GAPartNet: Cross-Category Domain-Generalizable Object Perception and Manipulation via Generalizable and Actionable Parts

Haoran Geng [1,2,3*]   Helin Xu [4*]   Chengyang Zhao [1*]
Chao Xu [5]   Li Yi [4]   Siyuan Huang [3]   He Wang[1,2†]
[1]CFCS, Peking University   [2]School of EECS, Peking University
[3]Beijing Institute for General Artificial Intelligence
[4]Tsinghua University   [5]University of California, Los Angeles
https://pku-epic.github.io/GAPartNet

Figure 1. **Overview.** We propose to learn generalizable object perception and manipulation skills via **G**eneralizable and **A**ctionable **Parts**, and present **GAPartNet**, a large-scale interactive dataset with rich part annotations. We propose a domain generalization method for cross-category part segmentation and pose estimation. Our GAPart definition boosts cross-category object manipulation and can transfer to real.

Abstract

*For years, researchers have been devoted to generalizable object perception and manipulation, where cross-category generalizability is highly desired yet underexplored. In this work, we propose to learn such cross-category skills via **G**eneralizable and **A**ctionable **Parts** (**GAParts**). By identifying and defining 9 GAPart classes (lids, handles, etc.) in 27 object categories, we construct a large-scale part-centric interactive dataset, GAPartNet, where we provide rich, part-level annotations (semantics, poses) for 8,489 part instances on 1,166 objects. Based on GAPartNet, we investigate three cross-category tasks: part segmentation, part pose estimation, and part-based object manipulation. Given the significant domain gaps between seen and unseen object categories, we propose a robust 3D segmentation method from the perspective of domain generalization by integrating adversarial learning techniques. Our method outperforms all existing methods by a large margin, no matter on seen or unseen categories. Furthermore, with part segmentation and pose estimation results, we leverage the GAPart pose definition to design part-based manipulation heuristics that can generalize well to unseen object categories in both the simulator and the real world.*

1. Introduction

Generalizable object perception and manipulation are at the core of building intelligent and multi-functional robots. Recent efforts on generalizing the vision have been devoted to category-level object perception that deals with perceiving novel object instances from known object categories, including object detectors from RGB images [17, 21, 46], point clouds [5, 19], and category-level pose estimation works on rigid [4, 53] and articulated objects [27, 59]. On the front of generalizable manipulation, complex tasks that involve interacting with articulated objects have also been proposed in a category-level fashion, as in the recent challenge on learning category-level manipulation skills [38]. Additionally, to boost robot perception and manipulation

---

with indoor objects, researchers have already proposed several datasets [37, 57, 61, 66, 68] with part segmentation and motion annotations, and have devoted work to part segmentation [37, 68] and articulation estimation [27].

However, these works all approach the object perception and manipulation problems in an intra-category manner, while humans can well perceive and interact with instances from unseen object categories based on prior knowledge of functional parts such as buttons, handles, lids, *etc*. In fact, parts from the same classes have fewer variations in their shapes and the ways that we manipulate them, compared to objects from the same categories. We thus argue that part classes are more elementary and fundamental compared to object categories, and generalizable visual perception and manipulation tasks should be conducted at part-level.

Then, what defines a part class? Although there is no single answer, we propose to identify part classes that are generalizable in both recognition and manipulation. After careful thoughts and expert designs, we propose the concept of *Generalizable and Actionable Part (GAPart)* classes. Parts from the same GAPart class share similar shapes which allow generalizable visual recognition; parts from the same GAPart class also have aligned actionability and can be interacted with in a similar way, which ensures minimal human effort when designing interaction guidance to achieve generalizable and robust manipulation policies.

Along with the GAPart definition, we present GAPartNet, a large-scale interactive part-centric dataset where we gather 1,166 articulated objects from the PartNet-Mobility dataset [61] and the AKB-48 dataset [32]. We put in great effort in identifying and annotating semantic labels to 8,489 GAPart instances. Moreover, we systematically align and annotate the GAPart poses, which we believe serve as the bridge between visual perception and manipulation. Our class-level GAPart pose definition highly couples the part poses with how we want to interact with the parts. We show that this is highly desirable – once the part poses are known, we can easily manipulate the parts using simple heuristics.

Based on the proposed dataset, we further explore three cross-category tasks based on GAParts: part segmentation, part pose estimation, and part-based object manipulation, where we aim at recognizing and interacting with the parts from novel objects in both known categories and, moreover, unseen object categories. In this work, we propose to use learning-based methods to deal with perception tasks, after which, based on the GAPart definition, we devise simple heuristics to achieve cross-category object manipulation.

However, different object categories may contain different kinds of GAParts and provide different contexts for the parts. Each object category thus forms a unique domain for perceiving and manipulating GAParts. Therefore, all three tasks demand domain-generalizable methods that can work on unseen object categories without seeing them during training, which is very challenging for existing vision and robotic algorithms. We thus consult the generalization literature [12, 13, 25] and propose to learn domain-invariant representation, which is often achieved by domain adversarial learning with a domain classifier. During training, the classifier tries to distinguish the domains while the feature extractor tries to fool the classifier, which encourages domain-invariant feature learning. However, it is highly non-trivial to adopt adversarial learning in our domain-invariant feature learning, due to the following challenges. 1) *Handling huge variations in part contexts across different domains.* The context of a GAPart class can vary significantly across different object categories. For example, in training data, round handles usually sit on the top of lids for the CoffeeMachine category, whereas for the test category Table, round handles often stand to the front face of the drawers. To robustly segment GAParts in objects from unseen categories, we need the part features to be context-invariant. 2) *Handling huge variations in part sizes.* Parts from different GAPart classes may be in different sizes, *e.g.*, a button is usually much smaller than a door. Given that the input is a point cloud, the variations in part sizes will result in huge variations in the number of points across different GAParts, which makes feature learning very challenging. 3) *Handling the imbalanced part distribution and part-object relations.* Object parts in the real world distribute naturally unevenly and a particular part class may appear with different frequencies throughout various object categories. For example, there can be more buttons than doors on a washing machine while the opposite is true in the case of a storage furniture. This imbalanced distribution also adds difficulties to the learning of domain-invariant features.

Accordingly, we integrate several important techniques from domain adversarial learning. To improve context invariance, we propose a part-oriented feature query technique that mainly focuses on foreground parts and ignores the background. To handle diverse part sizes, we propose a multi-resolution technique. Finally, we employ the focal loss to handle the distribution imbalance. Our method significantly outperforms previous 3D instance segmentation methods and achieves 76.5% AP50 on seen object categories and 37.2% AP50 on unseen categories.

