hat{y}_t^i = g_\phi(x_t^i, z),\ \forall x_t^i \in T$. M_t is the number of target inputs, g denotes the decoder, a neural network parametrized by ϕ.

The ability to extract meaningful latent representation from very few samples renders CNP well-suited for our purposes. Due to the fact that each distinct object comes with different predefined keypoints, prior keypoint-based methods for 6D pose estimation do not generalize well to novel objects. Meta-training CNP to extract latent keypoint representations from object features, however, allows us to overcome this difficulty.

4.4 Approach

In this paper, we propose a keypoint-based meta-learning approach for 6D pose estimation on unseen objects. Given an RGB-D image, the goal of 6D pose estimation is to calculate the rigid transformation $[R; t]$ from the object coordinates to the camera coordinates, where $R \in SO(3)$ represents the rotation matrix and $t \in \mathbb{R}^3$ represents the translation vector. We build on keypoint-based methods, that first predict the location of keypoints in camera coordinates from input RGB-D images and then regress the transformation between these and predefined keypoints in the object coordinates. The predefined keypoints in canonical object coordinates are thereby fixed beforehand, e.g., using the Farthest Point Sampling (FPS) algorithm on the object mesh.

4.4.1 Overview

We consider 6D pose estimation in three stages: feature extraction, keypoint detection and pose fitting. At the first stage, we employ the feature extractor FFB6D (He et al., 2021) to extract representative features from RGB-D images. For the second stage we use a CNP-based meta-learning approach. The flow of context and target samples through our model is shown in Fig. 4.3, where the context inputs for each task x_c, i.e., the features extracted from the context RGB-D images, and the correpsonding labels y_c are used jointly to distill a task representation. This representation serves as prior knowledge

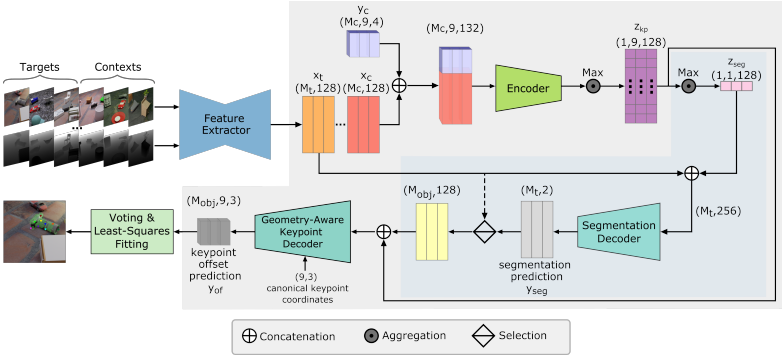


Figure 4.3: **Overview of the three stages of our method.** a) The feature extractor takes RGB-D images as inputs and produces point-wise features for a set of M_c (M_t) points subsampled from the input context (target) image. b) The meta-learner (grey shaded area) encodes and aggregates the features of several context images into two latent variables z_{kp} and z_{seg}. The segmentation module (blue shaded area) predicts a binary semantic label for each of the M_t feature points of a target image conditioned on the latent representations z_{seg}, indicating whether the respective point belongs to the queried object. The keypoint decoder predicts per-point offsets for each keypoint based on the segmented features and the keypoint latent variables z_{kp}. c) Lastly, 6D pose parameters are computed via voting and least-squares fitting.

for the subsequent prediction on target inputs x_t. We use two decoders in our meta-learning framework, predicting semantic labels and 3D keypoint offsets respectively. Furthermore, we propose a novel geometry-aware decoder using a GNN for the keypoint offsets prediction, which explicitly models the spatial constraints between the keypoints. Finally, the 6D pose parameters are regressed by least-squares fitting at the third stage.

4.4.2 Feature Extraction

For feature extraction we rely on the fusion network FFB6D (He et al., 2021) which combines appearance and geometry information from RGB-D images and extracts representative features for a subset of seed points sampled from the input depth images. Therefore, the output is a set of per-point features corresponding to the sampled seed points.

4.4.3 Meta-Learner for Keypoint Detection

Two steps are involved in the keypoint estimation procedure: segmentation of the queried object and keypoint detection, which both rely on a preceding extraction of latent representations.

Extraction of latent representations. Identifying and distinguishing a novel object from a multi-object scene and extracting its keypoints requires modules, which are conditioned on the latent representation of the queried object. In order to obtain such a latent representation, we need a set of context samples $\{(x_{c,i}, y_{c,i})\}_{i=1}^{M_c}$. Here $x_{c,i}$ denotes the per-point features extracted in the first stage from context images and $y_{c,i} = \{y_{c,i}^u\}_{u=1}^{M_K}$ is the ground-truth label where $y_{c,i}^u = \{y_{of}^u, y_{seg}\}_{c,i}$ includes the 3D keypoint offsets y_{of}^u between the seed point and predefined keypoint p_u, and semantic label $y_{seg} \in \{0, 1\}$ indicating whether the seed point belongs to the queried object. Given a context sample as input, an encoder generates per-seed-point embeddings for each of the M_k keypoints to be predicted:

$$r_i^u = h_\theta(x_{c,i} \oplus y_{c,i}^u),\ i = 1, ..., M_c,\ u = 1, ..., M_k, \tag{4.1}$$

where M_c denotes the number of seed points selected from each context image; M_k is the number of selected keypoints which in our case is 9. $\oplus$ stands for the concatenation operation, where the inputs are first broadcast to the same shape, if necessary. The obtained embeddings are next aggregated by max aggregation to first obtain a latent representation z_{kp}^u for each keypoint. A second aggregation over these keypoint representations is then applied in order to extract a representation z_{seg} for the segmentation task:

$$z_{kp}^u = \max_{i=1}^{M_c}(r_i^u),\ u = 1, ..., M_k, \tag{4.2}$$

$$z_{seg} = \max_{u=1}^{M_k}(z_{kp}^u). \tag{4.3}$$

Conditional Segmentation. In the step described above, the model encapsulates relevant information (e.g., shape and texture attributes) into the latent variable z_{seg}. This can then be used to identify and locate the queried object in the target images. The segmentation decoder g_S takes the latent variable z_{seg}

and features x_t extracted from the target images (see Fig. 4.3) and predicts a semantic label for each seed point via a multi-layer perceptron (MLP):

$$y_{seg,i} = g_{\mathcal{S}}(x_{t,i} \oplus z_{seg}),\ i = 1, ..., M_t, \tag{4.4}$$

where M_t is the number of seed points sampled from each target image, $x_{t,i}$ denotes the corresponding extracted features. These per-point segmentation predictions y_{seg} are then used to select only the seed point features x_{obj} belonging to the queried object from x_t for the subsequent keypoint prediction.

Conditional Keypoint Offset Prediction. The keypoint offsets decoder $g_{\mathcal{K}}$ takes the features extracted by the segmentation module along with the latent variables z_{kp} as input and predicts translation offsets y_{of} for each keypoint:

$$\begin{aligned} y^u_{of,i} &= g_{\mathcal{K}}(x_{obj,i} \oplus z^u_{kp}), \\ &i = 1, ..., M_{obj},\ u = 1, ..., M_k, \end{aligned} \tag{4.5}$$

where M_{obj} denotes the number of selected seed points on the queried object, $x_{obj,i}$ denotes the object features of i-th seed point. The decoder $g_{\mathcal{K}}$ can be any appropriate module in Eq. (4.5). In the vanilla version of our framework, it is given by a trivial MLP. However, we use a GNN for $g_{\mathcal{K}}$ in our final version, the details of which will be given in Section 4.4.4.

Pose Fitting. Similar as in He et al. (2021), we adopt MeanShift (Kobayashi and Otsu, 2010) to obtain the final keypoint prediction $\{p^*_i\}^{M_k}_{i=1}$ in the camera coordinates, based on keypoint candidates output by the keypoint decoder. Given predefined 3D keypoints in object coordinates $\{p_i\}^{M_k}_{i=1}$, 6D pose estimation can be converted into a least-squares fitting problem (Arun et al., 1987) where the optimized pose parameters $[R; t]$ are calculated by minimizing the squared loss using singular value decomposition(SVD):

$$L_{lsf} = \sum_{i=1}^{M_k} \left\| p^*_i - (R \cdot p_i + t) \right\|^2. \tag{4.6}$$

4.4.4 Geometry-Aware Keypoint Decoder

Similar to prior methods (He et al., 2020; He et al., 2021), we rely on predefined object keypoints for the final pose fitting. However, we also utilize them as an additional input to the keypoint decoder. Since they contain useful prior

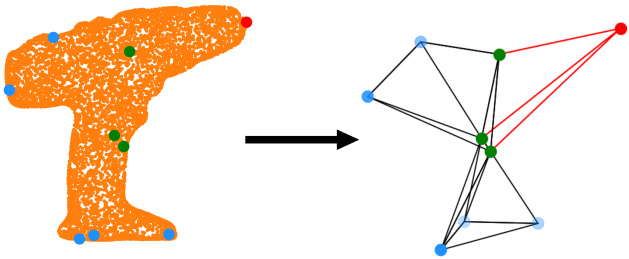


Figure 4.4: **Example of the graph generation.** The node positions are determined by the predefined keypoints in object coordinates. By applying the K Nearest Neighbor (KNN) algorithm, we find the k closest adjacent nodes of each parent node and connect them by edges. For instance, the given graph is generated with $k = 3$, where the red node is selected as the parent node and green nodes are the three nearest neighbors. The driller is sampled from LineMOD.

knowledge of the object's geometric structure, they can significantly improve keypoint detection. In order to highlight the additional input to our decoder, we rewrite Eq. (4.5) as follows:

$$y^u_{of,i} = g_{\mathcal{K}}(x_{obj,i},\, z^v_{kp},\, p_v),\ v \in \mathcal{N}(u), \tag{4.7}$$

where $\mathcal{N}(u)$ denotes the neighbor set of keypoint u including u itself and p_v are the 3D object coordinates of keypoint v. To leverage the geometric information contained in the relation among the keypoints, we propose a GNN-based decoder $g_{\mathcal{K}}$ instead of the trivial MLP in Eq. (4.5). For this purpose, we create a graph over the keypoints of each object. The nodes are given by the keypoints which share edges with their k nearest neighbours. Fig. 4.4 illustrates an example with $k = 3$. Internally, Eq. (4.7) is split into the following two steps involved in message passing along the graph:

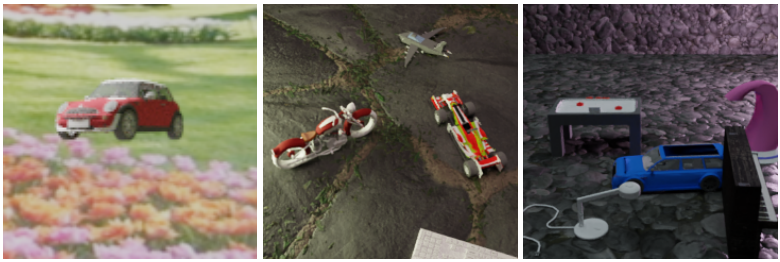
$$\alpha^{u,v}_i = f^l(x_{obj,i} \oplus z^v_{kp}, p_u - p_v),\ \forall v \in \mathcal{N}(u), \tag{4.8}$$

$$y^u_{of,i} = f^g(\max_{v \in \mathcal{N}(u)} \alpha^{u,v}_i). \tag{4.9}$$

$g_{\mathcal{K}}$ is correspondingly composed of two sub-networks, f^l and f^g. These correspond to updating the messages $\alpha^{u,v}_i$ sent along all edges, aggregating the messages arriving at each node u to update the corresponding node features and decoding them into keypoint offsets $y^u_{of,i}$.

4.5 Experiments

4.5.1 Datasets



(a) Toy (b) PBR (c) Occlusion

Figure 4.5: Samples from MCMS dataset.

LineMOD. LineMOD (Hinterstoisser et al., 2013) is a widely used dataset for 6D pose estimation which comprises 13 different objects in 13 scenes. Each scene contains multiple objects, but only one of them is annotated with a 6D pose and instance mask.

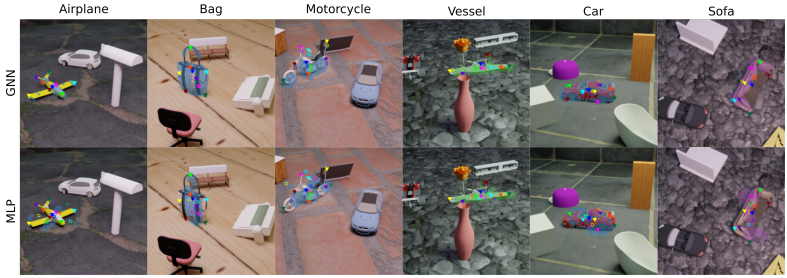


Figure 4.6: **Qualitative comparison between GNN and MLP decoder for keypoints prediction on PBR-MCMS.** Triangles and circles are the projected ground-truth and predicted keypoints respectively. The keypoint predictions of the MLP decoder are randomly shifted without considering geometric constrains between keypoints. By contrast, the predictions by the GNN decoder are more accurate. The example of the motorcycle shows that though the keypoints predicted by the GNN are slightly shifted here, the geometric constraints are met, resulting in a uniform shift of all keypoints.

MCMS dataset. Due to the unavailability of datasets for cross-category level 6D pose estimation, we generate two fully-annotated synthetic datasets using objects from ShapeNet (Chang et al., 2015), which contain various objects from **Multiple Categories** in **Multiple Scenes (MCMS)**. The simple version of MCMS, named Toy-MCMS, is composed of images containing a single object with backgrounds randomly sampled from the real-world image dataset SUN (Song et al., 2017). Our second dataset can be further divided into a non-occluded and an occluded version, called PBR-MCMS and Occlusion-MCMS. To create these datasets, we extend the open-source physics-based rendering (PBR) pipeline (Denninger et al., 2019) with functionalities such as online truncation and occlusion checks. For each image, five objects are placed in a random scene with textured planes and varying lighting conditions. Images are then photographed with a rotating camera from a range of distances. PBR-MCMS contains images without occlusion while Occlusion-MCMS contains images with $5\% - 20\%$ occlusion of the queried object. Fig. 4.5 shows an example for each dataset using an object from the car category as the queried object.

4.5.2 Evaluation Metrics

We use the average distance metrics ADD (Hinterstoisser et al., 2013) for evaluation. Given the predicted 6D pose $[R; t]$ and the ground-truth pose $[R^*; t^*]$, the ADD metric is defined as:

$$\text{ADD} = \frac{1}{m} \sum_{x \in O} \|(Rx + t) - (R^* x + t^*)\|, \tag{4.10}$$

where O denotes the object mesh and m is the total number of vertices on the object mesh. This metric calculates the mean distance between the two point sets transformed by predicted pose and ground-truth pose respectively. Similar to other works (Xiang et al., 2018; Peng et al., 2019; He et al., 2021), we report the ADD-0.1d accuracy, which indicates the ratio of test samples, where the ADD is less than 10% of the object's diameter.

4.5.3 Implementation and Training Details

For each object, we define 9 keypoints, where 8 keypoints are sampled from the 3D object model using FPS, and the other one is the object center. The

nearest neighbors used for each keypoint is set to $k = 8$ in our geometry-aware decoder. To train the meta-learner, we use the Focal Loss (Lin et al., 2017) to supervise the segmentation module and a L1 loss for per-point translation offset prediction. The overall loss is weighted sum of both terms, with a weight 2.5 for segmentation and 1.0 for keypoint offsets. During training, for each iteration, we arbitrarily sample 18 objects and 12 images per object. The number of context images is randomly chosen between 2 and 8 per object while the remaining images are used as target set.

Training setup. For the LineMOD dataset, we use iron, lamp, and phone as novel objects for testing and the 10 remaining objects for training. Since LineMOD contains only a very limited number of objects, we only evaluate the keypoint offset prediction module using the ground-truth segmentation for selecting the points belonging to the queried object. For Toy- and PBR-MCMS, we use 20 and 19 categories for training respectively, with 30 objects per category and 50 images per object. During evaluation, 30 novel objects of each training category are tested for intra-categorical performance and 5 novel categories for cross-category performance. All experiments are conducted on NVIDIA V100-32GB GPU.

	FFB6D		Ours	
Object	L1 Loss	ADD	L1 Loss	ADD
Ape	0.06	**100**	0.02	**100**
Holepuncher	0.07	**100**	0.02	**100**
Iron*	1.39	0.6	0.26	**36.2**
Lamp*	1.52	0.5	0.38	**22.4**
Phone*	0.89	0.0	0.17	**17.8**

Table 3: Evaluation results on LineMOD dataset.

4.5.4 Evaluation Results

We evaluate our approach using the LineMOD and MCMS datasets at intra- and cross-category levels.

LineMOD. The LineMOD dataset (Hinterstoisser et al., 2013) is split into 10 training objects and 3 unseen test objects, where iron, lamp and phone

are the novel test objects. Fig. 4.7 show the qualitative comparison between FFB6D (He et al., 2021) and the proposed model on training objects. It can be observed that our model can predict keypoints more accurately. From Fig. 4.8, we can see that our model achieves better performance on novel objects. It should be noted that we only train one model for all objects, rather than train one model for each object respectively. Table 3 shows training and test results following He et al. (2021). Note that the segmentation ground-truth is used for these results and we only evaluate the performance and generalization ability of the keypoint offset prediction module. Our model not only performs better on training objects, but also generalizes well to new objects even though it is trained on a limited number of objects and tested on new objects with large variations in appearance and geometry.

Training objects	L1 loss[m]	ADD-0.1d[%]	Time[h]
1100	0.128	98.7	159
80	0.245	96.7	16

Table 4.2: Single category - car evaluation on Toy-MCMS dataset

Toy-MCMS Dataset Table 4.2 provides the quantitative results of inter-category 6D pose estimation on the car category. We use 50 images per object for training and vary the number of training objects. From the experimental results, 80 car objects can achieve a similar ADD accuracy as 1100

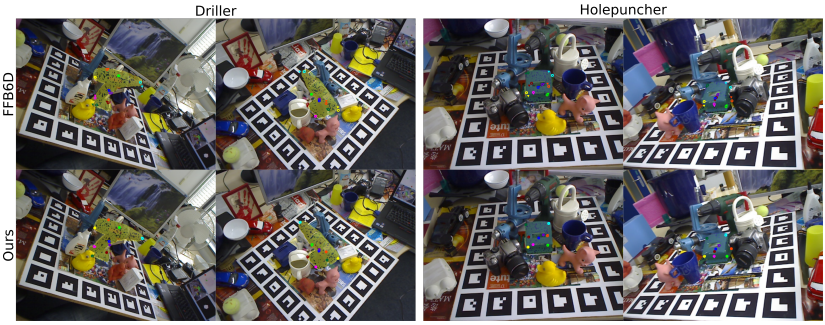


Figure 4.7: **Qualitative comparison on trained LineMOD objects.** Triangles and circles are the projections of ground-truth and predicted keypoints respectively. It can be observed that keypoint predictions of our method are more accurate.

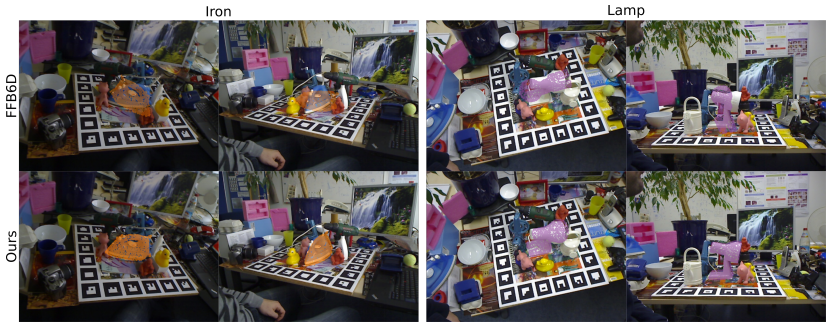


Figure 4.8: **Qualitative comparison on new LineMOD objects.** Compared with FFB6D, the pose estimation on new objects of our GAML model is more accurate.

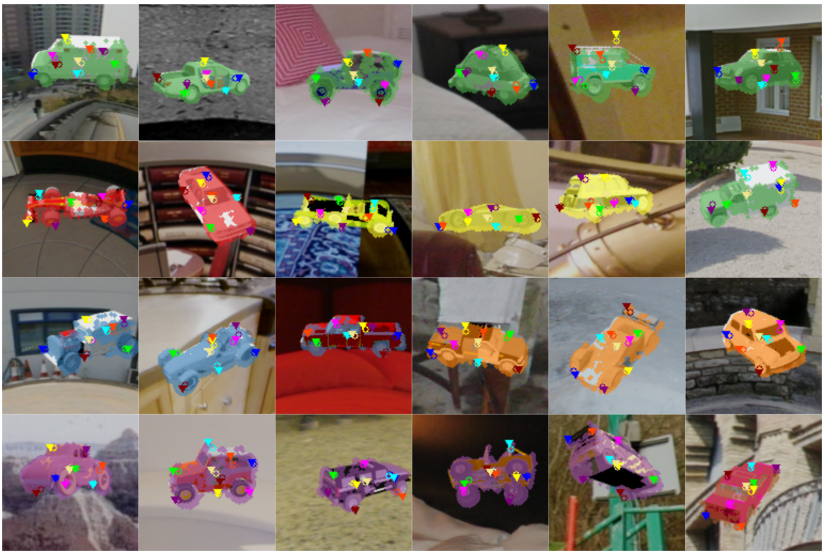


Figure 4.9: **Qualitative results on Toy-MCMS.** Our model can handle large intra-category variations. The car category is illustrated as an example.

objects, while the training time is reduced evidently. Overall, this represents a good comprise between prediction performance and training overhead. Fig. 4.9 shows the qualitative results on novel test objects using the model trained with

	FFB6D			Vanilla-ML			GAML		
Category	L1 Loss	ADD	ADD-S	L1 Loss	ADD	ADD-S	L1 Loss	ADD	ADD-S
Airplane	1.51	9.1	85.7	0.11	**90.4**	96.8	0.11	89.8	98.8
Bag	1.98	5.1	48.1	0.41	40.0	85.0	0.47	**42.7**	87.1
Bathtub	2.22	2.7	41.4	0.55	43.3	86.7	0.60	**45.2**	90.8
Bed	2.31	2.9	33.3	0.31	**72.3**	90.4	0.41	58.5	90.8
Bench	2.26	2.9	43.0	0.39	62.1	91.4	0.35	**69.8**	91.7
Bookshelf	2.28	2.4	32.6	0.36	**55.0**	85.4	0.41	50.2	77.9
Bus	1.94	3.5	56.5	0.51	41.5	89.6	0.36	**69.8**	92.7
Cabinet	2.38	2.2	24.0	0.43	53.5	73.7	0.34	**67.7**	83.5
Camera	1.93	2.1	51.5	0.38	46.3	86.3	0.34	**54.8**	85.8
Cap	1.75	3.1	68.5	0.19	79.2	98.5	0.19	**80.8**	98.8
Chair	2.27	1.1	26.1	0.18	**80.0**	93.1	0.19	**80.0**	89.6
Earphone	1.79	4.2	62.8	0.38	34.0	86.2	0.43	**49.2**	97.1
Motorcycle	1.65	12.7	85.1	0.16	**90.2**	98.5	0.21	85.6	98.1
Mug	2.08	0.9	43.7	0.12	**86.8**	97.9	0.14	84.2	94.4
Table	2.38	2.0	19.2	0.61	33.1	73.5	0.65	**39.2**	93.1
Train	1.67	12.9	71.9	0.46	38.5	85.8	0.49	**47.7**	90.6
Vessel	1.64	11.0	68.9	0.35	**57.7**	94.1	0.37	56.0	90.4
Washer	2.48	4.1	29.1	0.33	54.8	85.4	0.30	**68.1**	89.0
Printer	2.21	1.0	33.1	0.41	47.9	83.9	0.43	**55.2**	80.0
Birdhouse*	2.09	0.8	21.0	0.39	**35.6**	59.4	0.43	35.4	64.6
Car*	1.79	2.4	70.5	0.44	52.5	97.1	0.43	**56.9**	96.7
Laptop*	2.22	1.3	10.5	0.32	54.0	82.9	0.20	**85.0**	93.1
Piano*	2.04	2.0	39.3	0.43	**45.8**	77.5	0.44	**45.8**	80.0
Sofa*	2.17	2.6	27.1	0.34	68.1	84.6	0.34	**69.8**	79.0
Intra-Categ.	2.03	4.53	48.7	0.35	58.2	88.5	0.36	**62.9**	90.0
Cross-Categ.	2.06	1.81	33.7	0.38	51.2	80.2	0.37	**58.6**	82.7
All	2.04	3.96	45.5	0.36	56.7	86.8	0.36	**62.0**	88.4

Table 4.3: Multi-category evaluation on PBR-MCMS dataset

80 objects. Note that even within the car category, the colors and shapes of novel objects still vary a lot.

Table 4.4 shows the quantitative results of the multi-category evaluation on the Toy-MCMS dataset. The vanilla meta-learner (Vanilla-ML) using MLP decoder is compared with the proposed geometry-aware meta-learner (GAML). It is obvious that GAML outperforms the Vanilla-ML by a large margin.

PBR-MCMS Dataset We compare FFB6D, Vanilla-ML and GAML on intra- and cross-category levels. The full statistical summary can be found in Table 4.3. In general, the ADD metric is used for non-symmetric objects and ADD-S (Hinterstoisser et al., 2013) for symmetric objects. Since the matching

	Vanilla-ML		GAML	
Category	L1 Loss	ADD-0.1d	L1 Loss	ADD-0.1d
Airplane	0.41	80.6	0.33	**87.2**
Bag	0.34	85.2	0.30	**87.1**
Basket	0.83	49.7	0.62	**65.1**
Bathtub	0.46	79.3	0.34	**88.6**
Bed	0.72	60.1	0.57	**72.0**
Bench	0.95	56.7	0.63	**72.3**
Birdhouse	0.35	83.2	0.28	**89.7**
Bookshelf	0.48	79.5	0.42	**80.9**
Cabinet	0.39	83.5	0.31	**89.7**
Car	0.31	**92.9**	0.28	**92.9**
Camera	0.57	65.0	0.46	**73.4**
Chair	0.57	62.8	0.42	**80.9**
Helmet	0.46	69.8	0.43	**80.9**
Motorcycle	0.29	92.6	0.25	**94.7**
Mug	0.25	91.9	0.24	**93.3**
Pillow	0.76	54.7	0.58	**81.5**
Table	0.80	61.1	0.54	**76.2**
Train	0.41	80.6	0.36	**86.2**
Vessel	0.62	66.9	0.53	**70.9**
Washer	0.36	85.4	0.28	**91.4**
Bus*	0.46	83.0	0.42	**85.4**
Cap*	0.71	46.6	0.64	**54.2**
Laptop*	1.02	18.8	0.73	**48.8**
Piano*	0.86	47.1	0.75	**50.7**
Remote*	0.53	53.5	0.52	**56.1**
Intra-Categ.	0.52	74.2	0.41	**81.9**
Cross-Categ.	0.72	50.3	0.62	**59.0**
All	0.56	69.4	0.45	**77.2**

Table 4.4: Multi-category evaluation on Toy-MCMS dataset

between points is ambiguous for some poses, ADD-S computes the mean distance based on the minimum point distance:

$$\text{ADD} - \text{S} = \frac{1}{m} \sum_{x_1 \in O} \min_{x_2 \in O} \|(Rx + t) - (R^* x + t^*)\| . \tag{4.11}$$

	Airplane	Chair	Car	Laptop	Sofa
FFB6D	60.0	52.0	36.7	48.0	49.3
GAML	**89.8**	**80.0**	**56.9**	**85.0**	**69.8**

Table 4.5: Comparison between GAML and fine-tuned FFB6D on PBR-MCMS using ADD metric.

Fig. 4.6 visualizes some test examples for qualitative comparison. Next, we compare our meta-learner to FFB6D on the PBR dataset. Table 4.3 shows that our model generalizes well while FFB6D cannot directly transfer to novel objects. For a fair comparison, we further train FFB6D on the PBR dataset and fine-tune the pretrained model on each specific novel object with the same context images as given to GAML. Table 4.5 shows that our model still outperforms the fine-tuned FFB6D reliably and requires no trade-off between new and preceding tasks, whereas fine-tuning normally leads to a performance decrease on the previous tasks.

Occlusion-MCMS. Quantitative and qualitative results on Occlusion-MCMS are presented in Table 4.6 and Fig. 4.11. Strikingly, our approach achieves consistent and robust performance on occluded scenes even though training is conducted on non-occluded PBR-MCMS.

Network Architecture The detailed architecture model is shown in Table 4.7. We use ReLU as activation function after each FC layer except the output layer of segmentation decoder and global GNN decoder for keypoint offset prediction.

4.5.5 Ablation Study

Effect of K Neighbors in GNN. In Table 4.8, we study the effect of the k neighbors in the GNN. We run tests using five seeds and calculate the mean. Compared to $k = 3$, using all keypoints as neighbors can improve the robustness. We find this to be more crucial when training on a single category with limited object variations, where involving all keypoints gives more expressive spatial representation.

Effect of the Aggregation Module in CNP. In our work, CNP uses max aggregation instead of mean as used in the original paper (Garnelo et al.,

2018a). We further compare max aggregation with the cross-attention module proposed in Attentive Neural Processes (ANPs) (Kim et al., 2019) removing the self-attention part. Table 4.9 illustrates that CNP generalizes slightly better to novel tasks on both intra- and cross-category levels.

Robustness to Occlusion. To further illustrate the benefits coming from the geometry-aware estimator, we compare GAML with Vanilla-ML. The results



Figure 4.10: **Qualitative comparison between GNN and MLP decoder on Occlusion-MCMS.** Triangles and circles are the projected ground-truth and predicted keypoints respectively.

	Vanilla-ML		GAML	
Category	L1 Loss	ADD-0.1d	L1 Loss	ADD-0.1d
Airplane	0.93	43.2	0.69	**46.6**
Bag	0.46	29.4	0.54	**39.6**
Bathtub	0.66	29.6	0.76	**34.0**
Bed	0.48	**53.3**	0.71	44.2
Bench	0.66	40.4	0.67	**49.0**
Bookshelf	0.38	42.5	0.49	**42.7**
Bus	0.67	37.5	0.55	**47.5**
Cabinet	0.48	35.2	0.44	**55.2**
Camera	0.45	39.4	0.45	**40.4**
Cap	0.39	42.1	0.45	**43.1**
Chair	0.36	54.4	0.32	**55.6**
Earphone	0.44	24.1	0.56	**35.6**
Motorcycle	0.37	**64.8**	0.40	54.4
Mug	0.38	40.2	0.34	**52.9**
Table	0.62	23.3	0.72	**29.4**
Train	0.52	32.3	0.53	**36.2**
Vessel	0.60	**40.3**	0.90	33.4
Washer	0.46	37.7	0.44	**55.0**
Printer	0.56	27.5	0.57	**39.6**
Birdhouse*	0.44	23.5	0.51	**28.0**
Car*	0.52	42.9	0.50	**44.4**
Laptop*	0.47	26.0	0.41	**47.7**
Piano*	0.51	27.3	0.66	**32.5**
Sofa*	0.48	45.4	0.44	**57.9**
Intra-Categ.	0.52	38.7	0.55	**43.9**
Cross-Categ.	0.48	33.0	0.51	**42.1**
All	0.51	37.6	0.54	**43.5**

Table 4.6: Multi-category evaluation on Occlusion-MCMS dataset

in Table 4.3 show that our purposed GNN decoder significantly improves the performance and robustness on occluded scenes.

