Bongard-HOI: Benchmarking Few-Shot Visual Reasoning for Human-Object Interactions

Huaizu Jiang[1*], Xiaojian Ma[2*], Weili Nie[3] , Zhiding Yu[3] ,
Yuke Zhu[3,4], Song-Chun Zhu[2], Anima Anandkumar[3,5]
[1]Northeastern University [2]UCLA [3]NVIDIA [4]UT Austin [5]Caltech
h.jiang@northeastern.edu, xiaojian.ma@ucla.edu, {wnie,zhidingy}@nvidia.com,
yukez@cs.utexas.edu, sczhu@stat.ucla.edu, anima@caltech.edu

Abstract

A significant gap remains between today's visual pattern recognition models and human-level visual cognition especially when it comes to few-shot learning and compositional reasoning of novel concepts. We introduce ***Bongard-HOI****, a new visual reasoning benchmark that focuses on compositional learning of human-object interactions (HOIs) from natural images. It is inspired by two desirable characteristics from the classical Bongard problems (BPs): 1) few-shot concept learning, and 2) context-dependent reasoning. We carefully curate the few-shot instances with hard negatives, where positive and negative images only disagree on action labels, making mere recognition of object categories insufficient to complete our benchmarks. We also design multiple test sets to systematically study the generalization of visual learning models, where we vary the overlap of the HOI concepts between the training and test sets of few-shot instances, from partial to no overlaps. Bongard-HOI presents a substantial challenge to today's visual recognition models. The state-of-the-art HOI detection model achieves only 62% accuracy on few-shot binary prediction while even amateur human testers on MTurk have 91% accuracy. With the Bongard-HOI benchmark, we hope to further advance research efforts in visual reasoning, especially in holistic perception-reasoning systems and better representation learning. Code is available.*[1]

1. Introduction

In recent years, great strides have been made on visual recognition benchmarks, such as ImageNet [8] and COCO [34]. Nonetheless, there remains a considerable gap between machine-level pattern recognition and human-level cognitive reasoning. Current image understanding models typically require a large amount of training data yet struggle to generalize beyond the visual concepts seen during training. In contrast, humans can reason about new visual concepts in a compositional manner from just a few examples [22]. To march towards human-level visual cognition, we need to depart from conventional benchmarks on closed-vocabulary recognition tasks and aim to systematically examine compositional and few-shot learning of novel visual concepts.

While existing benchmarks such as miniImageNet [47], Meta-Dataset [45], and ORBIT [45] have been dedicated

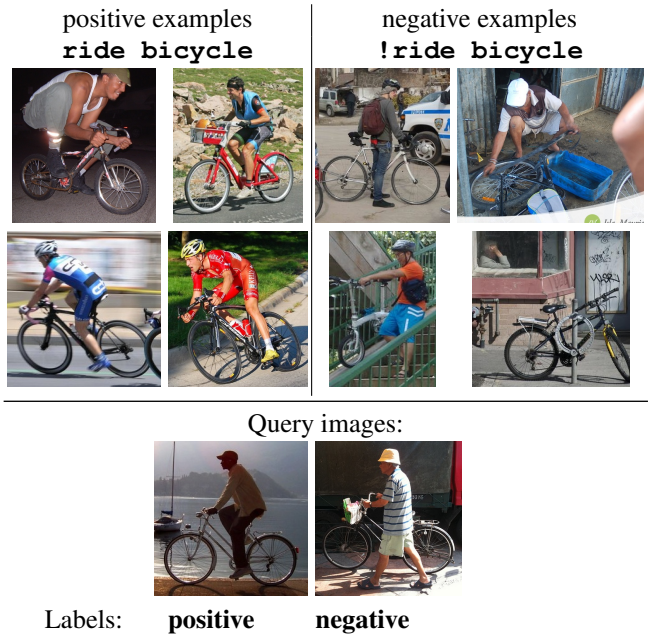


Figure 1. **Illustration of a few-shot learning instance from our Bongard-HOI benchmark.** The positive images in the top left part follow the visual relationship of riding a bike between the person and objects while such a relationship does not exist in the negative examples. Note that an actual problem in Bongard-HOI contains 6 images of positive examples, 6 negative examples, and 1 query image, which is different from the illustration here.