To summarize, our main contributions are as follows:

1. We provide the concept of *GAPart* and present a large-scale interactive dataset, *GAPartNet*, with rich part semantics and pose annotations that facilitates generalizable part perception and part-based object manipulation.

2. We propose a first-ever pipeline for domain-generalizable 3D part segmentation and pose estimation via learning domain-invariant features, which significantly outperforms the baselines.

3. We provide a new solution to generalizable object manipulation by leveraging the concept of GAParts. Thanks

to innate generalizability and actionability, minimal human effort is needed when designing interaction guidance to achieve generalizable and robust manipulation policies.

2. Related Work

Part Instance Segmentation from Point Cloud Observations. Large-scale datasets of 3D shapes are fundamental to 3D part segmentation works, *e.g.*, ShapeNet (2∼5 parts per object) [3, 66] and PartNet (15 parts per object on average) [37]. Based on such datasets, much progress has been made on unified architectures for point cloud learning [28, 43, 44, 58], specialized supervised segmentation networks [56, 67], shape abstraction and part discovery [35, 40, 63, 65], *etc*. However, these works all approach object perception in an intra-category manner. We instead tackle 3D part instance segmentation in a cross-category way, based on our newly proposed GAPartNet dataset.

Domain Generalization. To tackle the out-of-distribution problems, domain generalization methods try to learn from multiple source domains to generalize to the unseen domains, which can be divided into the following three categories [55]: 1) data manipulation methods (*e.g.*, data augmentation [72], data generation [45, 71]); 2) learning strategy design (*e.g.*, ensemble learning [64, 72], meta learning [23, 24], automated machine learning [7]); 3) domain-invariant representation learning (*e.g.*, explicit feature alignment [42, 70], domain adversarial learning [12, 13, 25, 29, 39]). However, works on domain generalization mainly focus on 2D tasks (*e.g.*, image classification), whose techniques are not suitable to be directly used in our 3D multi-stage part segmentation and pose estimation tasks. [33, 34] try to discover parts in a category-agnostic manner, but their task settings are also different from ours. In our tasks, we need to tackle irregular point cloud representation and take the multi-stage, multi-part setting into account.

Category-level Object Pose Estimation. Pose estimation has been studied at instance-level as well as category-level. Instance-level object pose estimation works [18, 22, 30, 41, 47, 52, 62] assume known CAD models and thus have their limitations. Other works, on the other hand, deal with 3D bounding boxes prediction and 6D pose estimation at category-level, including single-frame pose estimation such as NOCS [53], FS-Net [6], CASS [4, 54], and category-level tracking such as 9-PACK [51], CAPTRA [59]. Wang *et al.* [53] innovates Normalized Object Coordinate Space (NOCS), a unified coordinate space where objects from the same category are normalized, canonicalized and share an identical orientation. CASS [4] learns a canonical latent shape space for certain object categories, while [48] leverages category shape priors and models shape deformations to handle intra-class shape variations. FS-Net [6] designs

a fast shape-based network that extracts efficient category-level pose features. [54] uses a cascaded relation network to relate 2D, 3D, shape priors, and proposes a recurrent reconstruction network to make iterative improvements.

Generalizable Object Manipulation. On the front of object manipulation, Mu *et al.* proposes [38] a challenge to learn generalizable manipulation skills for articulated objects from known categories. Although some previous methods [14, 15, 36] have certain generalizability, robotic manipulation in a novel environment still calls for the ability to handle novel object categories. Although, for simple rigid objects, there is existing literature on robust and object-agnostic object grasping [2, 9, 20] and planar pushing [26, 69] algorithms, while very few works have been devoted to interacting with articulated objects that contain movable parts. Recently, Mo *et al.* [36] and Wu *et al.* [60] tackle this problem by leveraging low-level generalizability. The most related work to us is Gadre *et al.* [11] which proposes an interactive perception pipeline learning to touch, watch, then segment the object into movable parts. However, this work does not consider the consistent geometry and actionability patterns behind parts from the same class and can only deal with simple objects with up to three parts on the table surfaces, *e.g.*, scissors and eyeglasses.
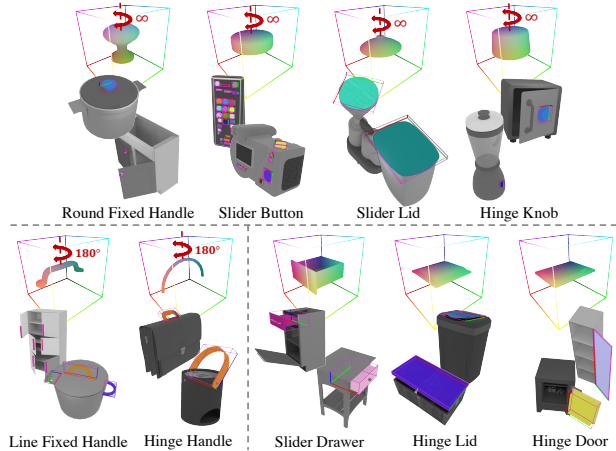

Figure 2. **GAPart Classes.** Here we highlight the parts from 9 GAPart classes along with their normalized part coordinate spaces. On the top, we show the four GAPart classes that have continuous rotation symmetry along the z axis, denoted with the red-dashed line and the ∞ remark; the bottom-left shows the two GAPart classes that have $180°$ mirror symmetry along the z axis; and the bottom-right shows the rest three asymmetric GAPart classes.

3. GAPart Definition and GAPartNet Dataset

3.1. GAPart Definition

Different from previous works, we give a rigorous definition to the GAPart classes, which not only are Generalizable to visual recognition but also share similar Action-
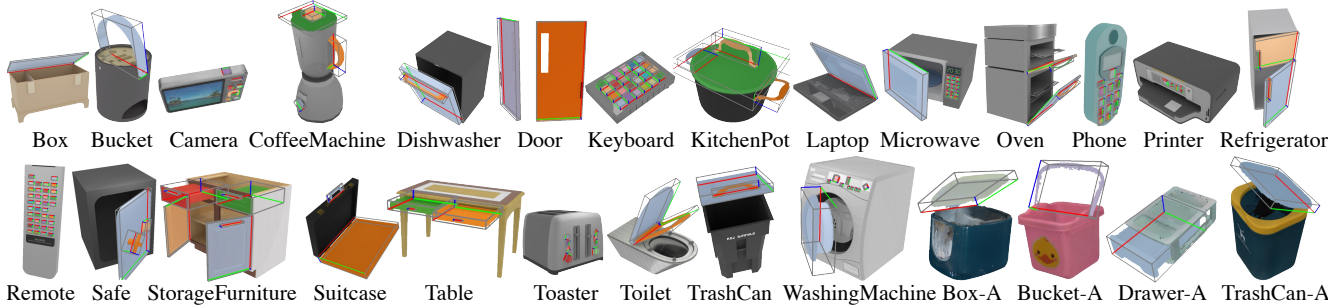
Figure 3. **GAPartNet Objects.** Objects collected from AKB-48 [32] end with '-A', while the others are from PartNet-Mobility [61].