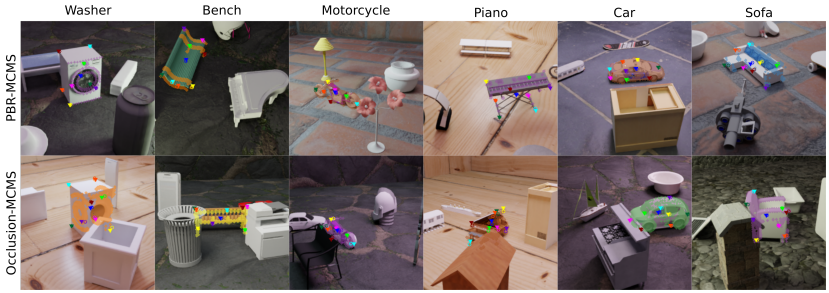


Figure 4.11: **Qualitative results on PBR- and Occlusion-MCMS datasets.** Triangles and circles are the projections of ground-truth and predicted keypoints respectively. Note that our model is trained only on PBR-MCMS but shows robust performance on Occlusion-MCMS.

Component	Layer	Output Size
Encoder	FC	128
	FC	128
	FC	128
Seg. Decoder	FC	128
	FC	128
	FC	128
	FC	2
Local GNN	FC	128
	FC	128
	FC	128
Global GNN	FC	128
	FC	128
	FC	3

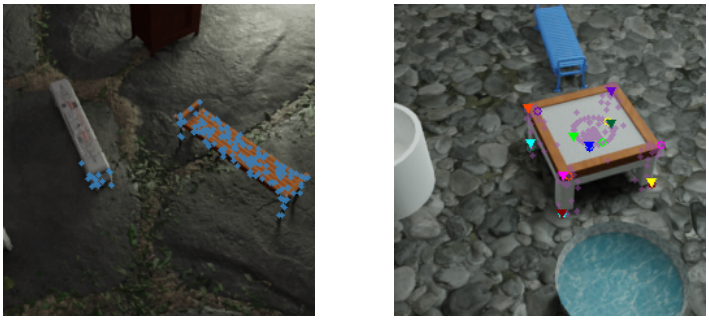
Table 4.7: GAML network architecture.

	k = 3	k = 8
Multi-Categ.	60.1	**61.9**
Single-Categ.	77.9	**83.3**

Table 4.8: ADD Results on PBR-MCMS using different number k of neighbors in GNN decoder.

	Intra-Categ.	Cross-Categ.	All
CNP	**81.9**	**59.0**	**77.2**
ANP	80.8	58.1	76.3

Table 4.9: ADD Results of CNP and ANP on Toy dataset.



(a) Feature ambiguity. (b) Symmetry ambiguity.

Figure 4.12: Limitations of the proposed method.

4.5.6 Limitations

We find two limitations of our method. First, we observe that in rare cases, our model suffers from *Feature Ambiguity* by struggling to disentangle feature variations, e.g., textures, shapes and lighting conditions. Sometimes it can be fooled by two similar objects which results in inaccurate segmentation (see Fig. 4.12a). Second, keypoint-based approaches suffer from *Symmetry Ambiguity*, especially on novel objects where the symmetric axis is unknown. Consequently, keypoint predictions around the symmetric axis can be mismatched and hamper the training (see Fig. 4.12b).

4.6 Conclusion

In this paper, we present a CNP-based meta-learner for cross-category level 6D pose estimation, which is capable of extracting and transferring latent representation on unseen objects from only a few samples. Besides, we

propose a simple yet effective geometry-aware keypoint detection module using GNN, which leverages the spatial connections between keypoints and improves generalization on unseen objects and robustness on occluded scenes. Furthermore, we create fully-annotated synthetic datasets called MCMS with various objects and categories, aiming to fill the vacancy for cross-category pose estimation.

5 SA6D: Self-Adaptive Few-Shot 6D Pose Estimator for Novel and Occluded Objects

5.1 Introduction

Accurately estimating the 6D pose of novel objects is critical for robotic grasping, especially for the tabletop setup. Prior work has investigated instance-level 6D pose estimation (Wang et al., 2019a; Peng et al., 2019; He et al., 2020; He et al., 2021), where the objects are predefined. Although achieving satisfying performance, these methods are prone to overfit to specific objects and suffer from poor generalization. Due to the high variety of objects with different colors and shapes in the real-world, it is impractical to retrain the model every time new objects come in, which is time-consuming and data inefficient. Recently, several approaches (Wang et al., 2019b; Chen et al., 2020b; Wang et al., 2020a; Chen et al., 2021b; Chen et al., 2020d; Chen and Dou, 2021; Fu and Wang, 2022; Zhang et al., 2022) have been proposed for category-level 6D pose estimation instead of specific objects, where they map different instances of each category into a unified representational space based on RGB or RGB-D features. However, conditioning on specific object categories limits the generalization to objects from novel categories with strong object variations. Meanwhile, some approaches (Li et al., 2022; Liu et al., 2022; Gao et al., 2022b; Park et al., 2020; He et al., 2022b; Shugurov et al., 2022) investigate generalizable 6D pose estimation as a few-shot learning problem, i.e., predicting the 6D pose of novel and category-agnostic objects given a few labeled reference images with the known pose of the novel object to define the object canonical coordinates. Although achieving promising results, these methods so far only work well on non-occluded and object-centric images, i.e., without the interference of other objects. This limits the

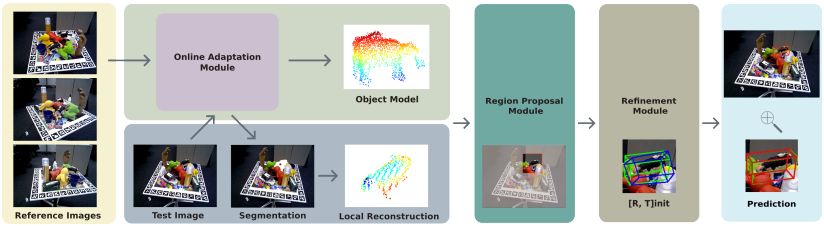


Figure 5.1: We present a generalizable and category-agnostic few-shot 6D object pose estimator using a small number of posed RGB-D images as reference. Compared to existing methods, our approach provides robust and accurate predictions on novel objects against occlusions without requiring retraining or any object information.

generalization to real-world scenarios with multiple objects in cluttered and occluded scenes. Furthermore, additional object information is required such as object diameter (Liu et al., 2022), mesh model (Shugurov et al., 2022; Li et al., 2022), object 2D bounding box (He et al., 2022b) or ground-truth mask (Park et al., 2020; Lin et al., 2021b), which is not always available for novel object categories. Our method aims to enable a fully generalizable few-shot 6D object pose estimation (FSPE) model.

In summary, we identify four primary challenges that are not adequately addressed by the current state-of-the-art methods (Liu et al., 2022; Park et al., 2020; He et al., 2022b; Shugurov et al., 2022): i) The category-agnostic 6D pose estimation in cluttered scenes with heavy occlusions is performing poorly. ii) The object-centric reference images from cluttered scenes are cropped by ground-truth segmentation or bounding box of the target object, which limits the generalization in real-world scenarios. iii) The prediction should not require any form of prior information from the unseen object such as diameter, mask or object mesh model. iv) The requirement of extensive reference images covering all different view-points is not practical.

To address the aforementioned challenges, we propose a robust self-adaptive 6D pose estimation approach called SA6D. As shown in Fig. 5.1, SA6D uses RGB-D images as input since i) depth images are normally easy to obtain along with RGB images in robotic setup, and ii) depth images reveal additional geometric features and can improve the robustness of prediction against occlusion. SA6D employs an online self-adaptative segmentation module to contrastively learn a distinguishable representation of the novel target object from the reference images of cluttered scenes. Meanwhile, a

canonical point cloud model of the object is constructed from the depth images. After the online adaptation, the segmentation module is capable to segment the target object from new images and construct the local point cloud from depth. Incorporating geometric features from the extracted point cloud, a region proposal module crops the test image by localizing the target object. With the cropped test image and the reference images, we employ Gen6D (Liu et al., 2022) to first predict an initial pose using visual input, followed by a refinement module using the induced geometric features. The use of Gen6D refiner before the usage of ICP alleviates the well-known local optima problem of ICP as the initial pose is already close to the ground-truth. Moreover, in contrast to Gen6D, SA6D does not require object diameter or object-centric reference images as input. Furthermore, our method is robust to occlusion since our model leverages the information from both the image and the constructed object geometry. As a final step, we employ the ICP algorithm again using the output of the refiner as an initial guess. Since the initial guess is in many cases already close to the ground truth, the second call of ICP does not suffer so much from local optima and therefore yields a more accurate estimation.

To the best of our knowledge, we are the first to provide a few-shot category-agnostic pose estimation approach that is capable to deal with heavy occlusions without requiring prior knowledge of novel objects. Our work focuses on the scenario with tabletop objects used for robotic manipulation. Our primary contributions are summarized as follows:

- SA6D is fully generalizable to new datasets without requiring any object or category information such as ground-truth segmentation, mesh model, or object-centric image. Instead, only a limited number of RGB-D reference images with the ground-truth 6D pose of the predicted object are needed.
- A self-adaptive segmentation module is proposed to learn a distinguishable representation of novel objects during inference.
- SA6D significantly outperforms current state-of-the-art methods against occlusion in real-world scenarios while trained entirely on synthetic data.

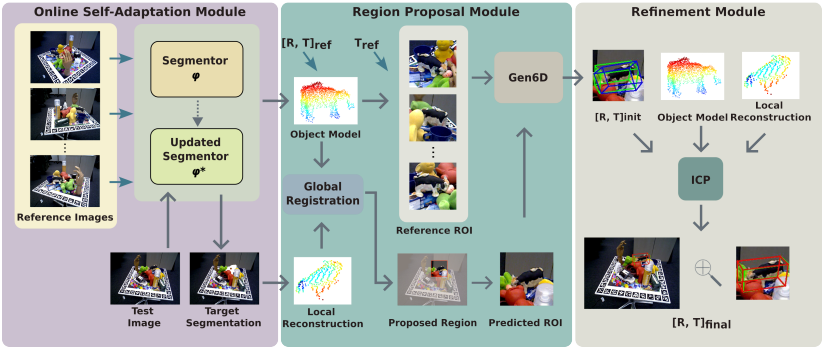


Figure 5.2: **Overview.** SA6D includes three modules: i) The *online self-adaptation module* discovers and segments the target object (*milk cow*) from a cluttered scene giving a few posed RGB-D images as reference. Subsequently, the canonical object point cloud model from the reference images and the local model from the test image are constructed based on the segments. ii) The *region proposal module* outputs a robust region of interest (ROI) of the target object against occlusion by incorporating visual and geometric features. A coarse 6D pose is then estimated by comparing the cropped test and reference images using Gen6D (Liu et al., 2022) and iii) further fine-tuned using ICP (Rusinkiewicz and Levoy, 2001).

5.2 Related Work

Category-Level 6D Object Pose Estimation. Methods for generalizable 6D object pose estimation can be divided into category-specific and category-agnostic models. For the category-specific estimation, Wang et al.(Wang et al., 2019b) first propose a canonical representation for all possible object instances within a category using Normalized Object Coordinate Space (NOCS). However, inferring the object pose by predicting only the NOCS representation is non-trivial given large object variations (Fan et al., 2021a). To tackle this problem, Tian et al.(Tian et al., 2020) account for intra-categorical shape variations by explicitly modelling the deformation from shape prior to the object model, while CASS (Chen et al., 2020b) generates 3D point clouds in the canonical space using a variational autoencoder (VAE) (Kingma and Welling, 2014). FS-Net (Chen et al., 2021b) proposes a shape-based model using 3D graph convolutions and a decoupled rotation mechanism to further reduce the feature sensitivity to the color variations. Wang et al.(Wang et al., 2020a) predict the relative 6D pose between two consecutive images using a category-based keypoint matching model. Chen et al.(Chen et al., 2020d)

employ a VAE-based generator to learn a categorical prior and update the prior with online rendering w.r.t. the test image. Recently, Fu et al.(Fu and Wang, 2022) facilitate the generalization by collecting a large-scale dataset with object-centric RGB-D videos called Wild6D. Based on Wild6D, Zhang et al.(Zhang et al., 2022) propose to learn the dense 2D-3D correspondences between the 2D image pixels and the categorical shape prior while the final 6D pose is computed by the least-square-fitting algorithm (Umeyama, 1991). Similar to our work, UDA-COPE (Lee et al., 2021b) employs self-supervised training while TTA-COPE (Lee et al., 2023) addresses the source-to-target domain gap using test time adaptation. Nevertheless, these methods require a manually defined categorical prior for training and therefore are limited to generalize across categories. In contrast, our method learns 6D pose estimation in a category-agnostic manner.

Category-Agnostic 6D Object Pose Estimation. Category-agnostic pose estimation can be formulated as a few-shot learning problem (Gao et al., 2022b): During inference, the model can generalize and predict the pose of novel objects given a few images with known poses as reference. LatentFusion (Park et al., 2020) and iNeRF (Lin et al., 2021b) employ the neural rendering technique (Mildenhall et al., 2020) to refine the predicted pose based on a latent representation obtained from the reference images while a segmentation of the object is required as input. FS6D (He et al., 2022b) extracts features from both the reference images and test images followed by a prototype matching algorithm to obtain the point-wise correspondences. OnePose (Sun et al., 2022) and OnePose++ (He et al., 2022a) build an object model from a single RGB video and employ feature mapping between the test image and the object model, which are not end-to-end and deviate from the few-shot domain. Furthermore, all the aforementioned methods require object-centric images for either reference or test images. In contrast, Gen6D (Liu et al., 2022) is applicable for cluttered scenes where both reference and test images contain multiple objects, although it struggles with occlusion. Our work is inspired by Gen6D and exploits the geometric information of the target object to enable robust prediction against occlusion. This geometric information allows us to be more robust against clutter and occlusion. In addition, it enables us to lift the requirement of Gen6D to know the object diameter as this information can be directly estimated from the geometry of the object.

Unseen Object Segmentation. Recently, several approaches have been proposed to close the gap between learning unseen object segmentation from synthetic datasets and real world datasets (Gouda et al., 2022; Danielczuk

et al., 2019; Xie et al., 2019; Gao et al., 2023a). Xie et al.(Xie et al., 2021) propose to learn a two-stage segmentation model by separately leveraging RGB and depth information in a hierarchical way, where the model is fully trained on the synthetic data. UCN (Xiang et al., 2020) proposes to learn from RGB and depth images jointly and generate pixel-wise feature embeddings. To enable a generalizable 6D pose estimation in cluttered scenes, we design a self-adaptive module to generate target-object-oriented segmentation model using UCN as a base segmentor.

5.3 Preliminaries

In this section, we present an overview of Gen6D (Liu et al., 2022) and highlight the key enhancements of our proposed method over Gen6D, which consists of three components, namely detector, selector, and refiner. Given RGB images, the detector initially identifies the target object in the test image by comparing the visual similarity between the patches of test image and the object-centric reference images cropped based on the known pose and object diameter. Afterwards, the detector outputs a score map based on the visual similarity along with a predicted region of interest (ROI). Consequently, the selector selects the most similar reference image based on the predicted ROI and uses the reference pose as an initial guess, which will be updated by a learning-based refiner to obtain the final pose. However, since detector and selector rely purely on the visual similarity between test and reference images, Gen6D requires large and diverse reference images to cover the appearance of all different viewpoints and cannot deal with occlusions as the authors claimed. In contrast, SA6D solves the aforementioned problems by incorporating depth information to reveal the geometric features which are used to improve the scene understanding against occlusions. Notably, SA6D improves the efficiency with fewer reference images compared with Gen6D.

5.4 Method

SA6D is comprised of three parts, i.e., an online self-adaptation module (OSM) for target object segmentation from cluttered scenes, a region proposal module (RPM) to infer the region of interest (ROI) for the target object against

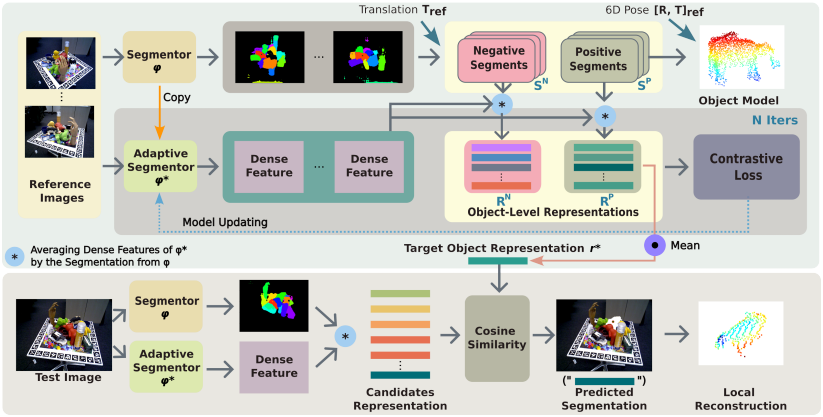


Figure 5.3: **Online self-adaptation module**. A pretrained segmentor φ is first applied on reference images to predict segmentations. Meanwhile, an adaptive segmentor $\varphi*$ is initialized from φ. With the ground-truth translation of the target object in the reference images $\mathrm{T}_{\mathrm{ref}}$, the object center can be reprojected to the image. For each reference image, one segment is chosen as a positive sample if it includes the reprojected object center while the remaining segments are considered as negative samples. Subsequently, an object-level representation of each segment is computed by averaging the pixel-wise dense features from $\varphi*$. A contrastive loss is then applied over the positive and negative object representations and updates $\varphi*$ iteratively. After adaptation, $\varphi*$ generates the target object representation $r*$ by averaging over all positive representations from reference images. Given a test image, we obtain the representation of each candidate segment in the same way and compute the cosine similarity between each candidate and $r*$, where the most similar candidate is chosen as the segment of the target object. Meanwhile, the canonical and local object models are computed based on the segments and depth images.

occlusion, and a refinement module (RFM) to refine the predicted 6D pose of the target object using both visual and the inferred geometric features. The proposed pipeline is shown in Fig. 5.2.

5.4.1 Online Self-Adaptation Module

To alleviate the dependence on prior object information and object-centric reference images, and improve the prediction against occlusions, it is essential to build a model which can discover the objects from the cluttered scene and identify the occluded target object from other objects. To achieve this, we design a self-adaptive segmentation module which is updated only during

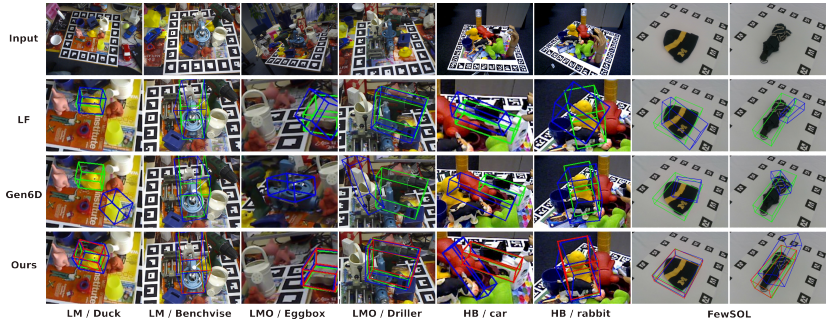


Figure 5.4: **Qualitative results.** The green bounding box denotes the ground-truth pose and blue denotes the prediction. In SA6D, blue denotes prediction before refinement while red is the final prediction.

inference in a self-supervised manner given posed reference images, where no retraining is needed.

In particular, we employ the segmentation model from Xiang et al.(Xiang et al., 2020) as our base segmentor φ, which segments all instances of each image by clustering the pixel-wise features using the mean-shift algorithm (Kobayashi and Otsu, 2010). In our work, we only use the coarse segmentor without the zoom-in refinement. Examples of predicted segmentation are shown in Fig. 5.3. Given the ground-truth translation $\mathrm{T}^{i}_{\mathrm{ref}} \in \mathbb{R}^3$ of the target object in the i-th reference image, we can reproject the object center on the image plane and select the segment, which includes the reprojected object center, as a positive target segment s^P_i, while the remaining segments are considered as negative segments $S^N_i = \{s^N_1, ..., s^N_K\}$. K denotes the number of negative segments in each reference image. Given M reference images, we obtain a positive set of target object segments $S^P = \{s^P_1, ..., s^P_M\}$ and a negative set of segments $S^N = \cup^M_{i=1} S^N_i$.

The adaptive segmentor φ^* is initialized by copying the parameters of φ and used to generate distinguishable representations between the target object and the remaining objects, not for generating the segmentation. The positive and negative object-level representations, R^P and R^N are computed by averaging the pixel-wise dense features of φ^* while grouping by each segment from S^P and S^N. Based on the positive and negative representation sets R^P and R^N, φ^*

is updated iteratively using a contrastive loss (Chen et al., 2020c). Specifically, for each positive pair $r_i^P, r_j^P \in R^P$, the loss is computed as

$$l_{ij} = -\log \frac{\exp(\mathrm{sim}(r_i^P, r_j^P)/\tau)}{\sum_{r' \in R^N \cup \{r_j^P\}} \exp(\mathrm{sim}(r_i^P, r')/\tau)}, \tag{5.1}$$

where τ is a hyper-parameter and set to 0.07, sim denotes the cosine similarity between two representations. The loss is summed over all combinations of the positive pairs from R^P and back-propagated through φ^*. After adaptation, φ^* generates the target object representation $r*$ by averaging over all positive representations $r* = \mathrm{mean}(r_1^P, ..., r_M^P)$. Note that R^P and R^N are updated along with φ^* simultaneously.

Given a test image, the candidate segments are obtained in the same way from φ followed by computing the representations from $\varphi*$. By comparing the cosine similarity between the candidates and the target object representation r^*, the candidate with the highest similarity score is selected as the segment of the target novel object.

Object Model Reconstruction. We reconstruct the object model from the reference images by computing the partial point clouds for each positive segment and transfer them to the canonical coordinates given the ground-truth 6D pose $[R, T]_{\mathrm{ref}}$. The combination of partial point clouds obtained from the reference images assembles a coarse geometric model of the object. For inference, we obtain a partial point cloud model (local reconstruction) using the predicted target segment from the test image.

5.4.2 Region Proposal Module

The region proposal module combines 2D image features and the geometric features from induced point cloud model. The idea is to improve the robustness against clutter and occlusion in comparison to Gen6D (Liu et al., 2022). The region of interest (ROI) denotes a squared area where the target object is located. While Gen6D can predict the ROI of novel objects, it suffers from occlusion since the prediction depends purely on the visual similarity between the reference and test images. For cluttered scenes as shown in Fig. 5.2, the reference and test ROI candidates include other objects due to occlusion. Hence, the predicted similarity score map used by Gen6D exhibits several

spurious peaks as the target object can not be differentiated from other objects contained in the ROI. Thus, selecting the highest score from the entire score map is prone to selecting a wrong target object (an example on LM/duck is shown in Fig. 5.4).

Instead of requiring the object diameter in Gen6D, we estimate the object diameter $\hat{d}$ from the reconstructed object model. Furthermore, in contrast to Gen6D, which first predicts the ROI of test image to estimate the in-depth distance T^z, we proceed reversely by utilizing the geometric features from the partial point cloud model. With the reconstructed object point cloud model from inference images and the partial point cloud model from the test image, we estimate an initial translation T_{init} using global registration, i.e., using RANSAC to first match the points between the two models followed by the fast point registration proposed by Zhou et al.(Zhou et al., 2016). With $\hat{d}$ and the estimated depth T^z_{init}, the ROI scale is computed by $s = \hat{d} * f/T^z_{init}$ where f is the camera focal length. Meanwhile, the ROI position $[u, v]$ is calculated by $u = T^x_{init} * f/T^z_{init}$ and $v = T^y_{init} * f/T^z_{init}$. As shown in Fig. 5.2, we can accurately predict the ROI against occlusion using the geometric information without interference from the environment. Similar to the test image, the cropped reference images are obtained given the reconstructed object model and ground-truth pose $[R, T]_{\text{ref}}$. Subsequently, we employ the pretrained Gen6D detector to predict an initial pose based on the visual input. The predicted rotation of ICP in this step is inaccurate as we found that ICP often gets stuck in local optimum, which yielded poor results in our preliminary experiments.

5.4.3 Refinement Module

Based on the reconstructed geometric features of the object, we employ Iterative Closest Point (ICP) (Rusinkiewicz and Levoy, 2001) and use the output of Gen6D as an initial pose, which helps alleviate the problem of local optimal in ICP. In our experiments, we found this is particularly useful for rotation estimation where the global point registration struggles.

Method	GT-Mask	Ref. Num	LineMOD (Hinterstoisser et al., 2013)							LineMOD-OCC (Brachmann et al., 2014)							HomeBrewedDB (Kaskman et al., 2019)				
			eggbox	duck	benchvise	cam	cat	glue	avg.	driller	eggbox	duck	glue	ape	can	avg.	cow	flange	car	rabbit	avg.
Gen6D (Liu et al., 2022)	✗	20	0.63	0.30	0.45	0.29	0.25	0.26	0.36	0.09	0.02	0.07	0.03	0.12	0.21	0.09	0.36	0.15	0.15	0.52	0.30
SA6D (ICP only)	✗	20	0.53	0.31	0.37	0.25	0.21	0.17	0.31	0.17	0.16	0.10	0.08	0.14	0.22	0.14	0.23	0.17	0.20	0.44	0.26
SA6D (wo/ RPM)	✗	20	0.63	0.47	0.50	0.37	0.36	0.38	0.45	0.19	0.15	0.13	0.10	0.17	0.28	0.17	0.37	0.19	0.21	0.61	0.35
SA6D (wo/ RFM)	✗	20	0.57	0.36	0.45	0.34	0.29	0.26	0.38	0.15	0.08	0.09	0.04	0.10	0.28	0.12	0.39	0.12	0.20	0.55	0.32
SA6D	✗	20	**0.73**	**0.73**	**0.55**	**0.50**	**0.47**	**0.72**	**0.62**	**0.45**	**0.26**	**0.30**	**0.21**	**0.32**	**0.53**	**0.35**	**0.62**	**0.35**	**0.33**	**0.78**	**0.52**
Gen6D	✗	64	0.74	0.40	**0.73**	0.65	0.65	0.53	0.62	0.27	0.09	0.23	0.03	0.11	0.50	0.21	0.38	0.06	0.49	**0.78**	0.43
SA6D	✗	64	**0.80**	**0.84**	**0.73**	**0.80**	**0.84**	**0.75**	**0.79**	**0.44**	**0.41**	**0.38**	**0.31**	**0.33**	**0.66**	**0.42**	**0.72**	**0.49**	**0.72**	0.69	**0.66**
LF (Park et al., 2020)	✓	20	0.61	**0.61**	0.68	0.65	**0.72**	**0.78**	0.67	0.28	0.01	0.00	0.18	**0.45**	0.17	0.18	0.33	0.00	0.00	0.16	0.12
SA6D (wo/ RFM)	✓	20	0.56	0.32	0.54	0.30	0.26	0.29	0.38	0.10	0.06	0.08	0.04	0.14	0.24	0.11	0.41	0.13	0.15	0.54	0.31
SA6D	✓	20	**0.68**	0.58	**0.80**	**0.73**	**0.72**	**0.78**	**0.72**	**0.33**	**0.26**	**0.29**	**0.30**	0.19	**0.45**	**0.30**	**0.58**	**0.17**	**0.44**	**0.76**	**0.49**

Table 5.1: Evaluation of ADD-0.1d on LineMOD, LineMOD-OCC, and HomeBrewedDB datasets against category-agnostic baselines.

Method	ADD-0.1d	ADD-0.3d	ADDs-0.1d	ADDs-0.3d
LF (Park et al., 2020)	0.1162	0.1738	0.1299	0.1907
Gen6D (Liu et al., 2022)	0.3571	0.6399	0.6399	0.7530
SA6D (wo/ RFM)	0.4018	0.7292	0.6964	**0.8780**
SA6D	**0.5595**	**0.7887**	**0.8393**	**0.8780**

Table 5.2: Evaluation on FewSOL (Padalunkal et al., 2022) dataset over 336 objects using 8 reference images.

Method	$IOU_{0.5}$	5°2cm	5°5cm	10°5cm
CASS (Chen et al., 2020b)	0.01	0.0	0.0	0.0
Shape-Prior (Tian et al., 2020)	0.33	0.03	0.04	0.14
DualPoseNet (Lin et al., 2021a)	0.70	0.18	0.23	0.37
RePoNet (Fu and Wang, 2022)	**0.71**	0.30	0.34	**0.43**
SA6D	0.65	**0.37**	**0.40**	0.42

Table 5.3: Evaluation on Wild6D (Fu and Wang, 2022) dataset against category-level baselines.

5.5 Experiments

We employ two baselines that are most relevant to our work on category-agnostic unseen objects, namely LatentFusion (LF) (Park et al., 2020) and Gen6D (Liu et al., 2022). Besides the input image, LatentFusion requires ground-truth segmentation of the target object while Gen6D requires the object diameter as input. FS6D (He et al., 2022b) also requires object-centric reference images with ground-truth segmentations for cluttered scenes. In contrast, our method does not require any additional information. We also compare SA6D against category-level SOTA methods which use RGB-D as input. It is good to note that SA6D is not trained on any specific category

while all category-level baselines are trained and tested on the objects within the same category. We exclude another related work, namely FS6D in our comparisons since its code is not published. Furthermore, FS6D does not work on cluttered image without ground-truth segmentations.

5.5.1 Datasets and Metrics

Evaluation datasets. We use four datasets for evaluation against category-agnostic methods, namely LineMOD (Hinterstoisser et al., 2013), LineMOD-OCC (Brachmann et al., 2014), HomebrewedDB (Kaskman et al., 2019) and FewSOL (Padalunkal et al., 2022). None of these datasets is used during the training phase. LineMOD (LM) includes annotations of 15 test objects in cluttered scenes without occlusion while LineMOD-OCC (LMO) and HomebrewedDB (HB) have heavy occlusion. FewSOL includes 336 real-world objects and 9 RGB-D images for each object from different viewpoints, where we randomly select 8 images as references and test on the remaining image. FewSOL includes large-scale object variations but without occlusion or clutter. Furthermore, we compare against category-level methods on Wild6D (Fu and Wang, 2022) which is a RGB-D video dataset including 5 object categories.

Training datasets. The base segmentor is trained fully on the synthetic Tabletop Object Dataset (TOD) generated by Xie et al. (2019) and the pretrained Gen6D model uses the rendered images from ∼2000 ShapeNet (Chang et al., 2015) models and 1023 Google Scanned Object by Wang et al.(Wang et al., 2021b). Note that only synthetic datasets are used for training.

