PartAfford: Part-level Affordance Discovery from Cross-category 3D Objects

Chao Xu[1] Yixin Chen[1] He Wang[2,4]
Song-Chun Zhu[2,3,4] Yixin Zhu[2,4] Siyuan Huang[4]

[1] University of California, Los Angeles [2] Peking University
[3] Tsinghua University [4] Beijing Institute for General Artificial Intelligence

Abstract. Understanding what objects could furnish for humans–*viz.*, learning object *affordance*–is the crux to bridge perception and action. In the vision community, prior work has primarily focused on learning object affordance with dense (*e.g.*, at a per-pixel level) supervision. In stark contrast, we humans learn the object affordance *without* dense labels. As such, the fundamental question to devise a computational model is: What is the natural way to learn the object affordance from geometry with humanlike sparse supervision? In this work, we present the new task of **part-level affordance discovery (PartAfford)**: Given only the affordance labels for each object, the machine is tasked to (i) decompose 3D shapes into parts and (ii) discover how each part of the object corresponds to a certain affordance category. We propose a novel learning framework for *PartAfford*, which discovers part-level representations by leveraging only the affordance set supervision and geometric primitive regularization, without dense supervision. To learn and evaluate PartAfford, we construct a part-level, cross-category 3D object affordance dataset, annotated with 24 affordance categories shared among $> 25,000$ objects. We demonstrate through extensive experiments that our method enables both the abstraction of 3D objects and part-level affordance discovery, with generalizability to difficult and cross-category examples. Further ablations reveal the contribution of each component.

1 Introduction

The human vision system could swiftly locate the functional part upon using an object for specific tasks [27]. Such a critical capability in object interaction requires fine-grained object *affordance* understanding. *Affordance*, coined and originally theorized by Gibson [14,13], characterizes how humans interact with human-made objects and environments. As such, affordance understanding of objects and scenes has a significant influence on bridging visual perception and holistic scene understanding [22,21,3] with actionable information [45]. It is considered as one of the critical ingredients for the artificial general intelligence [60].

Object affordances have two main characteristics. First, object affordances are not defined in terms of conventional categorical labels in computer vision; instead, they are defined by the associated actions for various tasks and are

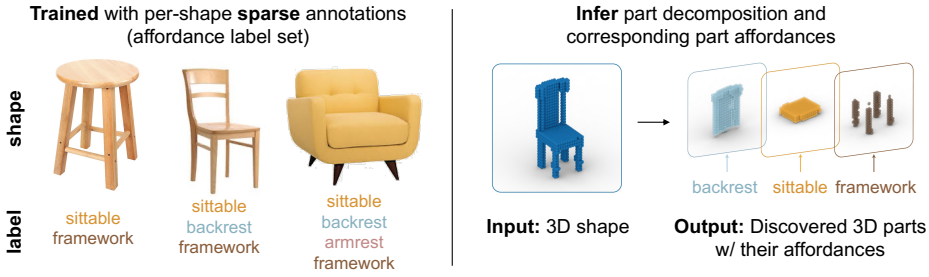


Fig. 1: **The proposed PartAfford: discover 3D object part affordances by learning contrast in affordance compositions.** During training (**left**), given sparse annotations (per-shape affordance label set), a learning framework is devised to ground affordance (*e.g.*, backrest) to 3D part (*e.g.*, sofa back) through learning cross-category, affordance-related shapes (*e.g.*, chair, sofa, *etc.*) with various affordance compositions. At test time (**right**), the learned model decomposes the 3D object into parts and infers the part-level affordances.

naturally *cross-category*. For example, both chair and sofa can be sat on, which indicates they share the *sittable* affordance. Similarly, desktop and bookshelf share the *support* affordance. Second, object affordances are intrinsically *part-based*. We could easily associate *sittable* affordance with the seats of chairs and sofas, and *support* with the boards of desktop and bookshelf. As such, the ability to learn **part-based**, **cross-category** affordance is essential to demonstrate the general object affordance understanding.

In passive affordance learning, prior literature follows the supervised learning paradigm, in which dense affordance annotation on the objects is fed as supervised signals [6]. However, this line of thought depends heavily on the quality of dense annotation, significantly deviated from how we humans learn to understand affordance. A humanlike supervision would be: "you can sit on this chair and rest your arm," "you can open the lid and hold water with the cup." In this paper, we try to answer: How to distinguish each object part while recognizing corresponding affordances with such sparse and natural supervisions?

To tackle this problem, we present *PartAfford*, a new task of part-level affordance discovery, which learns the object affordance with natural supervision of affordance set. As shown in Fig. 1, by providing only the set of affordance labels for each object, the algorithm is tasked to decompose the 3D shapes into parts and discover how each part corresponds to a certain affordance category. which is challenging and under-explored in the area of generalizable part-level object understanding and affordance learning.

To address this, we propose a novel method that discovers part-level representations with self-supervised 3D reconstruction, affordance set supervision and primitive regularization. The proposed approach consists of two main components. The first component is an encoder with slot attention for unsupervised clustering and abstraction. Specifically, we encode the 3D object into visual features and abstract the low-level features into a set of *slot* variables [29]. The second component is a decoder built upon the learned slot features. It has three

output branches that jointly reconstruct the 3D parts and object, predict the affordance labels, and regularize the learned part-level shapes with cuboidal primitives. Our method does not rely on dense supervision but instead learns from the sparse set supervision. It discovers the part-level affordance by learning the correspondence between affordance labels and abstracted 3D object parts.

Learning and evaluating *PartAfford* demands collections of 3D objects and their affordance labels for object parts. Prior work on visual affordance learning [20] either focuses on 2D objects and scenes or lacks part-based annotation [6]. Hence, we construct a part-level, cross-category 3D object affordance dataset annotated with 24 affordance categories shared among over 25,000 3D objects. The 3D objects are collected from PartNet dataset [36] and the PartNet-Mobility dataset [52]. The 24 part affordance categories are defined in terms of adjectives (*e.g.*, "sittable") or nouns (*e.g.*, "armrest"); they describe how object parts could afford human daily actions and activities. We annotate the part-level object affordances by manually mapping the fine-grained object part defined in Mo *et al*. [36] to the part affordances defined in this work.