	All	Bo	Bu	Ca	Co	Di	Do	Ke	Ki	La	Mi	Ov	Ph	Pr	Ref	Rem	Sa	St	Su	Ta	Toa	Toi	Tr	Wa	Bo-A	Bu-A	Dr-A	Tr-A
Object	1,166	25	31	32	41	41	14	31	20	48	16	23	15	28	37	49	29	324	10	77	19	66	52	17	40	37	22	22
Ln.F.Hl.	922	2	–	–	10	28	2	–	40	–	5	29	–	–	43	–	1	667	–	60	–	–	35	–	–	–	–	–
Rd.F.Hl.	151	–	–	–	8	–	9	–	14	–	–	–	–	–	–	–	–	54	–	65	–	–	1	–	–	–	–	–
Hg.Hl.	78	–	31	–	–	–	–	–	–	–	–	–	–	–	–	–	–	10	–	–	–	–	–	–	–	37	–	–
Hg.Ld.	260	49	–	–	1	–	–	–	–	48	–	1	–	–	–	–	1	7	–	–	–	55	31	5	40	–	–	22
Sd.Ld.	89	–	–	1	19	–	–	–	20	–	–	–	–	–	–	–	–	–	–	–	–	44	5	–	–	–	–	–
Sd.Bn.	5,526	–	–	208	140	5	–	2,934	1	–	39	17	227	311	–	1,433	89	–	2	–	24	15	–	81	–	–	–	–
Sd.Dw.	546	1	–	–	–	1	–	–	–	–	–	–	–	6	–	–	–	333	–	161	–	–	7	–	–	–	37	–
Hg.Dr.	678	–	–	–	–	41	18	–	–	–	16	28	–	–	60	–	29	433	–	24	–	–	15	14	–	–	–	–
Hg.Kb.	239	–	–	11	47	2	–	–	–	–	8	77	–	1	–	1	37	–	–	–	28	–	–	27	–	–	–	–

Table 1. **GAPartNet Statistics.** We show how the GAPart instances distribute across all object categories, where object categories titles are in the same order as in Fig. 3, *e.g.*, **Bo** = Box, **Bu** = Bucket, *etc.* The first row **Object** shows the object number for each object category. The following rows, *i.e.*, line fixed handle, round fixed handle, hinge handle, hinge lid, slider lid, slider button, slider drawer, hinge door, and hinge knob, show the number of GAPart instances in each object category.

ability, corresponding to the G and A in GAPartNet. Our main purpose of such a definition is to bridge the perception and manipulation, to allow joint learning of both vision and interaction. Accordingly, we propose two principles to follow: **firstly, geometric similarity within part classes,** and **secondly, actionability alignment within part classes.**

GAPart Semantics. Based on such principles, we identify 9 common GAPart classes across 27 object categories: *line fixed handle, round fixed handle, hinge handle, hinge lid, slider lid, slider button, slider drawer, hinge door, hinge knob*. Note that based on different actionability, handles are split into *fixed* handles and *hinge* handles, while lids are split into *hinge* lids and *slider* lids. We further identify *line* fixed handles and *round* fixed handles according to their difference in geometry.

GAPart Poses. Following previous works [27,53], we define the canonicalized part position and orientation in Normalized Part Coordinate Space (NPCS) for each GAPart class. We illustrate our pose definition in Fig. 2. Note that some of the GAPart classes have innate symmetry, which should be taken care of when dealing with their poses.

Based on the rigorous and manipulation-oriented definition, simple heuristics can be designed to achieve generalizable part-based manipulation across different object categories, once we know the part classes and the part poses.

3.2. GAPartNet Dataset

Following the GAPart definition, we construct a large-scale part-centric interactive dataset, GAPartNet, with rich, part-level annotations for both perception and interaction tasks. Our 3D object shapes come from two existing datasets, PartNet-Mobility [61] and AKB-48 [32], which are cleaned and provided with new uniform annotations based on our GAPart definition. The final GAPartNet has 9 GAPart classes, providing semantic labels and pose annotations for 8,489 GAPart instances on 1,166 objects from 27 object categories. On average, each object has 7.3 functional parts. Each GAPart class can be seen on objects from more than 3 object categories, and each GAPart class is found in 8.8 object categories on average, which lays the foundation for our benchmark on generalizable parts.

Tab. 1 and Fig. 3 show the statistics and selected examples of GAPartNet.

3.3. Data Annotation

We direct systemic works to guarantee cross-category generalizable part semantics and pose annotations. We follow the steps below to clean and annotate our data: 1) Fixing imperfect meshes and re-merging the meshes into new parts. The average fixing time per object is 15 minutes, while the average re-merging time per part is 5 minutes.

Over 100 object instances are fixed and over 1,000 GAPart instances are newly merged. 2) Annotating cross-category semantic labels. 3) Aligning and Annotating poses. We spend more than 200 hours building a whole pipeline as well as several manually designed rules to align and annotate the poses of all GAParts.

More dataset visualizations and annotation details can be found in the supplementary materials.

4. Problem Formulation

Given the GAPart definition and the proposed GAPart-Net dataset, we investigate the problems of cross-category generalizable object perception and manipulation.

Perception. The input to our pipeline is a *partial* colored point cloud observation of the object $\mathbf{P} \in \mathbb{R}^{N \times 3}$, where N denotes the number of points. Assume the object contains L GAParts and the i-th part is with a class label $p_i \in \{1, ..., 9\}$. Then the goal for perception is as follows: for each individual GAPart, locating its segmentation masks C_i and recognizing its part pose, *i.e.*, a rotation $\mathbf{R}_i \in \mathrm{SO}(3)$, a translation $\mathbf{t}_i \in \mathbb{R}^3$, and a size $\mathbf{s}_i \in \mathbb{R}^3$.