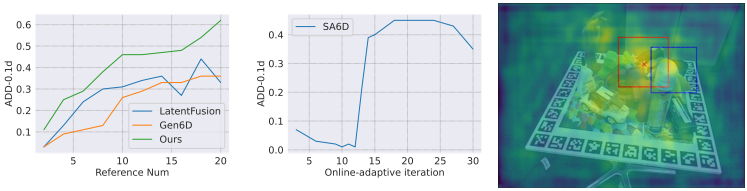
Evaluation metrics We use the average distance (ADD) (Hinterstoisser et al., 2013) to evaluate the object points after being transformed by the ground-truth and predicted pose. ADD-0.1d (ADD-0.3d) denotes the accuracy of the predictions with an average distance below 10% (30%) of the object diameter. ADD-S is used in FewSOL dataset due to the large amount of symmetric objects, where the average distance is computed based on the closest point. For comparison against category-level methods, we employ the same BOP (Sundermeyer et al., 2023) metrics as used in RePoNet (Fu and Wang, 2022). n°, $m\ cm$ denotes the accuracy when both rotation error is less than n° and translation error is less than $m\ cm$.

5.5.2 Results and Discussion

Comparison against category-agnostic methods. As shown in Table 5.1, although baselines show promising results on LineMOD dataset, they perform poorly and cannot generalize on occluded datasets (LineMOD-OCC and HomeBrewedDB). In contrast, without requiring ground-truth segmentation or object diameter, SA6D significantly increases the performance over all datasets, especially under the circumstances where fewer reference images are given or the objects are occluded. Furthermore, without ground-truth segmentation, SA6D still outperforms LatentFusion on occluded datasets by a large margin. Table 5.2 shows SA6D is able to generalize on large object variations while LatentFusion cannot generalize even without occlusion. We find that LatentFusion requires high-quality depth images and more reference images to reconstruct the latent representation, and works poorly on flat objects. Examples are shown in Fig. 5.4. Furthermore, SA6D demonstrates superior performance against Gen6D even without using the geometric features in the refinement module (RFM). The reason is, Gen6D struggles with localizing the target object in FewSOL dataset since the evaluated objects in FewSOL dataset are close to the camera and occupy a larger area than the dataset used for training, indicating a poor generalization of Gen6D on out-of-distribution data. In contrast, the region proposal module (RPM) used in SA6D alleviates the problem.

Ablation study on components of SA6D. To evaluate the importance of different components in SA6D, we conduct an ablation study by removing the region proposal module (*wo/ RPM*), the refinement module (*wo/ RFM*), and remove both by only using the global point cloud registration between the reconstructed global and local object model (*ICP only*). The results are shown in Table 5.1. The performance decrease of *ICP only* indicates that classic point cloud registration between partial and global point clouds is often stuck at a suboptimal position. The performance drop on *wo/ RFM* demonstrates the importance of the induced geometric features and the notable performance drop on LineMOD-OCC and HomeBrewedDB without using RPM shows its crucial role against occlusion. For cluttered scenes as shown in Fig. 5.2, the reference and test ROI candidates include other objects due to occlusion. In Fig. 5.5c, we show an example of a test image and visualize the pixel-wise visual similarity between reference and test images on top, where higher brightness indicates higher visual similarity. RPM is capable to localize the target object (cow) while Gen6D depends purely on visual similarity and selects a wrong object.

Accuracy vs Reference Number. We report the ADD-0.1d w.r.t. the number of reference in Fig. 5.5a on HomebrewedDB/cow. Increasing the number of reference images overall benefits all methods except LatentFusion, which sometimes shows degradation in performance because a heavily occluded reference image can drastically alter the implicit representation in the latent space due to the employed online rendering. Notably, SA6D performs consistently better than baselines and shows reasonable prediction with a one-shot reference image.



(a) Acc. vs reference number (b) Acc. vs online iterations (c) Region proposal

Figure 5.5: Analysis of the number of (a) reference and (b) online iterations. (c) An example of proposed ROI from SA6D (red) and Gen6D (blue), the red cross denotes the target object.

Analysis of Online Self-Adaptation. The performance of SA6D w.r.t. the number of iterations in the online self-adaptation module is shown in Fig. 5.5b on LineMOD-OCC/driller. At the beginning, SA6D performs poorly since the segmentor $\varphi*$ cannot learn and differentiate the representation of the target object from others, which also leads to a performance decrease. After 12 iterations, with a learned distinguishable target object representation, SA6D gains significant improvement. With more iterations, the performance decreases again since the updated segmentor $\varphi*$ starts overfitting to the reference images. We prevent overfitting by automatically stopping updating $\varphi*$ with a defined threshold to the contrastive loss in Eq. (5.1). In our experiments, we set the threshold to 0.01 over all datasets.

Comparison against category-level methods. Table 5.3 demonstrates the comparison against category-level SOTA methods on Wild6D dataset. Although SA6D is not trained specifically on each category, it overall achieves competitive performance and even outperforms baselines using the strict criteria (*5°2cm*), which indicates SA6D can predict more accurate poses than all baselines. In the appendix, we also visualize the predictions of SA6D and RePoNet (Fu and Wang, 2022) for comparison.

Discussions. We find that our online self-adaptation module is robust against false positive samples and is able to learn a correct target-oriented representation and shows remarkable performance against severe occlusion and truncation. Moreover, SA6D provides explainable confidence scores by computing the cosine similarity among the candidate segments. We also tried replacing ICP with a learning-based method, namely RPM-Net (Yew and Lee, 2020). However, the prediction is always stuck at the sub-optimum. Nevertheless, we believe SA6D can be further improved with future development in the area of segmentation and learning-based point could registration, which is not the main focus of this work. The inference running time on a single image costs ~0.93s in total on Nvidia Tesla V100 for SA6D.

Gen6D without ground-truth object diameter. In Table 5.4, we demonstrate that using the object diameter as input is a strong prior knowledge which limits the generalization of Gen6D, by fixing the diameter over all objects with two different values, namely $10\ cm$ and $50\ cm$. Based on the diameter, Gen6D generates different ROI scales, i.e., the small diameter generates a small ROI focusing only on part of the object while the large diameter leads to a large ROI which includes several objects. We conduct the evaluation for each dataset and report the averaged results over the objects in Table 5.4. Without ground-truth diameter, Gen6D cannot generalize well on any of the datasets.

Diam. (m)	Dataset			
	LM	LMO	FewSOL	HB
0.1	0.06	0.06	0.04	0.10
0.5	0.16	0.05	0.00	0.19
GT	**0.35**	**0.08**	**0.36**	**0.30**

Table 5.4: Evaluation on Gen6D with different object diameters as prior knowledge. Results are averaged over objects for each dataset.

Compare ICP with learning-based point cloud registration algorithm We show a few predicted examples of a state-of-the-art learning-based point cloud registration model, namely RPM-Net, on the LineMOD-OCC/driller in Fig. 5.6. RPM-Net is prone to the local optimal position for 6D pose estimation, especially for rotation. In our experiments, ICP is more robust to unseen objects.

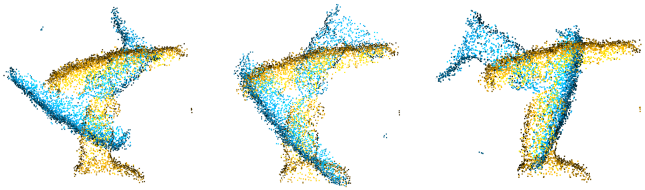


Figure 5.6: Examples of using RPM-Net for point cloud registration instead of ICP. The yellow point cloud denotes the reconstructed object point cloud model and the blue one denotes the prediction after transformation using the predicted pose from RPM-Net. Better overlapping between two point clouds indicates better performance. RPM-Net cannot generalize on unseen objects and is prone to get stuck in local optima.

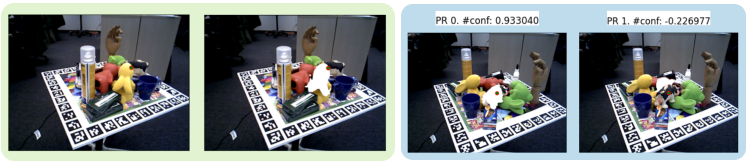
Results on dynamic scenes. We have conducted the evaluation on dynamic scenes where we sample 20 images from LineMOD-OCC (LMO) as reference images and evaluate query images from LineMOD (LM) which includes different background and objects. Thus, the reference and query images are from different scenes with changing light conditions and configuration of surrounding objects. The results are shown in Table 5.5. SA6D achieves competitive results compared to the original setup (first line in the table). Note that this is still not a fair comparison for our method since we use reference images with occluded target objects, which makes it more difficult to reconstruct the object model. However, this experiment demonstrates the capability of our method on query images from novel scenes.

Ref. image	Query image	eggbox	duck	cat	glue	avg.
LM	LM	0.73	0.73	0.47	0.72	0.66
LMO	LM	0.70	0.71	0.46	0.72	0.64

Table 5.5: Evaluation on LineMOD using LineMOD-OCC as reference.

SA6D is robust to false positive samples in reference Using reprojected object center to select positive segments sometimes leads to a false positive sample given the target object center is occluded. An example is shown in Fig. 5.7a, in which a wrong segment (*yellow rabbit*) is selected as a positive sample for the target object (*milk cow*). However, we find that our online self-adaptation module is robust against false positive samples and is able to learn a correct target-oriented representation. Moreover, SA6D provides

explainable confidence scores by computing the cosine similarity between each segment representation and the target object representation. Fig. 5.7b shows an example of the predicted target (*milk cow*) segments with reasonable induced confidence score though wrong positive samples are given in the reference set.

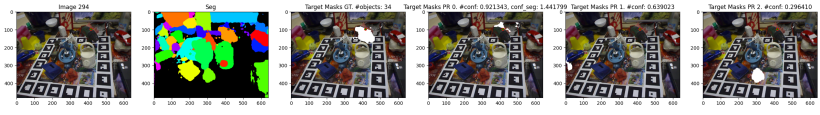


(a) False positive sample. (b) Robust prediction.

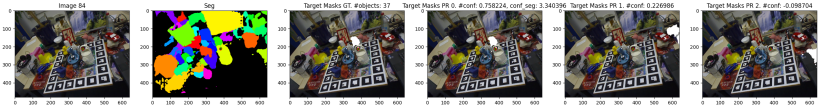
Figure 5.7: Discussion. (a) A false positive sample is selected given the reprojected center of the target object (*milk cow*) is occluded by another object (*yellow rabbit*). Nevertheless, (b) SA6D provides robust prediction with explainable confidence scores.

SA6D demonstrates remarkable performance against severe occlusion and truncation We show superior performance of SA6D on challenging scenes with severe occlusion and truncation in Fig. 5.8, where the input images, predicted segmentations from the base segmentor φ, ground-truth segmentation of target object based on the reprojected object center, and three predicted candidates with the highest predicted confidence scores are given on each column from left to right. The selected segments are marked in white color. The confidence score *conf* denotes the cosine similarity between the candidate segment representation and the target object representation r^*. The *conf_seg* is computed by dividing the confidence scores between the first and second most similar segment candidates w.r.t. the target object representation. Thus, it can be used in crucial scenarios if the prediction is uncertain among different segments. Note that in Fig. 5.8a, our method is able to differentiate the target object segment while the provided ground-truth segmentation points to a wrong segment due to the center of target object is occluded.

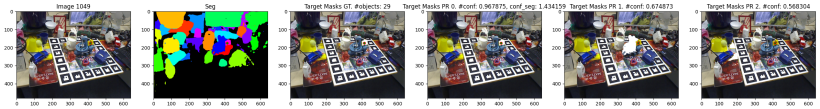
Robust and explainable confidence score of the online self-adaptation module We show more results on the predicted segmentation of the online self-adaptation module in Fig. 5.9 on LineMOD dataset, Fig. 5.10 on LineMOD-OCC, and Fig. 5.11 on HomebrewedDB. Some candidates in Fig. 5.11 with white background indicate the background segments are selected.



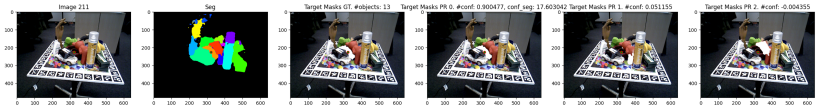
(a) cat



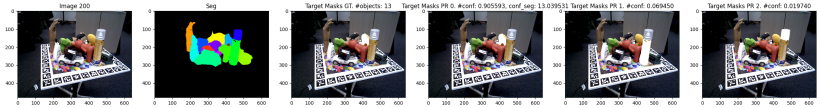
(b) driller



(c) duck



(d) flange

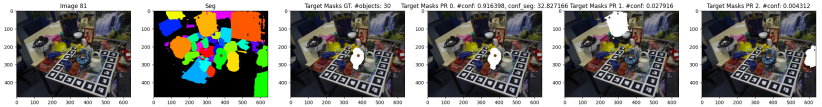


(e) yellow rabbit

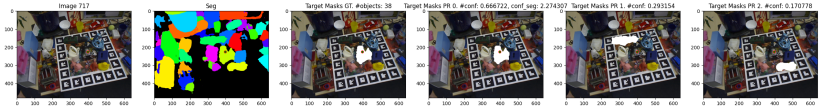
Figure 5.8: Online-Adaptation results on challenging scenes against severe occlusion and truncation. Three candidates with the highest confidence scores are visualized in order.

More Qualitative Results We show more qualitative results of the 6D pose prediction and compare our method with Gen6D on LineMOD dataset in Fig. 5.13, LineMOD-OCC in Fig. 5.14, HomebrewedDB in Fig. 5.15 and FewSOL in Fig. 5.16. The comparison on Wild6D dataset between SA6D and category-level SOTA method RePoNet is shown in Fig. 5.17.

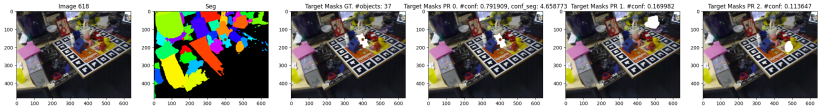
Failure Cases We show the examples in Fig. 5.12 where using ICP leads to a worse prediction than without using ICP in the refinement module. Results are evaluated on the FewSOL dataset, indicating future work on generalizable and learnable point cloud registration is essential to further improve the performance.



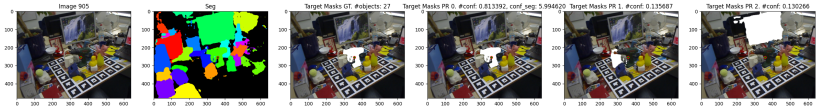
(a) benchvise



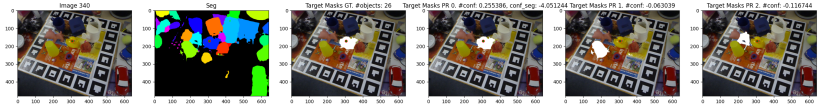
(b) cam



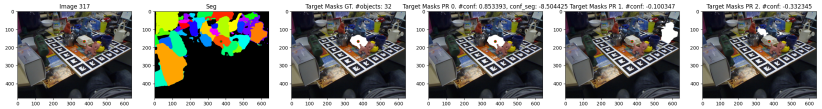
(c) cat



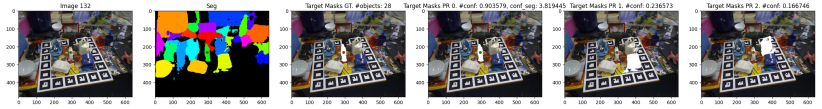
(d) driller



(e) duck

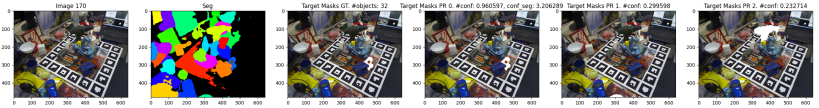


(f) eggbox

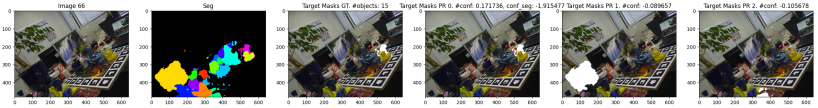


(g) glue

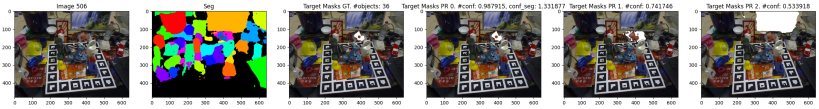
Figure 5.9: Robust prediction of target segmentation on LineMOD. Three candidates with the highest scores are visualized in order.



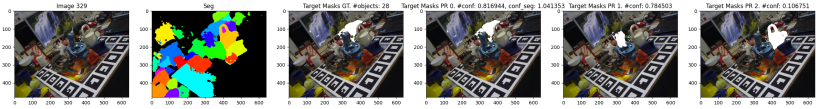
(a) ape



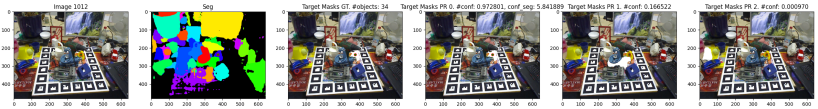
(b) can



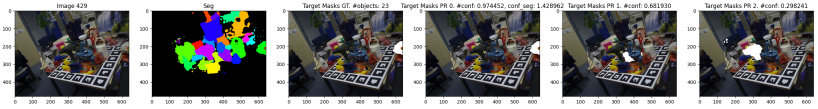
(c) cat



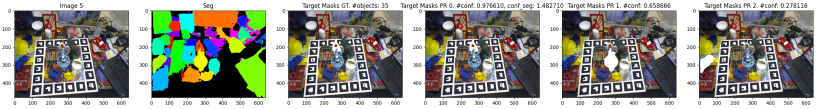
(d) driller



(e) duck

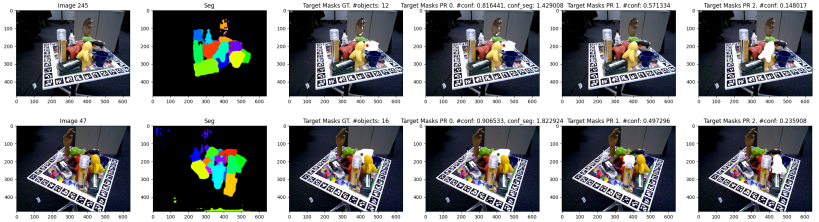


(f) eggbox

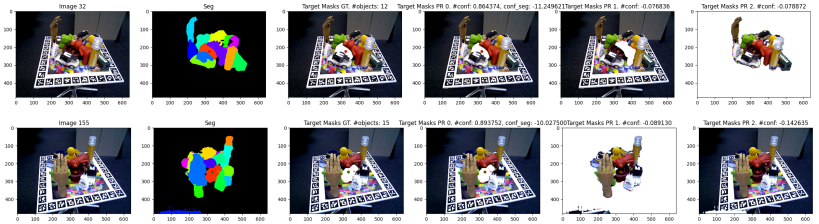


(g) glue

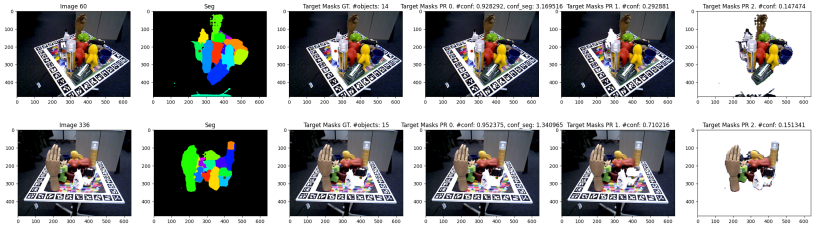
Figure 5.10: Robust prediction of target segmentation on LineMOD-OCC. Three candidates with the highest scores are visualized in order.



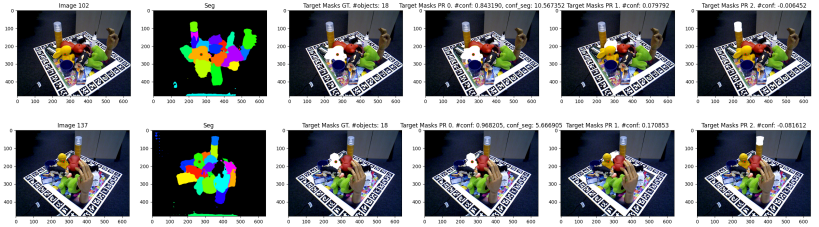
(a) cow



(b) flange



(c) car



(d) yellow rabbit

Figure 5.11: Robust prediction of target segmentation on HomebrewedDB. Three candidates with the highest scores are visualized in order.

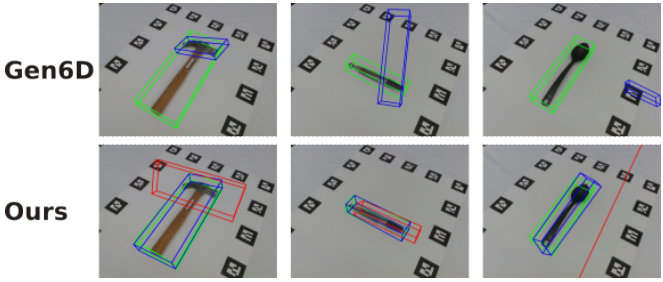


Figure 5.12: Failure cases. Using ICP in the refinement module leads to a worse prediction than the initial prediction. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

Selection of Reference Images Regarding the selection of the reference images on the LM, LM-O and HB datasets, the original Gen6D selects 64 reference images from a predefined set of images with farthest point sampling (FPS) to make sure that the view distributes evenly among the reference images. We follow the same setup when all models are evaluated with 64 reference images. However, it is not efficient to sample 64 images and it is often the case that the reference images are not distributed evenly in the real-world. Therefore, we also evaluate all methods by randomly selecting 20 reference images from the dataset, which significantly increases the task difficulty but is more realistic and plausible because it is not always obtainable to collect reference images that could cover all viewpoints.

Comparison with FS6D and Model-Based Models Similar to LatentFusion, FS6D (He et al., 2022b) also requires object-centric reference images with ground-truth segmentations for cluttered scenes. Considering that its code is not published and we could not reproduce its results, we hence excluded FS6D in our comparisons. Meanwhile, we cannot add the model-based methods (He et al., 2021; He et al., 2020; Peng et al., 2019; Wang et al., 2019a) into comparison due to their limitation, i.e., the model-based methods can only be applied on the specifically trained object and cannot work in our setup where the results are evaluated on new objects. Also, it is unfair to compare them with our work if we train the model-based methods on the new objects. Moreover, the FewSOL dataset contains only 9 images for each object, which is insufficient to train the model-based methods. Considering all these limitations

of the model-based methods, it is also one of our motivations to work on this paper.

Effort of Annotation Compared with Prior Work The annotation of a limited number of reference images requires human effort. However, the effort of annotation is also essential in prior work (Chen et al., 2021b; He et al., 2021; He et al., 2020; Peng et al., 2019; Tian et al., 2020; Wang et al., 2019a; Wang et al., 2019b) where thousands of annotated images are required for every single object or category. Category-agnostic methods such as our method tremendously reduce human effort by requiring only a small number of annotations. Still, similar to Gen6D and LatentFusion, it is necessary to have a small number of posed reference images for an unseen object to set the canonical object coordinates to further determine the object rotation w.r.t. the camera. Importantly, our method does not require any additional effort compared to existing methods.

Practical Use Case Our method can be used in the lifelong robot item picking/sorting in industry. Each time when a new product comes in, the robot only needs to sample a small number of images with ground-truth 6D pose between the new product and the camera by moving the robot arm around the new product where the camera is mounted on the robot arm and the other objects together with the new product are placed on a calibrated picking plate. The pose between the camera and the new product is easily obtainable since the pose of the camera and new product w.r.t. the robot base coordinates are known. Thus, the whole system can be fully automatic and does not require further training for new products.

5.5.3 Limitations

Our work does not consider deformable or articulated objects, especially for cases where reference and test images have drastic shape diversity. Another notorious concern is predicting transparent objects where sensors are often failed to capture depth information. Recent work on depth completion for transparent such as Zhu et al.(Zhu et al., 2021) can alleviate the problem. Furthermore, our method requires an accurate and generalizable base segmentor φ. Although SA6D achieves promising results in most cases for tabletop objects, the under- and over-segmentation behaviors still limit the performance. Moreover, a more generalizable learning-based registration method between partial and global point clouds would be an interesting direction to replace ICP. In our work, we

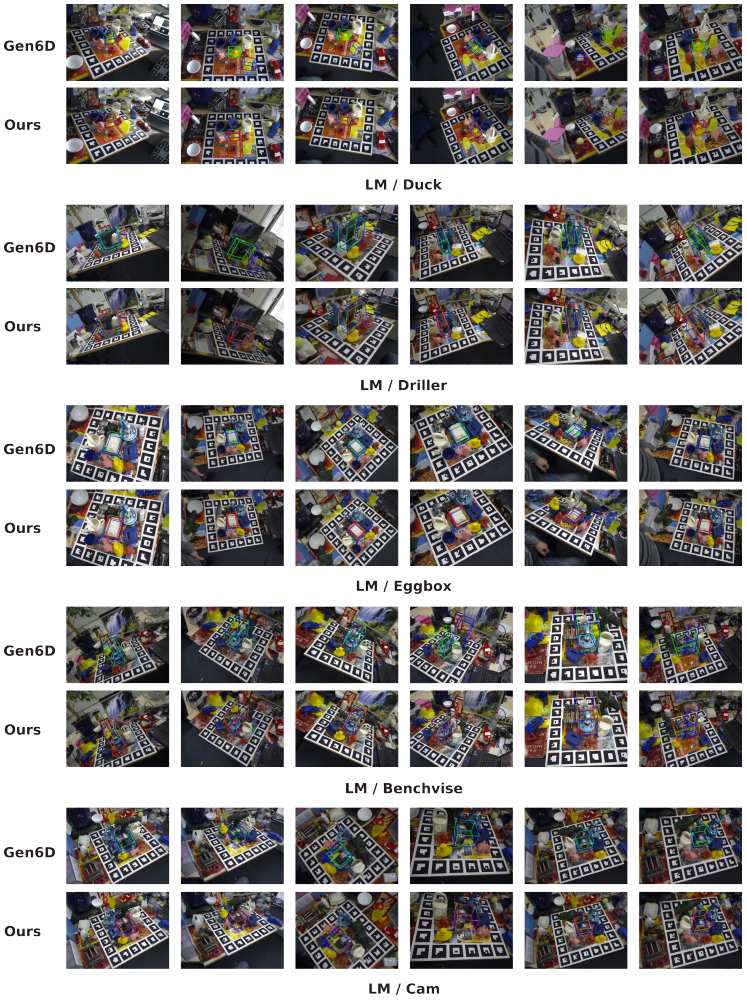


Figure 5.13: Prediction on LineMOD dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

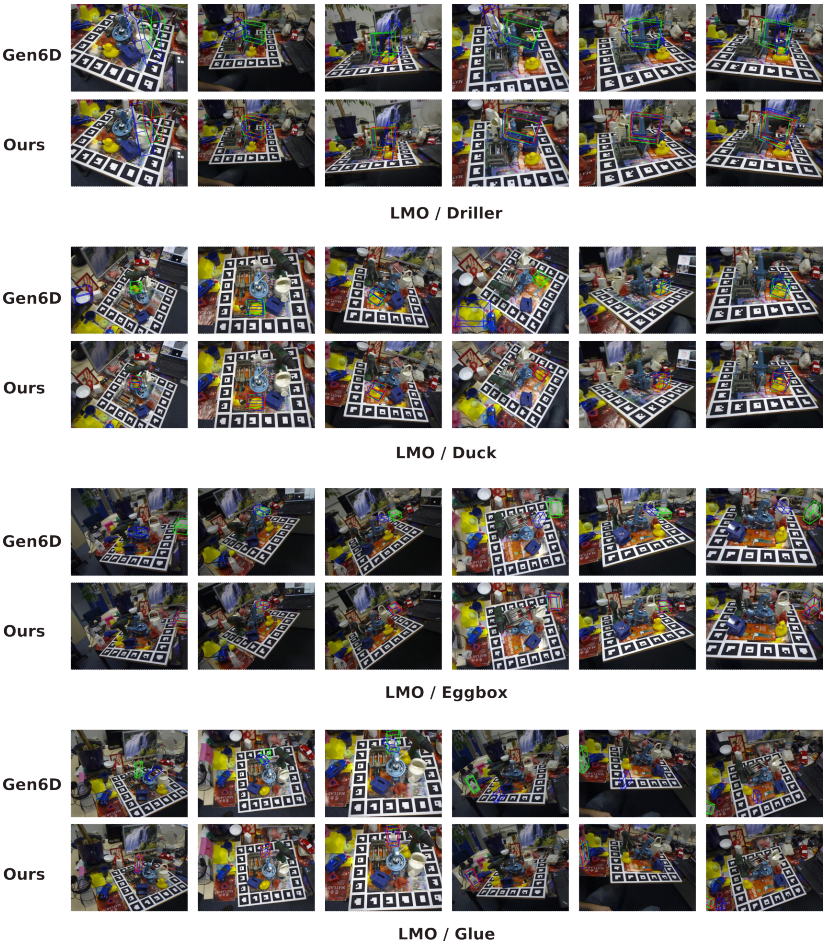


Figure 5.14: Prediction on LineMOD-OCC dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

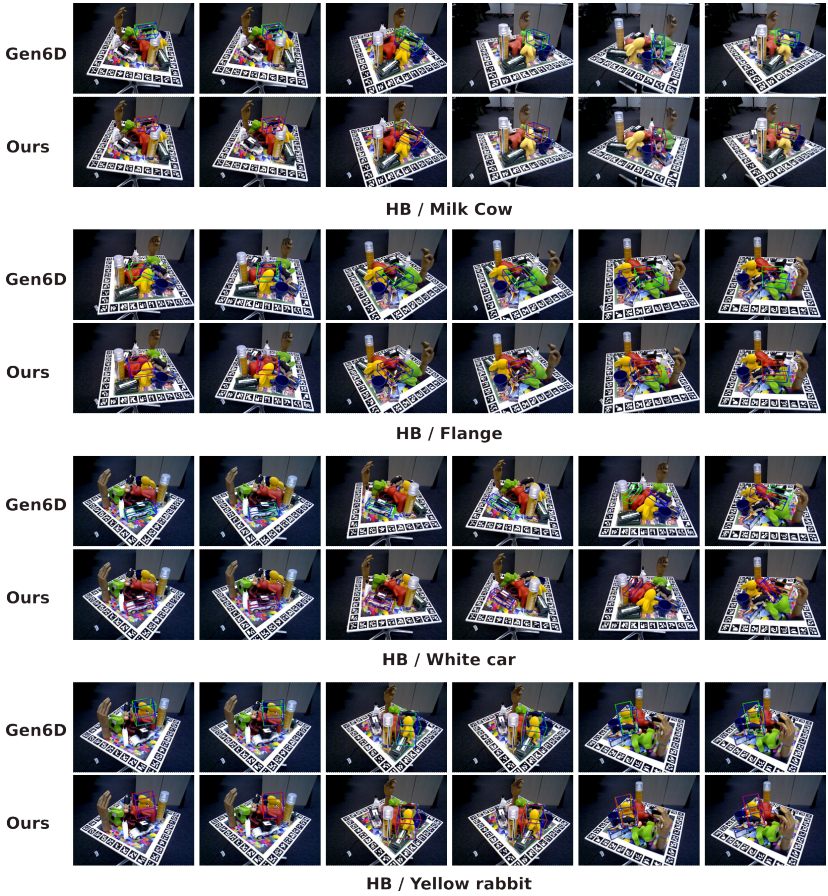


Figure 5.15: Prediction on HomebrewedDB dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

use a simple downsampling strategy to remove the noise based on the point density since noisy points are less often sampled over multiple reference images. Moreover, enforcing a ground-truth segmentation does not always gain additional benefits as it destroys the feature space of the base segmentor φ in the self-online adaptation module. For instance, the ground-truth segmentation

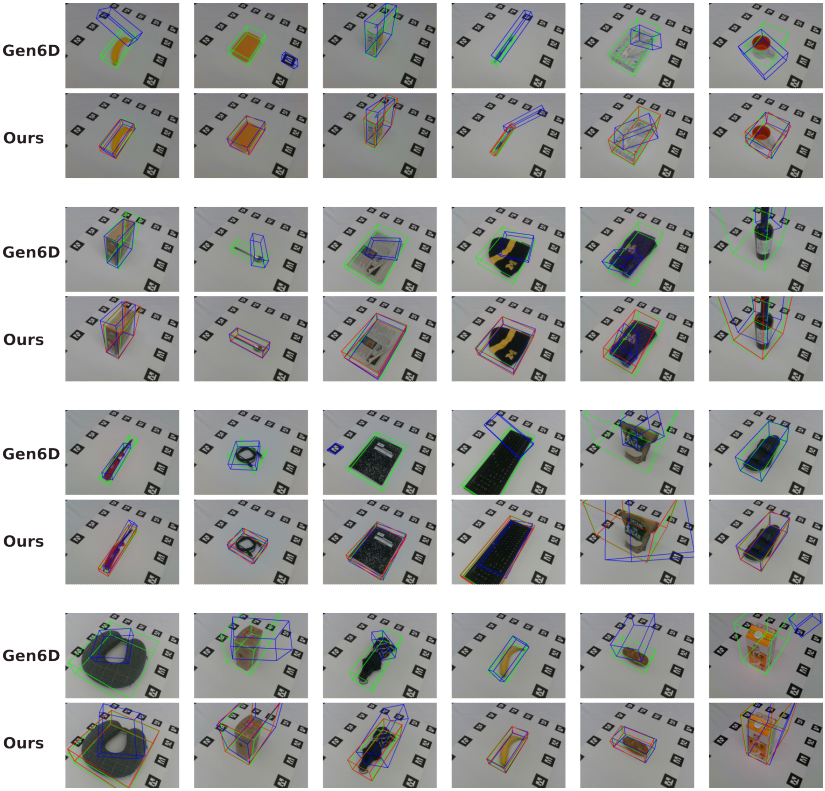


Figure 5.16: Prediction on FewSOL dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

might be considered as background or different components due to occlusion, which will be forced as a positive sample to calculate the contrastive loss if ground-truth segmentation is given.