By experimenting on this newly constructed *PartAfford* dataset, we empirically demonstrate that our method jointly enables the abstraction of 3D objects and part-level affordance discovery. Our model also shows strong generalizability on hard and cross-category objects. Further experiments and ablations analyze each component's contribution and point out future directions.

In summary, our work makes four main contributions:

- We present a new *PartAfford* task for part-level affordance discovery. Compared to the prior densely-supervised learning paradigm, *PartAfford* learns the visual object affordance in a more natural manner.
- We propose a novel learning framework for tackling *PartAfford*, which jointly abstracts 3D objects into part-level representations and discovers the affordance by learning the affordance correspondence.
- We build the benchmark for learning and evaluating *PartAfford* by curating a dataset consisting of 3D objects and annotate part-level affordances.
- We empirically demonstrate the efficacy and generalization capability of the proposed method and analyze each component's significance via a suite of ablation studies. Code and data will be released for research purposes.

2 Related Work

Affordance Learning Affordance learning is a multidisciplinary research field of vision, cognition, and robotics. In general, "affordance" is first perceived from images [17,62,44] or videos [53,61,10,37,39], followed by cognitive reasoning [62,60], and finally serves for task and motion planning in robotics [38,54,30,35]. Prior work tackles affordance at various scales and representations. Although affordance has been studied at the scene level [59,18,44], object level [40,34,12], and associated with generated human poses [24,62,51], few attempts study affordance as a 3D shape analysis task [58,28,61,50] since it would normally require large-scale, high-quality 3D data. A notable recent work [6] benchmarks several affordance estimation tasks on PartNet [36] with dense affordance heatmap

supervisions, annotated by densely selecting keypoints without considering affordance compositionality. In comparison, *PartAfford* studies affordance in a weakly supervised manner, such that the affordance discovery will be guided by affordance set matching and geometry abstraction. The new part affordance dataset we construct provides fine-grained, part-level 3D affordance annotations, tailored for the sparse supervision setting and affordance compositionality study.

Unsupervised Object-centric Learning Object discovery has been studied in an iterative end-to-end fashion [16,49,1,9,15,7]. Recently, [29] presents the slot attention module, an efficient and generic framework for object-centric representation extraction. It is capable of modeling compositional nature in synthetic scenes with multiple simple geometry shapes [23]. Subsequently, Stelzner *et al.* [46] and Yu *et al.* [57] apply slot attention on unsupervised 3D-aware scene decomposition, integrating NeRF [32] as object representations. They demonstrate that slot-based bottleneck could perform reasonably on synthetic multi-view RGB datasets with a textureless background. Our work takes one step further to tackle the challenge of *part-level* affordance discovery of 3D objects; part discovery is more complex than object discovery, primarily due to the ambiguity in the object part segmentation without applying additional constraints. Fortunately, for man-made objects, affordances are attached to objects at the part level. This observation implies the possibility of combining part discovery and affordance learning with minimal supervision. In this work, we integrate part discovery with affordance estimation, hoping that affordance information would help discover object parts sharing similar affordances.

Unsupervised Geometric Primitive Modeling Whereas supervised geometric primitive abstraction methods [33,55] require dense hierarchical annotations, unsupervised frameworks using cuboid-based [48,47], superquadrics-based [43,41], or other genus-zero-shape [5,42] primitives discover structural information naturally embedded in the geometry. Recently, Yang *et al.* [56] unsupervisedly learn the cuboid-based shape abstraction with shape co-segmentation. Yet, it relies heavily on the ground-truth point normals for accurate abstraction and lacks semantic representation for object understanding. In our affordance discovery framework, we leverage the cuboidal regularization to refine the reconstructed affordance part, which distinguishes densely connected 3D parts by providing geometric prior, thus improving the affordance part discovery.

3 Task Definition

We formulate the new task *PartAfford* as discovering the part-level object affordance with the affordance set supervision. We define $K = 24$ common affordance categories $\mathcal{S} = \{s_k\}_{k=1}^{K}$, such as "sittable" and "openable," for object understanding. Input is given as a collection of N objects $\{o_i\}_{i=1}^{N}$ and their corresponding affordance set labels $\{\mathcal{A}_i\}_{i=1}^{N}$, where $\mathcal{A}_i = \{a_i^j\}_{j=1}^{J_i}$. $a_i^j \in \mathcal{S}$, and J_i represents the number of distinct assigned affordances for each object i. *PartAfford* requires an algorithm to decompose each object into parts and discover the affordance corresponding to each object part. Fig. 1 illustrates the *PartAfford* task.

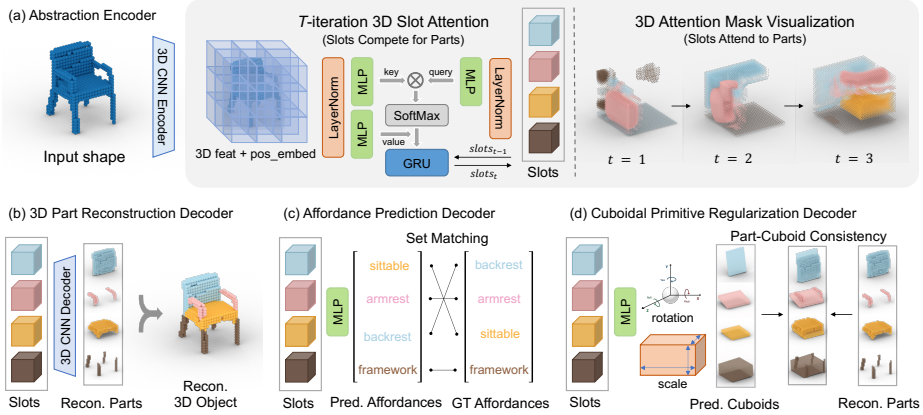


Fig. 2: **Illustration of the proposed method for *PartAfford***. Our model contains two main components: abstraction encoder and affordance decoder. (a) **Abstraction encoder** takes a 3D object as input, extracts features with 3D convolutional neural networks, and abstracts it into several slots. **Affordance decoder** with three branches jointly (b) reconstructs the 3D parts, (c) predicts affordance labels, and (d) regularizes cuboidal primitives.