Note that the perception tasks are carried out in a cross-category, domain-generalizable fashion, *i.e.*, the perception network is trained on objects from a set of seen object categories $\{O_j^S\}_j$ (*i.e.*, seen domains $\{D_j^S\}_j$), and is expected to generalize to unseen object categories $\{O_j^U\}_j$ (*i.e.*, unseen domains $\{D_j^U\}_j$).

Manipulation. We need to develop a pose-based interaction policy π for generalizable part-based object manipulation. Given a single partial point cloud observation $\mathbf{P} \in \mathbb{R}^{N \times 3}$, the robot needs to manipulate the target part using the previous understanding for the GAParts, *e.g.*, open a door on an object from a previous unseen object category.

5. Method

Our proposed pipeline for domain-generalizable 3D part segmentation and pose estimation is shown in Fig. 4.

5.1. Domain-generalizable 3D Part Segmentation

Architecture Overview. Following the previous works [19, 50], with the input point cloud $\mathbf{P}$, our 3D part segmentation network leverages a Sparse UNet [16] as the backbone to extract point-wise feature $\mathbf{F}$ with K channels, followed by a *Dual Set Grouping* module introduced by [19] to generate M' mask proposals $\mathcal{C}' = \{C_1', C_2', \dots, C_{M'}'\}$. The proposals are then passed through a *Scoring* module which predicts confidence scores $\mathbf{S}$, with $\mathbf{S}_i$ as the score for the proposal i, followed by Non-Maximum Suppression (NMS) to output final M segmentation masks $\mathcal{C} = \{C_1, C_2, \dots, C_M\}$. Most importantly, to enable domain-invariant feature extraction for mask proposals and tackle

the aforementioned challenges, we introduce a domain adversarial training strategy for 3D part segmentation to help learn domain-invariant features.

Domain-invariant GAPart Feature Learning. Inspired by [12, 13, 25], we introduce a domain classifier $\mathcal{D}$ and a Gradient Reverse Layer (GRL) at the training time for domain adversarial training, as shown in Fig. 4. Specifically, the classifier $\mathcal{D}$ takes the features as input and tries to distinguish the different domains, while the GRL passes the negative gradients of the classification back to the feature extractor, which encourages domain-invariant feature extraction during this adversarial training procedure.

Furthermore, to address the challenges mentioned in Sec. 4, we consider 1) how to process the feature from the part segmentation pipeline to make the GAPart feature domain-invariant, 2) where to place the domain classifier to better tackle parts with different sizes, and 3) how to do domain adversarial training to deal with the distribution-imbalance. The designed techniques are as follows.

1) Part-oriented Feature Query ($\mathbf{Q}$). To better handle the huge variations in part contexts across different domains, the part features need to be context-invariant and contain less domain-relative information. An intuitive design is to make the domain classifier $\mathcal{D}$ part-oriented (*i.e.*, taking foreground part features as input and domain labels as output), which can help the feature extractor focus on the foreground (*i.e.*, the GAParts) rather than the background (*i.e.*, the rest of the object bodies). Specifically, we query the features of mask proposals $\{\mathbf{F}_{C_i'}\}_i$ with scores above the threshold $\mathbf{s}_{thre}$ from the feature $\mathbf{F}$ and pass them to the domain classifier. The domain discrimination loss is

$$\mathcal{L}_{\mathbf{Q}-\mathrm{adv}}(\mathbf{F}) = \frac{1}{M_s'} \sum_{i=1}^{M'} \mathbb{1}_{\{\mathbf{S}_i > \mathbf{s}_{thre}\}} \mathcal{L}_{\mathrm{cls}}^{\mathrm{adv}}(\mathcal{D}(\mathbf{F}_{C_i'}), d_i),$$

where M_s' indicates the number of proposals with scores above the threshold, d_i is the domain label (*i.e.*, object category) of the mask proposal C_i', and $\mathcal{L}_{\mathrm{cls}}^{\mathrm{adv}}(\cdot, \cdot)$ denotes the domain classification loss.

2) Multi-resolution ($\mathbf{R}$). Part instances come in significantly different sizes, *e.g.*, a door can be an order of magnitude larger than a handle. We thus propose to extract the mask proposal features from different UNet layers in different resolutions, so that the size variances of GAParts can be taken care of. In the implementation, we choose three hidden layers from the UNet decoder and query proposal features from the three features $\{\mathbf{F}^l\}_l$ respectively.

Combined with multi-resolution, $\mathcal{L}_{\mathbf{Q}-\mathrm{adv}}$ can be rewritten as follows:

$$\mathcal{L}_{\mathbf{QR}-\mathrm{adv}}(\{\mathbf{F}^l\}_l) = \sum_{l=1}^{3} w_l \mathcal{L}_{\mathbf{Q}-\mathrm{adv}}(\mathbf{F}^l),$$

where $\mathcal{L}_{\mathbf{Q}-\mathrm{adv}}(\mathbf{F}^l)$ indicates the domain discrimination loss for features queried from the lth layer and w_l is the corresponding weight for each layer.
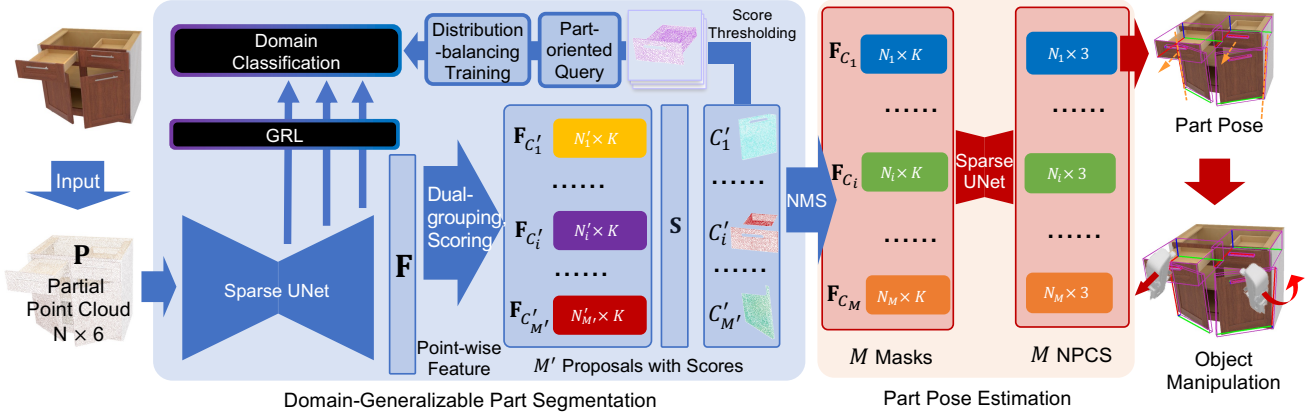
Figure 4. **An Overview of Our Domain-generalizable Part Segmentation and Pose Estimation Method.** We introduce a part-oriented domain adversarial training strategy that can tackle multi-resolution features and distribution imbalance for the domain-invariant GAPart feature extraction. The training strategy tackles the challenges in our tasks and dataset, significantly improving the generalizability of our method for part segmentation and pose estimation.