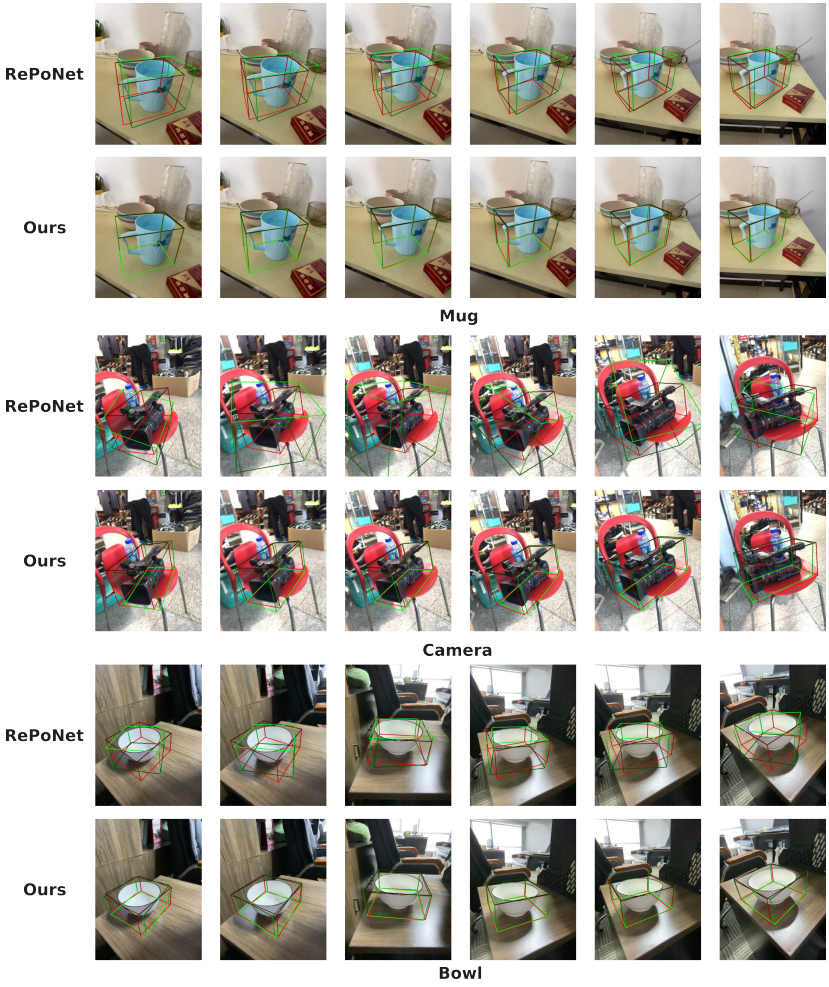


Figure 5.17: Prediction on Wild6D dataset with 20 reference images. The red bounding box is the ground-truth pose and the green bounding box denotes the prediction.

5.6 Conclusion

We propose an approach that can efficiently and robustly predict the 6D pose of novel objects with heavy occlusions while not requiring any object information or object-centric images. We hope our approach can facilitate generalizable 6D object pose estimation in robotic applications.

6 Enhancing Interpretable Object Abstraction via Clustering-based Slot Initialization

6.1 Introduction

Object-centric representations using slots have shown good performance in object detection (Li et al., 2021a; Locatello et al., 2020), segmentation (Kabra et al., 2021; Greff et al., 2019) and tracking (Wu et al., 2021; Kipf et al., 2022; Li et al., 2020b) tasks. Slots are a set of latent variables. The common approach is to frame disentangled and structured slot representations of the compositional scene with some iterative refinement mechanisms in a self-supervised manner, e.g., using softmax-based attention (Locatello et al., 2020) or variational inference (Greff et al., 2019). The idea is to improve the sample efficiency and generalization of capturing the structured environment to unseen compositions or objects. However, most slot-based approaches have difficulties in representing complex scenes. Moreover, the number of slots needs to be specified beforehand on each dataset, which limits the generalization across datasets. In addition, a random slot initialization from a common distribution is widely used in prior works, which lacks consideration between the slots and the perceptual input. Consequently, the quality of the following iterative slot refinement is also affected by the sub-optimal initialization.

To overcome these challenges, instead of random sampling, it is intuitive to sample the initial slot representations conditioned on the perceptual input (see Fig. 6.1). Hence, we employ the k-means clustering algorithm on the convolutional features of the input image. A set of cluster centers are specified based on the features. Afterwards, a set of slots are initialized given the cluster centers as input. Since the order of cluster centers changes randomly, we extend this idea with a permutation-invariant mechanism, where the initial slot

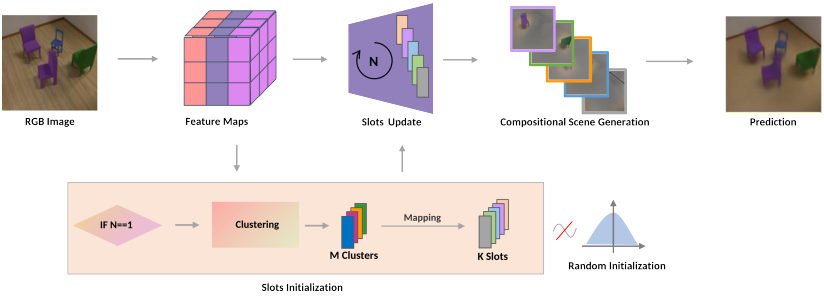


Figure 6.1: The network architecture. Instead randomizing slot initialization from a common distribution widely used in prior work, we initialize slot representations conditioned on the input features. A clustering algorithm and a mapping layer are adopted.

representations remain invariant w.r.t. the order of clusters. To further evaluate the effect of permutation symmetry for slot representations, we employ another permutation equivariant model with mean-shift clustering algorithm, where the slot representations change accordingly with respect to the permutation of the clusters. Mean-shift identifies the number of clusters automatically based on each perceptual input, followed by an injective mapping where each slot is considered as an output of each cluster individually. Thus, it does not require a fixed number of slots based on the whole dataset as prior works.

Our proposed method can be easily placed on top of existing slot-based approaches and trained in an end-to-end manner. In this work, we consider object discovery and novel view synthesis as downstream tasks. To evaluate the improvement and versatility of our method, we choose Slot Attention (Locatello et al., 2020) and IODINE (Greff et al., 2019) as baselines for object discovery task, and uORF (Yu et al., 2022) for novel view synthesis. The experiments are conducted on various datasets.

Our main contributions are as follows: i) We propose the conditional slot initialization using clustering algorithms instead of random initialization. ii) We analyze the effect of permutation symmetry including invariance and equivariance on the object-centric slot representations. iii) We apply mean-shift clustering on the perceptual features which allows to generate flexible number of slots. iv) We demonstrate that, our proposed idea achieves significant improvement over all baselines, while the permutation equivariant mean-shift model presents notable advances especially for complex scenes.

6.2 Guiding Slot Initialization using Clustering

In this section, we will introduce i) the conditional slot initialization with k-means clustering (KM) in Section Section 6.2.1, ii) the permutation invariant version named *Pseudoweights* (PW) in Section 6.2.2, iii) and the permutation equivariant version with variable slot generation using the mean-shift clustering (MS) in Section 6.2.3. More details about implementations and architectures are shown in appendix A.1.

6.2.1 Image-Dependent Slot Initialization

Most slot-based methods typically sample from a standard Gaussian as the random initialization for the slot latent variables (see 6.1). Although the slots are updated by the refinement mechanism incorporating the features from the perceptual input, it is inefficient to start from a random initialization and also limits the final accuracy. Since the perceptual input includes a strong inductive bias about the represented scene, it is straightforward to incorporate the perceptual input directly from the beginning. We first implement a non-permutation symmetric model using k-means clustering. K-means is applied on the pixel-wise convolutional perceptual feature $\mathbf{x} \in \mathbb{R}^{N \times D}$ to get the feature-based cluster centers: $\mathbf{c} = \text{K-means}(\mathbf{x}) \in \mathbb{R}^{M \times D}$ where N is number of pixels from the feature input, M is the number of clusters and D is the feature dimension. Afterwards, the cluster centers are flattened and mapped to the K slots using multi-layer perceptrons (MLPs): $\mathbf{z}_{slots} = \text{MLP}(\mathbf{c}.\text{flat}()).\text{reshape}(K, D)$. Therefore, the number of slots is fixed beforehand like in prior works, as well as the amount of cluster centers. Some cluster centers can vanish during the iterative updates of k-means. For this reason, we resample new clusters during iteration, such that the output of k-means is consistently in $\mathbb{R}^{M \times D}$ and compatible with the MLPs mapping.

6.2.2 Permutation-Invariant Slot Initialization

A good slot representation respects the permutation symmetry (Locatello et al., 2020). In our case, the order of the predicted slots should either remain the same (permutation invariance) w.r.t. the permutation of the cluster centers or change correspondingly in the same order as the cluster centers (permutation

equivariance). Such symmetric behavior enables good generalization of slot representations to unseen world and objects. However, a simple mapping between M cluster centers and K slots as shown in Section 6.2.1 breaks the permutation symmetry and cannot generalize to more slots during evaluation for the scenes with more objects. To address this issue, we propose a permutation invariant model named *Pseudoweights*. To identify different slots, we use a sine-cosine positional encoding $\mathbf{p}_k$ for the k-th slot as follows:

$$\mathbf{p}_k = \left(\sin\left(\frac{\pi}{D'}k\right), \cos\left(\frac{\pi}{D'}k\right), \sin\left(\frac{2\pi}{D'}k\right), \cos\left(\frac{2\pi}{D'}k\right), ..., \cos(\pi k)\right), \tag{6.1}$$

where $k = 1, ..., K$, $D' = \frac{D}{2}$ and D denotes the embedding length. Afterwards, the cluster centers are broadcasted along the slot dimension $\mathbf{c} \in \mathbb{R}^{K\times M\times D}$ and are concatenated with the broadcast of the positional encoding $\mathbf{p} \in \mathbb{R}^{K\times M\times D}$ to predict the weights $\mathbf{w} = \mathrm{MLPs}([\mathbf{c}, \mathbf{p}]) \in \mathbb{R}^{K\times M\times D}$, which allocate the importance of the cluster centers to the different slots. We use a soft-max layer such that the weights allocated for each slot are normalized as follows:

$$\sum_{m=1}^{M} w_{k,m,d} = 1,\ w_{k,m,d} \in [0, 1],\ k = 1, ..., K,\ m = 1, ..., M,\ d = 1, ..., D. \tag{6.2}$$

The slots are then initialized as the weighted sum over the cluster centers by $\mathbf{w}$:

$$\mathbf{z}_k = \sum_{m=1}^{M} \mathbf{w}_{k,m} \cdot \mathbf{c}_{k,m}. \tag{6.3}$$

Thus, the *Pseudoweights* mapping applies a permutation invariant assignment of cluster centers into the slots. Moreover, since the slots are identified by the positional encoding, it enables generalization on increasing objects during test by changing the defined number of slots K without increasing the model parameters. A detailed visualization of the architecture is depicted in appendix A.1.

6.2.3 Automatic Tuning of the Number of Slots using Mean-Shift

Both models introduced in Section 6.2.1 and Section 6.2.2 still require a fixed number of slots beforehand. Therefore, it is essential to apply an unsupervised

clustering mechanism to determine the number of slots conditioned on the input features while keeping the permutation symmetry. Consequently, we perform the mean-shift clustering algorithm (Kobayashi and Otsu, 2010) over the feature space to determine the cluster centers. Mean-shift is an iterative procedure to approximate different modes of a distribution using kernel density estimation. Each mode is represented as a cluster which does not need to be determined beforehand. In our model, we use a Gaussian kernel $k(x, y) = \exp(-\frac{1}{\sigma^2}||x - y||^2)$ for the density estimation. σ is a hyperparameter which affects the granularity of the modes. A shared mapping layer is utilized to initialize the slots based on each cluster respectively $\mathbf{z}_i = \mathrm{MLP}_{\mathrm{shared}}(\mathbf{c}_i)$ where $i \in \{1, ..., K\}$. Thus, it holds the permutation equivariance but requires to have the same number of slots as the number of the predicted cluster centers $K = M$. Since the Gaussian kernel is predefined by a hyperparameter, an expressive learned convolutional feature space is crucial to output distinctive modes.

6.3 Related Work

Object-centric slot representations. Slot representations have been widely used in static scenes (Locatello et al., 2020; Greff et al., 2019; Carion et al., 2020; Burgess et al., 2019; Engelcke et al., 2020) and videos (Li et al., 2021b; Yang et al., 2021a; Kipf et al., 2022; Veerapaneni et al., 2020; Weis et al., 2021). Each slot represents a corresponding object in the scene. This can be achieved either by accumulating the evidence over time to maintain the consistent object slot (Weis et al., 2021) if a variational auto-encoder (Kingma and Welling, 2014) is employed, or using softmax-based attention mechanism (Locatello et al., 2020; Bao et al., 2022). However, all of these approaches require a fixed set of slot variables. The set size needs to be strictly equal or larger than the number of objects in the scene, which limits the generalization on real-world applications since the number of objects is changing dynamically over time and cannot be determined in advance.

Scene decomposition. Most works formulate scene decomposition as compositional generative model (Greff et al., 2019; Eslami et al., 2016; Kügelgen et al., 2020) or a mixture of components (Locatello et al., 2020; Burgess et al., 2019; Engelcke et al., 2020). Recently, some works (Stelzner et al., 2021; Yu et al., 2022; Bing et al., 2022) extend 2D scene decomposition to 3D with the

advances of Neural Radiance Field (NeRF) (Mildenhall et al., 2020). Chen et al. (2020a) and MULMON (Li et al., 2020b) infer 3D scenes from multiple reference images and textureless background. In contrast, uORF (Yu et al., 2022) infer from a single image and test on complex objects with diverse textured background.

Object discovery. Object discovery requires to differentiate the objects and background in an unsupervised way. These methods typically model objects as a set of latent embeddings (Carion et al., 2020) and adopt topic modelling (Russell et al., 2006), group image patches (Tuytelaars et al., 2009; Grauman and Darrell, 2006) or clustering-based deep learning algorithms (Li et al., 2019; Vo et al., 2020). Some methods (Zhao and Wu, 2019; Vo et al., 2021) also apply saliency detection and region proposals on the entire image to group and localize the objects.

Novel view synthesis. Novel view synthesis aims to generate novel views of the given scene from a single (Greff et al., 2019; Eslami et al., 2018; Yu et al., 2022) or multiple (Li et al., 2020b; Mildenhall et al., 2020) source views. Liu et al. (2021a) employ a token-transformation module to synthesize the novel views from a single image without requiring the pose information. Chen et al. (2021c) extend GQN (Eslami et al., 2018) with a Spatial Transformation Routing (STR) mechanism without requiring explicit camera intrinsic information. Lochmann et al. (2016) enable the real-time novel view inference with the advantage of volume rendering. Recently, FWD (Cao et al., 2022) replace the expensive computation of volumetric sampling in NeRF-like methods by pixel-wise depth prediction and a differentiable point cloud renderer.

Deep clustering. Clustering is central to many data-driven research and unsupervised learning. In particular, it helps analyze unstructured and high-dimensional data into meaningful and low-dimensional representations, which has been improved with deep learning techniques in recent years (Xie et al., 2016). Guo et al. (2017) propose an iterative optimization of learning low-dimensional representations from an auto-encoder by minimizing the Kullback-Leibler divergence between the pixel-wise features to each cluster center. Ghasedi Dizaji et al. (2017) extend it with a classifier on top which predicts the probability over the k classes where k is the number of cluster centers. Yang et al. (2017) employ the objective of k-means as clustering loss in the feature space while Fard et al. (2020) relax the cluster assignment problem by using a soft-assignment which can fully benefit from the efficiency of stochastic gradient decent (SGD). Genevay et al. (2019) propose a fully

differentiable version with the cluster parameters while Cai et al. (2022) reduce the computational time by introducing a subspace-based clustering and improve the scalability of deep clustering.

6.4 Experiments

To evaluate our method, we choose two object-centric tasks: object discovery in Section 6.4.4 and novel view synthesis in Section 6.4.5. We employ our idea on top of three state-of-the-art methods: Slot Attention (Locatello et al., 2020), IODINE (Greff et al., 2019) and uORF (Yu et al., 2022). We show more details about implementations in appendix A.1 and qualitative results in appendix A.2 and A.3.

6.4.1 Baselines

In the object discovery task, we use Slot Attention and IODINE as baselines and build our method on top of them. Both baselines use slot representations but with different procedures to refine the slots: Slot Attention uses simple but effective softmax-based attention mechanism while IODINE considers slots as probabilistic latent variables and employs variational inference to accumulate the evidence during iterations. For the novel view synthesis task, we choose uORF as baseline which uses softmax-based attention module to update slots and generate slot-based compositional scenes with Neural Radiance Field (NeRF). Note that all these models use random slot initialization. In addition, we also design two ablated models where the slot initialization is conditioned on the input features. First, we employ the k-means initialization directly as slot representations without any mapping layers in between (*direct* model). Second, we design a simple and permutation equivariant model using shared MLPs to map the k-means cluster centers of the input features to the slots (*shared MLPs* model).

6.4.2 Datasets and Metrics

Datasets. We use three datasets for the object discovery task: Multi-dSprites (MDS), CLEVR and Chairs datasets. Each dataset contains multiple objects in

the scene. Similar as Slot Attention, we extract the CLEVR dataset to have maximum 4, 6, and 10 objects respectively and denote them as CLEVR4, CLEVR6 and CLEVR10. The Chairs dataset originates from uORF (Yu et al., 2022), which includes 4 chairs in each scene. The dataset includes 1200 different shapes of chairs sampled from ShapeNet (Chang et al., 2015) and 50 different floor textures as background. To train the Slot Atttention related models, we use 5k images for CLEVR4 and 10k for MDS, CLEVR6 and Chairs. To train the IODINE related models, we use the same datasets except 13.9k images for MDS. Each dataset contains another 500 images for evaluation. For the novel view synthesis task: We only use the Chairs dataset but it includes 5k scenes for training and 500 scenes for testing, where each scene includes 4 images with different camera viewpoints. Therefore, there are in total 20k images for training and 2k images for testing.

Metrics. As prior works (Greff et al., 2019; Locatello et al., 2020; Burgess et al., 2019), for the object discovery task, we adapt the Adjusted Rand Index (ARI) score to be evaluated only on the pixels of the foreground objects and evaluate the predicted segmentation with the groundtruth mask. For the novel view synthesis, we follow uORF and adopt ARI on the fully reconstructed image, the foreground regions (Fg-ARI) and the synthesized novel view images (NV-ARI). Furthermore, we use LPIPS (Zhang et al., 2018), SSIM (Wang et al., 2004) and PSNR (Horé and Ziou, 2010) as perceptual metrics for both tasks.

6.4.3 Implementation details

We show the implementation details and the architectures of different variants here. **Slot Attention:** This subsection provides a detailed explanation of all the methods presented in chapter 2. The Slot Attention architecture in Fig. 6.2 is extended by a clusterization algorithm, that can be either k-means or mean shift, and by a mapping algorithm, being one of *Direct*, *Small MLP*, *Large MLP* or *Pseudoweights*. The encoder can be a U-Net or a size preserving convolution network. The extension initializes slots conditioned on the perceptual input and not like the original Slot Attention architecture from random gaussian distributions. During the iterative slot attention process, the initialized slots are updated to attend to certain feature pixels, while ignoring others. This is described by the bright yellow markings in the attention masks in Fig. 6.2. The Slot attention uses three iterations to update the slots. Each slot is decoded

into a rgb-image and an α-mask. The renderer calculates, with a weighted sum, the output according to the slotwise rgb-image and the α-mask.

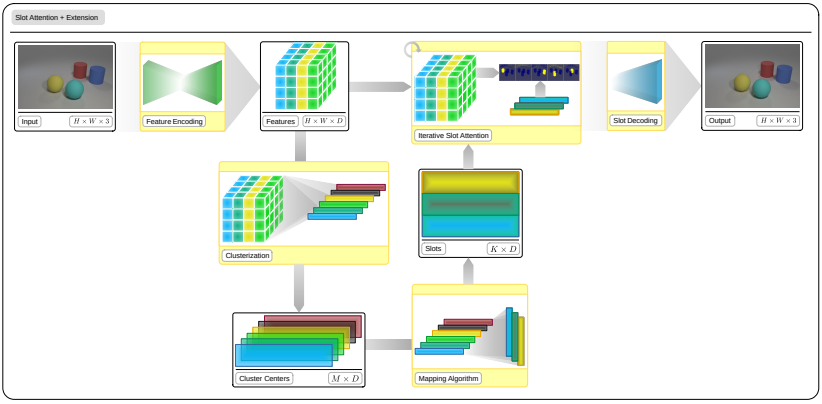


Figure 6.2: The framework architecture for slot initialization for slot attention. The top row is the original architecture.

IODINE: The extension for IODINE resemble the same structure as in the slot attention architecture in Fig. 6.3. The only difference is that the mapping algorithm has to map between the cluster centers of dimension D to two parameters μ, σ of the Gaussian distribution. That is why *Direct* mapping is impossible for IODINE. Slot initializations are now drawn out of the perceptual conditioned gaussian distribution and have dimension D. A decoder calculates, in the same fashion as for slot attention, for each slot a rgb-image and an α-mask. The render outputs the reconstructed image, that will be compared to the groundtruth image to produce a loss. The loss is used in a refinement network, with auxiliary inputs, to update the gaussian parameters μ, σ. This process is repeated five times.

Direct mapping: This simple permutation equivariant approach depicted in Fig. 6.4 directly injects the cluster centers determined by the clusterization algorithms into the slots. Since there is no mapping network involved, this approach can not be used for IODINE, because the cluster centers have to be mapped to two gaussian parameters μ, σ.

Small MLPs: This mapping extends *Direct*-mapping with a non linear network between the cluster centers and the slots, that is shared between all slots, as depicted in Fig. 6.5. The *Direct*- and *Small MLPs*-mapping are used for

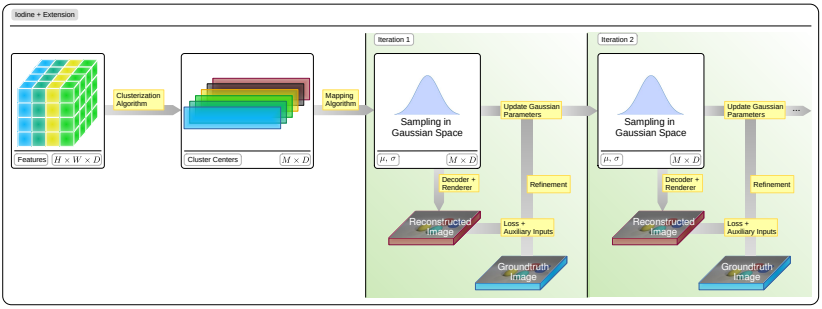


Figure 6.3: The framework architecture for IODINE based extensions. The original starts directly at iteration 1 with slots drawn out of the standard gaussian distribution with $(\mu, \sigma) = (0, 1)$.

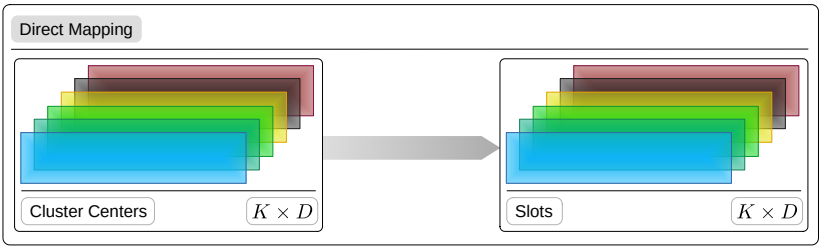


Figure 6.4: The Direct mapping approach. Slots are identical to the cluster centers chosen by the clusterization algorithm.

their simplicity and the permutation equivariance. But they can only translate between the same number of cluster centers and slots.

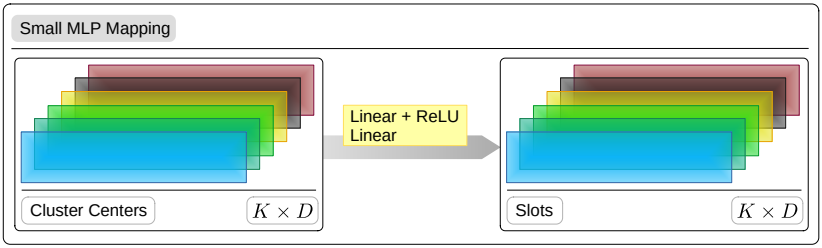


Figure 6.5: The Small MLPs mapping approach. It extends the direct mapping approach by a nonlinear network between cluster centers and slots.

Large MLPs: This network maps between a different number of cluster centers and slots, as provided in Fig. 6.6. The reason for this is to increase the sampling amount of cluster centers from the perceptual input without increasing the model size noticeably, which scales linear with the amount of slots. It is not shared between the slots and thus it is not permutation symmetric. The cluster centers are flattened into one large vector and then mapped to a flattened representation of the slots. These slots are then reshaped to $M \times D$. A drawback of this design is, that it can not generalize to more slots, like all other mapping networks, because of the fixed large MLPs.

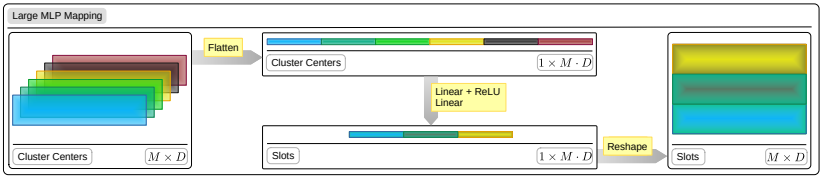


Figure 6.6: The Large MLPs approach.

Pseudoweights: This algorithm incorporates several concepts into one mapping approach. It can map between a different amount of cluster centers and slots, while being able to generalize to more slots and keeping permutation invariance. It has to be permutation invariant, because it is ambiguous to define permutation equivariance between two not equally large sets. This mapping sorts cluster centers into slots. It is aware in which slot it is, because of the position encoding of the K slots. Thus the segregation network before the pseudoweights tensor can decide, if a cluster center should be sorted into a particular slot, then the weights in the pseudoweights tensor will be high, other wise the weights will be low. This segregation network does the decision conditioned only on one cluster center and one position code for all possible $M \times K$ pairs. The last step calculates the weighted sum with the pseudoweights tensor and returns the initialized slots. An explanation of this process and a visual proof of permutation invariance is provided in Fig. 6.8.