4 Method

We propose a novel framework for affordance discovery from 3D objects. It integrates unsupervised part discovery with affordance set prediction and geometric primitive abstraction; see Fig. 2. Given a 3D shape represented by voxel grids $\mathcal{V}$ of resolution 32^3, our method first encodes the 3D shape into visual features and abstracts it into M slots; each slot represents an abstracted high-level feature for downstream tasks. Next, we utilize a decoder with three branches to jointly (i) decode the features into parts, (ii) predict the affordance label, and (iii) regularize the parts with cuboidal primitives. By composing the 3D parts from the slots with the 3D reconstruction as self-supervision, we ensure the slots combined can depict the entire 3D object. With the set matching loss of affordance prediction, the model discovers the correspondence between parts in slots and the affordance labels. Fitting the reconstructed 3D parts into cuboid representation further regularizes the shape of the part discovery. Below, we describe in detail how each module is constructed and the loss design.

4.1 Abstraction Encoder

The encoder takes 3D shape as input and abstracts part-centric latent codes in an unsupervised manner. It consists of a feature extraction module and a 3D slot attention module; see Fig. 2a.

Feature Extraction The feature embedding backbone encodes the input voxels and generates a $D = 64$ dimensional feature for each voxel. Following [31], voxels are encoded by five layers of 3D convolutional neural networks. The embedded feature is then augmented with absolute positional embedding [29].

3D Slot Attention The 3D slot attention architecture, adapted from [29], serves as the part-centric representational bottleneck between the 3D feature embedding network and the downstream decoders. The encoded feature of a 3D shape is fed into an iterative attention module, where M randomly initialized slots are updated for $T = 3$ iterations through a Gated Recurrent Unit [4]. During each iteration, the attention coefficients are calculated by applying softmax normalization over the slots on the dot-product similarity between queries (*i.e.*, linearly-mapped 3D slot features) and keys (*i.e.*, linearly-mapped input features). The attention coefficients are then applied as the weight for aggregating the values (*i.e.*, linearly-mapped input feature) and updating the slots.

Since the inputs-to-slots attention assignment is normalized over the 3D slots, those slots compete to attend to a clustering of the input 3D shape. Such clusterings are similar to human-defined parts on common objects. 3D slot attention masks naturally segment the object through iterations. An example of the learned 3D attention masks are shown in Fig. 2a.

4.2 Affordance Decoder

Shown in Fig. 2b-d, the affordance decoder takes part-centric slot features as input, followed by three branches for 3D part reconstruction, affordance prediction and primitive regularization. The decoder parameter is shared across slots.

3D Part Reconstruction We design a 3-layer 3D transposed convolutional decoder followed by a single MLP layer to reconstruct voxel values $\hat{\mathcal{V}}^m$ and a voxel mask for each slot. The mask is normalized across slots with softmax, which generates a normalized mask $\hat{\Lambda}^m \in \mathbb{R}^{32\times32\times32}$. It is then used to compute the weighted sum of voxel values across slots and combine the reconstructed parts $\{\hat{\mathcal{V}}^m\}_{m=1}^M$ into a full 3D shape $\hat{\mathcal{V}}$:

$$\hat{\mathcal{V}} = \sum_{m=1}^{M} \hat{\Lambda}^m \hat{\mathcal{V}}^m \tag{1}$$

The 3D part reconstruction branch is self-supervised by the reconstruction loss between original voxels $\mathcal{V}$ and reconstructed voxels $\hat{\mathcal{V}}$; we use the binary cross-entropy (BCE):

$$\mathcal{L}_{\text{recon}} = \texttt{BCE}(\mathcal{V}, \hat{\mathcal{V}}) \tag{2}$$

Affordance Prediction We predict a one-hot affordance label for each slot with a two-layer MLP with sharing weights across slots for classification.

The affordance prediction branch is weakly-supervised as we do not provide affordance labels for each voxel. Instead, only the affordance label set for the entire object is used as the supervision signal. The model is tasked to learn the alignment between the abstracted parts and the affordance labels from set supervision.

As defined in Sec. 3, the ground-truth set of affordance labels for an input 3D object is denoted as $\mathcal{A}$. We denote $\hat{\mathcal{A}}$ as the set of slot affordance predictions. $\hat{\mathcal{A}}_\sigma$ is a permutation of elements in $\hat{\mathcal{A}}$, where $\sigma \in \mathfrak{G}$ and $\mathfrak{G}$ represent all $M!$ possible

permutations. $\mathcal{L}_{\text{match}}$ is the pairwise matching cost between two sets, which can be calculated by mean square error (MSE) or cross-entropy:

$$\mathcal{L}_{\text{pred}} = \min_{\sigma \in \mathfrak{S}} \mathcal{L}_{\text{match}}(\mathcal{A}, \hat{\mathcal{A}}_\sigma). \tag{3}$$

Due to the order-invariant nature of slot modules, we apply the Hungarian matching algorithm [26], with Huber loss as the pair-wise matching cost, to calculate the set-based [2] affordance prediction loss in Eq. (3).

Cuboidal Primitive Regularization As a generalized soft k-means algorithm, the slot attention mechanism heavily relies on visual cues, such as the clustering of pixel colors on the image. As such, it cannot perform precisely in a crowded scene with overlapping objects even on a toy image dataset [29]. In 3D voxel regime, segmenting object into parts is challenging since every voxel is connected to neighboring voxels without distinguishable visual appearances.

Therefore, we introduce the cuboidal primitive regularization module, providing geometric prior for segmentation: Human-made objects usually have geometric regularity, and cuboid is a concise structural representation for abstraction.

From each slot embedding, the cuboid abstraction module predicts a cuboid parametrized by two vectors [56]: a scale vector $\boldsymbol{s} \in \mathbb{R}^3$ and a quaternion vector $\boldsymbol{r} \in \mathbb{R}^4$ for 3D rotation. Of note, we calculate the cuboid center from the weighted mean of voxel positions in the slot.