*First two authors contributed equally.
[1]https://github.com/nvlabs/Bongard-HOI

Figure 2. **Examples of different actions with the same object.** From top to bottom, left to right: washing, walking, and feeding dogs; eating, squeezing, and peeling oranges. To differentiate these images, we need compositional understanding on both the actions and the objects. We exploit this to select *hard negatives* in Bongard-HOI: negative images contain the same object as the positives, but the actions are different.

to studying few-shot visual learning, they focus on recognizing object categories instead of the compositional structures of visual concepts, *e.g.*, visual relationships. A parallel line of research aims at building benchmarks for abstract reasoning by taking inspiration from cognitive science such as RPM (Raven-style Progressive Matrices) [2, 44] and Bongard-LOGO [3, 33]. In these benchmarks, a model has to learn concept induction rules from a few examples and the concepts are context-dependent in each task. However, they use simple synthetic images [2, 33] or focus on basic object-level properties, such as shapes and categories [44].

Our new benchmark: In this paper, we introduce **Bongard-HOI**, a new benchmark for compositional visual reasoning with natural images. It studies human-object interactions (HOIs) as the visual concepts, requiring explicit compositional reasoning of object-level concepts. Our Bongard-HOI benchmark inherits two important characteristics of the classic Bongard problems (BPs) [3]: 1) *few-shot binary prediction*, where a visual concept needs to be induced from just six positive and six negative examples and 2) *context-dependent reasoning*, where the label of an image may be interpreted differently under different contexts.

Furthermore, Bongard-HOI upgrades the original BPs from synthetic graphics to natural images. Our benchmark contains rich visual stimuli featuring large intra-class variance, cluttered background, diverse scene layouts, etc. In Bongard-HOI, a single few-shot binary prediction instance, referred to as BP, contains a set of six positive images and a set of six negative images, along with query images (see Fig. 1 for examples). The task is making binary predictions on the query images.

We construct the few-shot instances in Bongard-HOI on top of the HAKE dataset [24, 25]. To encourage the explicit reasoning of visual relationships, we use *hard negatives* to construct few-shot instances. The hard negatives consist of negatives that contain objects from the same categories as those contained in the positive sets but with different action labels. Fig. 2 presents some examples of these images. Since both positive and negative examples contain object instances from the same categories, mere recognition of object categories is insufficient to complete the tasks. Rather, reasoning about visual relationships between person and objects is required to solve these few-shot binary prediction problems. The existence of such hard negatives distinguishes our benchmark from existing visual abstract reasoning counterparts [2, 33, 44]. Comparisons with different benchmarks can be found in Table 1.

We carefully curate the annotations in HAKE when constructing the few-shot instances. Recall the visual concept contained in the positive images should not appear in any of the negative ones. Thus, we have to carefully select the images in both sets. We employ high-quality annotators from the Amazon Mechanical Turk platform to curate the test set to further remove ambiguously and wrongly labeled few-shot instances. In this process, 2.5% of the few-shot instances in the test set are discarded. We end up with 23K and 15K few-shot instances in disjoint training and test sets, respectively.

An important goal of the Bongard-HOI benchmark is to *systematically* study the generalization of machine learning models for real-world visual relationship reasoning. To this end, we introduce four separate test sets to investigate different types of generalization, depending on whether the action and object classes are seen in the training set. Fig. 3 illustrates their design. This way, we have full control of the overlap between the concepts (*i.e.*, HOIs) between training and test of few-shot instances. It enables us to carefully examine the generalization of visual learning models. Ideally, a learning model should be able to generalize beyond the concepts it has seen during training. Even for unseen HOI concepts, the model should be able to learn *how to induce* the underlying visual relationship from just a few examples.