Note that these multi-resolution features only serve domain-adversarial learning for parts with different sizes and are not involved in the grouping for mask proposals.

3) Distribution-balancing (**B**). As is often the case in the real world, part instances on different objects can be extremely imbalanced. We thus introduce a part-level domain discrimination focal loss inspired by [31] for adversarial training to tackle this problem. Combining with distribution-balancing, $\mathcal{L}_{\mathbf{Q}-\mathrm{adv}}$ can be modified as follows:

$$\mathcal{L}_{\mathbf{QB}-\mathrm{adv}}(\mathbf{F}) = \frac{1}{M'_s} \sum_{i=1}^{M'} \mathbb{1}_{\{\mathbf{S}_i > \mathbf{s}_{thre}\}} w_{d_i}^{p_i} \mathcal{L}_{\mathrm{cls}}^{\mathrm{adv}}(\mathcal{D}(\mathbf{F}_{C'_i}), d_i),$$
$$w_{d_i}^{p_i} = -\alpha_{d_i}^{p_i}(1 - acc_{d_i}^{p_i})^{\gamma},$$

where for the specific part class p_i and the domain d_i of a proposal C'_i, the loss weight $w_{p_i}^{d_i}$ is determined by a hyper-parameter $\alpha_{d_i}^{p_i}$, negatively correlated with the domain distribution, $acc_{d_i}^{p_i}$, the mean accuracy of the classification for the domain d_i in the part class p_i, and γ, a hyper-parameter.

With the three techniques introduced, our proposed domain adversarial training method is part-oriented and can tackle multi-resolution features as well as distribution imbalance, which better encourages domain-invariant GAPart feature learning. The final domain adversarial loss is

$$\mathcal{L}_{\mathbf{QRB}-\mathrm{adv}}(\{\mathbf{F}^l\}_l) = \sum_{l=1}^{3} w_l \mathcal{L}_{\mathbf{QB}-\mathrm{adv}}(\mathbf{F}^l),$$

and the total loss for domain-generalizable part segmentation is as follows:

$$\mathcal{L}_{\mathrm{seg}}^{\mathrm{DG}} = \mathcal{L}_{\mathrm{seg}} + \mathcal{L}_{\mathbf{QRB}-\mathrm{adv}},$$

where $\mathcal{L}_{\mathrm{seg}}$ is the part segmentation loss without domain adversarial training.

5.2. Part Pose Estimation

NPCS Map Prediction and Pose Fitting. For each predicted part segmentation mask C_i, we query its mask feature $\mathbf{F}_{C_i}$ from the feature $\mathbf{F}$. Then the $NPCS\text{-}Net$ is used for point-wise NPCS coordinates regression. Applying RANSAC [10] for outlier removal and Umeyama algorithm [49], we estimate the 7-dimensional rigid transformation and obtain the pose of the predicted part. Based on the domain-invariant feature, thanks to our domain adversarial training, the prediction of NPCS values in our pipeline can be independent of the context, color, etc. of the part. This can significantly improve the generalizability of our part pose estimation method.

Symmetry-aware Pose Estimation and Joint Prediction. To tackle the symmetries naturally existing in some GAPart classes, we design a symmetry-aware NPCS regression loss that can tolerate different symmetry patterns for different part classes. We then follow our GAPart pose definition to simplify the joint prediction procedure. For each GAPart class, the part pose definition contains a wealth of information, including the joint position and direction, where we can directly get the joint position and direction instead of relying on an additional network for estimation like [27].

5.3. Interaction Policy

Given part segmentation and pose estimation, based on the proposed GAPart pose definition where actionability information is included, we design part-pose-based, effective interaction policies for part-based object manipulation, which provide the community with a novel approach to cross-category robotic manipulation and interaction tasks.

More design and implementation details of our method can be found in the supplementary materials.

6. Experiments

6.1. Data Preparation

With our dataset described in Sec. 3, we render RGB-D images of objects with annotations using the SAPIEN en-

		Ln.F.Hl.	Rd.F.Hl.	Hg.Hl.	Hg.Ld.	Sd.Ld.	Sd.Bn	Sd.Dw.	Hg.Dr.	Hg.Kb.	Avg.AP	Avg.AP50
Seen (%)	PG [19]	86.1	23.0	84.6	80.01	88.3	49.3	62.6	92.8	34.6	57.3	66.8
	SG [50]	57.8	**93.6**	81.2	76.0	89.3	25.2	50.8	93.9	51.5	58.5	68.8
	AGP [33]	86.8	20.3	87.7	79.7	89.4	62.3	61.6	92.5	16.7	57.2	66.3
	Ours	**89.2**	54.9	**90.4**	**84.8**	**89.8**	**66.7**	**67.2**	**94.7**	**52.9**	**67.6**	**76.5**
Unseen (%)	PG [19]	32.44	9.8	2.1	26.8	0.0	42.6	57.0	63.9	1.7	21.9	26.3
	SG [50]	25.8	5.0	0.4	33.9	0.6	**51.5**	51.2	69.0	12.1	22.0	27.7
	AGP [33]	**45.6**	4.8	**3.1**	34.3	0.0	47.8	64.1	63.1	11.5	25.7	30.5
	Ours	**45.6**	**40.0**	**3.1**	**40.2**	**5.0**	49.1	**64.2**	**69.1**	**23.4**	**32.0**	**37.2**

Table 2. **Results of Part Segmentation on Seen Object Categories and Unseen Object Categories in terms of Per-part-class AP50 (%), Average AP50 (%), and Average AP (%).** Ln.=Line. F.=Fixed. Rd.=Round. Hl.=Handle. Ld.=Lid. Bn.=Button. Dw.=Drawer. Dr.=Door. Kb.=Knob. PG=PointGroup [19]. SG=SoftGroup [50]. AGP=baseline modified from AutoGPart [33].

use adv	use **Q**-adv	use **R**-adv	use **B**-adv	Seen (%) Avg. AP	Seen (%) Avg. AP50	Unseen (%) Avg. AP	Unseen (%) Avg. AP50
✗				61.1	71.1	22.2	28.1
✓				61.0	70.6	23.2	29.8
✓	✓			62.8	71.6	27.1	32.3
✓	✓	✓		64.9	73.7	29.6	35.0
✓	✓	✓	✓	**67.6**	**76.5**	**32.0**	**37.2**

Table 3. **Ablation Studies for Domain-generalizable Part Segmentation.** The left four columns stand for using adversarial learning, part-oriented feature query technique, multi-resolution technique, distribution-balancing technique, respectively.