To evaluate how the predicted cuboid fits the reconstructed part in the m-th slot, we first compute the Euclidean distance d_i^m from each voxel $\boldsymbol{p}_i^m$ to its closest cuboid face. Next, we calculate the weighted sum distance for all voxels, where the weight $v_i^m \in \hat{\mathcal{V}}^m$ is the reconstructed voxel value within $[0, 1]$. Additionally, we designed a binary surface mask $f(i)$ that masks out internal voxels in the loss. Thus, the loss encourages a cuboid to tightly wrap a solid object. The regularization loss for all the slots is defined as:

$$\mathcal{L}_{\text{cuboid}} = \sum_m \sum_i f(i) v_i^m d_i^m. \tag{4}$$

4.3 Total Loss

Taking together, the total training loss is the sum of 3D reconstruction loss, affordance prediction loss, and the primitive regularization loss:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{recon}} \mathcal{L}_{\text{recon}} + \lambda_{\text{pred}} \mathcal{L}_{\text{pred}} + \lambda_{\text{cuboid}} \mathcal{L}_{\text{cuboid}}, \tag{5}$$

where λ_{recon}, λ_{cuboid}, and λ_{pred} are balancing coefficients.

Of note, with the current architecture design, for the first time, we demonstrate the capability of part-level affordance discovery from set labels. The exploration of more complex and practical modules are left for future work.

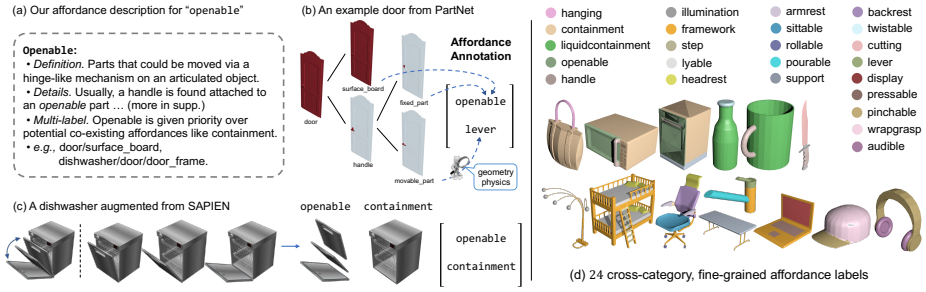


Fig. 3: **Part affordance dataset.** (a) Description for the "openable" affordance to construct the mapping. (b) Given the part hierarchy of a door from PartNet [36], we annotate its affordance labels by manual mapping and inspection. (c) Given a dishwasher from PartNet-Mobility (SAPIEN) [52] and its kinematics, we rotate the door frame to include 3D objects with different articulation states. (d) Some exemplar 3D object models with color-coded affordance visualization.

5 Part Affordance Dataset

To benchmark *PartAfford* and facilitate the research in affordance understanding, we construct a part-level 3D object affordance dataset. We focus on 24 cross-category, fine-grained affordance labels as shown in Fig. 3. The dataset is annotated with over 25,000 3D CAD models from the PartNet dataset [36] and 625 articulated objects among 9 categories from the PartNet-Mobility dataset in SAPIEN [52]. Below, we describe how to define part affordances and the generate affordance annotation. See *supplementary materials* for more details.

5.1 Affordance Definition & Dataset Construction

Part affordances in our dataset are defined in terms of adjectives (*e.g.*, *sittable*) or nouns (*e.g.*, *armrest*), which describe how object parts could afford human daily actions and activities. We adopt certain common affordance categories from a comprehensive survey of visual affordance [20], *e.g.*, *containment*, *sittable*, *support*, *openable*, *rollable*, *display*, and *wrapgrasp*. However, they are coarse-grained–either at the object-level or scene-level. For example, "*openable*" only indicates whether an object can be opened, but is unclear about which object part can *afford* the object to be opened.

To pursue a fine-grained understanding of object affordance, we manually construct a one-to-multiple mapping from 479 kinds of object part labels defined at the finest granularity in Mo *et al.* [36] to 24 potential affordance labels, given the detailed affordance description. We provide expert-defined descriptions for 24 affordances to guarantee the quality and consistency of the mapping construction. An example is shown in Fig. 3a. Given the part hierarchy of a 3D object, we can get the corresponding affordance annotation by the mapping. We also perform manual inspection to correct the affordance labels, especially for

some fine-grained parts, according to their specific geometry and physics property. For instance, different door handles will be mapped to different affordance label (twistable, lever, *etc.*) according to how they should be operated (Fig. 3b).

The PartNet dataset does not contain articulation information, making affordances such as *openable* not geometrically distinguishable. Therefore, we generate a set of shapes with *openable* affordance from the PartNet-Mobility dataset by capturing 3D shapes with various opening angles (Fig. 3c).

As can be seen from Fig. 3d, each affordance type–due to its cross-category nature–may be found on a variety of object part instances. For example, *openable* is usually afforded by rotatable doors for unobstructed access. Under such criteria, the door frame of a dishwasher and the surface board of a door are both mapped to *openable*. Please refer to the *supplementary materials* for a full list of all affordance categories, descriptions and mapping examples.

6 Experiments

In this section, we design and conduct comprehensive experiments to evaluate the proposed method. Fig. 4 visualizes our main results. We present both quantitative and qualitative comparisons of baseline models and our model variants. To illustrate the generalizability of our approach, we evaluate the model generalization on novel objects. We further analyze the failure cases and propose potential improvement directions. Please refer to the *supplementary materials* for additional experimental results and analyses.

6.1 Experimental Settings

Benchmarks To benchmark *PartAfford*, we curate different subsets of samples from our constructed dataset. Specifically, we study the subsets related to the most representative affordance categories "sittable," "support," and "openable" separately, where the subsets are created by collecting all cross-category objects that have the corresponding affordance label in our dataset. For each subset, we learn to distinguish all the affordance labels appear in the 3D objects. Note that although a part can have multiple affordances as mentioned in Sec. 5, we only keep the most prioritized affordance for each part to ease the ambiguities in learning. Below, we describe the statistics of objects and the related affordances for these subsets; we discuss detailed reasons about why we choose these three subsets in Sec. 7.2.

- "Sittable": We collect all object instances that have affordance "sittable"; most of them are chairs and sofas. Their part-level affordances belong to the set $\{sittable, backrest, armrest, framework\}$. We split the training, validation, and test set in the ratio of $7:1:2$. In total, we have $5,093$ instances for training and $1,457$ for test.
- "Support": We collect objects with affordance "support", mainly from categories table and cabinet. Their affordances belong to $\{support, framework\}$. There are $7,974$ instances for training and $2,279$ instances for test.

