# Beyond Object Recognition: A New Benchmark towards Object Concept Learning

Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, Yuan Yao, Siqi Liu, Cewu Lu[*]
Shanghai Jiao Tong University
{yonglu_li, silicxuyue, xuxinyu2000, mxh1999, yaoyuan2000, magi-yunan, lucewu}@sjtu.edu.cn

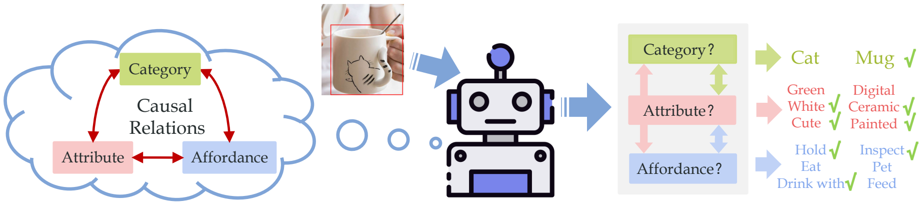


Figure 1: For embodied agents, understanding daily objects requires the ability to perceive not only **category** but also **attribute** and **affordance**. In OCL, we try to reveal object concept learning in both three levels and explore their profound causal relations.

## Abstract

*Understanding objects is a central building block of AI, especially for embodied AI. Even though object recognition excels with deep learning, current machines struggle to learn higher-level knowledge,* e.g.*, what attributes an object has, and what we can do with it. Here, we propose a challenging **Object Concept Learning** (OCL) task to push the envelope of object understanding. It requires machines to reason out affordances and simultaneously give the reason: what attributes make an object possess these affordances. To support OCL, we build a densely annotated knowledge base including extensive annotations for three levels of object concept (category, attribute, affordance), and the clear causal relations of three levels. By analyzing the causal structure of OCL, we present a baseline, Object Concept Reasoning Network (OCRN). It leverages concept instantiation and causal intervention to infer the three levels. In experiments, OCRN effectively infers the object knowledge while following the causalities well. **Our data and code are available at `https://mvig-rhos.com/ocl`**.*

## 1. Introduction

Object understanding is essential for intelligent robots. Recently, benefiting from deep learning and large-scale datasets [10, 36], category recognition [30, 55] has made tremendous progress. But to close the gap between human and machine perception, machines need to pursue deeper understanding, *e.g.*, recognizing higher-level attributes [27] and affordances [16], which may help it establish object concept [41] when interacting with contexts.

Category `apple` is a symbol indicating its referent (real apples). In line with symbol grounding [22], machines should learn knowledge beyond category to approach concept understanding. According to cognition studies [58, 41], attribute depicting objects from the physical/visual side plays an important role in object understanding. Thus, many works [32, 68, 13] studied to ground objects with attributes, *e.g.*, a `hammer` consists of a `long` handle and a `heavy` head. Moreover, attributes can depict object states [27]. An elegant characteristic of attributes is *cross-category*: objects of the same category can have various states (`big` or `fresh apple`), whilst various objects can have the same state (`sliced orange` or `apple`). If the category is the **first** level of object concept, the attribute can be seen as the **second** level closer to the physical fact.

However, recognizing attributes is still far away from concept understanding. Given a `hammer`, we should know it can be `held` to `hit` nails, *i.e.*, requiring machines to infer affordance [16] indicating what actions humans can perform with objects. Thus, we refer to affordance as the **third** level, which is closely related to common sense and causal inference [16]. Though affordance has been

[*] Corresponding author.

studied in robotics [12, 24] and vision [8, 73] communities for decades, it is still challenging. First, previous works [45, 15] often focus on recognizing affordance solely. But we usually infer affordance based on attribute observation. If we need to knock in a nail without a `hammer` at hand, we may find other `hard` or `heavy` objects instead, *e.g.*, a `thick` book. This profoundly reveals the **causality** between attribute and affordance. Second, previous works are designed for scale/scene-limited tasks, *e.g.*, in [73], 40 objects and 14 affordances are included; Hermans *et al.* [24] collect 375 indoor images of 6 objects, 21 attributes, and 7 affordances; a recent dataset [45] contains 10 indoor objects and 9 affordances. Thus, they cannot afford general affordance reasoning for large-scale applications.

To reshape object learning, we believe it is essential to look at the above three levels in a **unified** and **causal** way based on an extensive knowledge base. Hence, we move a step forward to propose the object concept learning (OCL) task: given an object, machines need to infer its category, attributes, and further answer "*what can we do upon it and why*", as shown in Fig. 1. In a nutshell, machines need to reason affordance based on object appearance, category, and attributes. To this end, we build a large-scale and dense dataset consisting of **381** categories, **114** attributes, and **170** affordances. It contains **80,463** images of diverse scenes and **185,941** instances in different states. Different from previous works [6, 24, 73], OCL offers a more subtle angle. It includes: (1) **category**-level attribute ($A$) and affordance ($B$) labels; (2) **instance**-level attribute ($\alpha$) and affordance ($\beta$) labels. Besides, we annotate the *causal relations* between three levels to evaluate the reasoning ability of models and keep the follow-up methods from fitting data only. Accordingly, based on the causal structure of OCL, we propose a *neuro-causal* method, **O**bject **C**oncept **R**easoning **N**etwork (**OCRN**), as the future baseline. It leverages concept instantiation (from category-level to instance-level) and causal intervention [50] to infer attributes and affordances. OCRN outperforms a host of baselines and shows impressive performance while following the causal relations well.

In summary, our contributions are threefold:

(1) Introducing the object concept learning task poses challenges and opportunities for object understanding and knowledge-based reasoning.

(2) Building a benchmark consisting of diverse objects, elaborate attributes, and affordances, together with their clear causal relations.

(3) An object concept reasoning network is introduced to reason three levels with concept instantiation performing well on OCL.

## 2. Related Work

**Object Attribute** depicts the physical properties like color, size, shape, *etc*. It usually plays the role of intermedia between pixels and higher-level concepts, *e.g.*, prompting object recognition [13], affordance learning [24], zero-shot learning [32], and object detection [31]. Recently, several large-scale datasets [13, 68, 38, 49, 27, 29, 25] are released. For attribute recognition, besides direct attribute classification [32, 48, 68, 49] and leveraging the correlation between attribute-attribute and attribute-object [26, 7, 40], intrinsic properties (compositionality, contextuality [42, 43], symmetry [34, 35]) of attribute-object are also proven useful. [42] uses the model weight space to encode the attributes to model the compositionality and contextuality. [43] uses the attributes as linear operators to transform object embeddings. [34] leverages the symmetry property to model the attribute changes within attribute-object coupling and decoupling.

**Object Affordance.** is introduced by [16]. Affordance learning has two canonical paradigms: direct mapping [15] or indirect method [73, 71, 67, 59] with intermediates like object category, attribute, and 3D contents. Some works learned affordance from human-object interactions (HOI) to encode the relation between object and action [18, 70, 28]. Visual Genome [29] provides relations between objects, including actions instead of affordances. However, these relations cover limited and sparse affordances. Differently, we use easily accessible object images as the knowledge source and densely annotate all attributes/affordances for all objects. Besides the vision community, the robot community pays much attention to affordance [53, 64, 52] for grasping and manipulation. For instance, [53] utilized the robot to discover the object affordance via self-supervised learning. Recently, several datasets [45, 6, 8] have been proposed. IIT-AFF [45] collected ten daily indoor objects and provided nine common affordances to construct a dataset for robot applications. Zhu *et al.* [73] built a dataset containing attribute, affordance, human pose, and HOI spatial configuration. But labeling pose and HOI are costly. Chao *et al.* [6] proposed a *semantic* category-level affordance dataset including 91 objects [36] and 957 affordances.