Clusterization Algorithms: The k-means algorithm used in the presented methods uses the k-means++ initialization, where the first center is randomly chosen and all other centers are initialized iteratively at the data point being the farthest away from all current initialized centers. If k-means is used with the *Large MLP*, it requires a cluster dying prevention, because sometimes a cluster center will vanish, if all data points are closer to other cluster centers. In that

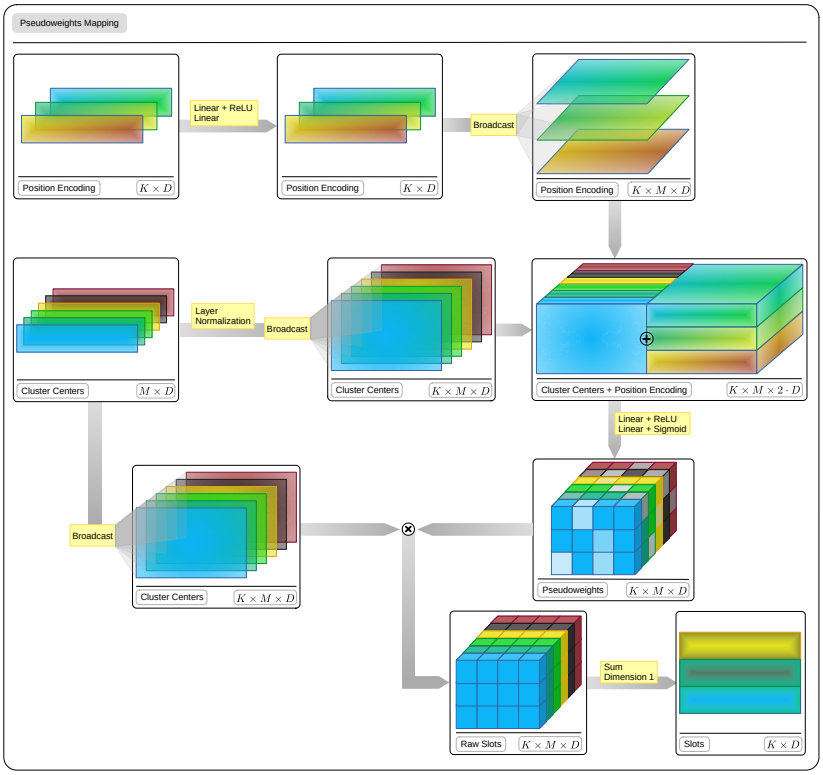


Figure 6.7: The permutation invariant Pseudoweights mapping.

case, a new cluster center is initialized with the k-means++ initialization. A pseudo code is provided in Algo. 2. The amount of cluster centers used in k-means is always initialized with the double amount of the maximum objects count in the dataset. So for CLEVR6, where there are up to six foreground objects and one background object, we initialize always 14 cluster centers at the start of k-means. The mean shift algorithm is initialized with 20 cluster centers for all datasets, because after mean shift converges an algorithm called *connected-components* is used to merge clusters centers, that are very close to each other in to one vector. This ability lets mean shift to determine the amount of slotsflexible. The hyper parameter ϵ is used to determine the radius in the *connected-components*, where all cluster centers within the ϵ-sphere are

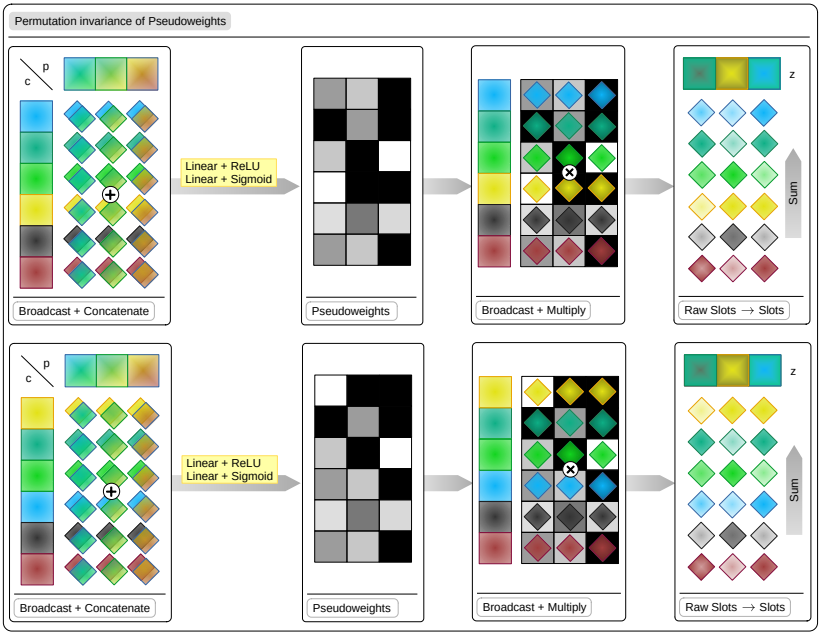


Figure 6.8: The permutation invariant mapping between 6 cluster centers and 3 slots. For this example all slots and cluster centers are of dimension D=1, to keep it simple. The pseudoweights tensor has high values in black squares and low values in white squares. If the blue and yellow slot change their position, the slots won't change their initialization.

merged to one vector. Another hyper parameter used in mean-shift is σ and is used to determine the bandwidth of the gaussian kernel. A detailed pseudo code is provided in Algo.1. We determine the hyperparameters dependent on the weight initialization of the network, so that from the beginning of the training, the output amount of slots fluctuates between 1 and 20, but will never be always 20 or always 1. This happens if σ or ϵ are too small, then mean shift will converge into every little mode, or if the hyperparameters are too large, then all cluster centers can merge into the same spot.

Algorithm 1 K-means algorithm with cluster dying prevention, that reinitializes a new cluster center as soon as one vanishes.

$c_i \leftarrow$ k-means++ initialization; $i \leq N$
repeat
  for each c_i **do**
    $C_i = \{x_j : d(x_j, c_i) \leq d(x_j, c_k); \forall x_j \wedge \forall k \neq i\}$
  end for
  for each C_i **do**
    if $C_i = \emptyset$ **then**
      $c_{inew} \leftarrow$ k-means++ reinitialization
    else
      $c_{inew} = \sum_{c_i \in C_i} \frac{c_i}{|C_i|}$
    end if
  end for
  if $d(c_i, c_{inew}) \leq tolerance \, \forall i$ **then**
    Return c_{inew}
  end if
until max iterations
Return c_{inew}

Algorithm 2 Mean shift algorithm, with the hyperparameters ϵ used in the connected-components algorithm and σ used in the gaussian kernel function.

for $n \in 1, ..., N$ **do**
  $x \leftarrow x_n$
  repeat
    $\forall n : p(n|x) \leftarrow \frac{exp(-0.5||\frac{x-x_n}{\sigma}||^2)}{\sum_{n'=1}^{N} exp(-0.5||\frac{x-x_n}{\sigma}||^2}$)
    $x \leftarrow \sum_{n'=1}^{N} p(n|x) \cdot x_n$
  until stop
  $z_n \leftarrow x$
end for
connected-components($\{z_n\}_{n=1}^{N}, \epsilon$)

6.4.4 Object Discovery

Visualizations on object discovery task We show some qualitative evaluation examples here for the object discovery task.

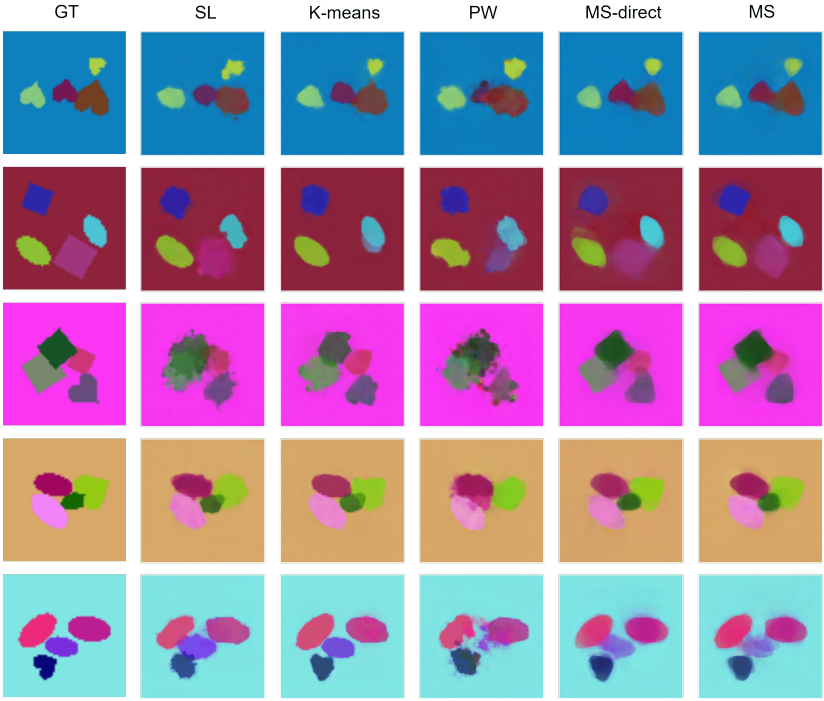


Figure 6.9: Qualitative results on MDS dataset.

Training. We follow the same training setup of Slot Attention and IODINE. We use Adam optimizer (Kingma and Ba, 2015) with a learning rate of 4×10^{-4} for Slot Attention based models and 3×10^{-4} for IODINE related models. We train the Slot Attention related models with 2 NVIDIA Tesla V100-32GB GPUs and a batch size of 32 on each GPU. For IODINE related models, we use 4 GPUs since IODINE requires more computation and gpu memory. We train each model for 1000 epochs with a warm-up training strategy (Goyal et al., 2017) and an exponential learning rate decay. We use $K = 5$ for MDS, CLEVR4 and Chairs datasets since there are maximum 4 objects in each scene,

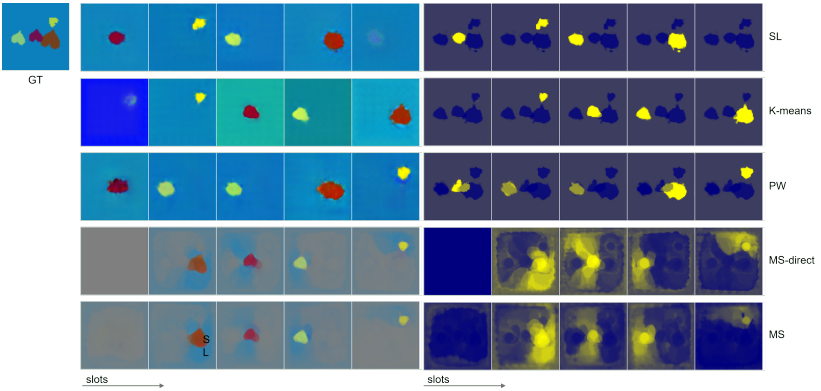


Figure 6.10: The slot-wise predicted masks and reconstructed scenes on MDS dataset.

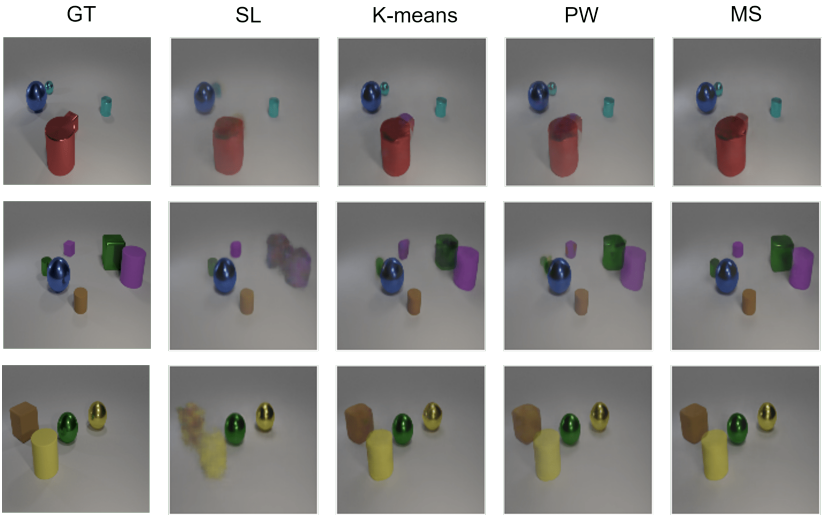


Figure 6.11: The original Slot Attention model struggles with overlapped objects.

and $K = 7$ for CLEVR6. The cluster number is set to $M = 2K$ except for the *mean-shift*, *direct* and *shared MLPs* versions which require $M = K$.

Results. Quantitative results are shown in Table 6.1 and qualitative results in Fig. 6.17. In general, learning inductive slot initialization from input features

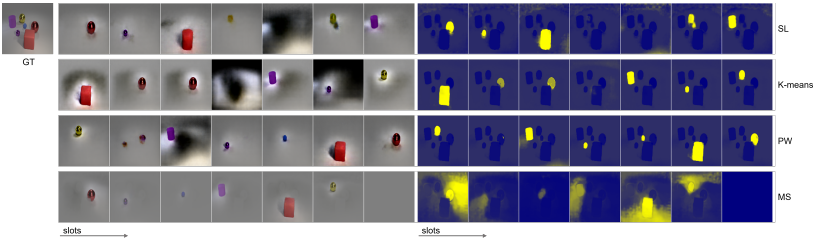


Figure 6.12: The slot-wise predicted masks and reconstructed scenes on CLEVR6 dataset.

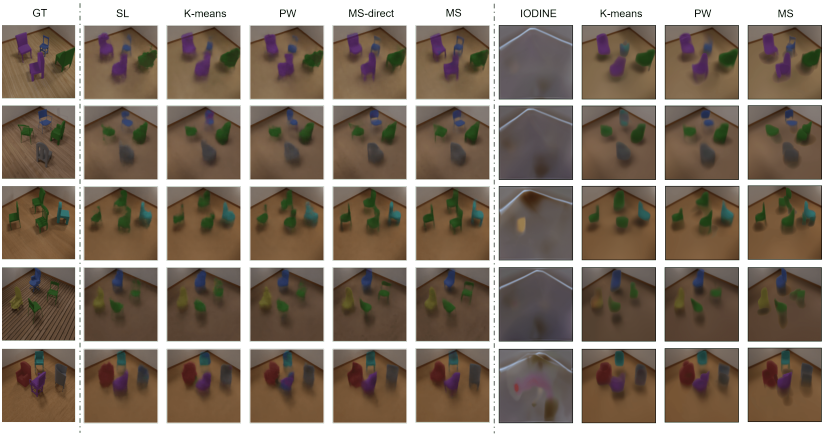


Figure 6.13: Qualitative results on Chairs dataset.

improve the performance on both baselines, where *mean-shift* models achieve the best performance consistently over all datasets. **Well-recovered details:** Surprisingly, all our IODINE-based variants achieve higher resolution even than the groundtruth image for MDS dataset, while the original IODINE is struggled with the data prior and cannot reconstruct the shape of objects. Furthermore, in Fig. 6.17, we observe that only the *mean-shift* models can capture the details of objects for Slot Attention based models. For example, it captures the "heart" objects in MDS while others struggle with the data prior. In particular, our models (especially for *mean-shift models*) can reconstruct the appearance in very good details, e.g., the small blue sphere in CLEVR6 and the legs and rims of various chairs in Chairs dataset. **Slots disentanglement:** We also visualize the slot-wise reconstructed scenes and masks in Fig. 6.18. From

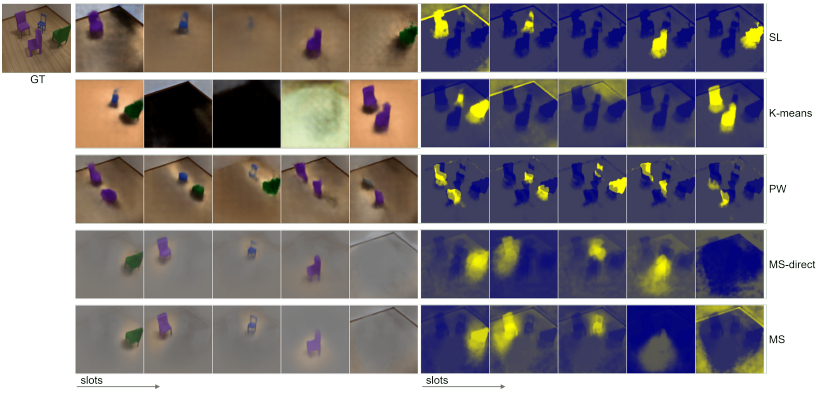


Figure 6.14: The slot-wise predicted masks and reconstructed scenes on Chairs dataset.

Table 6.1: Quantitative results on the object discovery task.

	MDS					CLEVR6					Chairs				
Model	ARI ↑	LPIPS_A ↓	LPIPS_V ↓	PSNR ↑	SSIM ↑	ARI ↑	LPIPS_A ↓	LPIPS_V ↓	PSNR ↑	SSIM ↑	ARI ↑	LPIPS_A ↓	LPIPS_V ↓	PSNR ↑	SSIM ↑
SL	0.9671	0.0693	0.1351	27.43	0.9237	0.9815	0.0748	0.1486	32.11	0.8908	0.9982	0.3144	0.4362	24.49	0.6035
SL + kmeans (direct)	0.9223	0.1074	0.1606	26.13	0.9095	0.9963	0.0381	0.1097	34.22	0.9161	0.9963	0.2971	0.4273	24.17	0.6024
SL + kmeans	0.9837	0.0519	0.1149	28.88	0.9417	0.9970	0.0313	0.1032	34.98	0.9255	0.6271	0.2948	0.4274	24.31	0.6034
SL + kmeans (shared MLPs)	0.9043	0.1174	0.1672	25.84	0.9019	0.9989	0.0320	0.1041	34.82	0.9255	0.9974	0.3173	0.4297	25.01	0.6199
SL + PW	0.9605	0.0834	0.1526	26.25	0.9104	0.9937	0.0371	0.1056	34.04	0.9251	0.9523	0.3052	0.4363	24.82	0.6104
SL + MS (direct)	0.9893	0.0448	0.1059	31.39	0.9559	0.6114	0.1098	0.1957	29.43	0.8555	0.9999	0.2757	0.3997	26.02	0.6341
SL + MS	**0.9944**	**0.0393**	**0.0919**	**32.17**	**0.9613**	**1.0000**	**0.0306**	**0.1022**	**35.32**	**0.9301**	**1.0000**	**0.2693**	**0.3774**	**26.03**	**0.6444**
ID	0.9362	0.0504	0.0888	30.91	0.9591	0.8990	0.0224	0.0500	37.5	0.9661	0.2185	0.2757	0.3843	24.27	0.6299
ID + kmeans (direct)	0.9910	0.0193	0.0492	36.03	0.9833	0.8791	0.0254	0.0559	36.86	0.9619	0.6881	0.2666	0.3842	24.25	0.6322
ID + kmeans	0.9962	0.0166	0.0415	37.06	0.9861	0.8325	0.0198	0.0479	37.725	0.9667	0.7281	0.2559	0.3744	24.31	0.6314
ID + PW	0.9930	0.0207	0.0440	36.42	0.9834	0.9818	0.0190	0.0483	37.725	0.9667	0.8792	0.2192	0.3712	29.025	0.6362
ID + MS	**0.9970**	**0.0143**	**0.0401**	**38.12**	**0.9921**	**0.9909**	**0.0141**	**0.0361**	**38.90**	**0.9753**	**0.9991**	**0.1645**	**0.3219**	**31.07**	**0.6995**

the masks, we observe that only the *mean-shift* models can fully disentangle the objects and background where the highlighted area indicates large attention. In contrast, original Slot Attention mixes the background and a chair in slot #1 while IODINE cannot even work with textured background. *Pseudoweights* and *k-means* models also entangle the chairs into one slot even though the overall reconstructed performance is still better than the baselines (Table 6.1 and Fig. 6.17). The slot-wise reconstructed scenes also reveal our conclusion that *mean-shift* models contain more appearance details with fully disentangled slots. **Mapping between clusters and slots:** Furthermore, our ablation studies demonstrate that the *k-means* models using non-linear mapping layers between the clusters and slots gain additional benefits compared to the *direct* models (in Table 6.1). Additionally, the permutation equivariant model (*shared MLPs*)

Table 6.2: Evaluation with different number of iterations (5 iterations are used for training). In particular, our models achieve significant improvement already at the first iteration.

	Iter 1				Iter 3				Iter 7			
Model	$\text{LPIPS}_A \downarrow$	$\text{LPIPS}_V \downarrow$	PSNR ↑	SSIM ↑	$\text{LPIPS}_A \downarrow$	$\text{LPIPS}_V \downarrow$	PSNR ↑	SSIM ↑	$\text{LPIPS}_A \downarrow$	$\text{LPIPS}_V \downarrow$	PSNR ↑	SSIM ↑
ID	0.4415	0.6071	12.72	0.3820	0.4477	0.5804	16.33	0.4908	0.4363	0.5646	19.53	0.5001
ID + kmeans	0.2108	0.3768	27.05	0.6202	0.1956	0.3607	28.75	0.6533	0.1884	0.3545	29.33	0.6656
ID + PW	0.2269	0.3734	27.57	0.6297	0.1973	0.3531	29.33	0.6642	0.1885	0.3461	29.92	0.6768
ID + MS	**0.1798**	**0.3545**	**28.39**	**0.6467**	**0.1602**	**0.3343**	**30.16**	**0.6828**	**0.1528**	**0.3273**	**30.68**	**0.6951**

performs better than the non-permutation symmetric model (*k-means*) on CLEVR6 and Chairs datasets, indicating the benefits of permutation symmetry on complex scenes, though it is not as good as the *mean-shift* models especially on MDS dataset. **Generalization on increasing objects:** In addition, we evaluate the generalization on more objects and slots (CLEVR10) while the models are trained on CLEVR6. The qualitative results are shown in Fig. 6.19. We observe that the original baselines struggle with closed or overlapped objects by missing, mixing or predicting wrong color of objects, while our models (especially the *mean-shift* models) can detect the overlapped objects perfectly without missing any object even for extremely difficult scenes. For example, i) the *mean-shift* Slot Attention model (ID + MS) can recognize all the objects in the first example with right colors and shapes, ii) in the second example, both *mean-shift* models (SL + MS and ID + MS) and *k-means* IODINE (ID + KM) can detect the red small cylinder in front of the red cube, though the objects are overlapped and with the same color, and iii) both *mean-shift* models and *Pseudoweights* IODINE (ID + PW) can reconstruct the yellow cylinder in the third example. We believe the benefits come from the inductive slot initialization conditioning on the perceptual input features, which gives expressive slot representations used in the following slot refinement. Note that *k-means* models can merely detect 6 objects from the scene since the slot number is by design not scalable. **Generalization on increasing iterations:** Furthermore, Table 6.2 shows the evaluation with increasing number of iterations up to 7 while the models are trained with 5 iterations. All models are capable of generalizing on more iterations with performance gains. In particular, using inductive slot initialization enables notable improvement at the first iteration, which indicates the efficiency of the learned inductive slot initialization. **Failure cases:** We further investigate the cases when *k-means* and *Pseudoweights* are failed to disentangle objects in Chairs dataset. Examples are shown in Fig. 6.20. Interestingly, we find

they learned structured slot representations not always based on the objects. The slot representations of *k-means* model are not generalize due to the non-permutation symmetry. Thus, it always uses the same slot to represent specific area, e.g., the first slot to represent the objects in the top right area, the second and third slots for walls. On the other hand, *Pseudoweights* outputs the same slot representations while changing the object positions due to the permutation invariance. As a result, it neglects the object-centric spatial features in the scene. Thus, the model tends to reconstruct the scenes by assigning fixed spatial area to each individual slot. Such undesirable behaviors occur especially on Chairs dataset where each scene includes 4 images with changing viewpoints. In contrast, a good permutation equivariant model such as *mean-shift* can alleviate this issue and decouple the objects (as shown in Fig. 6.18).

6.4.5 Novel View Synthesis

Setup. The Chairs dataset contains 4 images from different viewpoints of each scene. During training, we randomly pick one image from each scene as input and reconstruct the images for the other 3 viewpoints. We use the same training loss functions and strategies as uORF (Yu et al., 2022). uORF is a memory-extensive model which only works with a batch size of 1 on NVIDIA Tesla V100-32GB. Meanwhile, mean-shift also consumes large memory for the intermediate tensors due to its iterative optimizations. Therefore, we cannot build a mean-shift algorithm on top of uORF with our available hardware. We consider this as a limitation of our *mean-shift* model.

Table 6.3: Results of novel view synthesis on Chairs-diverse.

Model	ARI ↑	Fg-ARI ↑	NV-ARI ↑	LPIPS ↓	SSIM ↑	PSNR ↑
uORF	0.4974	0.5347	0.4291	0.2417	0.6862	24.9712
uORF + kmeans	**0.651**	**0.7346**	**0.5304**	**0.1894**	**0.7176**	**26.1833**
uORF + PW	0.5784	0.6943	0.4773	0.221	0.703	25.6277

Results. We show quantitative results in Table 6.3 and qualitative results in appendix A.3. Overall, our models outperform the original uORF consistently over all metrics. In particular, our models can better reconstruct the chairs pointed to the right direction while original uORF cannot build a clear shape for most chairs.

Visualizations on novel view synthesis task We visualize the examples of novel view synthesis tasks in Fig. 6.21.

6.5 Conclusion

We propose to learn an inductive slot initialization from the input instead of using a random initialization which is widely used in the prior works for the slot-based methods. To evaluate the importance of permutation symmetry over slots, we design various models with non-permutation symmetry, permutation invariance and permutation equivariance into consideration. In particular, our proposed permutation equivariant mean-shift model enables additional flexibility without requiring a fixed number of slots in advance, while it achieves notable improvements on the reconstructed perception details.

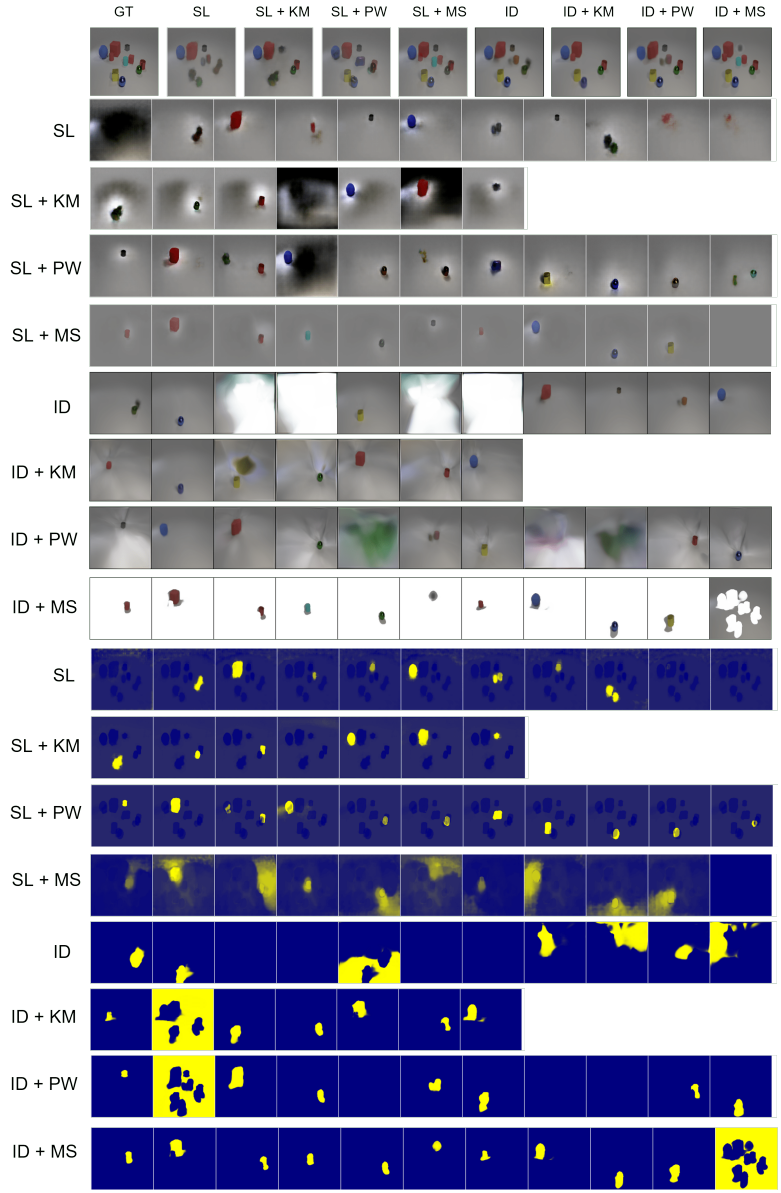


Figure 6.15: Qualitative comparison of generalization on CLEVR10 while the models are trained with CLEVR6.

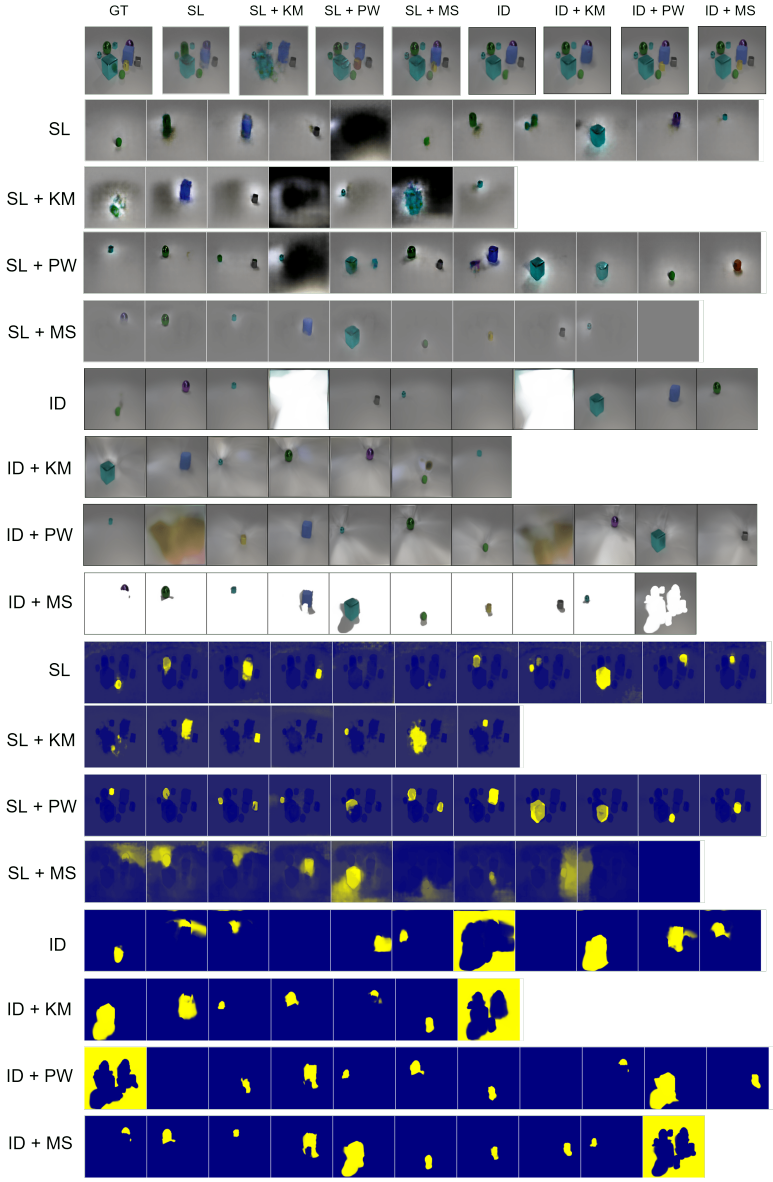


Figure 6.16: Another qualitative comparison of generalization on CLEVR10.