– "Openable": This subset contains objects from frige, dishwasher, washing machine, and microwave. Their affordances belong to {*openable*, $framework$, *handle*}. There are 807 instances for training and 232 instances for test.

Data Augmentation To enrich affordance compositions, we augment the training data by randomly removing certain object parts with corresponding affordance labels.

Evaluation Metrics In *PartAfford*, we evaluate the performances of part discovery (clustering similarity), 3D reconstruction, and affordance prediction.

– Part Discovery: We use the Intersection over Union (IoU) to evaluate the part similarity. Specifically, we employ Hungarian matching to find the best matches between the reconstructed parts and ground truth parts using voxel IoU as the matching score. Then we compute the mean IoU (mIoU) by averaging the IoU between best matches.
– 3D Reconstruction: We evaluate the overall 3D shape reconstruction quality using mean squared error (MSE).
– Affordance Prediction: Following [29], we use Average Precision (AP) to evaluate the affordance set prediction accuracy. A correct prediction means an exact match of the affordance label set.

Baselines and Ablations Since we are the first to propose and formulate *PartAfford*, there is no previous work for us to make direct comparisons. Therefore, we compare with two designed baseline models and three variants of our method to evaluate the efficacy of the proposed method and its components:

– Slot MLP: a simple MLP-based baseline where we replace Slot Attention with an MLP that maps from the learned feature maps (resized and flattened) to the (now ordered) slot representation.
– IODINE: a baseline where we replace the Slot Attention with an object-centric learning method IODINE [15] to abstract and cluster the encoded feature.
– Ours w/o Afford & Cuboid: our model variant that only keeps the 3D part reconstruction branch.
– Ours w/o Afford: our model variant without the affordance prediction branch.
– Ours w/o Cuboid: our model variant that disgards the cuboidal primitive regularization branch.
– Ours Full: our full model with all branches.

6.2 Implementation Details

Learning Strategy To stabilize the training, we split the training into two stages. In the first stage, we train the decoder only with 3D part reconstruction and affordance prediction branches. In the second stage, we add the cuboidal primitive regularization branch into joint training with a lower learning rate.

Hyperparameter We set learning rate as 4×10^{-4} for the first stage, 2×10^{-4} for the second stage, and apply Adam optimizer [25] for optimization. It takes $5+9$ hours on 4 RTX A6000 GPUs for two-stage full-model training of "sittable"-related objects. For slot attention, we empirically set the number of GRU iterations $T=3$. We set the number of slots to the maximal number of affordance labels that appear in each training subset. For example, we learn

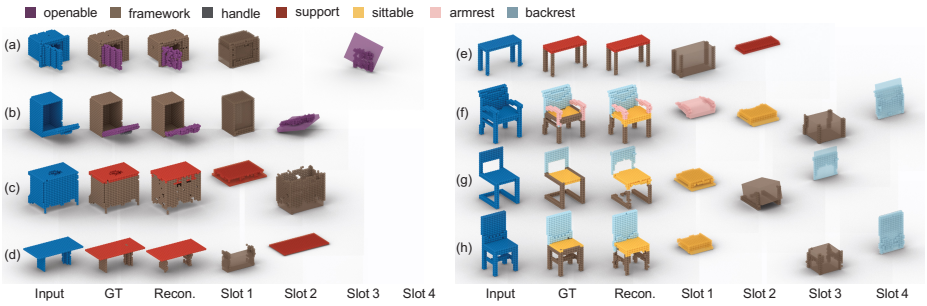
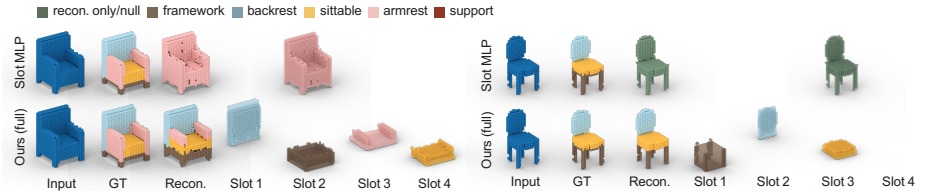
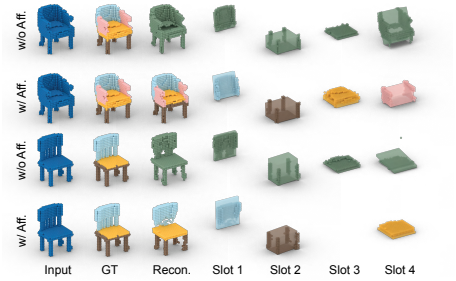


Fig. 4: Qualitative results on three curated subsets: "openable" (rows a-b), "support" (rows c-e), and "sittable" (rows f-h).

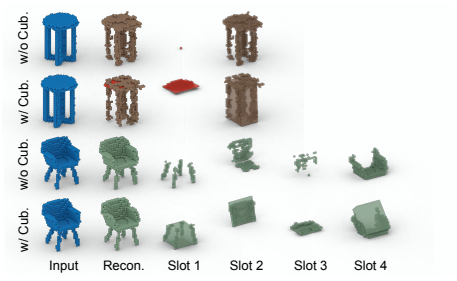
the "sittable" with 4 slots, "support" with 2 slots, and "openable" with 3 slots. Sec. 7.1 further discusses choices of the number of slots. For the joint loss weight, we set $\lambda_{\text{recon}} = 1.0, \lambda_{\text{pred}} = 0.5, \lambda_{\text{cuboid}} = 0.1$.



(a) Comparisons between the Slot MLP and our model.



(b) Comparisons between the models without and with affordance prediction branch. The affordance prediction facilitates the abstraction of 3D shape (rows 1-2) and elimination of spare slots (rows 3-4).



(c) Comparisons between the models without and with cuboidal primitive regularization branch. The cuboidal primitive regularization helps to guide and improve the segmentation.

Fig. 5: Qualitative comparison results.