**Causal Inference.** There is increasing literature on exploiting causal inference [50] in machine learning, especially with causal graphical models [62, 50], including feature selection [21] and learning [4], video analysis [51, 33], reinforcement learning [44, 9], *etc*. Recently, Wang *et al.* [66] studied the causal relation between objects in images and used intervention [50] to alleviate the observation bias. Atzmon *et al.* [1] analyze the causal generative model of compositional zero-shot learning and disentangle the representations of attributes and objects. Here, we explore the causal relations between three object levels and apply backdoor adjustment [50] to alleviate the existing bias.

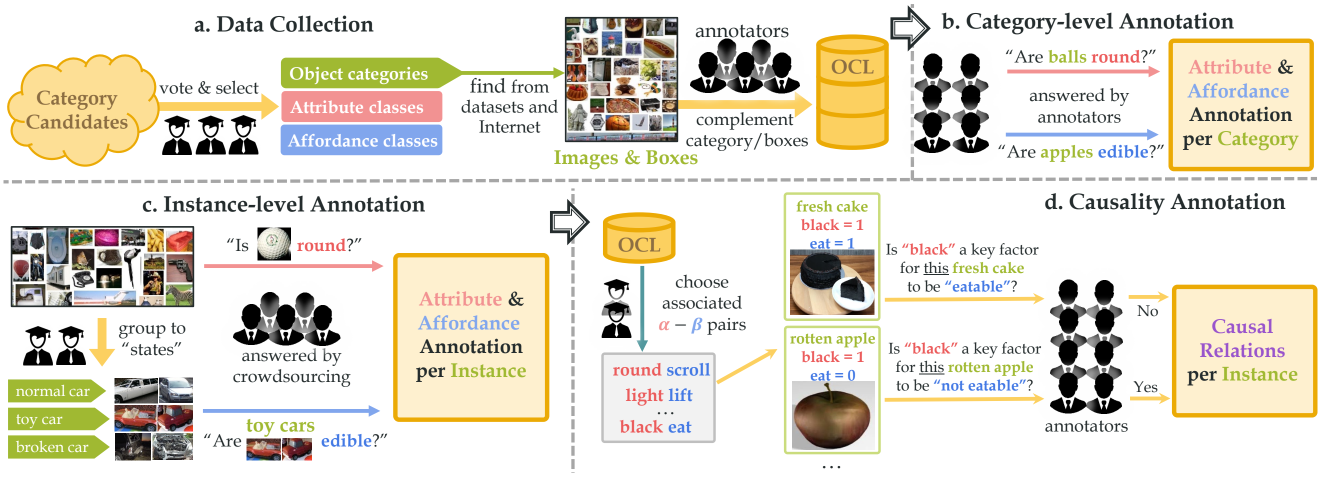

a. Data Collection
Category Candidates
vote & select
Object categories
Attribute classes
Affordance classes
find from datasets and Internet
Images & Boxes
annotators
complement category/boxes
OCL
b. Category-level Annotation
"Are balls round?"
answered by annotators
"Are apples edible?"
Attribute & Affordance Annotation per Category
c. Instance-level Annotation
group to "states"
normal car
toy car
broken car
"Is round?"
answered by crowdsourcing
toy cars
"Are edible?"
Attribute & Affordance Annotation per Instance
OCL
choose associated α − β pairs
round scroll
light lift
…
black eat
d. Causality Annotation
fresh cake
black = 1
eat = 1
Is "black" a key factor for this fresh cake to be "eatable"?
rotten apple
black = 1
eat = 0
Is "black" a key factor for this rotten apple to be "not eatable"?
…
No
Yes
annotators
Causal Relations per Instance

Figure 2: OCL construction. a) Data collection. b) Annotating category-level attributes and affordances. c) Annotating instance-level attributes and affordances. d) Finding direct and clear instance-level causal relations.

## 3. Constructing OCL Benchmark

We construct a benchmark to characterize abundant object knowledge following Fig. 2.

### 3.1. Fine-Grained Object Knowledge Base

**Data Collection.** We briefly introduce the collection of affordances, categories, and attributes classes and image sources here.

(1) **Affordance**: We collect 170 affordances out of 1,006 candidates from widely-used action/affordance datasets [6, 17, 5, 20, 73, 45] given generality and commonness.

(2) **Category**: Considering the taxonomy (WordNet [14]) and diversity, we collect 381 objects out of 1,742 candidates from object datasets [13, 68, 36, 49, 10, 38].

(3) **Attribute**: We manually filter the 500 most frequent attributes from attribute datasets [13, 68, 36, 49, 10, 38, 29] and choose 114 attributes, covering colors, deformations, supercategories, surface, geometrical and physical properties.

(4) **Image**: We extract 75,578 images from object datasets [13, 68, 36, 49, 10, 38, 29], together with Ground Truth (GT) boxes. To ensure diversity, we also manually collected 4,885 Internet images of selected categories. Then, we annotate the missing box and category labels for all instances. Finally, **185,941** instances of **381** categories from **80,463** images are collected: average of 488 instances per category and 2.31 boxes per image. Details are given in the supplementary. OCL is long-tail distributed, where the head categories have over 5,000 instances each, but the rarest categories have only 9 instances, which challenges the robustness of machines greatly.

| Dataset | # Image | # Instance | # Object | # Attribute | # Affordance |
|---|---|---|---|---|---|
| APY [13] | 15,339 | 15,339 | 32 | 64 | / |
| SUN [68] | 14,340 | 14,340 | 717 | 102 | / |
| COCO-a [49] | 84,044 | 188,426 | 29 | 196 | / |
| ImageNet150k [38] | 150,000 | 150,000 | 1,000 | 25 | / |
| Chao *et al.* [6] | / | / | 91 | / | 957 ($B$) |
| Hermans et.al. [24] | 375 | - | 6 | 21 | 7 |
| Zhu *et al.* [73] | 4,000 | 4,000 | 40 | 57 | 14 |
| OCL | 80,463 | 185,941 | 381 | 114 | 170 |

Table 1: *Dense annotated* datasets. OCL provides category- and instance- level attributes ($A$, $\alpha$), affordances ($B$, $\beta$).

**Annotating Attribute** in two levels of granularity: (1) **Category-level** attribute ($A$) contains common sense. For each category, we annotate its *most common* attributes. In concept learning, the usage of the category-level labels as common knowledge can date back to [46]. Following [46], to avoid bias, annotators are given *category-attribute pairs* (category names instead of images) and multiple annotators vote to build the binary $A$ matrix $M_A$ in size of $[381, 114]$. (2) **Instance-level** attribute ($\alpha$) is the individual attributes of *each instance*. The annotation unit is an *attribute-instance pair* and each pair is labeled by multiple annotators.