		$R_e\downarrow$	$T_e\downarrow$	$S_e\downarrow$	$\theta_e\downarrow$	$d_e\downarrow$	mIoU↑	A_5↑	A_{10}↑
Seen	PG [19]	14.3	0.034	0.039	7.947	0.020	49.4	24.4	47.0
	AGP [33]	14.4	0.036	0.039	7.955	0.021	48.7	26.8	49.1
	Ours	**9.9**	**0.024**	**0.035**	**7.4**	**0.014**	**51.2**	**28.3**	**53.1**
Unseen	PG [19]	18.2	0.056	0.073	12.0	0.031	36.2	19.2	42.9
	AGP [33]	18.2	0.57	0.076	11.9	0.029	36.3	20.8	46.5
	Ours	**14.8**	**0.047**	**0.067**	**11.3**	**0.024**	**40.6**	**23.4**	**51.6**

Table 4. **Results of Part Pose Estimation on Seen and Unseen Object Categories in terms of R_e (°), T_e (cm), S_e (cm), θ_e (°), d_e (cm), mIoU=3D mIoU (%), A_5=5°5cm accuracy (%), A_{10}=10°10cm accuracy (%).** PG=baseline modified from PointGroup [19]. AGP=baseline modified from AutoGPart [33].
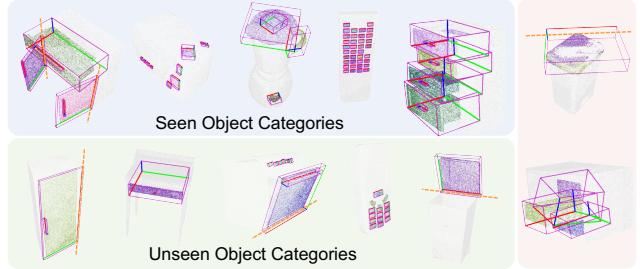

Figure 5. **Qualitative Results of Perception.** Left two figures show the results of cross-category part segmentation and pose estimation on seen and unseen categories, while the right shows failure cases. Here we only show the revolute joint estimation results.

vironment [61] and obtain point cloud observations from back-projection. To study the cross-category generalizability of our method, we split the 27 object categories into 17 seen and 10 unseen categories, ensuring that all GAPart classes exist in both seen and unseen object categories. We train the network on seen categories and evaluate its GAPart understanding on unseen categories.

6.2. Cross-category Part Segmentation

Evaluation Metrics. Following the previous 3D semantic instance segmentation benchmarks in ScanNetV2 [8] and S3DIS [1], we use the widely-adopted metric average precision to evaluate the performance of part segmentation. Specifically, AP50, the average precision with Intersection over Union (IoU) threshold 50%, is used to evaluate the performance on each part class and the overall performance. As a complementary, we also use AP, the average precision

averages over IoU thresholds from 50% to 95% with a step of 5%, to evaluate the overall performance.

Main Results. Tab. 2 shows the quantitative comparisons between our method and previous state-of-the-art methods of 3D semantic instance segmentation (*i.e.*, PointGroup [19], SoftGroup [50]). We also set up a baseline modified from AutoGPart [33], whose task is different from ours thus we directly combine their methods with the original PointGroup [19] pipeline for comparison.

In both seen and unseen object categories, our method shows significant improvement compared to the others. For AP50, our method achieves 76.5% in seen categories, which beats the second-runner by absolutely 7.7% and relatively 11.2%. In unseen categories, our method achieves 37.2%, absolutely 6.7% and relatively 22.0% better than the second-runner, which shows significant relative improvement in unseen categories. It shows that our method could extract better domain-invariant features for parts and thus have great generalizability across object categories.

Ablation Studies. We conduct sufficient comparisons to demonstrate that our techniques contribute significantly to the generalizability across object categories, as shown in Table 3. Comparing the top two rows, we show that the domain adversarial training with the object global features as input helps the generalization to unseen categories, but somewhat at the expense of performance in seen categories. With our part-oriented feature query technique

(rows 2,3), the performance improves no matter on seen or unseen categories. The multi-resolution technique also contributes to the performance in the two areas (rows 3,4). The distribution-balancing technique (rows 4,5) takes the performance of our method a step further and achieves strong precision and generalizability.

6.3. Cross-Category Part Pose Estimation

Evaluation Metrics. We use the following metrics to evaluate the performance of part pose estimation: R_e (°), average rotation error; T_e (cm), average translation error; S_e (cm), average scale error; θ_e (°) average rotation error of part interaction axis; d_e(cm) average translation error of part interaction axis; **3D mIoU** (%), the average 3D IoU between ground-truths and predicted bounding boxes; **5°5cm accuracy** (%), the percentage of pose predictions with rotation error $< 5°$ and translation error < 5cm; **10°10cm accuracy** (%), the percentage of pose predictions with rotation error $< 10°$ and translation error < 10cm. We evaluate part pose only when the part is detected.

Main Results. Tab. 4 shows the results of our method and the baselines on part pose estimation. We modify Point-Group [19] and AutoGPart [33] as baselines. Our method outperforms the baselines on most of the metrics in seen categories, and on all of the metrics in unseen categories, which shows the value of our domain-invariant feature extraction. With our domain adversarial training strategy and the three techniques introduced, the performance of part pose estimation improves a lot, especially in unseen categories. Qualitative results are shown in Fig. 5.

6.4. Cross-category Part-based Manipulation

We showcase the usefulness of the concept of GAPart by performing cross-category, part-based object manipulation on four basic tasks. We use SAPIEN [61] environment for simulation and set up four tasks based on SAPIEN Manipulation Skill Benchmark [38], *i.e.*, opening drawers, opening doors, using handles, pressing buttons.

Task Setting. These four tasks exemplify robot manipulation under the motion constraint of a prismatic or a revolute joint, where a gripper is used on seen and unseen categories. The success of object manipulation is defined as opening up the target part for 90% of the motion range within 1,000 time-steps. and coming to a stable stop at the end.