6.3 Results and Analysis

Tab. 1 and Tab. 2 tabulate the quantitative performances of all the models under different settings. Fig. 4 qualitatively shows the capability of our method and Fig. 5 compares different models. Below, we summarize some key findings:

1. The proposed method achieves the best overall performance on the *PartAfford* task, especially in the part discovery (mean IoU) where it outperforms the baselines by a large margin. This demonstrates the outstanding abstraction capability of our approach given the sparse supervision. From Fig. 4, we can see our method can discover the detailed part-level representation with their aligned affordances for the 3D objects.
2. The most challenging part of *PartAfford* lies in the part discovery, and it's also where our model differentiates with other baselines. For example, Tab. 1 shows that Slot MLP achieves the best affordance prediction performance (AP) but fails in the part discovery (mean IoU) and 3D reconstruction (MSE). As also shown in Fig. 5a, the Slot MLP cannot segment the object input to parts due to the lack of abstraction capability.
3. Affordance prediction branch significantly escalates the part discovery performance since it helps to learn part-affordance correspondence from the affordance composition, which provides contrasts to distinguish different parts among the training objects. Our qualitative results also show that the affordance prediction facilitates the abstraction of 3D shape (*e.g.*, rows 1-2 of Fig. 5b) and elimination of spare slots (*e.g.*, rows 3-4 of Fig. 5b);
4. Cuboidal primitive regularization branch also boosts the part discovery, especially when affordance prediction is not available. This demonstrates that geometric priors play a crucial role in segmentation when data are not diverse enough. From Fig. 5c, we can see the cuboidal primitive regularization helps to segment better primitives and avoid scattered voxels.

6.4 Model Generalization

With the cross-category nature of affordance, we qualitatively test how the learned model can be generalized to novel objects and unseen categories. We conduct model generalization experiments by testing hard examples or objects from other categories. Examples from Fig. 6 demonstrates the learned model could be generalized to objects with diverse shapes. We show the results of testing the learned model on a novel object shape (bean bag) (a) from [11] and unseen categories (b-d). For example, (b) shows the result of learning with "support" and test on an "openable" object (*i.e.*, a microwave). Although the reconstructions may not be perfect, partly due to reconstruction bottleneck's impact on disentanglement quality [8], the learned model can successfully identify the functional parts given novel objects.

6.5 Failure Cases

We show some failure cases of our method in Fig. 7. For "sittable" and "support," the failures are commonly caused by (i) the difficulties to reconstruct the fine-grained details of 3D objects with novel shapes; (ii) certain parts that violate the cuboid assumption, and thus hurts other components.

For objects in "openable" category, our model cannot discover and reconstruct "handle," as shown in Fig. 4. This is because the objects with related

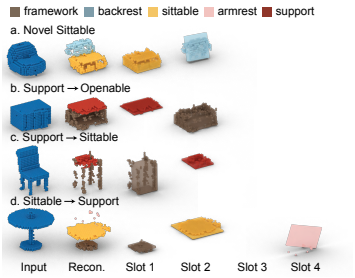


Fig. 6: Generalization results.

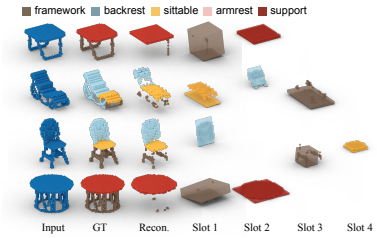


Fig. 7: Failure cases.

Table 1: Quantitative results on "sittable." We evaluate the mean IoU (mIoU), mean squared error (MSE), and average precision (AP) on included objects.

Model	mIoU (%) ↑	MSE ↓	AP (%)↑
Slot MLP	21.5	0.0150	**94.5**
IODINE	49.2	0.0102	92.5
Ours w/o Afford & Cuboid	31.5	0.0112	N/A
Ours w/o Afford	39.4	0.0100	N/A
Ours w/o Cuboid	55.3	0.0102	92.7
Ours (full)	**57.3**	**0.0097**	92.9

Table 2: Quantitative results on "support" and "openable."

	Model	mIoU (%) ↑	MSE ↓	AP (%) ↑
support	Ours w/o Afford	34.8	0.0092	N/A
	Ours w/o Cuboid	51.3	0.0087	**95.2**
	Ours (full)	**52.7**	**0.0085**	95.1
openable	Ours w/o Afford	19.9	0.0104	N/A
	Ours w/o Cuboid	46.7	0.0097	55.8
	Ours (full)	**47.6**	**0.0093**	**60.4**

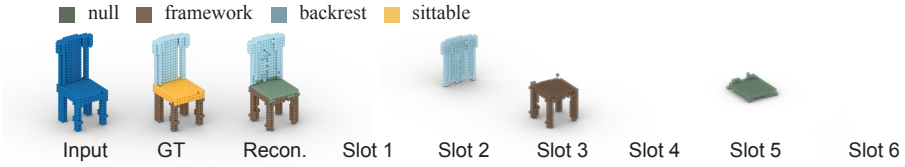


Fig. 8: Model performance when the number of slots increases. The model learns the "sittable" part in a "null" slot.

affordances come from various object categories with diverse shapes, making it challenging for the model to capture such complex mixtures of distributions and reconstruct the fine-grained 3D shapes, especially tiny parts (*i.e.*, "handle"). This points out future directions to better understand object parts (*e.g.*, segment, reconstruct), and potentially an interactive learning framework to learn beyond geometry and appearance.

7 Further Analysis & Discussion

7.1 Number of Slots

We set the number of slots as the maximal number of affordance labels that appear in one subset, which is different from previous object-centric learning algorithm [29], where the number of slots could be arbitrary. This is because when we increase the number of slots, we also increase the ambiguities of the affordance composition and set matching at the same time. It prevents the model from learning accurate correspondence between affordance labels and parts. As shown in Fig. 8, the model learns the "sittable" in a "null" slot.

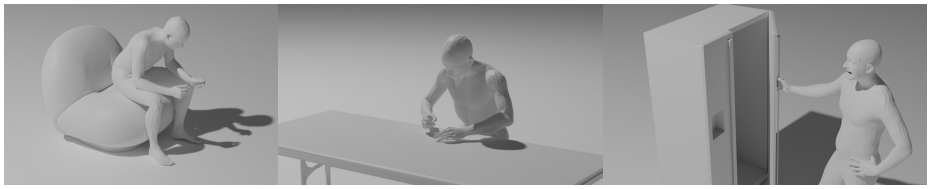


Fig. 9: 3D Human synthesis conditioned on inferred part affordance using [19].