**Annotating Affordance** in two levels of granularity: (1) **Category-level** affordance $B$, similar to $A$, is annotated in *category-affordance pairs*, indicating the common affordances of each category. Following [6], the annotators label $B$ matrix $M_B$ in size of $[381, 170]$. (2) **Instance-level** affordance $\beta$ is annotated for *each instance* with the help of object *state*. As $B$ is defined by common states, objects in specific states may have different affordances from $B$: if a service robot finds a `broken cup`, it may infer

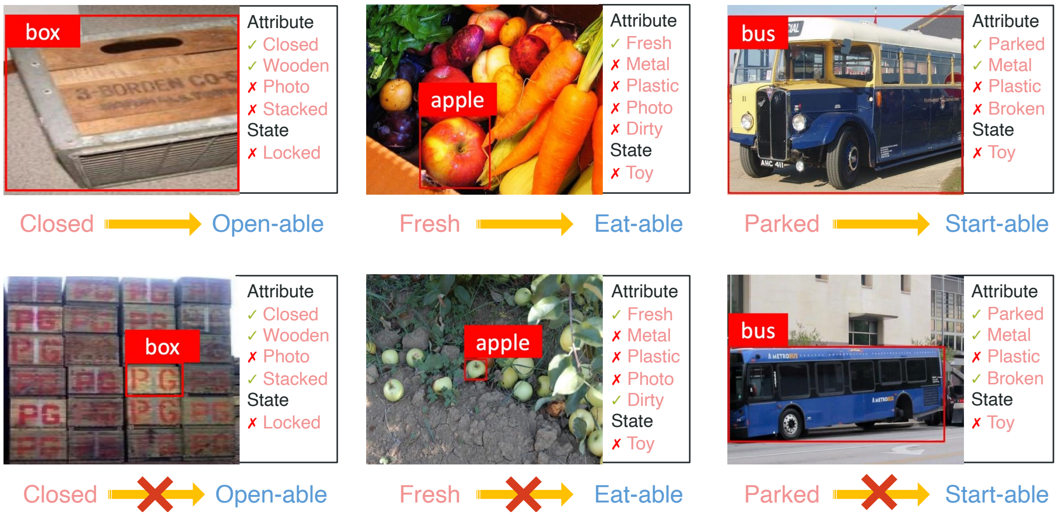


Figure 3: OCL samples including category, $\alpha$ (red), $\beta$ (blue), and their causal relations in various contexts.

that the `cup` can still hold water as it is trained with $B$ labels. Thus, we need detailed $\beta$ beyond $B$. $\beta$ exhibits evident similarities for objects in similar status forming "state" aligning with commonsense, thus we use them to streamline annotation and reduce the annotator discrepancy. A state is defined as an `[category, description` (*e.g.*, a set of attributes)] pair, and instances in a state usually possess similar affordances, *e.g.*, `fresh, juicy, clean oranges` are `eatable`. First, six experts conclude the states by scanning *all* instances of a category and listing all states according to affordance. Then these states were merged manually. In total, **1,376** states are defined, and each category has 3.6 states on average. Next, $\beta$ is annotated for *each state*, and the instances are first assigned with the state-level $\beta$. Bext, the instance-level $\beta$ is **detailed** based on the state-level $\beta$ according to the visual content of each instance. Note that the state is category-dependent and can not be transferred among object categories, which is different from attribute and affordance. Besides, the composition of attributes makes the state space huge and there can be many *unseen* states. Thus, we only use them in annotation but not in our method.

Fig. 3 shows some examples of OCL. We compare OCL with previous dense datasets in Tab. 1. More details, figures, and tables are given in the supplementary.

### 3.2. Causal Graph Definition

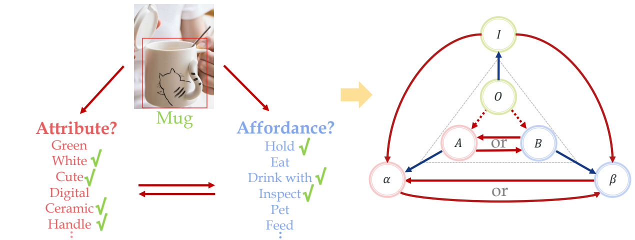
We use a causal graph to shed light on the subtle causalities of our knowledge base in Fig. 4. Causal graph [50] indicates the underlying causalities based on components:

- $O$: object category
- $I$: object instance in an image
- $A$: category-level attribute
- $\alpha$: instance-level attribute
- $B$: category-level affordance
- $\beta$: instance-level affordance

According to the prior knowledge about the causalities between three levels, a hierarchical structure is depicted: **(a)** the **inner** triangle with dotted lines is the **category**-level: object category $O$, category-level attributes $A$, and affordances $B$; **(b)** the **outer** triangle is the **instance**-level: instance visual appearance $I$, instance-level attributes $\alpha$, and affordances $\beta$. Each directed *possible* arc in the graph indicates the *possible* causality between two nodes.

Here, besides the red arcs indicating the common causal relations (*e.g.*, $I \rightarrow \alpha$, $I \rightarrow \beta$ as attribute/affordance recognition from images), we define some special arcs given our category-level attribute and affordance settings: (1) $O \rightarrow A$, $O \rightarrow B$ (dotted arcs): Given $O$, $A$, $B$ are strictly *determined* within labels. (2) $O \rightarrow I$, $A \rightarrow \alpha$, $B \rightarrow \beta$ (blue arcs): The category-level $O$, $A$, and $B$ are direct causes of instance-level $I$, $\alpha$, and $\beta$ during the concept *instantiation*. Note that, according to the previous analysis, we focus on the $A \rightarrow B$ and $\alpha \rightarrow \beta$ but sometimes the opposite can also happen: $A \leftarrow B$ and $\alpha \leftarrow \beta$ ("or" in Fig. 4). In annotation and experiments, we observe that $\alpha \rightarrow \beta$ is stronger and more common and natural to human perception, so we focus more on $\alpha \rightarrow \beta$ in our causal benchmark (Sec. 3.4).

In this work, we focus on $\alpha, \beta$ perception ($I \rightarrow \alpha$, $I \rightarrow \beta$) and visual reasoning (with $I$, inferring $\beta$ given $\alpha$) for embodied AI. Thus, Fig. 4 is simplified. Our knowledge base can support more tasks such as attribute/affordance conditioned image generation ($\alpha \rightarrow I$, $\beta \rightarrow I$) [57]. However, they are beyond the scope of this paper (Suppl. Sec. 3).



(a) **OCL: O**bject **C**oncept **L**earning (b) Causal Graph

Figure 4: Causal graph of our OCL task. "or" indicates that either $A \leftarrow B$ or $A \rightarrow B$ ($\alpha \rightarrow \beta$ or $\alpha \leftarrow \beta$) exists.

### 3.3. Causal Inference Benchmark on $\alpha \rightarrow \beta$

We annotate *instance*-level (considering the context of each instance) causality of $\alpha \rightarrow \beta$ to answer "*which attribute(s) are the critical and direct causes of a certain affordance?*" in two phrases:

**Filtering:** Initially, we need to make binary decisions on all *instance*-$\alpha$-$\beta$ triplets, which is far beyond handleable. Fortunately, we find that **most** $\alpha$-$\beta$ classes (*e.g.*, `shiny` and `kick`) are meaningless and always of no causality. Thus, we exclude the most impossible pairs and only annotate existing rules without ambiguity, meanwhile, guaranteeing the completeness of causality. For each of the 114×170 $\alpha$-$\beta$ pairs, we attach 10 samples for reference and 3 experts vote *yes/no/not-sure*. We take the majority vote and the *not-sure* and controversial pairs are rechecked. The *not-sure* and *no* pairs are removed, and so do the ambiguous pairs. Finally, we obtain about 10% $\alpha$-$\beta$ classes as candidates. The left 90% pairs may hold value, we plan to use LLMs to mine new rules in future work, especially from ambiguous pairs.

**Instance-level Causality**: we adopt object states as a reference. Multiple annotators have been involved for each *state*-$\alpha$-$\beta$ triplet and are asked whether the specific attribute is the *clear* and *direct* cause of this affordance in this state. The answers are combined and checked for all instances of a state. Finally, we obtain about 2 M *instance*-$\alpha$-$\beta$ triplets of causal relations. As we have labeled all $\alpha$ and $\beta$ for all instances, the causal relations would be in four situations: [0,0], [1,1]; [0,1], [1,0]. The former two are "positive", *e.g.*, `fresh`(1/0)→`eat`(1/0) for an `apple`. While the last two are "negative", *e.g.*, `broken`(1/0)→`drive`(0/1) for a *car*.

Fig. 3 shows some causal examples. These causalities are not thoroughly studied in previous datasets [73, 45, 15, 24]. For more details, please refer to the supplementary.