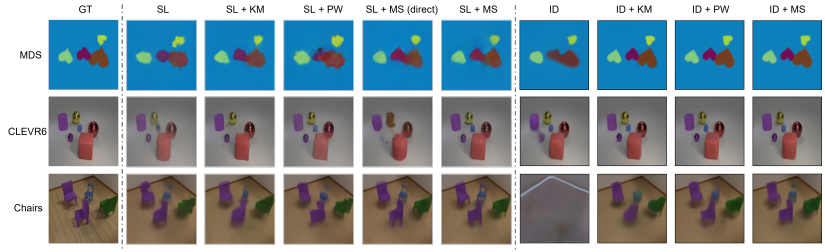


Figure 6.17: Qualitative results on the object discovery task. Notably, the *mean-shift* (MS) versions can recover detailed appearance over all datasets with even better quality than original input for IODINE-based models in MDS dataset.

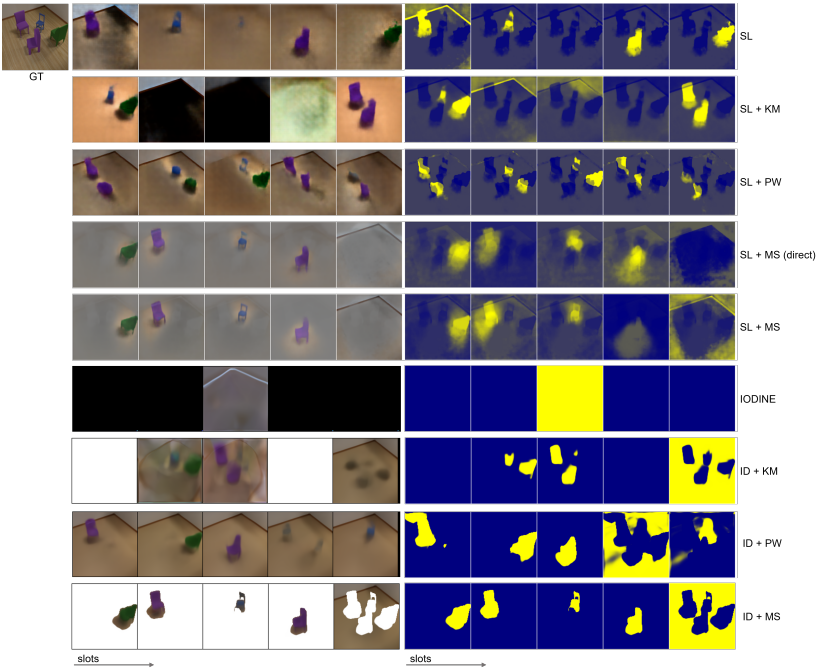


Figure 6.18: Qualitative results of slot-wise reconstructed scenes (left) and masks (right). *Mean-shift* models disentangle the objects better than others and recover more details.

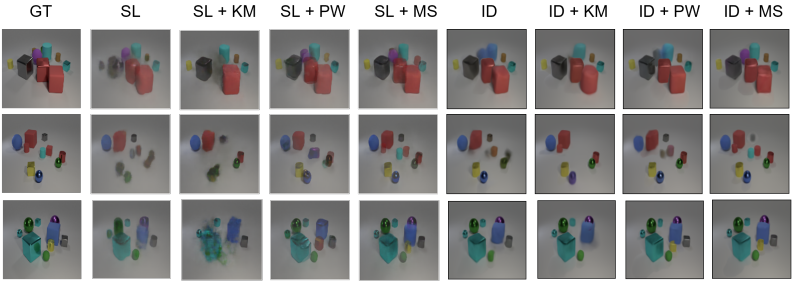


Figure 6.19: Qualitative results on increasing objects. The models are trained on CLEVR6 but evaluated on CLEVR10 with larger number of objects.

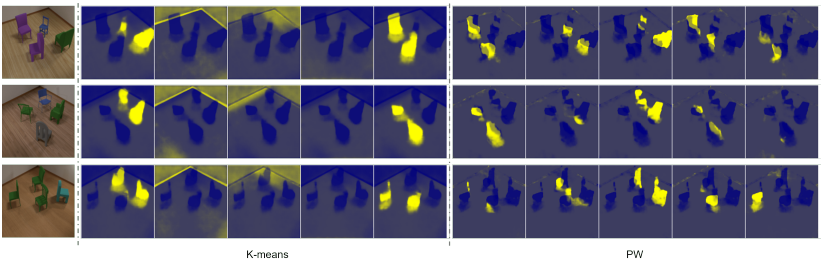


Figure 6.20: Failure cases on Chairs dataset where *k-means* and *Psuedoweights* (PW) cannot disentangle the objects and use each individual slot for specific areas.

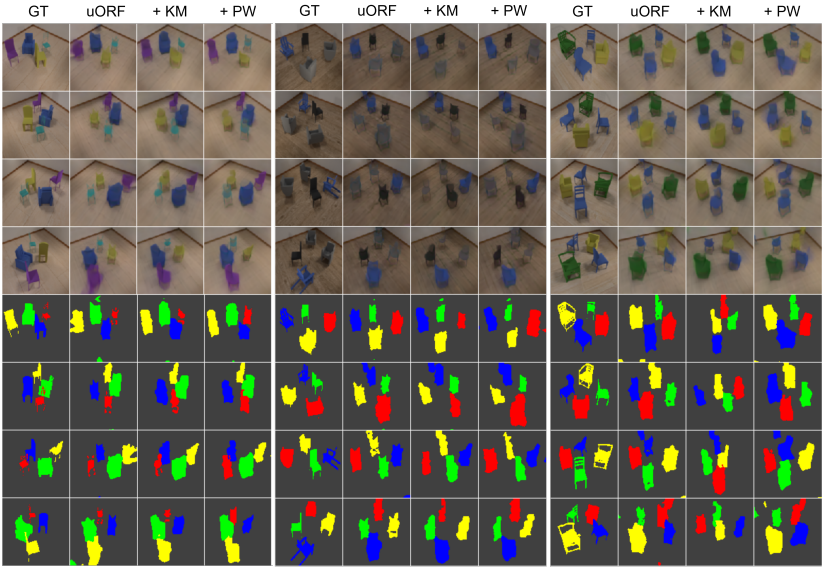


Figure 6.21: Qualitative results on novel view synthesis. Our models can represent the chairs with more details than the original uORF.

7 Meta-Learning Regrasping Strategies for Physical-Agnostic Objects

7.1 Introduction

Grasp detection is one of the fundamental problems in robotic manipulation, which has led to great progress in recent years thanks to the advancements of deep learning techniques. Grasp detection is normally formed as finding the stable grasp position w.r.t. the geometry of the object, the configuration of the end-effector, and the specific manipulation tasks (Newbury et al., 2022). Extensive studies have investigated various end-effector such as parallel jaw (Mahler et al., 2016; Mahler et al., 2017a), suction gripper (Eppner et al., 2016; Mahler et al., 2017b) and multi-finger gripper (Mayer et al., 2022) using RGB-D (Jiang et al., 2011; Lenz et al., 2013) or depth (Morrison et al., 2019) images from synthetic (Schaub and Schöttl, 2020) or real-scene (Song et al., 2019; Fang et al., 2020; Levine et al., 2016) datasets, with the purpose of increasing the generalization on unseen objects (Kalashnikov et al., 2018), closing the sim-to-real gap (Kleeberger et al., 2020; Quillen et al., 2018), and increasing the robustness against occlusion (Breyer et al., 2021). Recently, Huang et al. (2022), Farias et al. (2022), and Rho et al. (2021) further improve the performance on grasping deformable objects.

Although these methods have achieved promising results on various open datasets, it is noteworthy that these datasets predominantly adhere to a homogeneous assumption of the physical properties. The 3D objects (Chang et al., 2015; Kasper et al., 2012; Depierre et al., 2018) in simulation and 3D printed objects (Morrison et al., 2020) are typically treated as entire entities, neglecting the consideration of diverse material properties and friction coefficients for individual components, while most real-world datasets (Çalli et al., 2015;

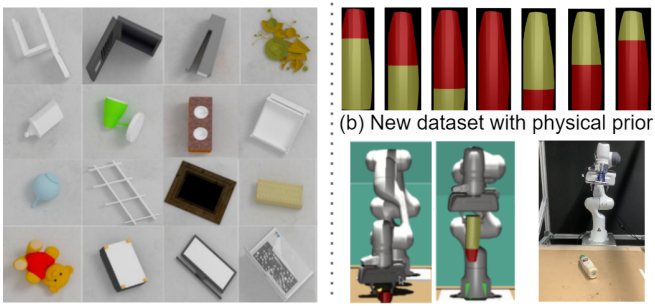


Figure 7.1: (a) Existing datasets such as Jacquard dataset (Depierre et al., 2018) exhibit a variety of textures and geometries. In contrast, (b) our research centers on heterogeneous physical properties (mass distribution and friction coefficients) across the object. For instance, the red part denotes higher mass density and friction while yellow denotes lower mass density. (c) We further evaluate our method in a sim-to-real scenario.

Singh et al., 2014; Lenz et al., 2013; Choi et al., 2008) frequently exhibit a variety of textures and geometries but tend to feature uniform distribution of mass. An example is shown in Fig. 7.1a. Crucially, neither the training nor evaluation stages explicitly incorporate physical properties. Thus, most vision-based grasp detection algorithms rely solely on geometries and textures. This limitation becomes evident in practical scenarios, exemplified by an instance where methods relying solely on object geometry and texture fail to effectively lift an object due to the oversight of variations in part density or friction coefficients.

In contrast to prior works, we explicitly construct various objects with distinct mass distributions and friction coefficients as shown in Fig. 7.1b and employ a shared model to acquire knowledge of these properties purely from depth images, emphasizing the significance of discerning physical attributes solely from visual input. We frame this challenge as a few-shot learning problem, i.e., wherein the physical properties of each object must be gleaned from contextual information derived from a limited number of grasp trials. In essence, our method aims to emulate human learning principles by: 1) Accumulatively acquiring knowledge of physical properties through previous experiences. 2) Facilitating learning during both online and offline inference processes. 3) Seamlessly integrating into existing grasp pipelines without compromising per-

formance. 4) Enhancing real-world performance while leveraging knowledge gained from simulation.

Conditional Neural Processes (CNP) (Garnelo et al., 2018a) has shown advances in few-shot classification and regression tasks (Gao et al., 2022b), characterized by rapid adaptation and inference capabilities. In our work, we integrate CNP into DexNet-2.0 (Mahler et al., 2017a) with minimal alterations to the original grasp frameworks and encode the object's physical properties as latent representations derived from contextual information. To address the limited availability of appropriate datasets, we create two synthetic datasets characterized by distinguishable physical properties in comparison to existing ones. Subsequently, we conduct performance evaluations using the Pybullet and Mujoco simulators, including novel objects from both intra-category (IC) and cross-category (CC). Furthermore, we extend our evaluation to real-world scenarios, where the model is exclusively trained in Mujoco, facilitating the investigation of the sim-to-real gap (shown in Fig. 7.1c).

In summary, our contributions can be summarized as follows:

- We introduce a novel meta-learning grasp framework aimed at addressing the relatively unexplored challenge of grasping objects characterized by diverse physical properties, relying solely on visual input.
- We introduce two innovative synthetic datasets that explicitly incorporate physical properties, making them compatible with a wide range of simulation frameworks.
- Our approach demonstrates substantial advantages in real-world object manipulation, despite being trained exclusively in a simulated environment.

7.2 Related Work

Few-Shot Grasp Detection. Few-shot learning is crucial to generic grasp detection, e.g., generalizing to unseen objects or layouts with grasp preference from only a few examples (Du et al., 2020). DemoGrasp (Wang et al., 2021a) reconstructs the object mesh and predicts the grasp pose from a sequence of RGB-D images with a human demonstration while Hélénon et al. (2020) learns the grasp point from demonstrations including both authorised and

prohibited locations. FSG-Net (Barcellona et al., 2023) and GAS (Kaynar et al., 2023) employ a few-shot semantic segmentation module to grasp a specific object from the clutter although the object has never shown during training. Meanwhile, IGML (Guo et al., 2022), LGPS (Fleytoux et al., 2022) and TACK (Vecerík et al., 2022) predict the grasp point of a novel object out of clutter given a few examples with the specified grasp point. Our work shares similarities with IGML by employing a meta-learning algorithm, however, our approach dispenses with the need for predefined grasp point labels and facilitates online adaptation with rapid inference capabilities.

Grasp Datasets. A variety of grasp datasets have been proposed over recent years. Berscheid et al. (2021), Choi et al. (2018), and Kasaei and Kasaei (2021) collect dataset using real robots while Fang et al. (2020), Kalashnikov et al. (2018), and Mahler et al. (2017a) combine the real-world and simulated data. For example, GraspNet-1Billion (Fang et al., 2020) captures real RGB-D images with generated grasp poses. Meanwhile, Eppner et al. (2021), Wang et al. (2023), and Wei et al. (2022) include purely the synthetic datasets based on simulation. Nonetheless, none of the previously mentioned datasets has provided explicit definitions of discernible physical attributes across distinct parts of individual objects. Conversely, the prevailing norm entails objects exhibiting uniform mass distribution and friction coefficients, whether within a simulated environment or in real-world scenarios. For instance, DVGG (Wei et al., 2022) shares identical friction coefficient (0.25) and density (1500 $\mathrm{kg/m^3}$) over all objects within the simulation. Similarly, widely-used large-scale object datasets like ShapeNet (Chang et al., 2015) and ObjectNet3D (Xiang et al., 2016) lack explicit physical configuration, whereas commonly employed object sets like YCB (Çalli et al., 2015), KIT (Kasper et al., 2012) and BigBIRD (Singh et al., 2014) predominantly feature household items or toys characterized by uniform mass distribution.

Meta-Learning. In the realm of meta-learning, a learning agent acquires meta-knowledge from previous learning episodes or different domains and then uses this acquired knowledge to improve the learning on future tasks (Hospedales et al., 2021). MAML (Finn et al., 2017) is an optimization-based meta-learning method where meta-knowledge is encapsulated within model parameters, and adaptation to new tasks is achieved through further optimization during inference. In contrast, Neural Processes (NPs) fall within the category of neural latent variable models and interpret meta-learning as conditional few-shot function regression (Garnelo et al., 2018b). Similar to Gaussian Processes, NPs model function distributions conditioned on contex-

tual information (Garnelo et al., 2018b; Kim et al., 2019; Gordon et al., 2019). Meta-learning algorithms have been applied in various domains, including low-dimensional function regression (Garnelo et al., 2018a; Garnelo et al., 2018b; Wang and Van Hoof, 2020), image completion (Gordon et al., 2020; Louizos et al., 2019a; Lee et al., 2020), and few-shot classification (Sung et al., 2018; Vuorio et al., 2019). Recent advancements (Yao et al., 2021; Gao et al., 2022b) have extended the application of meta-learning to pose estimation. Additionally, Gondal et al. (2021) and Gao et al. (2022b) enhance meta-learning by incorporating contrastive representation learning from disjoint context sets. CLNP (Kallidromitis et al., 2021) further extends this idea to time series data by combining contrastive learning with ConvNP (Gordon et al., 2020). In our work, we employ CNP (Garnelo et al., 2018a) to meta-learn a latent embedding to represent the physical properties of each object which facilitates online adaptation and expedites the inference process.

DexNet-2.0 DexNet-2.0 (Mahler et al., 2017a) is one of the most famous discriminative approaches using depth images as input. The Grasping Quality Convolutional Neural Networks (GQ-CNN) is trained with millions of synthetic images together with grasp depth and grasp success, generated from thousands of 3D models from DexNet-1.0 (Mahler et al., 2016). The goal of GQ-CNN is to learn a robustness function $Q_{\theta^*}(u, y) \in [0, 1]$ over the possible grasping candidates:

$$\theta^* = \arg\min_{\theta \in \Theta} E_{p(S,u,x,y)}[L(S, Q_\theta(u, y))] \tag{7.1}$$

where θ is the weight parameters, θ^* is the optimized weight parameters, $x = (O, T_o, T_c)$ denotes a state of the variable properties of the camera and objects, where O specifies the geometry and mass properties of an object, T_o, T_c are the 6D poses of the object and camera. u denotes a parallel-jaw (antipodal) grasping candidate in the 3D space. y is a 2.5D point cloud represented as a depth image, $S(u, x) \in \{0, 1\}$ is a binary-valued grasp success metric, $p(S, u, x, y)$ stands for the product of a state distribution, L is the cross-entropy loss function.

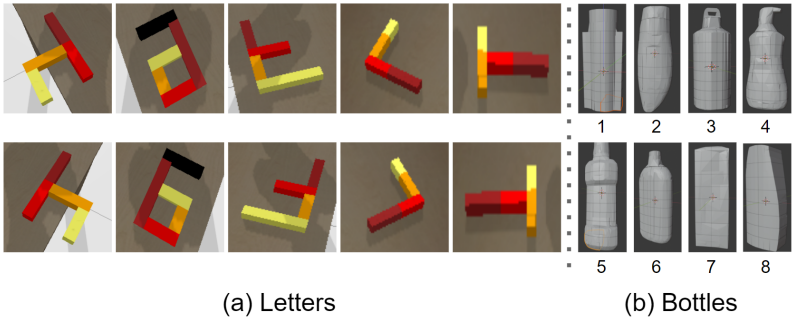


Figure 7.2: We generate two types of datasets. Each shape is considered as one category with numerous instances incorporating varying combinations of mass distribution, friction coefficient, and size. Different colors represent distinct physical properties for visualization.

7.3 Methodology

In this section, we present the creation of object assets characterized by diverse physical attributes, encompassing both *Letters* datasets and *Bottles* datasets. We introduce the process of gathering data from simulators utilizing Pybullet and Mujoco, as well as the meta-learner pipeline namely *ConDex* to discern these physical properties.

7.3.1 Datasets with Heterogeneous Physical Properties

***Letters* Dataset.** We adhere to the Unified Robot Description Format (URDF) for the synthesis of a diverse object dataset tailored for simulation. This synthesis is structured hierarchically, wherein multiple cubes combine to form a bar, and several bars merge to compose objects of distinct shapes, and each shape is considered as a category (as shown in Fig. 7.2a). Alterations within each shape are accomplished by varying the number of cubes within each bar, with the option to adjust the size of individual cubes. The *Letters* dataset encompasses 10 distinct shape categories, each containing $200 \sim 250$ object instances generated randomly, with varying physical properties by independently manipulating the mass, friction coefficient and size of each cube. For training purposes, we select 8 categories, reserving 5% of the objects

for intra-category (IC) evaluation. The remaining 2 categories are used for cross-category (CC) evaluation.

***Bottles* Dataset.** This dataset encompasses 8 distinct bottle shapes as shown in Fig. 7.2b, each of which is disassembled into multiple smaller components. These object configurations are exported in XML format, with specific mass density, friction coefficient and scale assigned to each component in order to generate numerous instances per shape. Notably, we adopt a systematic approach, consistently assigning higher mass density and friction coefficient to a random but dense area of the object to emulate realistic scenario. Scales of components range between 0.8 and 1.2 relative to the original object's size. Consequently, each object is treated as a category encompassing a total of 84 variants with different sizes and diverse physical property configurations.

7.3.2 Simulator

In our study, we utilize both Pybullet and Mujoco as integral components of our data collection and grasping behavior generation processes. This is motivated by our goal to comprehensively assess the distinctions between these simulators, thereby mitigating any potential biases stemming from simulator-specific effects. Specifically, we use Pybullet to generate grasp behaviors for *Letters* dataset and Mujoco for *Bottles* dataset. Our work involves the rigorous evaluation of these simulators to provide a more robust and balanced understanding of their respective performance characteristics and their implications for our research.

7.3.3 Meta-Learn Physical Properties as Task Embeddings

We now formally describe ConDex in the context of meta-learning grasping. We assume that all objects are sampled from the same distribution $p(T)$, each object T_i includes a context set of grasping observations $D_C^i = \{(x_{C,1}, z_{C,1}, y_{C,1}), \ldots, (x_{C,K}, z_{C,K}, y_{C,K})\}_i$ and a target set $D_T^i = \{(x_{T,1}, z_{T,1}, y_{T,1}), \ldots, (x_{T,M}$ where K and M are the number of samples in each set which could be varied at each iteration. Similar to DexNet-2.0 (Mahler et al., 2017a), variables x, z and y are i) the cropped depth image w.r.t. the grasp candidate, ii) the distance between the grasp point and gripper, and iii) the binary grasp label indicating if the grasp succeeds or not. The label of target set is used during training in

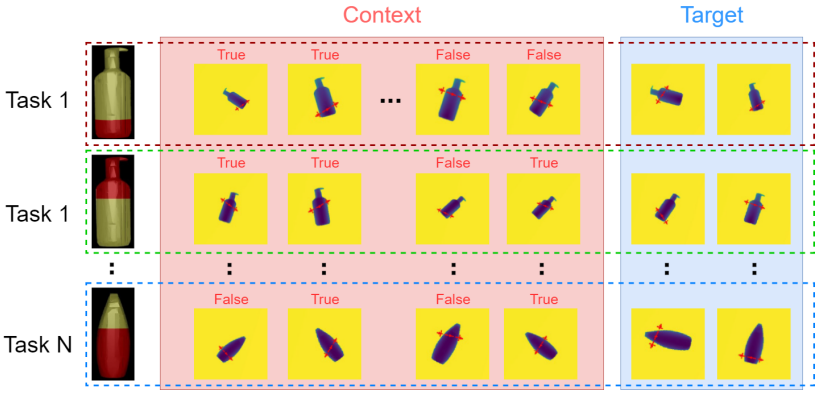


Figure 7.3: The dataset is split into context and target sets for each object. The context set includes the depth images x w.r.t. the grasp candidates, the distance z between the grasp candidates and the gripper, and the binary grasp labels indicating if the grasp succeeds. In contrast, the target set lacks labels during the inference phase. The data is split randomly between context and target sets for each training iteration.

loss calculation but not available during evaluation. An example is shown in Fig. 7.3.

The entire traning dataset is denoted as $D = \{D_C^i, D_T^i\}_{i=1}^N$ where N is the number of objects sampled for training. During inference, the model is tested on new sampled objects $T^* \sim p(T)$. Given a small context set of the new task, the model has to infer a new function $f^* : (D_C^*, (x_T^*, z_T^*)) \rightarrow \hat{y}_T^*$. In meta-learning, there are two types of learned parameters. The meta-parameters θ, which is learned during the training phase using D, and the task-specific parameters ϕ^* which is updated based on samples from each individual new task D_C^* conditioned on θ and context set. Predictions can be constructed as $\hat{y}_T^* = f_{\theta,\phi^*}(x_T^*, z_T^*)$, where f is the meta-model parameterized by θ and ϕ^*.

ConDex considers θ as the neural weights consisting of an encoder h_θ and a decoder g_θ. The structure of the encoder and decoder is shown in Fig. 7.4. ϕ^* is considered as the encoded task representation $\mathbf{r}$ of object properties predicted from the context set of the novel task. The encoder h takes each observation consisting of the cropped depth image $x_{(C,i)}$, the grasping distance $z_{(C,i)}$, and the grasp label $y_{(C,i)}$ as context input and extracts the feature embedding $\mathbf{r}_{(C,i)}$. A permutation invariant aggregator $\bigoplus$ merges all feature embeddings as a task representation $\mathbf{r}$ to represent the object properties: $\mathbf{r} = \bigoplus_{i=1}^K h_\theta(x_{C,i}^*, z_{C,i}^*, y_{C,i}^*)$,

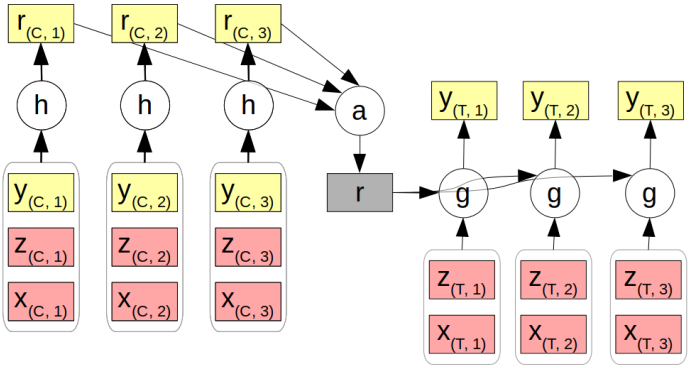


Figure 7.4: The structure of ConDex.

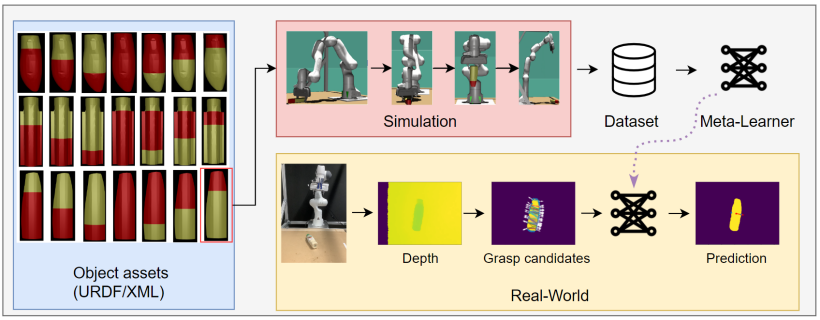


Figure 7.5: pipeline

where we follow CNP and use the same mean aggregator. Subsequently, a decoder g_θ takes $\mathbf{r}$ as an additional input and outputs the score of the grasp candidate $\hat{y_T^*} = g_\theta((x_T^*, z_T^*), \mathbf{r})$. Meta-parameter θ is fixed after training and only $\mathbf{r}$ is updated as a task representation. Thus, ConDex is able to adapt in an online fashion simultaneously as new contexts are acquired. The network architecture is subject to the design of DexNet-2.0 except for having additional input which needs to be concatenated, i.e., the context and the task representation. We show the details of the network architecture in the supplementary video. The goal of ConDex is to predict the confidence score $\hat{y}_T \in [0, 1]$ over the possible

grasping candidates for the target set by minimizing the cross-entropy loss function:

$$\theta^* = \arg\min_{\theta} \frac{1}{TM} \sum_{t=1}^{T} \sum_{i=1}^{M} L(y_{t,i}, \hat{y}_{t,i}) \tag{7.2}$$

where θ is the neural weights, y is the ground-truth binary grasp label, L is the cross-entropy loss, T is the batch size of sampled tasks and M is the size of the target set.

7.4 Experiments

In this section, we introduce metrics for evaluation and analyze the gathered dataset, offering insights into its characteristics. Subsequently, we present the outcomes of our experiments and provide a detailed comparative analysis of our proposed method alongside the established baselines. Finally, we evaluate the sim-to-real performance of ConDex in a real robot.

7.4.1 Evaluation Metrics

Grasp Error Rate. The inference can be considered as a binary classification. The grasping quality $Q \in [0, 1]$ indicates the confidence in each grasping candidate's success. If the robustness is higher than 50%, it will be predicted as success. Otherwise, it will be predicted as failure. The error rate is formulated as:

$$Error\ Rate = \frac{FP + FN}{P + N}, \tag{7.3}$$

where FP, FN are the number of false positive and false negative predictions, P, N denote the ground-truth positive and negative grasps in total.

Grasp Accuracy. In both simulation and real-world scenarios, a successful grasp is defined as when an object is lifted and successfully dropped in the target position as shown in Fig. 7.5. The grasping performance is indicated by the grasp accuracy:

$$Grasp\ Accuracy = \frac{Number\ of\ Successful\ GraspTrials}{Total\ Number\ of\ GraspTrials} \tag{7.4}$$

7.4.2 Statistics of the Collected Data

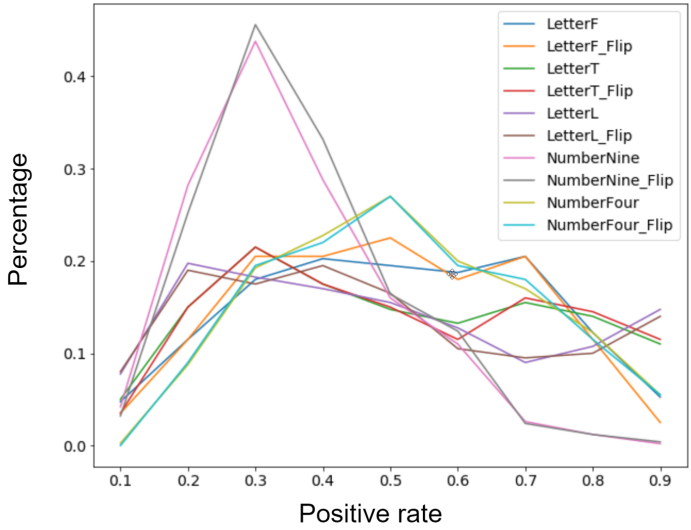


Figure 7.6: **Statistics of *Letters* dataset.** Each curve represents one category comprising numerous instances. The collected data exhibit a normal distribution centered around a positive rate of approximately 50%.

Letters dataset comprises 10 distinct shape categories, each containing either $200 \sim 250$ object instances. Every object instance undergoes 30 random grasps in Pybullet executed at positions within the robot arm's feasible range. We collect in total 63000 observed data on 2.100 objects. As shown in Fig. 7.6, the horizontal axis represents the positive rate resulting from 30 random grasps for each object instance while the vertical axis illustrates the percentage of object instances belonging to a specific shape category. The distribution of object instances across different positive rates demonstrates a notable diversity within our collected image dataset. This diversity is shape complexity-dependent, with more intricate shape categories, such as Category Nine, exhibiting significantly lower overall positive rates in contrast to other categories. Nevertheless, our collected image data remains balanced, with a normal distribution centered around a positive rate of approximately 50%.

7.4.3 Baselines

We employ the following baselines and two variants of ConDex for evaluation:

- **DexNet (Pretrained)** indicates a DexNet-2.0 model pretrained on an object dataset with homogeneous physical properties following Mahler et al. (2017a).
- **DexNet** indicates a DexNet-2.0 model trained only on our datasets.
- **IGML** is a meta-learning grasping method inspired by IGML (Guo et al., 2022) which employ MAML (Finn et al., 2017) to learn the distinguishable grasps from contexts. We adapt this model by using depth images as input and training it on our dataset for fair comparison.
- **ConDex (accumulated)** indicates a ConDex model trained on our dataset, where the collection of the next context is iteratively given and depends on previously acquired knowledge, which can be formed as :

$$(\mathbf{x}_{C,t+1}, z_{C,t+1}) \sim P(\mathbf{x}_{C,t+1}, z_{C,t+1} | (\mathbf{x}_{C,1}, z_{C,1}, y_{C,1}), \ldots, (\mathbf{x}_{C,t}, z_{C,t}, y_{C,t})) \tag{7.5}$$

- **ConDex** denotes a ConDex model trained on our dataset, where contexts are randomly collected.