7.2 Selection of Studied Affordances

Although we annotate objects with 24 affordance categories, we benchmark *PartAfford* only on three subsets with 7 kinds of affordances in this work. The main reason is that we find it much more challenging to discover part-level affordance for some other affordance categories. For example, "rollable" and "cutting" usually connect to tiny object parts that are challenging to be segmented from objects, "illumination" and "display" cannot be distinguished from the objects without a deeper understanding of the visual appearance and reflections, "pressable" and "pinchable" require richer interactions to be discovered. In summary, it is either especially challenging to segment or requires more than geometric information (*e.g.*, active interactions) to discover the other affordance categories. We hope to further explore the learning of these affordances in the future.

7.3 Ambiguities in Affordance Learning

Affordance is naturally ambiguous since the functions of object parts are rich. This work provides a well-defined benchmark to study how to learn affordance from accurate affordance definition and sparse set supervision. In another work, [6] models the multiple affordances with a mixture of distributions. We believe our current learning framework could also be extended to learn multiple affordances by switching the one-hot affordance label to multi-hot labels.

7.4 Potential Impacts for Vision & Robotics Community

By learning to discover the part-level affordance, our model could facilitate the understanding of human-object interaction and object manipulation. As shown in Fig. 9, the learned affordance could be applied to synthesize potential human actions and interactions with various 3D objects. It can also help the robot to learn cross-category object interaction and manipulation policies.

8 Conclusion

We present *PartAfford*, a new task in visual affordance research that aims at discovering part-level affordances from 3D shapes. We propose a novel learning framework that discovers part-level affordances by leveraging only the affordance set supervision and geometric primitive regularization. With comprehensive experiments and analyses, we point out potential directions for incorporating visual appearance to facilitate better shape abstraction and combining with an active learning approach for efficient affordance learning.

References

1. Burgess, C.P., Matthey, L., Watters, N., Kabra, R., Higgins, I., Botvinick, M., Lerchner, A.: Monet: Unsupervised scene decomposition and representation. arXiv preprint arXiv:1901.11390 (2019) 4
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: Proceedings of European Conference on Computer Vision (ECCV) (2020) 7
3. Chen, Y., Huang, S., Yuan, T., Qi, S., Zhu, Y., Zhu, S.C.: Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1
4. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014) 6
5. Deng, B., Genova, K., Yazdani, S., Bouaziz, S., Hinton, G., Tagliasacchi, A.: Cvxnet: Learnable convex decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
6. Deng, S., Xu, X., Wu, C., Chen, K., Jia, K.: 3d affordancenet: A benchmark for visual object affordance understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 3, 14
7. Du, Y., Li, S., Sharma, Y., Tenenbaum, J., Mordatch, I.: Unsupervised learning of compositional energy concepts. Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2021) 4
8. Engelcke, M., Jones, O.P., Posner, I.: Reconstruction bottlenecks in object-centric generative models. arXiv preprint arXiv:2007.06245 (2020) 12
9. Engelcke, M., Kosiorek, A.R., Jones, O.P., Posner, I.: Genesis: Generative scene inference and sampling with object-centric latent representations. arXiv preprint arXiv:1907.13052 (2019) 4
10. Fang, K., Wu, T.L., Yang, D., Savarese, S., Lim, J.J.: Demo2vec: Reasoning object affordances from online videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 3
11. Fu, H., Jia, R., Gao, L., Gong, M., Zhao, B., Maybank, S., Tao, D.: 3d-future: 3d furniture shape with texture. International Journal of Computer Vision (IJCV) **129**(12), 3313–3337 (2021) 12
12. Gadre, S.Y., Ehsani, K., Song, S.: Act the part: Learning interaction strategies for articulated object part discovery. In: Proceedings of International Conference on Computer Vision (ICCV). pp. 15752–15761 (2021) 3
13. Gibson, J.J.: The ecological approach to visual perception. Houghton, Mifflin and Company (1979) 1
14. Gibson, J.J., Carmichael, L.: The senses considered as perceptual systems, vol. 2. Houghton Mifflin Boston (1966) 1
15. Greff, K., Kaufman, R.L., Kabra, R., Watters, N., Burgess, C., Zoran, D., Matthey, L., Botvinick, M., Lerchner, A.: Multi-object representation learning with iterative variational inference. In: Proceedings of International Conference on Machine Learning (ICML) (2019) 4, 10
16. Greff, K., Van Steenkiste, S., Schmidhuber, J.: Neural expectation maximization. arXiv preprint arXiv:1708.03498 (2017) 4
17. Gupta, A., Satkin, S., Efros, A.A., Hebert, M.: From 3d scene geometry to human workspace. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011) 3

18. Gupta, S., Arbeláez, P., Girshick, R., Malik, J.: Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation. International Journal of Computer Vision (IJCV) **112**(2), 133–149 (2015) 3
19. Hassan, M., Ghosh, P., Tesch, J., Tzionas, D., Black, M.J.: Populating 3d scenes by learning human-scene interaction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 14
20. Hassanin, M., Khan, S., Tahtali, M.: Visual affordance and function understanding: A survey. ACM Computing Surveys (CSUR) **54**(3), 1–35 (2021) 3, 8
21. Huang, S., Qi, S., Xiao, Y., Zhu, Y., Wu, Y.N., Zhu, S.C.: Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In: Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2018) 1
22. Huang, S., Qi, S., Zhu, Y., Xiao, Y., Xu, Y., Zhu, S.C.: Holistic 3d scene parsing and reconstruction from a single rgb image. In: Proceedings of European Conference on Computer Vision (ECCV) (2018) 1
23. Kabra, R., Burgess, C., Matthey, L., Kaufman, R.L., Greff, K., Reynolds, M., Lerchner, A.: Multi-object datasets. https://github.com/deepmind/multi-object-datasets/ (2019) 4
24. Kim, V.G., Chaudhuri, S., Guibas, L., Funkhouser, T.: Shape2pose: Human-centric shape analysis. ACM Transactions on Graphics (TOG) **33**(4), 1–12 (2014) 3
25. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 10
26. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955) 7
27. Land, M., Mennie, N., Rusted, J.: The roles of vision and eye movements in the control of activities of daily living. Perception **28**(11), 1311–1328 (1999) 1
28. Liang, W., Zhao, Y., Zhu, Y., Zhu, S.C.: What is where: Inferring containment relations from videos. In: Proceedings of International Joint Conference on Artificial Intelligence (IJCAI) (2016) 3
29. Locatello, F., Weissenborn, D., Unterthiner, T., Mahendran, A., Heigold, G., Uszkoreit, J., Dosovitskiy, A., Kipf, T.: Object-centric learning with slot attention. Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020) 2, 4, 5, 6, 7, 10, 13
30. Mandikal, P., Grauman, K.: Learning dexterous grasping with object-centric visual affordances. In: Proceedings of International Conference on Robotics and Automation (ICRA) (2021) 3
31. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 5
32. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: Proceedings of European Conference on Computer Vision (ECCV) (2020) 4
33. Mo, K., Guerrero, P., Yi, L., Su, H., Wonka, P., Mitra, N.J., Guibas, L.J.: Structurenet: hierarchical graph networks for 3d shape generation. ACM Transactions on Graphics (TOG) **38**(6), 1–19 (2019) 4
34. Mo, K., Guibas, L.J., Mukadam, M., Gupta, A., Tulsiani, S.: Where2act: From pixels to actions for articulated 3d objects. In: Proceedings of International Conference on Computer Vision (ICCV) (2021) 3