### 3.4. Task Overview

Here, we formulate the OCL task formally. Given an instance $I$ (content in box $b_o$ representing an object instance), OCL aims to infer attribute $\alpha$ and affordance $\beta$ while following the causalities. Formally, OCL can be described as:

$$< P_\alpha, P_\beta >= \mathcal{F}(I, P(O|I)), \tag{1}$$

where $P_\alpha, P_\beta$ are the probabilities of $\alpha, \beta$, $P(O|I)$ is the predicted category probability from an object detector [55].

We aim at benchmarking the reasoning ability of machines, causal relations in Fig. 4 can all be candidates. However, annotating causal relations is usually ambiguous and it is impractical to cover all relations. In a user study, experts met significant divergence when annotating different arcs. For embodied AI, affordance $\beta$ is more important in robot-world interactions. Moreover, both the causal relation annotation and the ablations support that the causal effect of $\alpha \rightarrow \beta$ is more significant than the other alternatives. Thus, we only annotate the unambiguous $\alpha \rightarrow \beta$ (Sec. 3.3) and mainly measure the learning of $\alpha \rightarrow \beta$ here. Formally, the evaluation of $\alpha \rightarrow \beta$ learning follows

$$\Delta P_\beta = ITE[\mathcal{F}(I, P(O|I))], \tag{2}$$

where $\Delta P_\beta$ is the Individual Treatment Effect [60] of **affordance prediction change** after we operate $ITE[\cdot]$ on a model $\mathcal{F}(\cdot)$. $\Delta P_\beta$ is expected to follow the GT causal relation between $\alpha, \beta$ from humans. For example, when the attributes of an object change, then the causal-related affordances should also change accordingly. We will detail the ITE evaluation in Sec. 5. Note that $A, B$ are decided by $O$. Given $O$, we can get $A, B$ via querying the prior $M_A, M_B$ (Sec. 3). Thus, we do not evaluate $A \rightarrow B$ here.

We split images into the train, validation, and test sets with 56K:14K:9K images. The validation and test sets cover 221 of the 381 categories, and the train set covers all categories. OCL is a long-tailed recognition task [19, 69] and requires generalization to cover the whole object category-attribute-affordance space with imbalanced information. Thus, it is challenging for current machines without the reasoning ability to understand the causalities.

## 4. Object Concept Reasoning Network

Before proposing the OCRN, we first simplify the causal graph in Fig. 4 to facilitate the implementation. We focus on $\alpha \rightarrow \beta$ and omit $\beta \rightarrow \alpha$. Similarly, we omit $B \rightarrow A$. Besides, $I, \alpha, \beta$ are the *instantiations* of $O, A, B$ respectively and we use a $O'$ node to represent $O, A, B$. The adapted causal graph is shown in Fig. 5. OCRN implements the **instantiation** of attribute and affordance, corresponding to $A \rightarrow \alpha$, $B \rightarrow \beta$. Thus the model can propose a coarse estimation of attribute and affordance at category-level, then tune the results with the image patterns as a condition for a more accurate prediction. Besides, we exploit **intervention** to remove the causal relation between $I$ and $O$ to construct a category-agnostic model. It suffers less from category bias and is more capable of learning uncommon cases.

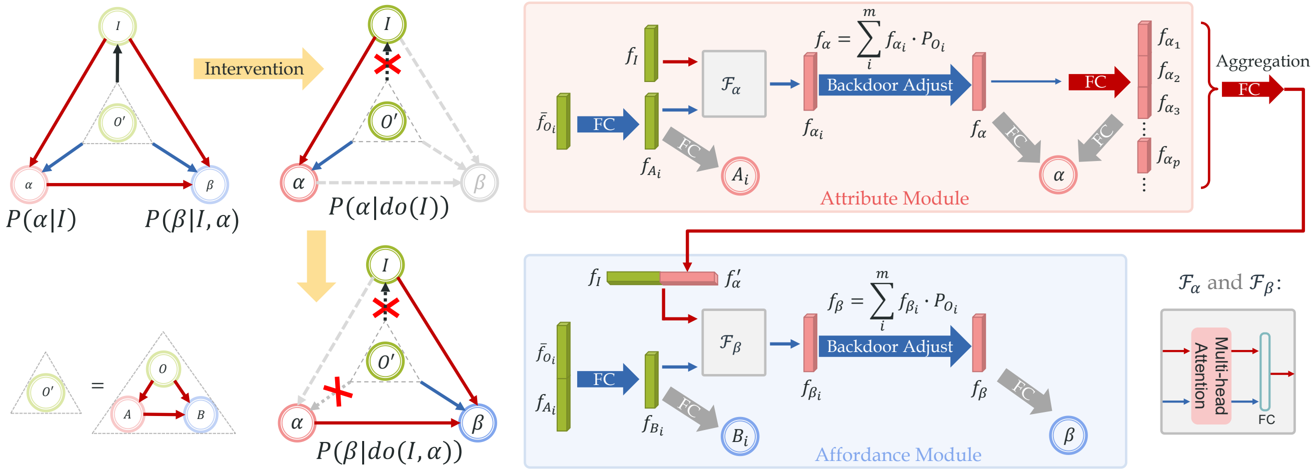


Figure 5: OCRN overview. The arc from $O$ to $I$ is deconfounded. Thus, we can eliminate the bias from the $O$ imbalance. Equations below the graphs are the original or deconfounded estimations of $\alpha, \beta$. Attribute and affordance modules are the **instantiations** of category-level features: categorical features $f_{A_i}$ or $f_{B_i}$ are obtained following the left-bottom-most causal graph and then instantiated via $\mathcal{F}_\alpha$ or $\mathcal{F}_\beta$ conditioned by the instance representations. $f_\alpha$ and $f_\beta$ after intervention are the expectations of instantiated $f_{\alpha_i}$ and $f_{\beta_i}$ w.r.t **prior** $P_{O_i}$. At last, linear-Sigmoid classifiers give the final predictions.

**Object Category Bias.** OCL can be depicted as $P(\alpha|I)$ and $P(\beta|I,\alpha)$. As the samples of different categories are usually imbalanced, conventional methods may suffer from severe *category bias* [66], *e.g.*, animal accounts for 22% instances in OCL, and home appliance only accounts for 3%. In $P(\alpha|I)$, category bias is imported following

$$P(\alpha|I) = \sum_i^m P(\alpha|I,O_i)P(O_i|I), \tag{3}$$

where $P(O_i|I)$ is the predicted category probability. That is, $O$ is a confounder [50] and pollutes attribute inference, especially for the *rare* categories.

**Causal Intervention.** To tackle this, we propose OCRN using intervention [50] to deconfound the confounder $O$ for $\alpha$ (Fig. 5). In $\alpha$ estimation, we use $do(\cdot)$ operation [50] to eliminate the arc from $O$ to $I$: $P(\alpha|do(I))$ is

$$\begin{aligned}&\sum_i^m P(\alpha|I,O_i)P(O_i)\\ =&\sum_i^m P(O_i)\sum_j^m P(\alpha|I,A_j)P(A_j|O_i)\\ =&\sum_i^m P(\alpha|I,A_i)P(O_i),\end{aligned} \tag{4}$$

where $m = 381$. $A_j$ is the category-attribute vector of $j^{th}$ category. As $A$ is decided by $O$, $P(A_j|O_i) = 1$ if $i = j$ and $P(A_j|O_i) = 0$ if $i \neq j$, where $O_i$ is the $i^{th}$ category and $A_j$ is the category-attribute of $j^{th}$ category. $P(O_i)$ is the **prior** probability of the $i$-th category (frequency in our train set). We apply the intervention to reduce the bias from $O$ recognition for an **category-agnostic** model.