7.4.4 Experimental Results

Fig. 7.7 illustrates that both ConDex variants exhibit superior performance compared to the baseline approaches on *Letters* dataset. Notably, DexNet (pretrained) performs even worse than random grasping, highlighting the limitation of training on large-scale homogeneous object datasets for our specific task. In contrast, these findings underscore the importance of incorporating physical properties during the training process. Fig. 7.8 shows the evaluation on *Bottles* dataset. Despite a decrease in performance observed when dealing with cross-category objects (object *2* and *6*) decreases, ConDex consistently outperforms the baseline methods by a substantial margin. Fig. 7.9 demonstrates the error rate with respect to the size of the context set for prediction on unseen categories. The dashed line corresponds to the performance of DexNet,

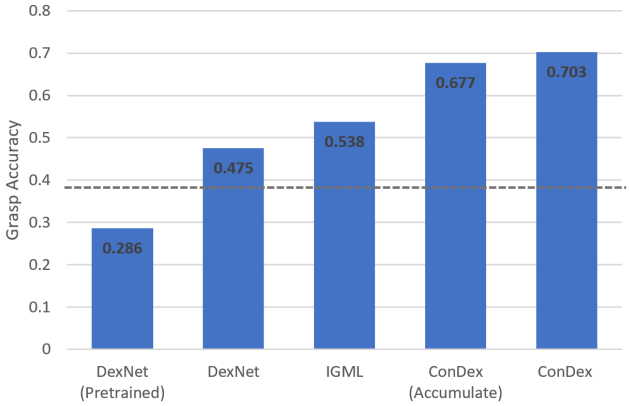


Figure 7.7: Results are evaluated on *Letters* dataset over 50 objects from intra- and cross-categories, each object is grasped 30 times. The dashed line denotes the performance with random grasping.

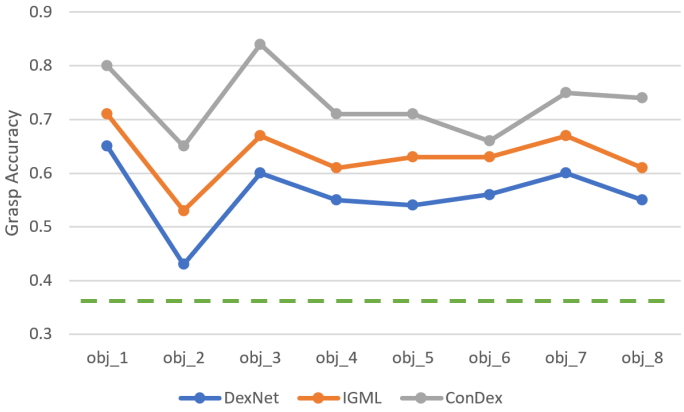


Figure 7.8: Evaluation on *Bottles* dataset from intra- and cross-categories (i.e., object *2* and *6*). The dashed line denotes the performance with random grasping.

which predicts independently of context information. Notably, ConDex consistently outperforms both DexNet and IGML over all different context numbers. Furthermore, the error rate of ConDex decreases as more context points are incorporated, which indicates that the aggregation of additional contexts yields

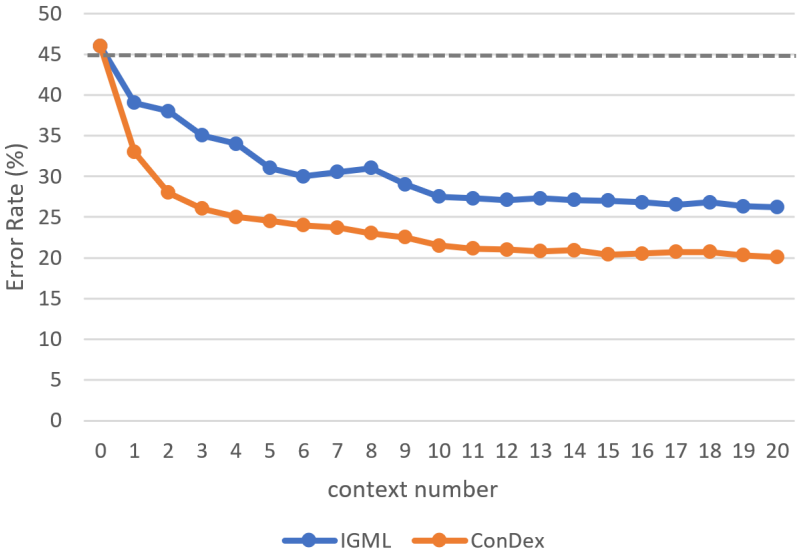


Figure 7.9: **Error rate vs. context number on cross-category.** 450 object instances are evaluated from two previously unseen categories from *Letters* dataset. Results are presented with a maximum of 20 context points during evaluation, while a maximum of 15 context points is provided during training. The dashed line denotes the performance of DexNet.

valuable information from diverse context pairs, enabling the model to adapt effectively to previously unseen tasks. Moreover, the model's performance can be further enhanced when the size of the context set exceeds the maximum number utilized during training, which is 15 in our case. Additionally, we observe a notable reduction in the error rate when the first 5 context points are provided, indicating a crucial factor in the initial stages to alleviate task ambiguity.

Experiments on sim-to-real. To assess the sim-to-real performance, we train ConDex using the *Bottles* dataset from Mujoco and test it on our real robot. The evaluation is conducted on two distinct types of bottles, each containing varying quantities of material. Given the differences in mass distribution among these objects, achieving a precise grasp at the correct position is crucial for successfully picking up and placing the bottle into the designated box. The experimental setup is depicted in Fig. 7.10. The manipulation is deemed as successful if the object can be effectively placed into the box. Fig. 7.11

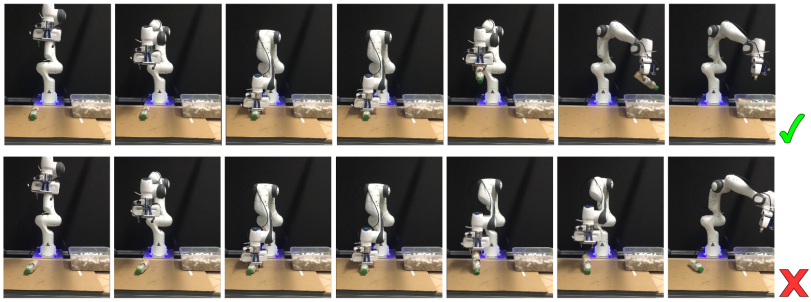


Figure 7.10: Experiments on a real robot involves evaluating the success of manipulation based on the criteria that a successful grasp results in the bottle being successfully placed inside the designated box. The bottle is filled with different quantities of material to acquire diverse mass distributions.

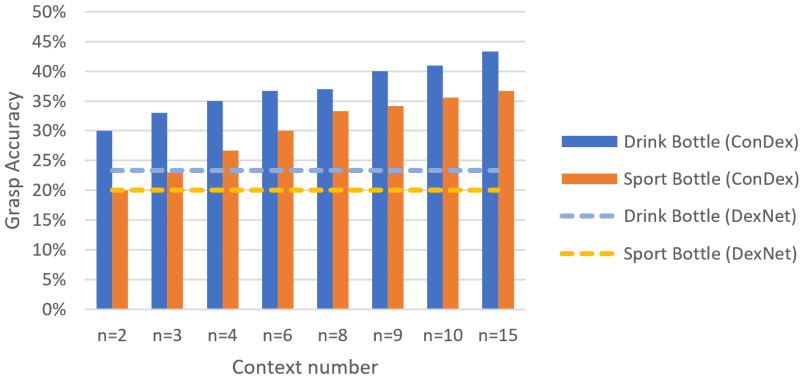


Figure 7.11: **Sim-to-real evaluation.** The results stem from experiments conducted on a real robot, with the models being entirely trained using the *Bottles* dataset obtained through rollouts within the Mujoco simulation.

presents the results with varying numbers of provided context points, along with 30 grasp executions for each trial. The enhanced performance of ConDex with the increasing number of context points signifies the successful transfer of knowledge, demonstrating the model's ability to efficiently extract valuable information from contexts. It is worth highlighting that ConDex outperforms DexNet, albeit with a slight decrease in performance compared to its per-

formance in the Mujoco simulation. In our supplementary video, we also show that ConDex can improve the grasp performance in an online fashion by accumulatively updating the context within the 30 trials on real robot.

7.4.5 Limitations

Static physical properties: Our approach is currently limited in its ability to handle dynamic physical properties, such as liquids with viscosity that change over time. This constraint arises from the limitations of the simulation techniques we employ. **Exclusion of transparent objects:** Transparent objects are not within the scope of our current study. Our work focuses on capturing heterogeneous properties across objects and utilizes depth as the primary input. Nonetheless, there is potential for future research to explore the inclusion of dynamic physical properties and expand the scope to encompass transparent objects as simulation techniques continue to evolve.

7.5 Conclusion

In this paper, we investigate grasping challenging objects with heterogeneous physical properties using meta-learning. Due to the lack of available datasets and the relatively unexplored nature of this field, we generate two datasets encompassing diverse mass distributions and friction coefficients, collecting data from both Pybullet and Mujoco simulation environments. These datasets are crucial in evaluating the effectiveness of our proposed model ConDex against baselines. Our study underscores the significance of leveraging contextual information, facilitating fast adaptation to complex objects and enabling seamless sim-to-real transfer. We hope that our research highlights the potential of this emerging direction and raises further attention in the field.

8 Conclusion

This thesis explores various advanced methodologies in the domain of meta-learning and object recognition, aiming to improve the accuracy, adaptability, and interpretability of machine learning models in robotic applications where the adaptation on novel tasks requires only a few examples (few-shot learning). Through extensive experimentation and innovative model designs, the work presented in this thesis contributes significantly to the field by proposing novel meta-learning algorithms, enhancing existing frameworks, and providing thorough evaluations across diverse datasets. The key contributions include advancements in vision regression tasks, the development of a self-adaptive 6D pose estimator, improved object abstraction techniques, and meta-learning strategies for regrasping. These contributions collectively push the boundaries of what can be achieved in robotic vision applications. We can attempt to answer the research questions chapter-wise:

Research Question 1: *What are the critical factors that influence the performance of meta-learning models in vision regression tasks?*

Gao et al. (2022b) (Chapter 3) investigates the importance of different factors such as data augmentation, domain randomization, task augmentation, and meta-regularization in meta-learning vision regression tasks. The study finds that augmenting the data and tasks significantly enhances the generalization ability of the models but meanwhile requries specific design. The introduction of Functional Contrastive Learning (FCL) further improves the model's performance by creating robust latent representations in the task level that can handle diverse and unseen tasks. Experiments demonstrate that FCL enhances task expressivity and reduces prediction errors.

Research Question 2: *How can we improve 6D pose estimation for novel and occluded objects using meta-learning and few-shot learning?*

Chapter 4 presents SA6D (Gao et al., 2023b), a self-adaptive few-shot 6D pose estimator designed to handle novel and occluded objects effectively. The

method incorporates an online self-adaptation module, a region proposal module, and a refinement module to enhance pose estimation accuracy. In particular, the onle self-adaptation module incorporates contrastive learning for latent task representation, optimization-based meta-learning, and metric learning. The experiments demonstrate that SA6D outperforms existing category-agnostic methods, particularly in cluttered and occluded scenarios, by leveraging geometric features and adaptive learning. SA6Ds robustness and accuracy make it suitable for practical robotic applications where minimal reference images are available.

Research Question 3: *How can we improve 6D pose estimation for novel and occluded objects using meta-learning and few-shot learning?*

Chapter 5 introduces a novel approach (Gao et al., 2023a) to slot initialization using clustering techniques, specifically mean-shift and k-means algorithms. The proposed models demonstrate significant improvements in object discovery and novel view synthesis tasks by using inductive slot initialization, which conditions on the perceptual input features. This method enhances the flexibility with various number of slots, and accuracy of slot-based models, allowing for better object representation and generalization across different datasets. Furthermore, this work discuss the geometric symmetries in the latent representations, i.e., the invariant and equivariant permutation. The experiments on image reconstruction and novel view-synthesis validate that clustering-based slot initialization leads to more interpretable and effective object abstraction. Furthermore, the experiments on extrapolation using more number of slots during evaluations than the number of slots used during training demonstrate the enhanced generalization and better object abstraction in our method using conditional model. However, how to employ slot representations on real and more complex scenes is still the critical for future investigations.

Research Question 4: *How can meta-learning be applied to improve regrasping strategies for objects with diverse physical properties?*

Chapter 6 explores the application of meta-learning to develop regrasping strategies for objects with varying physical properties, e.g., mass distributions and friction coefficients (Gao et al., 2022a). The proposed ConDex model leverages contextual information from prior knowledge to adapt quickly to new objects and demonstrates effective sim-to-real transfer. The research highlights the importance of context-aware grasping and shows that ConDex outperforms baseline methods in both simulated and real-world experiments. By learning task embeddings that capture the physical characteristics of objects implicitly,

ConDex enables robots to dynamically adjust their grasping strategies in an online fashion, showcasing the practical applicability of these methods. Future works should be investigated in more complex scenarios and more complex objects with more diverse properties.

The research presented in this thesis addresses several key challenges in the fields of meta-learning and object recognition in robotic vision applications. By proposing innovative solutions and thoroughly evaluating them, the work significantly advances our understanding and capabilities in these areas. The utilization of meta-learning techniques in this thesis brings several significant advantages, enhancing the efficiency, adaptability, and effectiveness of robotic vision systems. For example:

- **Reduction in human efforts for data collection**: Meta-learning allows models to generalize from a small number of examples, significantly reducing the need for extensive and labor-intensive data collection. This is particularly advantageous in real-world applications where acquiring large labeled datasets is impractical or expensive.
- **Time-efficient model adaptation**: Our methods can quickly adapt to new tasks with minimal data. This rapid adaptation capability is crucial for applications where quick deployment is necessary, such as robotic manipulation in dynamic environments.
- **Improved generalization**: Our models with techniques such as Functional Contrastive Learning (FCL) and task augmentation improve the robustness and the ability to generalize well to new, unseen tasks by learning meta-knowledge from a diverse set of tasks. This generalization is particularly beneficial in robotic applications where the environment and tasks are constantly changing.
- **Adaptive learning in Complex Scenarios**: Methods like SA6D demonstrate superior performance in dealing with occlusions and cluttered environments while ConDex model illustrates how meta-learning can be applied to develop regrasping strategies for objects with diverse physical properties, dynamically adjusting to new objects and scenarios. Both of them showcase the adaptability of meta-learning in complex and realistic scenarios.
- **Interpretable object abstraction**: The use of clustering techniques for slot initialization and the object-level representation with confidence score in SA6D enhance the interpretability and performance of object

abstraction models, leading to better object representation and scene understanding.

- **Sim-to-real transfer**: The research includes extensive simulation such as blender, mujoco and pybullet, and real-world experiments, demonstrating that meta-learning techniques can effectively bridge the gap between simulated and real-world environments, making the models practical for real-world robotic applications.

Overall, the application of meta-learning in this thesis offers substantial advantages in reducing human efforts for data collection, speeding up model adaptation, improving generalization to new tasks, and enhancing the robustness and adaptability of robotic vision systems. These benefits make meta-learning a powerful approach for advancing the capabilities of machine learning models in complex and dynamic environments. The methodologies developed here offer promising directions for future research and practical applications in machine learning and robotics. Specifically, the advancements in vision regression, 6D pose estimation, object abstraction, and regrasping strategies provide a solid foundation for building more adaptable, accurate, and interpretable robotic vision systems. The thesis emphasizes the potential impact of these contributions on advancing robotic vision systems and suggests future research directions, including integrating these techniques into more complex robotic systems and exploring additional applications in computer vision and robotics.

Bibliography

Antreas Antoniou, Harrison Edwards, and Amos Storkey (2019). "How to train your MAML". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HJGven05Y7.

K. S. Arun, 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.

Trapit Bansal, Shafiq Joty, and Dilek Hakkani-Tur (2020). "Self-supervised meta-learning for few-shot natural language classification tasks". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5222–5236.

Zhipeng Bao, Pavel Tokmakov, Allan Jabri, Yu-Xiong Wang, Adrien Gaidon, and Martial Hebert (2022). "Discovering Objects That Can Move". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11789–11798.

Leon di Barcellona, Alberto Bacchin, Alberto Gottardi, Emanuele Menegatti, and Stefano Ghidoni (2023). "FSG-Net: a Deep Learning model for Semantic Robot Grasping through Few-Shot Learning". In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1793–1799.

Lars Berscheid, Christian Friedrich, and Torsten Kröger (2021). "Robot Learning of 6 DoF Grasping using Model-based Adaptive Primitives". In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4474–4480.

Yanrui Bin, Zhao-Min Chen, Xiu-Shen Wei, Xinya Chen, Changxin Gao, and Nong Sang (2020). "Structure-aware Human Pose Estimation with Graph Convolutional Networks". In: *Pattern Recognition* 106, p. 107410. DOI: 10.1016/j.patcog.2020.107410.

Wang Bing, Lu Chen, and Bo Yang (2022). "DM-NeRF: 3D Scene Geometry Decomposition and Manipulation from 2D Images". In: *arXiv preprint arXiv:2208.07227*.

Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother (2014). "Learning 6D Object Pose Estimation Using 3D Object Coordinates". In: *The European Conference on Computer Vision (ECCV)*.

Michel Breyer, Jen Jen Chung, Lionel Ott, Roland Y. Siegwart, and Juan I. Nieto (2021). "Volumetric Grasping Network: Real-time 6 DOF Grasp Detection in Clutter". In: *Conference on Robot Learning*.

Christopher P. Burgess, Loïc Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matthew M. Botvinick, and Alexander Lerchner (2019). "MONet: Unsupervised Scene Decomposition and Representation". In: *ArXiv* abs/1901.11390.

Jinyu Cai, Jicong Fan, Wenzhong Guo, Shiping Wang, Yunhe Zhang, and Zhao Zhang (2022). "Efficient Deep Embedded Subspace Clustering". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–10.

Berk Çalli, Arjun Singh, Aaron Walsman, Siddhartha S. Srinivasa, P. Abbeel, and Aaron M. Dollar (2015). "The YCB object and Model set: Towards common benchmarks for manipulation research". In: *2015 International Conference on Advanced Robotics (ICAR)*, pp. 510–517.

Ang Cao, Chris Rockwell, and Justin Johnson (2022). "FWD: Real-Time Novel View Synthesis With Forward Warping and Depth". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15713–15724.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko (2020). "End-to-End Object Detection with Transformers". In: *European Conference on Computer Vision (ECCV)*, pp. 213–229.

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

Chang Chen, Fei Deng, and Sungjin Ahn (2020a). "Learning to Infer 3D Object Models from Images". In: *ArXiv* abs/2006.06130.

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

Kai Chen 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.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton (2020c). "A Simple Framework for Contrastive Learning of Visual Representations". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 1597–1607.

Tung-I Chen, Yueh-Cheng Liu, Hung-Ting Su, Yu-Cheng Chang, Yu-Hsiang Lin, Jia-Fong Yeh, and Winston H. Hsu (2021a). "Should I Look at the Head or the Tail? Dual-awareness Attention for Few-Shot Object Detection". In: *IEEE Trans. Multim.* 23. arXiv: 2102.12152. URL: https://arxiv.org/abs/2102.12152.

Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis (2021b). "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.

Wen-Cheng Chen, Min-Chun Hu, and Chu-Song Chen (2021c). "STR-GQN: Scene Representation and Rendering for Unknown Cameras Based on Spatial Transformation Routing". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5946–5955.

Xu Chen, Zijian Dong, Jie Song, Andreas Geiger, and Otmar Hilliges (2020d). "Category Level Object Pose Estimation via Neural Analysis-by-Synthesis". In: *European Conference on Computer Vision (ECCV)*. Cham: Springer International Publishing.

Changhyun Choi, Wilko Schwarting, Joseph DelPreto, and Daniela Rus (2018). "Learning Object Grasping for Soft Robot Hands". In: *IEEE Robotics and Automation Letters* 3.3, pp. 2370–2377.

Young Sang Choi, Travis Deyle, and Charles C. Kemp (2008). "A list of household objects for robotic retrieval prioritized by people with ALS". In: *IEEE International Conference on Rehabilitation Robotics*, pp. 510–517.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller (2021). "Rethinking Attention with Performers". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Ua6zuk0WRH.

Ignasi Clavera, Anusha Nagabandi, Simin Liu, Ronald S Fearing, Sergey Levine, and Pieter Abbeel (2018). "Learning to adapt in dynamic, real-

world environments through meta-reinforcement learning". In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Michael Danielczuk, 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: *2019 International Conference on Robotics and Automation (ICRA)*, pp. 7283–7290.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei (2009). "Imagenet: A large-scale hierarchical image database". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255.

Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam (2019). "BlenderProc". In: *arXiv* abs/1911.01911. URL: http://arxiv.org/abs/1911.01911.

Amaury Depierre, Emmanuel Dellandréa, and Liming Chen (2018). "Jacquard: A Large Scale Dataset for Robotic Grasp Detection". In: *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3511–3516.

Zi-Yi Dou, Zhou Yu, and Antonios Anastasopoulos (2019). "Domain attention with an ensemble of experts". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5990–5996.

Guoguang Du, Kai Wang, and Shiguo Lian (2019). "Vision-based Robotic Grasping from Object Localization, Pose Estimation, Grasp Detection to Motion Planning: A Review". In: *CoRR* abs/1905.06658. arXiv: 1905.06658. URL: http://arxiv.org/abs/1905.06658.

Guoguang Du, Kai Wang, Shiguo Lian, and Kaiyong Zhao (2020). "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review". In: *Artificial Intelligence Review* 54, pp. 1677–1734.

Thomas Elsken, Benedikt Staffler, Jan Hendrik Metzen, and Frank Hutter (2020). "Meta-Learning of Neural Architectures for Few-Shot Learning". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Martin Engelcke, Adam R. Kosiorek, Oiwi Parker Jones, and Ingmar Posner (2020). "GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations". In: *International Conference on Learning Representations (ICLR)*.

Clemens Eppner, Sebastian Höfer, Rico Jonschkowski, Roberto Martín-Martín, Arne Sieverling, Vincent Wall, and Oliver Brock (2016). "Lessons from the Amazon Picking Challenge: Four Aspects of Building Robotic Systems". In: *International Joint Conference on Artificial Intelligence*.

Clemens Eppner, Arsalan Mousavian, and Dieter Fox (2021). "ACRONYM: A Large-Scale Grasp Dataset Based on Simulation". In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6222–6227.

S. M. Ali Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, koray kavukcuoglu koray, and Geoffrey E Hinton (2016). "Attend, Infer, Repeat: Fast Scene Understanding with Generative Models". In: *Advances in Neural Information Processing Systems*. Vol. 29.

S. M. Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S. Morcos, Marta Garnelo, Avraham Ruderman, Andrei A. Rusu, Ivo Danihelka, Karol Gregor, David P. Reichert, Lars Buesing, Théophane Weber, Oriol Vinyals, Dan Rosenbaum, Neil C. Rabinowitz, Helen King, Chloe Hillier, Matthew M. Botvinick, Daan Wierstra, Koray Kavukcuoglu, and Demis Hassabis (2018). "Neural scene representation and rendering". In: *Science* 360, pp. 1204–1210.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman (2010). "The Pascal Visual Object Classes (VOC) Challenge." In: *International Journal of Computer Vision* 88, pp. 303–338.

Qi Fan, Wei Zhuo, Chi-Keung Tang, and Yu-Wing Tai (2020). "Few-Shot Object Detection With Attention-RPN and Multi-Relation Detector". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhaoxin Fan, Yazhi Zhu, Yulin He, Qi Sun, Hongyan Liu, and Jun He (2021a). "Deep Learning on Monocular Object Pose Detection and Tracking: A Comprehensive Overview". In: *ArXiv* abs/2105.14291.

Zhibo Fan, Yuchen Ma, Zeming Li, and Jian Sun (2021b). "Generalized Few-Shot Object Detection Without Forgetting". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4527–4536.

Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu (2020). "GraspNet-1Billion: A Large-Scale Benchmark for General Object Grasping". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11444–11453.

Maziar Moradi Fard, Thibaut Thonet, and Éric Gaussier (2020). "Deep k-Means: Jointly Clustering with k-Means and Learning Representations". In: *ArXiv* abs/1806.10069.

Cristiana Miranda de Farias, Brahim Tamadazte, R. Stolkin, and Naresh Marturi (2022). “Grasp Transfer for Deformable Objects by Functional Map Correspondence”. In: *2022 International Conference on Robotics and Automation (ICRA)*, pp. 735–741.

Chelsea Finn, Pieter Abbeel, and Sergey Levine (2017). “Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, pp. 1126–1135. URL: https://proceedings.mlr.press/v70/finn17a.html.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine (2019). “Online Meta-Learning”. In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 1920–1930.

Kai Fischer, Martin Simon, Florian Olsner, Stefan Milz, Horst-Michael Gross, and Patrick Mader (2021). “StickyPillars: Robust and Efficient Feature Matching on Point Clouds Using Graph Neural Networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 313–323.

Martin A. Fischler and Robert C. Bolles (1981). “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography”. In: *Commun. ACM* 24, pp. 381–395.

Yoann Fleytoux, Anji Ma, Serena Ivaldi, and Jean-Baptiste Mouret (2022). “Data-efficient learning of object-centric grasp preferences”. In: *International Conference on Robotics and Automation (ICRA)*, pp. 6337–6343.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil (2018). “Bilevel Programming for Hyperparameter Optimization and Meta-Learning”. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1568–1577. URL: https://proceedings.mlr.press/v80/franceschi18a.html.

Yang Fu and Xiaolong Wang (2022). “Category-Level 6D Object Pose Estimation in the Wild: A Semi-Supervised Learning Approach and A New Dataset”. In: *abs/arXiv:2206.15436*.

Ning Gao, Ruijie Chen, Jingyu Zhang, Ngo Anh Vien, Hanna Ziesche, and Gerhard Neumann (2022a). “Meta-Learning Regrasping Strategies for Physical-Agnostic Objects”. In: *International Conference on Robotics and Automation (ICRA) workshop on Scaling Robot Learning*.

Ning Gao, Bernard Hohmann, and Gerhard Neumann (2023a). "Enhancing Interpretable Object Abstraction via Clustering-based Slot Initialization". In: *34rd British Machine Vision Conference (BMVC)*.

Ning Gao, Vien Anh Ngo, Hanna Ziesche, and Gerhard Neumann (2023b). "SA6D: Self-Adaptive Few-Shot 6D Pose Estimator for Novel and Occluded Objects". In: *7th Annual Conference on Robot Learning*. URL: https://openreview.net/forum?id=gdkKi_F55h.

Ning Gao, Hanna Ziesche, Ngo Anh Vien, Michael Volpp, and Gerhard Neumann (2022b). "What Matters for Meta-Learning Vision Regression Tasks?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14776–14786.

Marta Garnelo, Dan Rosenbaum, Christopher Maddison, Tiago Ramalho, David Saxton, Murray Shanahan, Yee Whye Teh, Danilo Rezende, and S. M. Ali Eslami (2018a). "Conditional Neural Processes". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, pp. 1704–1713. URL: https://proceedings.mlr.press/v80/garnelo18a.html.

Marta Garnelo, Jonathan Schwarz, Dan Rosenbaum, Fabio Viola, Danilo J. Rezende, S. M. Ali Eslami, and Yee Whye Teh (2018b). "Neural Processes". In: *ICML Workshop on Theoretical Foundations and Applications of Deep Generative Models*.

Andreas Geiger, Philip Lenz, and Raquel Urtasun (2012). "Are we ready for autonomous driving? The KITTI vision benchmark suite". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3354–3361.

Aude Genevay, Gabriel Dulac-Arnold, and Jean-Philippe Vert (2019). "Differentiable Deep Clustering with Cluster Size Constraints". In: *ArXiv* abs/1910.09036.

Zigang Geng, Ke Sun, Bin Xiao, Zhaoxiang Zhang, and Jingdong Wang (2021). "Bottom-Up Human Pose Estimation via Disentangled Keypoint Regression". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14676–14686.

Kamran Ghasedi Dizaji, Amirhossein Herandi, Cheng Deng, Weidong Cai, and Heng Huang (2017). "Deep Clustering via Joint Convolutional Autoencoder Embedding and Relative Entropy Minimization". In: *International Conference on Computer Vision (ICCV)*.

Spyros Gidaris and Nikos Komodakis (2018). "Dynamic Few-Shot Visual Learning Without Forgetting". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Muhammad Waleed Gondal, Shruti Joshi, Nasim Rahaman, Stefan Bauer, Manuel Wuthrich, and Bernhard Schölkopf (2021). "Function Contrastive Learning of Transferable Meta-Representations". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 3755–3765. URL: https://proceedings.mlr.press/v139/gondal21a.html.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio (2014). "Generative adversarial nets". In: *Advances in Neural Information Processing Systems*, pp. 2672–2680.

Jonathan Gordon, John Bronskill, Matthias Bauer, Sebastian Nowozin, and Richard Turner (2019). "Meta-Learning Probabilistic Inference for Prediction". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HkxStoC5F7.

Jonathan Gordon, Wessel P. Bruinsma, Andrew Y. K. Foong, James Requeima, Yann Dubois, and Richard E. Turner (2020). "Convolutional Conditional Neural Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=Skey4eBYPS.

Anas Gouda, Abraham Ghanem, and Christopher Reining (2022). "Category-agnostic Segmentation for Robotic Grasping". In: *ArXiv* abs/2204.13613.

Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He (2017). "Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour". In: *ArXiv* abs/1706.02677.

Kristen Grauman and Trevor Darrell (2006). "Unsupervised Learning of Categories from Sets of Partially Matching Image Features". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 1, pp. 19–25.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner (2019). "Multi-Object Representation Learning with Iterative Variational Inference". In: *International Conference on Machine Learning*, pp. 2424–2433.

Weikun Guo, Wei Li, Ziye Hu, and Zhongxue Gan (2022). "Few-Shot Instance Grasping of Novel Objects in Clutter". In: *IEEE Robotics and Automation Letters* 7.3, pp. 6566–6573.

Xifeng Guo, Long Gao, Xinwang Liu, and Jianping Yin (2017). "Improved Deep Embedded Clustering with Local Structure Preservation". In: *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1753–1759.

Abhishek Gupta, Russell Mendonca, YuXuan Liu, Pieter Abbeel, and Sergey Levine (2018). "Meta-Reinforcement Learning of Structured Exploration Strategies". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2018/file/4de754248c196c85ee4fbdcee89179bd-Paper.pdf.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou (2022a). "OnePose++: Keypoint-Free One-Shot Object Pose Estimation without CAD Models". In: *Advances in Neural Information Processing Systems*.

Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun (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.

Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun (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)*.

Yisheng He, Yao Wang, Haoqiang Fan, Jian Sun, and Qifeng Chen (2022b). "FS6D: Few-Shot 6D Pose Estimation of Novel Objects". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6814–6824.

François Hélénon, Laurent Bimont, Eric Nyiri, Stéphane Thiery, and Olivier Gibaru (2020). "Learning prohibited and authorised grasping locations from a few demonstrations". In: *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1094–1100.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner (2017). "Beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework". In: *5th International Conference on Learning Representations (ICLR)*.

Stefan Hinterstoisser, Cedric Cagniart, Slobodan Ilic, Peter Sturm, Nassir Navab, Pascal Fua, and Vincent Lepetit (2012). "Gradient Response Maps for Real-Time Detection of Textureless Objects". In: *IEEE Transactions*

on Pattern Analysis and Machine Intelligence 34.5, pp. 876–888. DOI: 10.1109/TPAMI.2011.206.