Heuristics Design and Experiments in the Simulator. We first do cross-category part segmentation and pose estimation using our perception method. Based on the predictions of the part poses, we move the robot arm toward the target part, turn the gripper in the direction suitable for grabbing, and then close the gripper. Finally, we move the gripper along the proposed trajectories toward the target position, following our GAPart pose definition. The results
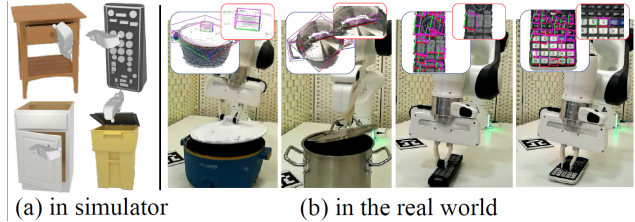

(a) in simulator   (b) in the real world

Figure 6. **Part-based Object Manipulation.** Left (a): In the simulator. Right (b): In the real world. For each subfigure in (b), the perception result is shown in the left box, while the target part is shown in the right box.

show that our perception model and manipulation heuristics can work well, achieve good performance on these tasks, and generalize to objects from unseen categories. Exemplar results are shown in Fig. 6 (a).

Real-world Experiments. Although trained on synthetic data, our method can be used in the real world. Experiments show that our method can successfully predict part segmentation and poses on real objects. We further show that cross-category part-based object manipulation can be successfully performed by robot arms using our method, as shown in Fig. 6 (b).

More experiment details, quantitative and qualitative results, and object manipulation demos can be found in the supplementary materials.

7. Conclusion

In this work, we reason that learning generalizable and actionable parts is the key to an intelligent agent capable of cross-category object perception and manipulation. We introduce the concept of GAPart and present the GAPart-Net dataset by annotating cross-category part semantics and poses. We explore three cross-category tasks based on GAParts: part segmentation, part pose estimation, and part-based object manipulation. Our proposed approach, adopting a domain generalization perspective, outperforms previous works in segmentation and pose estimation. Furthermore, we design part-pose-based interaction policies that enable effective and generalizable object manipulation in both the simulator and the real world, thanks to our GAPart definition and our domain-generalizable perception model.

Limitations. The cross-category tasks are challenging, and there is still room for improvement in generalizability. Our heuristic method for object manipulation relies on precise part pose predictions, which is an area for future research to achieve more robust manipulation strategies.

8. Acknowledgements

References

[1] Iro Armeni, Ozan Sener, Amir R. Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3d semantic parsing of large-scale indoor spaces. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2016. 7

[2] Michel Breyer, Jen Jen Chung, Lionel Ott, Siegwart Roland, and Nieto Juan. Volumetric grasping network: Real-time 6 dof grasp detection in clutter. In *Conference on Robot Learning*, 2020. 3

[3] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 3

[4] Dengsheng Chen, Jun Li, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 3

[5] Shaoyu Chen, Jiemin Fang, Qian Zhang, Wenyu Liu, and Xinggang Wang. Hierarchical aggregation for 3d instance segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15467–15476, 2021. 1

[6] Wei Chen, Xi Jia, Hyung Jin Chang, Jinming Duan, Linlin Shen, and Ales Leonardis. 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)*, pages 1581–1590, June 2021. 3

[7] Corinna Cortes, Xavier Gonzalvo, Vitaly Kuznetsov, Mehryar Mohri, and Scott Yang. Adanet: Adaptive structural learning of artificial neural networks. In *International conference on machine learning*, pages 874–883. PMLR, 2017. 3

[8] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017. 7

[9] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11444–11453, 2020. 3

[10] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 6

[11] Samir Yitzhak Gadre, Kiana Ehsani, and Shuran Song. Act the part: Learning interaction strategies for articulated object part discovery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15752–15761, 2021. 3

[12] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015. 2, 3, 5

[13] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016. 2, 3, 5

[14] Yiran Geng, Boshi An, Haoran Geng, Yuanpei Chen, Yaodong Yang, and Hao Dong. End-to-end affordance learning for robotic manipulation. *arXiv preprint arXiv:2209.12941*, 2022. 3

[15] Ran Gong, Yizhou Zhao, Xiaofeng Gao, Jiangyong Huang, Qingyang Wu, Wensi Ai, Baoxiong Jia, Zhou Ziheng, Song-Chun Zhu, and Siyuan Huang. ARNOLD: A benchmark for language-grounded task learning with continuous states in realistic scenes. In *Workshop on Language and Robotics at CoRL 2022*, 2022. 3

[16] Benjamin Graham and Laurens van der Maaten. Submanifold sparse convolutional networks. *arXiv preprint arXiv:1706.01307*, 2017. 5

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1

[18] Yisheng He, Haibin Huang, Haoqiang Fan, Qifeng Chen, and Jian Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3003–3013, June 2021. 3

[19] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4867–4876, 2020. 1, 5, 7, 8

[20] Zhenyu Jiang, Yifeng Zhu, Maxwell Svetlik, Kuan Fang, and Yuke Zhu. Synergies between affordance and geometry: 6-dof grasp detection via implicit representations. *Robotics: science and systems*, 2021. 3

[21] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 1

[22] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *European Conference on Computer Vision*, pages 574–591. Springer, 2020. 3

[23] Da Li, Yongxin Yang, Yi-Zhe Song, and Timothy Hospedales. Learning to generalize: Meta-learning for domain generalization. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 3

[24] Da Li, Jianshu Zhang, Yongxin Yang, Cong Liu, Yi-Zhe Song, and Timothy M Hospedales. Episodic training for domain generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1446–1455, 2019. 3

[25] Haoliang Li, Sinno Jialin Pan, Shiqi Wang, and Alex C Kot. Domain generalization with adversarial feature learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2018. 2, 3, 5

[26] Jue Kun Li, Wee Sun Lee, and David Hsu. Push-net: Deep planar pushing for objects with unknown physical properties. In *Robotics: Science and Systems*, volume 14, pages 1–9, 2018. 3

[27] Xiaolong Li, He Wang, Li Yi, Leonidas Guibas, A Lynn Abbott, and Shuran Song. Category-level articulated object pose estimation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 4, 6

[28] Yangyan Li, Rui Bu, Mingchao Sun, Wei Wu, Xinhan Di, and Baoquan Chen. Pointcnn: Convolution on x-transformed points. *NeurIPS*, 2018. 3