Similar to $\alpha$, in $\beta$ estimation, category bias also exists:

$$P(\beta|I,\alpha) = \sum_i^m P(\beta|I,\alpha,O_i)P(O_i|I,\alpha). \tag{5}$$

With Eq. 4, $\alpha$ is beforehand estimated and thus can be seen as "enforced" and deconfounded. For $I$, we again use the intervention [50]:

$$P(\beta|do(I,\alpha)) = \sum_i^m P(\beta|I,\alpha,B_i)P(O_i). \tag{6}$$

Similar to Eq. 4, $P(B_j|O_i) = 1$ if $i = j$, $P(B_j|O_i) = 0$ if $i \neq j$, we omit the process for clarity.

### 4.1. Model Implementation

We represent nodes $\{I, A, B, \alpha, \beta\}$ as $\{f_I, f_A, f_B, f_\alpha, f_\beta\}$ respectively in latent space. $f_I$ is the RoI pooling feature of an instance extracted by a COCO pre-trained ResNet-50 [23]. Following Eq. 4, we represent category-level attribute $A$ based on the *mean* object category feature $\bar{f}_{O_i}$, which is the mean of $f_I$ of all **training** samples in category $O_i$. We map $\bar{f}_{O_i}$ to the attribute latent space $f_{A_i}$ with fully-connected layers (FC) (Fig. 5). $f_{A_i}$ stands for the category-attribute representation for i$^{\text{th}}$ category.

**Attribute Instantiation.** Next, we obtain $\alpha$ representation following Eq. 4:

$$f_{\alpha_i} = \mathcal{F}_\alpha(f_I, f_{A_i}), \quad f_\alpha = \sum_i^m f_{\alpha_i} \cdot P_{O_i}, \tag{7}$$

where $P_{O_i}$ is the *prior* category probability ($P(O_i)$ in Eq. 4). Eq. 7 indicates the attribute *instantiation* from $A$ to $\alpha$ with $I$ as the *condition*. Hence, we can equally translate the $\alpha$ estimation problem into a **conditioned instantiation problem**. $\mathcal{F}_\alpha(\cdot)$ is implemented with multi-head attention [65] with two entries (Fig. 5). The attention output is compressed by a linear layer to the instantiated representation $f_{\alpha_i}$. The debiased representation $f_\alpha$ is the expectation of $f_{\alpha_i}$ w.r.t $P_{O_i}$ according to back-door adjustment in Eq. 4.

We also get the feature for specific attributes for ITE operation (Sec. 5). $f_\alpha$ is first separated to $f_{\alpha_p}$ for each attribute $p$ ($p \in [1, 114]$) by multiple independent FCs, then we can manipulate specific attributes by masking some certain $f_{\alpha_p}$. Next, the features are aggregated via concatenating-compressing by an FC to $f'_\alpha$ as shown in Fig. 5.

**Affordance Instantiation.** Similarly, FCs are used to obtain $f_B$ from $\bar{f}_{O_i}$ and $f_{A_i}$ and Eq. 6 is implemented as:

$$f_{\beta_i} = \mathcal{F}_\beta(f_I, f'_\alpha, f_{B_i}), \quad f_\beta = \sum_i^m f_{\beta_i} \cdot P_{O_i}. \tag{8}$$

$\mathcal{F}_\beta(\cdot)$ operates instantiation with conditions $\{f_I, f'_\alpha, f_{B_i}\}$.

### 4.2. Learning Objectives.

To drive the learning, we devise several objectives:

**Category-level loss** $L_C$. We input category-level $f_A, f_B$ to two linear-Sigmoid classifiers to classify $A, B$. The binary cross-entropy losses are $L_A$ and $L_B$. The total category-level loss is $L_C = L_A + L_B$.

**Instance-level loss** $L_I$. We input instance-level $f_\alpha$, $f_\beta$, together with $f_{\alpha_i}$, $f_{\beta_i}$ to linear-Sigmoid classifiers. The separated $f_{\alpha_p}$ are also sent to independent binary classifiers. The binary cross-entropy losses are represented as $L_\alpha, L_\beta$. The total instance-level loss is $L_I = L_\alpha + L_\beta$.

The total loss is $L = \lambda_C L_C + L_I$. We adopt a two-stage policy: first inferring attributes, then reasoning affordances.

## 5. Experiment

### 5.1. Metrics

$\alpha, \beta$ **Recognition**: we measure the correctness of model prediction $\hat{\alpha}$ and $\hat{\beta}$. For multi-label classification tasks, we use the mean Average Precision (mAP) metric.

**Reasoning**: we use **Individual Treatment Effect (ITE)** [60]. $ITE_i = Y_{i,T=1} - Y_{i,T=0}$ measures the causal effect $T \rightarrow Y$ of i$^{\text{th}}$ individual with the difference between outcomes ($Y$) with or without receiving the treatment ($T$). In OCL, we discuss the causal relation between p$^{\text{th}}$ attribute and q$^{\text{th}}$ affordance: $\alpha_p \rightarrow \beta_q$. So we interpret the treatment $T$ as the **existence of** $\alpha_q$ and the outcome $Y$ as the $\beta_q$ output. We measure the difference of $\beta_q$ output when the whole $\alpha_q$ feature is wiped out or not, which should be non-zero when the causal relation $\alpha_p \rightarrow \beta_q$ exists.

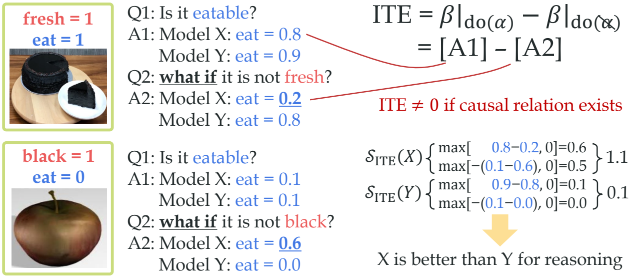

fresh = 1
eat = 1
Q1: Is it eatable?
A1: Model X: eat = 0.8
Model Y: eat = 0.9
Q2: what if it is not fresh?
A2: Model X: eat = 0.2
Model Y: eat = 0.8
ITE = $\beta|_{do(\alpha)} - \beta|_{do(\cancel{\alpha})}$
= [A1] – [A2]
ITE ≠ 0 if causal relation exists
black = 1
eat = 0
Q1: Is it eatable?
A1: Model X: eat = 0.1
Model Y: eat = 0.1
Q2: what if it is not black?
A2: Model X: eat = 0.6
Model Y: eat = 0.0
$\mathcal{S}_{\text{ITE}}(X)$ max[ 0.8−0.2, 0]=0.6, max[−(0.1−0.6), 0]=0.5 1.1
$\mathcal{S}_{\text{ITE}}(Y)$ max[ 0.9−0.8, 0]=0.1, max[−(0.1−0.0), 0]=0.0 0.1
X is better than Y for reasoning

Figure 6: Example of ITE reasoning benchmark.

In detail, given a model, for an instance with causal relation $\alpha_p \rightarrow \beta_q$ ($p \in [1, 114], q \in [1, 170]$), we first formulate ITE as the **affordance probability change** following Eq. 2:

$$ITE = \Delta\hat{\beta}_q = \hat{\beta}_q|_{do(\alpha_p)} - \hat{\beta}_q|_{do(\cancel{\alpha_p})}. \tag{9}$$

$\hat{\beta}_q|_{do(\alpha_p)}$ is the factual output of the affordance probability. $\hat{\beta}_q|_{do(\cancel{\alpha_p})}$ is the counterfactual output when the $\alpha_p$ is wiped out, which can be got by assign zero-mask [63] to the feature of $\alpha_p$ (*e.g.*, $f_{\alpha_p}$ in OCRN) and keep the other features.