Stefan Hinterstoisser, 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: *Computer Vision – ACCV 2012*. Ed. by Kyoung Mu Lee, Yasuyuki Matsushita, James M. Rehg, and Zhanyi Hu. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 548–562. ISBN: 978-3-642-37331-2.

Alain Horé and Djemel Ziou (2010). "Image Quality Metrics: PSNR vs. SSIM". In: *International Conference on Pattern Recognition*, pp. 2366–2369.

Timothy M Hospedales, Antreas Antoniou, Paul Micaelli, and Amos J. Storkey (2021). "Meta-Learning in Neural Networks: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1. DOI: 10.1109/TPAMI.2021.3079209.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei (2018). "Relation Networks for Object Detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Isabella Huang, Yashraj S. Narang, Clemens Eppner, Balakumar Sundaralingam, Miles Macklin, Ruzena Bajcsy, Tucker Hermans, and Dieter Fox (2022). "DefGraspSim: Physics-Based Simulation of Grasp Outcomes for 3D Deformable Objects". In: *IEEE Robotics and Automation Letters* 7, pp. 6274–6281.

Yun Jiang, Stephen Moseson, and Ashutosh Saxena (2011). "Efficient grasping from RGBD images: Learning using a new rectangle representation". In: *2011 IEEE International Conference on Robotics and Automation*, pp. 3304–3311.

Alexander B. Jung, Kentaro Wada, Jon Crall, Satoshi Tanaka, Jake Graving, Christoph Reinders, Sarthak Yadav, Joy Banerjee, Gábor Vecsei, Adam Kraft, Zheng Rui, Jirka Borovec, Christian Vallentin, Semen Zhydenko, Kilian Pfeiffer, Ben Cook, Ismael Fernández, François-Michel De Rainville, Chi-Hung Weng, Abner Ayala-Acevedo, Raphael Meudec, Matias Laporte, et al. (2020). *imgaug*. https://github.com/aleju/imgaug. Online; accessed 01-Feb-2020.

Rishabh Kabra, Daniel Zoran, Goker Erdogan, Loic Matthey, Antonia Creswell, Matt Botvinick, Alexander Lerchner, and Chris Burgess (2021). "SIMONe: View-Invariant, Temporally-Abstracted Object Representations via Unsupervised Video Decomposition". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 20146–20159.

Dmitry Kalashnikov, Alex Irpan, Peter Pastor Sampedro, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, and Sergey Levine (2018). "QT-Opt: Scalable Deep Reinforcement Learning for Vision-Based Robotic Manipulation". In: *The Conference on Robot Learning (CoRL)*.

Konstantinos Kallidromitis, Denis Gudovskiy, Kozuka Kazuki, Ohama Iku, and Luca Rigazio (2021). "Contrastive Neural Processes for Self-Supervised Learning". In: *Asian Conference on Machine Learning*. URL: http://www.acml-conf.org/2021/conference/accepted-papers/266/.

Michael Kampffmeyer, Yinbo Chen, Xiaodan Liang, Hao Wang, Yujia Zhang, and Eric P. Xing (2019). "Rethinking Knowledge Graph Propagation for Zero-Shot Learning". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seyed Hamidreza Mohades Kasaei and Mohammadreza Mohades Kasaei (2021). "MVGrasp: Real-Time Multi-View 3D Object Grasping in Highly Cluttered Environments". In: *ArXiv* abs/2103.10997.

Roman Kaskman, Sergey Zakharov, Ivan S. Shugurov, and Slobodan Ilic (2019). "HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects". In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*.

Alexander Kasper, Zhixing Xue, and Rüdiger Dillmann (2012). "The KIT object models database: An object model database for object recognition, localization and manipulation in service robotics". In: *The International Journal of Robotics Research* 31, pp. 927–934.

Furkan Kaynar, Sudarshan Rajagopalan, Shaobo Zhou, and Eckehard G. Steinbach (2023). "Remote Task-oriented Grasp Area Teaching By Non-Experts through Interactive Segmentation and Few-Shot Learning". In: *ArXiv* abs/2303.10195.

Hyunjik Kim, Andriy Mnih, Jonathan Schwarz, Marta Garnelo, Ali Eslami, Dan Rosenbaum, Oriol Vinyals, and Yee Whye Teh (2019). "Attentive Neural Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SkE6PjC9KX.

Diederik P. Kingma and Jimmy Ba (2015). "Adam: A Method for Stochastic Optimization". In: *International Conference on Learning Representations (ICLR)*.

Diederik P. Kingma and Max Welling (2014). "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations (ICLR)*.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and

Klaus Greff (2022). “Conditional Object-Centric Learning from Video”. In: *International Conference on Learning Representations (ICLR)*.

Kilian Kleeberger, Richard Bormann, Werner Kraus, and Marco F. Huber (2020). “A Survey on Learning-Based Robotic Grasping”. In: *Current Robotics Reports* 1, pp. 239–249.

Takumi Kobayashi and Nobuyuki Otsu (2010). “Von Mises-Fisher Mean Shift for Clustering on a Hypersphere”. In: *2010 20th International Conference on Pattern Recognition*, pp. 2130–2133.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov (2015). “Siamese neural networks for one-shot image recognition”. In: *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1–8.

Julius von Kügelgen, Ivan Ustyuzhaninov, Peter V. Gehler, Matthias Bethge, and Bernhard Schölkopf (2020). “Towards causal generative scene models via competition of experts”. In: *International Conference on Learning Representations (ICLR) Workshop: Causal learning for decision making*.

Loic Landrieu and Martin Simonovsky (2018). “Large-Scale Point Cloud Semantic Segmentation With Superpoint Graphs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Byung-Jun Lee, Seunghoon Hong, and Kee-Eung Kim (2020). “Residual Neural Processes”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04, pp. 4545–4552.

Chung-Wei Lee, Wei Fang, Chih-Kuan Yeh, and Yu-Chiang Frank Wang (2018). “Multi-Label Zero-Shot Learning With Structured Knowledge Graphs”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hayeon Lee, Eunyoung Hyung, and Sung Ju Hwang (2021a). “Rapid Neural Architecture Search by Learning to Generate Graphs from Datasets”. In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rkQuFUmUOg3.

Taeyeop Lee, Byeong-uk Lee, Inkyu Shin, Jaesung Choe, Ukcheol Shin, In-So Kweon, and Kuk-Jin Yoon (2021b). “UDA-COPE: Unsupervised Domain Adaptation for Category-level Object Pose Estimation”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14871–14880.

Taeyeop Lee, Jonathan Tremblay, Valts Blukis, Bowen Wen, Byeong-Uk Lee, Inkyu Shin, Stan Birchfield, In So Kweon, and Kuk-Jin Yoon (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.

Ian Lenz, Honglak Lee, and Ashutosh Saxena (2013). "Deep learning for detecting robotic grasps". In: *The International Journal of Robotics Research* 34, pp. 705–724.

Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen (2016). "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection". In: *The International Journal of Robotics Research* 37, pp. 421–436.

B. Li, Zhengxing Sun, Qian Li, Yunjie Wu, and Anqi Hu (2019). "Group-Wise Deep Object Co-Segmentation With Co-Attention Recurrent Neural Network". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8518–8527.

Hao Li, Pratik Chaudhari, Hao Yang, Michael Lam, Avinash Ravichandran, Rahul Bhotika, and Stefano Soatto (2020a). "Rethinking the Hyperparameters for Fine-tuning". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=B1g8VkHFPH.

Liangzhi Li, Bowen Wang, Manisha Verma, Yuta Nakashima, Ryo Kawasaki, and Hajime Nagahara (2021a). "SCOUTER: Slot Attention-based Classifier for Explainable Image Recognition". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1026–1035.

Nanbo Li, Cian Eastwood, and Robert Fisher (2020b). "Learning Object-Centric Representations of Multi-Object Scenes from Multiple Views". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 5656–5666.

Nanbo Li, Muhammad Ahmed Raza, Wenbin Hu, Zhaole Sun, and Robert Fisher (2021b). "Object-Centric Representation Learning with Generative Spatial-Temporal Factorization". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 10772–10783.

Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang (2020c). "FSS-1000: A 1000-Class Dataset for Few-Shot Segmentation". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yumeng Li, Ning Gao, Hanna Ziesche, and Gerhard Neumann (2022). "Category-Agnostic 6D Pose Estimation with Conditional Neural Processes". In: *Women in Computer Vision (WiCV) workshop, CVPR*.

Zhidong Liang, Ming Yang, Liuyuan Deng, Chunxiang Wang, and Bing Wang (2019). "Hierarchical Depthwise Graph Convolutional Neural Network for 3D Semantic Segmentation of Point Clouds". In: *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8152–8158. DOI: 10.1109/ICRA.2019.8794052.

Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li (2021a). "DualPoseNet: Category-level 6D Object Pose and Size Estimation Using Dual Pose Network with Refined Learning of Pose Consistency". In: *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3540–3549.

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

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). "Microsoft COCO: Common Objects in Context". In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 740–755.

Yen-Chen Lin, Peter R. Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin (2021b). "iNeRF: Inverting Neural Radiance Fields for Pose Estimation". In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1323–1330.

Bingzheng Liu, Jianjun Lei, Bo Peng, Chuanbo Yu, Wanqing Li, and Nam Ling (2021a). "Novel View Synthesis from a Single Image via Unsupervised learning". In: *ArXiv* abs/2110.15569.

Hanxiao Liu, Karen Simonyan, and Yiming Yang (2019). "DARTS: Differentiable Architecture Search". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=S1eYHoC5FX.

Lu Liu, William L. Hamilton, Guodong Long, Jing Jiang, and Hugo Larochelle (2021b). "A Universal Representation Transformer Layer for Few-Shot Image Classification". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=04cII6MumYV.

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

Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang (2022). "Gen6D: Generalizable Model-Free 6-DoF Object Pose Estimation from RGB Images". In: *The European Conference on Computer Vision (ECCV)*.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf (2020). "Object-Centric Learning with Slot Attention". In: *Advances in Neural Information Processing Systems*. Vol. 33, pp. 11525–11538.

Gerrit Lochmann, Bernhard Reinert, Arend Buchacher, and Tobias Ritschel (2016). "Real-time Novel-view Synthesis for Volume Rendering Using a

Piecewise-analytic Representation". In: *Vision, Modeling and Visualization (VMV)*.

Jianwu Long, Zeran yan, and Hongfa chen (2021). "A Graph Neural Network for superpixel image classification". In: *Journal of Physics: Conference Series* 1871.1, p. 012071. DOI: 10.1088/1742-6596/1871/1/012071. URL: https://doi.org/10.1088/1742-6596/1871/1/012071.

Christos Louizos, Xiahan Shi, Klamer Schutte, and Max Welling (2019a). "The Functional Neural Process". In: *Advances in Neural Information Processing Systems*.

– (2019b). "The Functional Neural Process". In: *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/db182d2552835bec774847e06406b Paper.pdf.

Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg (2017a). "Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics". In: *Robotics: Science and Systems (RSS)*.

Jeffrey Mahler, Matthew Matl, Xinyu Liu, Albert Li, David V. Gealy, and Ken Goldberg (2017b). "Dex-Net 3.0: Computing Robust Vacuum Suction Grasp Targets in Point Clouds Using a New Analytic Model and Deep Learning". In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8.

Jeffrey Mahler, Florian T. Pokorny, Brian Hou, Melrose Roderick, Michael Laskey, Mathieu Aubry, Kai J. Kohlhoff, Torsten Kröger, James J. Kuffner, and Ken Goldberg (2016). "Dex-Net 1.0: A cloud-based network of 3D objects for robust grasp planning using a Multi-Armed Bandit model with correlated rewards". In: *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1957–1964.

Vincent Mayer, Qian Feng, Jun Deng, Yunlei Shi, Zhaopeng Chen, and Alois Knoll (2022). "FFHNet: Generating Multi-Fingered Robotic Grasps for Unknown Objects in Real-time". In: *2022 International Conference on Robotics and Automation (ICRA)*, pp. 762–769.

Mateusz Michalkiewicz, Sarah Parisot, Stavros Tsogkas, Mahsa Baktashmotlagh, Anders P. Eriksson, and Eugene Belilovsky (2020). "Few-Shot Single-View 3-D Object Reconstruction with Compositional Priors". In: *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXV*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Vol. 12370. Lecture Notes in Com-

puter Science. Springer, pp. 614–630. DOI: 10.1007/978-3-030-58595-2_37. URL: https://doi.org/10.1007/978-3-030-58595-2%5C_37.

Ben Mildenhall, 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: *The European Conference on Computer Vision (ECCV)*.

Douglas Morrison, Peter Corke, and J. Leitner (2019). "Learning robust, real-time, reactive robotic grasping". In: *The International Journal of Robotics Research* 39, pp. 183–201.

– (2020). "EGAD! An Evolved Grasping Analysis Dataset for Diversity and Reproducibility in Robotic Manipulation". In: *IEEE Robotics and Automation Letters* 5, pp. 4368–4375.

Saeid Naderiparizi, Ke-Li Chiu, Benjamin Bloem-Reddy, and Frank D. Wood (2020). "Uncertainty in Neural Processes". In: *ArXiv* abs/2010.03753.

Rhys Newbury, Morris Gu, Lachlan Chumbley, Arsalan Mousavian, Clemens Eppner, Jürgen Leitner, Jeannette Bohg, Antonio Morales, Tamim Asfour, Danica Kragic, Dieter Fox, and Akansel Cosgun (2022). *Deep Learning Approaches to Grasp Synthesis: A Review*. arXiv: 2207.02556.

Renkun Ni, Micah Goldblum, Amr Sharaf, Kezhi Kong, and Tom Goldstein (2021). "Data Augmentation for Meta-Learning". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 8152–8161. URL: https://proceedings.mlr.press/v139/ni21a.html.

Alex Nichol, Joshua Achiam, and John Schulman (2018). "On First-Order Meta-Learning Algorithms". In: *ArXiv* abs/1803.02999.

Alexander Norcliffe, Cristian Bodnar, Ben Day, Jacob Moss, and Pietro Liò (2021). "Neural {ODE} Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=27acGyyI1BY.

Jishnu Jaykumar Padalunkal, Yu-Wei Chao, and Yu Xiang (2022). "FewSOL: A Dataset for Few-Shot Object Learning in Robotic Environments". In: *ArXiv* abs/2207.03333.

Ayyappa Kumar Pambala, Titir Dutta, and Soma Biswas (2021). "SML: Semantic meta-learning for few-shot semantic segmentation". In: *Pattern Recognition Letters* 147, pp. 93–99. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2021.03.036. URL: https://www.sciencedirect.com/science/article/pii/S0167865521001318.

Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox (2020). "LatentFusion: End-to-End Differentiable Reconstruction and Rendering for

Unseen Object Pose Estimation". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10707–10716.

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

Juan-Manuel Perez-Rua, Xiatian Zhu, Timothy M. Hospedales, and Tao Xiang (2020). "Incremental Few-Shot Object Detection". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Quang-Hieu Pham, Mikaela Angelina Uy, Binh-Son Hua, Duc Thanh Nguyen, Gemma Roig, and Sai-Kit Yeung (2020). "LCD: Learned cross-domain descriptors for 2D-3D matching". In: *the AAAI Conference on Artificial Intelligence*.

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

Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun (2017b). "3D Graph Neural Networks for RGBD Semantic Segmentation". In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Deirdre Quillen, Eric Jang, Ofir Nachum, Chelsea Finn, Julian Ibarz, and Sergey Levine (2018). "Deep Reinforcement Learning for Vision-Based Robotic Grasping: A Simulated Comparative Evaluation of Off-Policy Methods". In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6284–6291.

Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals (2019). "Rapid learning or feature reuse? Towards understanding the effectiveness of MAML". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pp. 2334–2345.

Janarthanan Rajendran, Alexander Irpan, and Eric Jang (2020). "Meta-Learning Requires Meta-Augmentation". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 5705–5715. URL: https://proceedings.neurips.cc/paper/2020/file/3e5190eeb51ebe6c5bbc54ee8950c548-Paper.pdf.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen (2019). "Efficient Off-Policy Meta-Reinforcement Learning via Probabilistic Context Variables". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov.

Vol. 97. Proceedings of Machine Learning Research. PMLR, pp. 5331–5340. URL: https://proceedings.mlr.press/v97/rakelly19a.html.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi (2016). "You Only Look Once: Unified, real-time object detection". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788.

Mengye Ren, Sachin Ravi, Eleni Triantafillou, Jake Snell, Kevin Swersky, Josh B. Tenenbaum, Hugo Larochelle, and Richard S. Zemel (2018). "Meta-Learning for Semi-Supervised Few-Shot Classification". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=HJcSzz-CZ.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun (2015). "Faster R-CNN: Towards real-time object detection with region proposal networks". In: *Advances in Neural Information Processing Systems*, pp. 91–99.

Eojin Rho, Daekyum Kim, Hochang Lee, and Sungho Jo (2021). "Learning Fingertip Force to Grasp Deformable Objects for Soft Wearable Robotic Glove With TSM". In: *IEEE Robotics and Automation Letters* 6, pp. 8126–8133.

Szymon M. Rusinkiewicz and Marc Levoy (2001). "Efficient variants of the ICP algorithm". In: *Proceedings Third International Conference on 3-D Digital Imaging and Modeling*, pp. 145–152.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Li Fei-Fei (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision* 115, pp. 211–252.

Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman (2006). "Using Multiple Segmentations to Discover Objects and their Extent in Image Collections". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 2, pp. 1605–1614.

Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy P. Lillicrap (2016). "Meta-Learning with Memory-Augmented Neural Networks". In: *International Conference on Machine Learning(ICML)*, pp. 1842–1850.

Henry Schaub and Alfred Schöttl (2020). "6-DOF Grasp Detection for Unknown Objects". In: *10th International Conference on Advanced Computer Information Technologies (ACIT)*, pp. 400–403.

Weijing Shi and Raj Rajkumar (2020). “Point-GNN: Graph Neural Network for 3D Object Detection in a Point Cloud”. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ivan Shugurov, Fu Li, Benjamin Busam, and Slobodan Ilic (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.

Mennatullah Siam, Boris N. Oreshkin, and Martin Jagersand (2019). “AMP: Adaptive Masked Proxies for Few-Shot Segmentation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Arjun Singh, James Sha, Karthik S. Narayan, Tudor Achim, and P. Abbeel (2014). “BigBIRD: A large-scale 3D database of object instances”. In: *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 509–516.

Jake Snell, Kevin Swersky, and Richard Zemel (2017). “Prototypical Networks for Few-shot Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2017/file/cb8da6767461f2812ae4290eac7cbc42-Paper.pdf.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan (2015). “Learning structured output representation using deep conditional generative models”. In: *Advances in Neural Information Processing Systems*, pp. 3483–3491.

Chen Song, Jiaru Song, and Qixing Huang (2020). “HybridPose: 6D Object Pose Estimation Under Hybrid Representations”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, Manolis Savva, and Thomas Funkhouser (2017). “Semantic Scene Completion From a Single Depth Image”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shuran Song, Andy Zeng, Johnny Lee, and Thomas A. Funkhouser (2019). “Grasping in the Wild: Learning 6DoF Closed-Loop Grasping From Low-Cost Demonstrations”. In: *IEEE Robotics and Automation Letters* 5, pp. 4978–4985.

Karl Stelzner, Kristian Kersting, and Adam R. Kosiorek (2021). “Decomposing 3D Scenes into Objects via Unsupervised Volume Segmentation”. In: *ArXiv* abs/2104.01148.

Jiaming Sun, 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 (CVPR)*, pp. 6825–6834.

Martin Sundermeyer, Tomas Hodan, Yann Labbe, Gu Wang, Eric Brachmann, Bertram Drost, Carsten Rother, and Jiri Matas (2023). *BOP Challenge 2022 on Detection, Segmentation and Pose Estimation of Specific Rigid Objects*. arXiv: 2302.13075 [cs.CV].

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

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip H.S. Torr, and Timothy M. Hospedales (2018). "Learning to Compare: Relation Network for Few-Shot Learning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1199–1208. DOI: 10.1109/CVPR.2018.00131.

Meng Tian, Marcelo H. Ang, and Gim Hee Lee (2020). "Shape Prior Deformation for Categorical 6D Object Pose and Size Estimation". In: *The European Conference on Computer Vision (ECCV)*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Cham: Springer International Publishing, pp. 530–546. ISBN: 978-3-030-58589-1.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel (2017). "Domain randomization for transferring deep neural networks from simulation to the real world". In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30. DOI: 10.1109/IROS.2017.8202133.

Hung-Yu Tseng, Hsin-Ying Lee, Jia-Bin Huang, and Ming-Hsuan Yang (2020). "Cross-Domain Few-Shot Classification via Learned Feature-Wise Transformation". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJl5Np4tPr.

Tinne Tuytelaars, Christoph H. Lampert, Matthew B. Blaschko, and Wray L. Buntine (2009). "Unsupervised Object Discovery: A Comparison". In: *International Journal of Computer Vision* 88, pp. 284–302.

Shinji. Umeyama (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. DOI: 10.1109/34.88573.

Mel Vecerík, Jackie Kay, Raia Hadsell, Lourdes de Agapito, and Jonathan Scholz (2022). "Few-Shot Keypoint Detection as Task Adaptation via Latent

Embeddings". In: *International Conference on Robotics and Automation (ICRA)*, pp. 1251–1257.

Rishi Veerapaneni, John D. Co-Reyes, Michael Chang, Michael Janner, Chelsea Finn, Jiajun Wu, Joshua Tenenbaum, and Sergey Levine (2020). "Entity Abstraction in Visual Model-Based Reinforcement Learning". In: *Conference on Robot Learning*, pp. 1439–1456.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Koray Kavukcuoglu, and Daan Wierstra (2016a). "Matching networks for one shot learning". In: *Advances in neural information processing systems*. Vol. 29.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, koray kavukcuoglu koray, and Daan Wierstra (2016b). "Matching Networks for One Shot Learning". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2016/file/90e1357833654983612fb05e3ec9148c-Paper.pdf.

Huy V. Vo, Patrick P'erez, and Jean Ponce (2020). "Toward unsupervised, multi-object discovery in large-scale image collections". In: *European Conference on Computer Vision (ECCV)*.

Van Huy Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce (2021). "Large-Scale Unsupervised Object Discovery". In: *Advances in Neural Information Processing Systems*. Vol. 34, pp. 16764–16778.

Michael Volpp, Fabian Flürenbrock, Lukas Grossberger, Christian Daniel, and Gerhard Neumann (2021). "Bayesian Context Aggregation for Neural Processes". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ufZN2-aehFa.

Michael Volpp, Lukas P. Fröhlich, Kirsten Fischer, Andreas Doerr, Stefan Falkner, Frank Hutter, and Christian Daniel (2020). "Meta-Learning Acquisition Functions for Transfer Learning in Bayesian Optimization". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=ryeYpJSKwr.

Risto Vuorio, Shao-Hua Sun, Hexiang Hu, and Joseph J Lim (2019). "Multimodal Model-Agnostic Meta-Learning via Task-Aware Modulation". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper/2019/file/e4da3b7fbbce2345d7772b0674a318d5-Paper.pdf.

Bram Wallace and Bharath Hariharan (2019). "Few-Shot Generalization for Single-Image 3D Reconstruction via Priors". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu (2020a). "6-PACK: Category-level 6D Pose Tracker with Anchor-Based Keypoints". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 10059–10066. DOI: 10.1109/ICRA40945.2020.9196679.

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

He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas (2019b). "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)*.

Jian Wang, Xiang Long, Yuan Gao, Errui Ding, and Shilei Wen (2020b). "Graph-PCNN: Two Stage Human Pose Estimation with Graph Pose Refinement". In: *Computer Vision – ECCV 2020*. Ed. by Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm. Cham: Springer International Publishing, pp. 492–508. ISBN: 978-3-030-58621-8.

Lei Wang, Yuchun Huang, Yaolin Hou, Shenman Zhang, and Jie Shan (2019c). "Graph Attention Convolution for Point Cloud Semantic Segmentation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pengyuan Wang, Fabian Manhardt, Luca Minciullo, Lorenzo Garattoni, Sven Meie, Nassir Navab, and Benjamin Busam (2021a). "DemoGrasp: Few-Shot Learning for Robotic Grasping with Human Demonstration". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5733–5740.

Qi Wang and Herke Van Hoof (2020). "Doubly Stochastic Variational Inference for Neural Processes with Hierarchical Latent Variables". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, pp. 10018–10028. URL: https://proceedings.mlr.press/v119/wang20s.html.

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser (2021b). "IBRNet: Learning Multi-View Image-Based Rendering". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4688–4697.

Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang (2023). "DexGraspNet: A Large-Scale Robotic Dexterous Grasp Dataset for General Objects Based on Simulation". In: *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11359–11366.

Xiaohan Wang and Haibo He (2019). "Active learning through density clustering and meta-learning for robotic knowledge acquisition". In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1728–1735.

Yongxin Wang, Kris Kitani, and Xinshuo Weng (2021c). "Joint Object Detection and Multi-Object Tracking with Graph Neural Networks". In: *Proceedings of (ICRA) International Conference on Robotics and Automation*.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli (2004). "Image quality assessment: from error visibility to structural similarity". In: *IEEE Transactions on Image Processing* 13.4, pp. 600–612.

Wei Wei, Daheng Li, Peng Wang, Yiming Li, Wanyi Li, Yongkang Luo, and Jun Zhong (2022). "DVGG: Deep Variational Grasp Generation for Dextrous Manipulation". In: *IEEE Robotics and Automation Letters* 7, pp. 1659–1666.

Ying Wei, Peilin Zhao, and Junzhou Huang (2021). "Meta-learning Hyperparameter Performance Prediction with Neural Processes". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 11058–11067. URL: https://proceedings.mlr.press/v139/wei21c.html.

Marissa A. Weis, Kashyap Chitta, Yash Sharma, Wieland Brendel, Matthias Bethge, Andreas Geiger, and Alexander S. Ecker (2021). "Benchmarking Unsupervised Object Representations for Video Sequences". In: *Journal of Machine Learning Research* 22.183, pp. 1–61.

Yizhe Wu, Oiwi Parker Jones, Martin Engelcke, and Ingmar Posner (2021). "APEX: Unsupervised, O'bject-Centric Scene Segmentation and Tracking for Robot Manipulation". In: *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3375–3382.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao (2015). "3D ShapeNets: A deep representation for volumetric shapes". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1912–1920.

Yu Xiang, Wonhui Kim, Wei Chen, Jingwei Ji, Christopher Bongsoo Choy, Hao Su, Roozbeh Mottaghi, Leonidas J. Guibas, and Silvio Savarese (2016).

"ObjectNet3D: A Large Scale Database for 3D Object Recognition". In: *European Conference on Computer Vision*.

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

Yu Xiang, Christopher Xie, Arsalan Mousavian, and Dieter Fox (2020). "Learning RGB-D Feature Embeddings for Unseen Object Instance Segmentation". In: *CoRL*.

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba (2010). "SUN database: Large-scale scene recognition from abbey to zoo". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492.

Christopher Xie, Yu Xiang, Arsalan Mousavian, and Dieter Fox (2019). "The Best of Both Modes: Separately Leveraging RGB and Depth for Unseen Object Instance Segmentation". In: *CoRL*.

– (2021). "Unseen Object Instance Segmentation for Robotic Environments". In: *IEEE Transactions on Robotics (T-RO)*.

Junyuan Xie, Ross Girshick, and Ali Farhadi (2016). "Unsupervised Deep Embedding for Clustering Analysis". In: *International Conference on Machine Learning (ICML)*, pp. 478–487.

Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong (2017). "Towards K-Means-Friendly Spaces: Simultaneous Deep Learning and Clustering". In: *International Conference on Machine Learning (ICML)*, pp. 3861–3870.

Charig Yang, Hala Lamdouar, Erika Lu, Andrew Zisserman, and Weidi Xie (2021a). "Self-Supervised Video Object Segmentation by Motion Grouping". In: *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7177–7188.

Yiding Yang, Zhou Ren, Haoxiang Li, Chunluan Zhou, Xinchao Wang, and Gang Hua (2021b). "Learning Dynamics via Graph Neural Networks for Human Pose Estimation and Tracking". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8074–8084.

Huaxiu Yao, Long-Kai Huang, Linjun Zhang, Ying Wei, Li Tian, James Zou, Junzhou Huang, and Zhenhui Li (2021). "Improving Generalization in Meta-learning via Task Augmentation". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, pp. 11887–11897. URL: https://proceedings.mlr.press/v139/yao21b.html.

Zi Jian Yew and Gim Hee Lee (2020). "RPM-Net: Robust Point Matching Using Learned Features". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11821–11830.

Mingzhang Yin, George Tucker, Mingyuan Zhou, Sergey Levine, and Chelsea Finn (2020). "Meta-Learning without Memorization". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BklEFpEYwS.

Jaesik Yoon, Taesup Kim, Ousmane Dia, Sungwoong Kim, Yoshua Bengio, and Sungjin Ahn (2018). "Bayesian Model-Agnostic Meta-Learning". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc.

Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu (2022). "Unsupervised Discovery of Object Radiance Fields". In: *International Conference on Learning Representations (ICLR)*.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine (2019). "Meta-World: A Benchmark and Evaluation for Multi-Task and Meta Reinforcement Learning". In: *CoRR* abs/1910.10897. URL: http://arxiv.org/abs/1910.10897.

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

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen (2020). "DeepEMD: Few-Shot Image Classification With Differentiable Earth Mover's Distance and Structured Classifiers". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kaifeng Zhang, Yang Fu, Shubhankar Borse, Hong Cai, Fatih Porikli, and Xiaolong Wang (2022). "Self-Supervised Geometric Correspondence for Category-Level 6D Object Pose Estimation in the Wild". In: *ArXiv* 2210.07199.

Lu Zhang, Shuigeng Zhou, Jihong Guan, and Ji Zhang (2021a). "Accurate Few-Shot Object Detection With Support-Query Mutual Guidance and Hybrid Loss". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14424–14432.

Penghao Zhang, Jiayue Li, Yining Wang, and Judong Pan (2021b). "Domain Adaptation for Medical Image Segmentation: A Meta-Learning Method". In: *Journal of Imaging* 7.2. ISSN: 2313-433X. DOI: 10.3390/jimaging7020031. URL: https://www.mdpi.com/2313-433X/7/2/31.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang (2018). "The Unreasonable Effectiveness of Deep Features as a Perceptual

Metric". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ting Zhao and Xiangqian Wu (2019). "Pyramid Feature Attention Network for Saliency Detection". In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun (2016). "Fast Global Registration". In: *European Conference on Computer Vision*.

Luyang Zhu, Arsalan Mousavian, Yu Xiang, Hammad Mazhar, Jozef van Eenbergen, Shoubhik Debnath, and Dieter Fox (2021). "RGB-D Local Implicit Function for Depth Completion of Transparent Objects". In: *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4647–4656.

Barret Zoph and Quoc V. Le (2017). "Neural Architecture Search with Reinforcement Learning". In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=r1Ue8Hcxg.