[29] Ya Li, Xinmei Tian, Mingming Gong, Yajing Liu, Tongliang Liu, Kun Zhang, and Dacheng Tao. Deep domain generalization via conditional invariant adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 624–639, 2018. 3

[30] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 683–698, 2018. 3

[31] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 6

[32] Liu Liu, Wenqiang Xu, Haoyuan Fu, Sucheng Qian, Qiaojun Yu, Yang Han, and Cewu Lu. Akb-48: A real-world articulated object knowledge base. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14809–14818, 2022. 2, 4

[33] Xueyi Liu, Xiaomeng Xu, Anyi Rao, Chuang Gan, and Li Yi. Autogpart: Intermediate supervision search for generalizable 3d part segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11624–11634, 2022. 3, 7, 8

[34] Tiange Luo, Kaichun Mo, Zhiao Huang, Jiarui Xu, Siyu Hu, Liwei Wang, and Hao Su. Learning to group: A bottom-up framework for 3d part discovery in unseen categories. *arXiv preprint arXiv:2002.06478*, 2020. 3

[35] Kaichun Mo, Paul Guerrero, Li Yi, Hao Su, Peter Wonka, Niloy Mitra, and Leonidas J Guibas. Structurenet: Hierarchical graph networks for 3d shape generation. *arXiv preprint arXiv:1908.00575*, 2019. 3

[36] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021. 3

[37] Kaichun Mo, Shilin Zhu, Angel X Chang, Li Yi, Subarna Tripathi, Leonidas J Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. In *CVPR*, 2019. 2, 3

[38] Tongzhou Mu, Zhan Ling, Fanbo Xiang, Derek Yang, Xuanlin Li, Stone Tao, Zhiao Huang, Zhiwei Jia, and Hao Su. ManiSkill: Generalizable Manipulation Skill Benchmark with Large-Scale Demonstrations. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3, 8

[39] Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *International Conference on Machine Learning*, pages 10–18. PMLR, 2013. 3

[40] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3215, 2021. 3

[41] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 3

[42] Xingchao Peng, Qinxun Bai, Xide Xia, Zijun Huang, Kate Saenko, and Bo Wang. Moment matching for multi-source domain adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1406–1415, 2019. 3

[43] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, pages 652–660, 2017. 3

[44] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. 3

[45] Mohammad Mahfujur Rahman, Clinton Fookes, Mahsa Baktashmotlagh, and Sridha Sridharan. Multi-component image translation for deep domain generalization. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 579–588. IEEE, 2019. 3

[46] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 1

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

[48] Meng Tian, Marcelo H Ang Jr, and Gim Hee Lee. Shape prior deformation for categorical 6d object pose and size estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 3

[49] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 13(04):376–380, 1991. 6

[50] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022. 5, 7

[51] Chen Wang, Roberto Martín-Martín, Danfei Xu, Jun Lv, Cewu Lu, Li Fei-Fei, Silvio Savarese, and Yuke Zhu. 6-pack: Category-level 6d pose tracker with anchor-based keypoints. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10059–10066. IEEE, 2020. 3

[52] Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3343–3352, 2019. 3

[53] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, pages 2642–2651, 2019. 1, 3, 4

[54] Jiaze Wang, Kai Chen, and Qi Dou. Category-level 6d object pose estimation via cascaded relation and recurrent reconstruction networks. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4807–4814. IEEE, 2021. 3

[55] Jindong Wang, Cuiling Lan, Chang Liu, Yidong Ouyang, Tao Qin, Wang Lu, Yiqiang Chen, Wenjun Zeng, and Philip Yu. Generalizing to unseen domains: A survey on domain generalization. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2022. 3

[56] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018. 3

[57] Xiaogang Wang, Bin Zhou, Yahao Shi, Xiaowu Chen, Qinping Zhao, and Kai Xu. Shape2motion: Joint analysis of motion parts and attributes from 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8876–8884, 2019. 2

[58] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E Sarma, Michael M Bronstein, and Justin M Solomon. Dynamic graph cnn for learning on point clouds. *ACM TOG*, 38(5):1–12, 2019. 3

[59] Yijia Weng, He Wang, Qiang Zhou, Yuzhe Qin, Yueqi Duan, Qingnan Fan, Baoquan Chen, Hao Su, and Leonidas J Guibas. Captra: Category-level pose tracking for rigid and articulated objects from point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13209–13218, 2021. 1, 3

[60] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. VAT-mart: Learning visual action trajectory proposals for manipulating 3d ARTiculated objects. In *International Conference on Learning Representations*, 2022. 3

[61] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, et al. Sapien: A simulated part-based interactive environment. In *CVPR*, 2020. 2, 4, 7, 8

[62] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 3

[63] Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *arXiv preprint arXiv:2202.13519*, 2022. 3

[64] Zheng Xu, Wen Li, Li Niu, and Dong Xu. Exploiting low-rank structure from latent domains for domain generalization. In *European Conference on Computer Vision*, pages 628–643. Springer, 2014. 3

[65] Kaizhi Yang and Xuejin Chen. Unsupervised learning for cuboid shape abstraction via joint segmentation from point clouds. *ACM Transactions on Graphics (TOG)*, 40(4):1–11, 2021. 3

[66] Li Yi, Vladimir G Kim, Duygu Ceylan, I-Chao Shen, Mengyan Yan, Hao Su, Cewu Lu, Qixing Huang, Alla Sheffer, and Leonidas Guibas. A scalable active framework for region annotation in 3d shape collections. *ACM TOG*, 35(6):1–12, 2016. 2, 3

[67] Li Yi, Wang Zhao, He Wang, Minhyuk Sung, and Leonidas J Guibas. Gspn: Generative shape proposal network for 3d instance segmentation in point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3947–3956, 2019. 3

[68] Fenggen Yu, Kun Liu, Yan Zhang, Chenyang Zhu, and Kai Xu. Partnet: A recursive part decomposition network for fine-grained and hierarchical shape segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9491–9500, 2019. 2

[69] Kuan-Ting Yu, Maria Bauza, Nima Fazeli, and Alberto Rodriguez. More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing. In *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 30–37. IEEE, 2016. 3

[70] Fan Zhou, Zhuqing Jiang, Changjian Shui, Boyu Wang, and Brahim Chaib-draa. Domain generalization with optimal transport and metric learning. *arXiv preprint arXiv:2007.10573*, 2020. 3

[71] Kaiyang Zhou, Yongxin Yang, Timothy Hospedales, and Tao Xiang. Learning to generate novel domains for domain generalization. In *European conference on computer vision*, pages 561–578. Springer, 2020. 3

[72] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021. 3