Then, based on ITE, we benchmark instances following:

**ITE**: If the causality $\alpha_p \rightarrow \beta_q$ exists on the instance, ITE should be non-zero when eliminating the effect of $\alpha_p$. And the direction of ITE depends on the affordance ground-truth $\beta_q$: if $\beta_q = 0$, the predicted $\hat{\beta}_q$ tend to be 1 after wiping out $\alpha_p$ so ITE should be a negative value; contrarily, ITE should be positive if $\beta_q = 1$. Hence we compute the ITE score as:

$$\mathcal{S}_{\text{ITE}} = \begin{cases} \max(\Delta\hat{\beta}_q,\ 0), & \beta_q = 1, \\ \max(-\Delta\hat{\beta}_q,\ 0), & \beta_q = 0, \end{cases} \tag{10}$$

so that larger $\mathcal{S}_{\text{ITE}}$ indicates the model infers more accurate ITE directions and has better reasoning performance. An example is given in Fig. 6.

$\alpha$**-**$\beta$**-ITE**: we combine recognition and reasoning performances. We multiply $\mathcal{S}_{\text{ITE}}$ with $P(\hat{\alpha_p} = \alpha_p)$ and $P(\hat{\beta_q} = \beta_q)$ as a unified metric $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$.

For all metrics, we compute AP for each $[\alpha_p, \beta_q]$ and average them to mAP. Non-existing pairs are not considered.

### 5.2. Baselines

Different methods exploit different causal paths including the sub-graphs with $\alpha \rightarrow \beta$ or $\alpha \leftarrow \beta$ based on Fig. 4. We implement a series of baselines following different sub-graphs to fully exert the potential of OCL and divide them into 3 folds w.r.t. $\alpha - \beta$ causal structure. We briefly list them here and detail them in the supplementary:

**Fold I.** No arc connecting $\alpha$ and $\beta$:

(1) Direct Mapping from $f_I$ to $P_\alpha, P_\beta$ via an MLP (DM-V): feeding $f_I$ into MLP-Sigmoids to predict $P_\alpha, P_\beta$.

(2) DM Linguistic feature (DM-L): replacing the $f_I$ of DM-V with linguistic feature $f_L$, which is the expectation of Bert [11] embeddings of category names w.r.t $P(O_i|I)$.

(3) Visual-Linguistic alignment, *i.e.*, Multi-Modality (MM): mapping $f_I$ to a latent space and minimizing the distance to $f_L$, feeding it to an MLP-Sigmoids to get $\alpha, \beta$.

(4) Linguistic Correlation of $O$-$\alpha$, $O$-$\beta$ (LingCorr): measuring the correlation between object and $\alpha$ or $\beta$ classes via their Bert [11] embedding cosine similarities. $P_\alpha$, $P_\beta$ are given by multiplying $P(O|I)$ to correlation matrices.

(5) Kernelized Probabilistic Matrix Factorization (KPMF) [72]: calculating feature similarity to all training samples as weights. Taking the weighted sum of GT $\alpha$ or $\beta$ of training samples as predictions.

(6) **A**&**B** Lookup: getting $P_A, P_B$ from $M_A, M_B$.

(7) Hierarchical Mapping (HMa): mapping $f_I$ to category-level attribute or affordance space by an MLP, then feeding it to an MLP-Sigmoids to predict $P_\alpha$ or $P_\beta$.

**Fold II.** $\beta \rightarrow \alpha$:

(8) DM from $\beta$ to $\alpha$ (DM-$\beta \rightarrow \alpha$): same as DM-V but using $f_\beta$ to infer $\alpha$.

(9) DM from $\beta$ and $I$ to $\alpha$ (DM-$\beta I \rightarrow \alpha$): same as DM-V but using both $f_I$ and $f_\beta$ to infer $\alpha$.

**Fold III.** $\alpha \rightarrow \beta$:

(10) DM from $\alpha$ to $\beta$ (DM-$\alpha \rightarrow \beta$): same as DM-V but using both $f_I$ and $f_\alpha$ to infer $\beta$.

(11) DM from $\alpha$ and $I$ to $\beta$ (DM-$\alpha I \rightarrow \beta$): same as DM-V but using both $f_I$ and $f_\alpha$ to infer $\beta$.

(12) Retrieving $\alpha$-$\beta$ relation by Ngram [37] (Ngram): adopting Ngram to retrieve the relevance of $\alpha$ & $\beta$. Then we use DM predicted $\alpha$ and the relevance to estimate $\beta$.

(13) Markov Logic Network [56] (MLN-GT): using **GT** $\alpha$ to infer $\beta$ with MLN.

(14) Instantiation with attention (Attention): feeding $[f_\alpha, f_I]$ to an MLP-Sigmoid to generate attentions and predicting $P_\beta$ by multiplying the attentions with $P_B$.

(15) DM with multi-head attention (DM-att): the $\alpha$ and $\beta$ features are sent to multi-head attention to learn their interaction, then use MLP-Sigmoids to get predictions.

(16) Vanilla CLIP: CLIP [54] trained from scratch.

### 5.3. ITE loss

Though machines are expected to learn the causalities given $\alpha, \beta$ labels only. We wonder how it would perform given *causal supervision*. We adopt an extra Hinge loss to maximize the ITE score of all $[\alpha_p, \beta_q]$. In detail, we intend the ITE of causal relations larger than a margin $\tau$ (= 0.1 in experiments), so the loss term is:

$$\begin{cases} \max\{0, \tau - \Delta\hat{\beta}_q\}, & \beta_q = 1, \\ \max\{0, \tau + \Delta\hat{\beta}_q\}, & \beta_q = 0. \end{cases} \tag{11}$$

We enumerate all *annotated* $[\alpha_p, \beta_q]$ of an instance to obtain $L_{ITE}$. Different from the default, the total loss here is $L = \lambda_C L_C + L_I + \lambda_{ITE} L_{ITE}$.

### 5.4. Implementation Details

For a fair comparison, all methods adopt a shared COCO [36] pre-trained ResNet-50 [23] (frozen) to extract $f_I$ and use the same object boxes in training and inference. In OCRN, the dimension of $f_I$ and all $f_{A_i}, f_{B_i}, f_\alpha, f_\beta$ is 1024. The individual features of each attribute category are 512d and aggregated to 1024d by an FC. We train the attribute module with a learning rate of 0.3 and batch size of 1024 for 470 epochs. Then the attribute module is frozen, and the affordance module is trained with a learning rate of 3.0e-3 and batch size of 768 for 20 epochs. In training, $\lambda_C = 0.03$, $\lambda_{ITE} = 3$.

### 5.5. Results

Tab. 2 presents the results. We can find that the causal structure of the models matters in OCL. Comparing DM methods implementing different causal graphs (including $\alpha \rightarrow \beta$, $\alpha \leftarrow \beta$), $\alpha$ as intermediate knowledge (DM-$\alpha \rightarrow \beta$ and DM-$\alpha I \rightarrow \beta$) could advance $\beta$ perception (DM-V). But when $\beta$ serves as intermediate (DM-$\beta \rightarrow \alpha$ and DM-$\beta I \rightarrow \alpha$), $\beta$ perception is comparable or even worse than DM-V. So the causal relation $\alpha \rightarrow \beta$ is more evident than $\beta \rightarrow \alpha$ in the realistic dataset, which supports our choice in Sec. 3.4 that we focus more on the $\alpha \rightarrow \beta$ arc and implement our model with only $\alpha \rightarrow \beta$.