35. Mo, K., Qin, Y., Xiang, F., Su, H., Guibas, L.: O2o-afford: Annotation-free large-scale object-object affordance learning. In: Conference on Robot Learning. pp. 1666–1677. PMLR (2022) 3
36. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3, 8
37. Nagarajan, T., Feichtenhofer, C., Grauman, K.: Grounded human-object interaction hotspots from video. In: Proceedings of International Conference on Computer Vision (ICCV) (2019) 3
38. Nagarajan, T., Grauman, K.: Learning affordance landscapes for interaction exploration in 3d environments. Proceedings of Advances in Neural Information Processing Systems (NeurIPS) (2020) 3
39. Nagarajan, T., Li, Y., Feichtenhofer, C., Grauman, K.: Ego-topo: Environment affordances from egocentric video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3
40. Nguyen, A., Kanoulas, D., Caldwell, D.G., Tsagarakis, N.G.: Object-based affordances detection with convolutional neural networks and dense conditional random fields. In: Proceedings of International Conference on Intelligent Robots and Systems (IROS) (2017) 3
41. Paschalidou, D., Gool, L.V., Geiger, A.: Learning unsupervised hierarchical part decomposition of 3d objects from a single rgb image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 4
42. Paschalidou, D., Katharopoulos, A., Geiger, A., Fidler, S.: Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 4
43. Paschalidou, D., Ulusoy, A.O., Geiger, A.: Superquadrics revisited: Learning 3d shape parsing beyond cuboids. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 4
44. Roy, A., Todorovic, S.: A multi-scale cnn for affordance segmentation in rgb images. In: Proceedings of European Conference on Computer Vision (ECCV) (2016) 3
45. Soatto, S.: Actionable information in vision. In: Machine Learning for Computer Vision, pp. 17–48. Springer (2013) 1
46. Stelzner, K., Kersting, K., Kosiorek, A.R.: Decomposing 3d scenes into objects via unsupervised volume segmentation. arXiv preprint arXiv:2104.01148 (2021) 4
47. Sun, C.Y., Zou, Q.F., Tong, X., Liu, Y.: Learning adaptive hierarchical cuboid abstractions of 3d shape collections. ACM Transactions on Graphics (TOG) **38**(6), 1–13 (2019) 4
48. Tulsiani, S., Su, H., Guibas, L.J., Efros, A.A., Malik, J.: Learning shape abstractions by assembling volumetric primitives. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 4
49. Van Steenkiste, S., Chang, M., Greff, K., Schmidhuber, J.: Relational neural expectation maximization: Unsupervised discovery of objects and their interactions. arXiv preprint arXiv:1802.10353 (2018) 4
50. Wang, H., Liang, W., Yu, L.F.: Transferring objects: Joint inference of container and human pose. In: Proceedings of International Conference on Computer Vision (ICCV) (2017) 3
51. Wang, X., Girdhar, R., Gupta, A.: Binge watching: Scaling affordance learning from sitcoms. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017) 3

52. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: SAPIEN: A simulated part-based interactive environment. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020) 3, 8
53. Xie, D., Todorovic, S., Zhu, S.C.: Inferring" dark matter" and" dark energy" from videos. In: Proceedings of International Conference on Computer Vision (ICCV) (2013) 3
54. Xu, D., Mandlekar, A., Martín-Martín, R., Zhu, Y., Savarese, S., Fei-Fei, L.: Deep affordance foresight: Planning through what can be done in the future. arXiv preprint arXiv:2011.08424 (2020) 3
55. Yang, J., Mo, K., Lai, Y.K., Guibas, L.J., Gao, L.: Dsg-net: Disentangled structured mesh net for controllable generation of fine geometry. arXiv preprint arXiv:2008.05440 (2020) 4
56. Yang, K., Chen, X.: Unsupervised learning for cuboid shape abstraction via joint segmentation from point clouds. ACM Transactions on Graphics (TOG) (2021) 4, 7
57. Yu, H.X., Guibas, L.J., Wu, J.: Unsupervised discovery of object radiance fields. arXiv preprint arXiv:2107.07905 (2021) 4
58. Yu, L.F., Duncan, N., Yeung, S.K.: Fill and transfer: A simple physics-based approach for containability reasoning. In: Proceedings of International Conference on Computer Vision (ICCV) (2015) 3
59. Zhao, Y., Zhu, S.C.: Scene parsing by integrating function, geometry and appearance models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2013) 3
60. Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y.N., et al.: Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. Engineering **6**(3), 310–345 (2020) 1, 3
61. Zhu, Y., Jiang, C., Zhao, Y., Terzopoulos, D., Zhu, S.C.: Inferring forces and learning human utilities from videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 3
62. Zhu, Y., Zhao, Y., Zhu, S.C.: Understanding tools: Task-oriented object modeling, learning and recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015) 3