OCRN outperforms the baselines and achieves decent improvements on all tracks. In terms of $\alpha$ recognition, with or without $L_{ITE}$, OCRN outperforms the second-best method with 1.7 and 2.5 mAP respectively. As for $\beta$ recognition, the improvements are 0.7 and 1.1 mAP with or without $L_{ITE}$. Comparatively, HMa utilizes the supervision of $A, B$, but it performs much worse. $A$&$B$ Lookup directly uses GT $A, B$ to infer $\alpha, \beta$, but its poor performance verifies the significant difference between $A, B$ and $\alpha, \beta$. Moreover, we find that all methods perform better on $\beta$ than $\alpha$, and the improvement of OCRN on $\alpha$ is larger too. This may be because $\alpha$ are more diverse than $\beta$, *e.g.*, we can `eat` lots of `foods`, but `foods` usually have various attributes (`fruit` vs. `pizza`). And OCL also has fewer attribute classes than affordance classes (114 vs. 170). Another reason is that the positive samples in $\beta$ labels (23.2%) are much more than the positives in $\alpha$ labels (9.4%). The different pos-neg ratio affects learning a lot and results in the above gap.

In ITE evaluation, without the guidance of $L_{ITE}$, all methods achieve unsatisfactory performances. However, OCRN still has an advantage. Only MLN-*GT* adopting the first-order logic and *GT* $\alpha$ labels is comparable with OCRN. If trained with $L_{ITE}$ and direct causality labels, all methods perform much better to learn the causalities, *e.g.* on OCRN, the ITE loss brings 10.8 and 7.7 mAP improvements on the two ITE tracks. Particularly, the typical deep learning

| Fold | Method | $\alpha$ | $\beta$ | $\mathcal{S}_{\text{ITE}}$ | $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$ |
|---|---|---|---|---|---|
| i N/A | DM-V | 29.9 | 51.8 | - | - |
| | DM-L | 21.2 | 47.5 | - | - |
| | MM | 23.8 | 48.9 | - | - |
| | LingCorr | 7.9 | 25.9 | - | - |
| | KPMF | 25.4 | 49.1 | - | - |
| | $A\&B$-Lookup | 18.9 | 30.9 | - | - |
| | HMa | 28.6 | 51.7 | - | - |
| | DM-att | 21.9 | 49.2 | - | - |
| | Vanilla CLIP | 23.6 | 49.6 | - | - |
| ii: $\beta \rightarrow \alpha$ | DM-$\beta \rightarrow \alpha$ | 30.0 | 52.0 | - | - |
| | DM-$\beta I \rightarrow \alpha$ | 29.5 | 51.8 | - | - |
| iii: $\alpha \rightarrow \beta$ | DM-$\alpha \rightarrow \beta$ | 28.7 | 52.6 | 7.6 | 6.7 |
| | DM-$\alpha I \rightarrow \beta$ | 29.0 | 52.6 | 8.1 | 7.0 |
| | Ngram | 22.6 | 50.8 | 8.3 | 7.6 |
| | MLN-*GT* | - | 33.4 | **9.5** | 9.1 |
| | Attention | 24.1 | 48.9 | 8.1 | 7.1 |
| | OCRN | **31.6** | **53.3** | **9.5** | **9.2** |
| $\alpha \rightarrow \beta$ | DM-$\alpha \rightarrow \beta$ w/ $L_{ITE}$ | 28.8 | 52.4 | 15.5 | 14.0 |
| | DM-$\alpha I \rightarrow \beta$ w/ $L_{ITE}$ | 29.0 | 52.5 | 15.4 | 13.6 |
| | Ngram w/ $L_{ITE}$ | 22.2 | 49.9 | 14.1 | 12.9 |
| | MLN-*GT* w/ $L_{ITE}$ | - | 33.7 | 12.3 | 11.8 |
| | Attention w/ $L_{ITE}$ | 23.9 | 49.0 | 17.8 | 15.5 |
| | OCRN w/ $L_{ITE}$ | **31.5** | **53.6** | **20.3** | **16.9** |

Table 2: OCL results. w/ $L_{ITE}$ means that training with ITE loss. The baselines in the upper block cannot operate ITE due to the model structure. Different $\alpha$-$\beta$ relations are exploited for causal graph comparison.

model Attention performs best in baselines, but MLN-*GT* no longer holds the advantage. Relatively, OCRN shows more improvements and outperforms Attention with 2.5 and 1.4 mAP on the two ITE tracks.

We provide more visualizations and discussions in the supplementary. In particular, we also apply OCRN to **Human-Object Interaction Detection** [5], where OCRN boosts the performances of multiple HOI models and verifies the generalization and application potential of OCL.

### 5.6. Ablation Study

We verify the components of OCRN on the validation set in Tab. 3.

**(1) Deconfounding.** OCRN w/o deconfounding is implemented following Eq. 3 and 5, where $P(O|I)$ and $P(O|I, \alpha)$ are the category predictions of pre-trained detectors [39]. All the $\alpha$, $\beta$ and ITE performances drop due to the object bias. For more bias analyses please refer to the supplementary.

**(2) Losses.** The performances slightly drop after removing category-level $L_{A_i}, L_{B_i}$, but significantly drop without instance-level $L_\alpha, L_\beta$ by over 20 mAP.

**(3) Feature dimension.** We compare different dimentionality for feature $f_{A_i}, f_{B_i}, f_\alpha, f_\beta$. Smaller and larger feature sizes than 1024 all have degrading effects.

**(4) ITE-related implementations.** We probe some different methods: (a) Mean aggregation: $f'_\alpha = \sum_i f_{\alpha_p}$; (b) Max-pooling aggregation: $f'_\alpha$ is the max value of $f_{\alpha_p}$ as each component; (c) Random counterfactual feature: assigned random vector as the counterfactual attribute feature (instead of zero vector) during ITE. These methods perform worse than the chosen setting on ITE performance but are comparable on $\alpha$ and $\beta$ performance.

| Method | $\alpha$ | $\beta$ | $\mathcal{S}_{\text{ITE}}$ | $\mathcal{S}_{\alpha\text{-}\beta\text{-ITE}}$ |
|---|---|---|---|---|
| OCRN | **32.4** | **52.2** | **20.5** | **17.0** |
| w/o deconfounding | 32.1 | 51.8 | 18.2 | 16.1 |
| w/o $L_{A_i}, L_{B_i}$ | 32.1 | 51.8 | 19.8 | 16.7 |
| w/o $L_\alpha, L_\beta$ | 10.0 | 27.0 | 16.6 | 16.4 |
| 128 Dims | 31.7 | 51.5 | 18.0 | 16.0 |
| 512 Dims | 32.3 | 52.1 | 19.9 | 16.7 |
| 2048 Dims | 32.2 | 51.5 | 19.1 | 16.3 |
| Mean aggregation | 32.2 | 51.3 | 18.9 | 16.7 |
| Max-pooling aggregation | 32.1 | 49.1 | 19.0 | 16.8 |
| Random counterfactual | **32.4** | 51.8 | 5.1 | 5.1 |

Table 3: Ablation study results (validation set).

### 5.7. Discussion

Overall, OCL poses extreme challenges to current AI systems. It expects representative learning to accurately recognize attributes and affordances from raw data meanwhile causal inference to capture the causalities within diverse instances and contexts, *i.e.*, both the *intuitive System 1 and logical System 2* [2]. From the experiments, we find that models struggle to achieve satisfying results on all tracks **simultaneously**. Notably, it is difficult to achieve a satisfying ITE score via data fitting. There is much room for improvement. For future studies, a harmonious performance on $\alpha$, $\beta$, and causality learning are encouraged to better capture object knowledge. Potential directions may include causal representation learning [61], neural-symbolic reasoning [3], and Foundation Models [47]. etc.

## 6. Conclusion

In this work, we introduce object concept learning (OCL) expecting machines to infer affordances and explain what attributes enable an object to possess them. Accordingly, we build an extensive dataset and present OCRN based on casual intervention and instantiation. OCRN achieves decent performance and follows the causalities well. However, OCL remains challenging and would inspire a line of studies on reasoning-based object understanding.

**Acknowledgment**: Supported by the National Key R&D Program of China (No.2021ZD0110704), Shanghai Municipal Science and Technology Major Project (2021SHZDZX0102), Shanghai Qi Zhi Institute, Shanghai Science and Technology Commission (21511101200).

## References

[1] Yuval Atzmon, Felix Kreuk, Uri Shalit, and Gal Chechik. A causal view of compositional zero-shot recognition. *arXiv preprint arXiv:2006.14610*, 2020. 2

[2] Yoshua Bengio. From system 1 deep learning to system 2 deep learning. In *Posner lecture at NeurIPS'2019*, 2019. 9

[3] Tarek R Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C Lamb, Daniel Lowd, Priscila Machado Vieira Lima, et al. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*, 2017. 9

[4] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. *arXiv preprint arXiv:1412.2309*, 2014. 2

[5] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huayi Zeng, and Jia Deng. Learning to detect human-object interactions. In *WACV*, 2018. 3, 9

[6] Yu-Wei Chao, Zhan Wang, Rada Mihalcea, and Jia Deng. Mining semantic affordances of visual object categories. In *CVPR*, 2015. 2, 3

[7] Chao-Yeh Chen and Kristen Grauman. Inferring analogous attributes. In *CVPR*, 2014. 2

[8] Ching-Yao Chuang, Jiaman Li, Antonio Torralba, and Sanja Fidler. Learning to act properly: Predicting and explaining affordances from images. In *CVPR*, 2018. 2

[9] Ishita Dasgupta, Jane Wang, Silvia Chiappa, Jovana Mitrovic, Pedro Ortega, David Raposo, Edward Hughes, Peter Battaglia, Matthew Botvinick, and Zeb Kurth-Nelson. Causal reasoning from meta-reinforcement learning. *arXiv preprint arXiv:1901.08162*, 2019. 2

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 3

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 8

[12] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 2

[13] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 1, 2, 3

[14] Christiane Fellbaum. Wordnet. *The encyclopedia of applied linguistics*, 2012. 3

[15] David F Fouhey, Xiaolong Wang, and Abhinav Gupta. In defense of the direct perception of affordances. *arXiv preprint arXiv:1505.01085*, 2015. 2, 5

[16] James J Gibson. The ecological approach to the visual perception of pictures. *Leonardo*, 11(3):227–235, 1978. 1, 2

[17] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*, 2018. 3

[18] Abhinav Gupta and Larry S Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007. 2

[19] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 5

[20] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3

[21] Isabelle Guyon, Constantin Aliferis, and André Elisseeff. Causal feature selection. *Computational methods of feature selection*, pages 63–82, 2007. 2

[22] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990. 1

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6, 8

[24] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *ICRA Workshop*, 2011. 2, 3, 5

[25] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 2

[26] Sung Ju Hwang, Fei Sha, and Kristen Grauman. Sharing features between objects and their attributes. In *CVPR*, 2011. 2

[27] Phillip Isola, Joseph J Lim, and Edward H Adelson. Discovering states and transformations in image collections. In *CVPR*, 2015. 1, 2

[28] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *ECCV*, 2018. 2

[29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2016. 2, 3

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[31] Krishna Kumar Singh, Santosh Divvala, Ali Farhadi, and Yong Jae Lee. Dock: Detecting objects by transferring common-sense knowledge. In *ECCV*, 2018. 2

[32] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, 2009. 1, 2

[33] Karel Lebeda, Simon Hadfield, and Richard Bowden. Exploring causal relationships in visual object tracking. In *ICCV*, 2015. 2

[34] Yong-Lu Li, Yue Xu, Xiaohan Mao, and Cewu Lu. Symmetry and group in attribute-object compositions. *CVPR*, 2020. 2

[35] Yong-Lu Li, Yue Xu, Xinyu Xu, Xiaohan Mao, and Cewu Lu. Learning single/multi-attribute of object with symmetry and group. *TPAMI*, 2021. 2

[36] Tsung Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 1, 2, 3, 8

[37] Y. Lin, J.-B. Michel, E. L. Aiden, J. Orwant, W. Brockman, and S. Petrov. Syntactic annotations for the google books ngram corpus. In *ACL*, 2012. 8

[38] H. Liu, R. Wang, S. Shan, and X. Chen. Learning multifunctional binary codes for both category and attribute oriented retrieval tasks. In *CVPR*, 2017. 2, 3

[39] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 9

[40] Dhruv Mahajan, Sundararajan Sellamanickam, and Vinod Nair. A joint learning framework for attribute models and object descriptions. In *ICCV*, 2011. 2

[41] Alex Martin. The representation of object concepts in the brain. *Annu. Rev. Psychol.*, 58:25–45, 2007. 1

[42] Ishan Misra, Abhinav Gupta, and Martial Hebert. From red wine to red tomato: Composition with context. In *CVPR*, 2017. 2

[43] Tushar Nagarajan and Kristen Grauman. Attributes as operators: factorizing unseen attribute-object compositions. In *ECCV*, 2018. 2

[44] Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019. 2

[45] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017. 2, 3, 5

[46] Daniel N Osherson, Joshua Stern, Ormond Wilkie, Michael Stob, and Edward E Smith. Default probability. *Cognitive Science*, 15(2):251–269, 1991. 3

[47] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022. 9

[48] Devi Parikh and Kristen Grauman. Relative attributes. In *ICCV*, 2011. 2

[49] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. In *ECCV*, 2016. 2, 3

[50] Judea Pearl, Madelyn Glymour, and Nicholas P Jewell. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016. 2, 4, 6

[51] Lyndsey C Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Scholkopf, and William T Freeman. Seeing the arrow of time. In *CVPR*, 2014. 2

[52] Lerrel Pinto, Dhiraj Gandhi, Yuanfeng Han, Yong-Lae Park, and Abhinav Gupta. The curious robot: Learning visual representations via physical interactions. In *ECCV*, 2016. 2

[53] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *ICRA*, 2016. 2

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 8

[55] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1, 5

[56] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006. 8

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 4

[58] B Ross. Category learning: Learning to access and use relevant knowledge. *Memory and mind: A Festschrift for Gordon H. Bower*, pages 229–246, 2008. 1

[59] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016. 2

[60] Donald B Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005. 5, 7

[61] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. 9

[62] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. 2

[63] Kaihua Tang, Yulei Niu, Jianqiang Huang, Jiaxin Shi, and Hanwang Zhang. Unbiased scene graph generation from biased training. In *CVPR*, 2020. 7

[64] Spyridon Thermos, Georgios Th Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. In *CVPR*, 2017. 2

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 7

[66] Tan Wang, Jianqiang Huang, Hanwang Zhang, and Qianru Sun. Visual commonsense r-cnn. In *CVPR*, 2020. 2, 6

[67] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 2

[68] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010. 1, 2, 3

[69] Yue Xu, Yong-Lu Li, Jiefeng Li, and Cewu Lu. Constructing balance from imbalance for long-tailed image recognition. In *ECCV*, 2022. 5

[70] Bangpeng Yao, Jiayuan Ma, and Li Fei-Fei. Discovering object functionality. In *ICCV*, 2013. 2

[71] Yibiao Zhao and Song-Chun Zhu. Scene parsing by integrating function, geometry and appearance models. In *CVPR*, 2013. 2

[72] Tinghui Zhou, Hanhuai Shan, Arindam Banerjee, and Guillermo Sapiro. Kernelized probabilistic matrix factorization: Exploiting graphs and side information. In *SDM*, 2012. 8

[73] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *ECCV*, 2014. 2, 3, 5