Establishing baselines: In our experiments, we first examine the state-of-the-art HOI detection models' performance on this new task, we trained an oracle model with HOITrans [55] on all the HOI categories, *including those in the test sets of our Bongard-HOI benchmark*, and output binary prediction on the query image via a majority vote based on HOI detections. Its accuracy is only 62.46% (with a chance performance of 50%), demonstrating the challenge of our visual reasoning tasks. We then evaluate state-of-the-art few-shot learning approaches, including non-episodic and meta-learning methods. We show that the current learning models struggle to solve the Bongard-HOI problems. Compared to amateur human testers' 91.42% overall accuracy, who have access to a few examples of visual relationships

Figure 3. **Illustration of our four separate test sets for different types of generalization.** We show a few HOI concepts in the training and test sets in the top and bottom row, respectively. We use the `red` fonts to denote an object or action class that is available in the training set and `blue` fonts indicate those held-on unseen ones in the test set.

before working on solving our problems, the state-of-the-art few-shot learning model [6] only has 55.82% accuracy.

The results above lead to this question: *why do they perform so poorly?* To this end, we offer a detailed analysis of the results and propose several conjectures. The first one is a lack of holistic perception and reasoning systems, since models that have only good pattern recognition performances, *e.g.* HOITrans, are likely to fail on our benchmarks. Moreover, we believe there is a need for additional representation learning, *e.g.* pre-training, since currently we only train on binary labels of few-shot instances. Nonetheless, we believe much effort is still needed to further investigate the challenges brought by our benchmark.

To sum up, this paper makes the following contributions:

- We introduce Bongard-HOI, a new benchmark for few-shot visual reasoning with human-object interactions, aiming at combining the best of few-shot learning, compositional reasoning, and challenging real-world scenes.
- We carefully curate Bonagrd-HOI with hard negatives, making mere recognition of object categories insufficient to complete our tasks. We also introduce multiple test sets to systematically study different types of generalization.
- We analyze state-of-the-art few-shot learning and HOI detection methods. However, experimental results show their inability on achieving good results on Bongard-HOI. Our conjectures suggest future research in models with holistic perception-reasoning systems and better representations.

2. Bongard-HOI Benchmark

For a few-shot binary prediction instance in Bonagrd-HOI, it has a set of positive examples $\mathcal{P}$, a set of negative samples $\mathcal{N}$, and a query image I_q. Images in $\mathcal{P}$ depict a certain visual concept (*e.g.*, `ride bicycle` in Fig. 1), while images in $\mathcal{N}$ do not. In each task, there are only six images in both $\mathcal{P}$ and $\mathcal{N}$. As a result, a human tester or machine learning model needs to induce the underlying concept from just a few examples. Given the query image I_q, a binary prediction needs to be made: whether the certain visual concept depicted in $\mathcal{P}$ is available in I_q or not. Later, we will detail how to construct these few-shot instances.

2.1. Constructing Bongard Problems

Few-shot instances in Bongard-HOI are constructed with natural images. We choose to use visual relationships as underlying visual concepts. In our early experiments, we also studied visual attributes to construct few-shot instances, for example, color and shape of bird parts [48], facial attributes [28]. But such visual attributes annotations either require too much domain knowledge for human annotators or are too noisy to curate. Another option we investigated is scene graph [19], which is a combination of both visual relationships and visual attributes. However, there could be too many convoluted visual concepts in a single image, resulting in ambiguous few-shot instances.

In this paper, we construct few-shot instances on top of the HAKE dataset [24, 25] focusing on human-object interactions. It provides unified annotations following the annotation protocol in HICO [4] for a set of datasets widely used for HOI detection, including HICO [4], V-COCO [10], Open-Images [20], HCVRD [53], and PIC [26]. HAKE has 80 object categories, which are consistent with the vocabulary defined in the standard COCO dataset [27]. It also has 117 action labels, leading to 600 human-object relationships[2].

Denote a concept $c = \langle s, a, o \rangle$ as a visual relationship triplet, where s, a, o are the class labels of subject, action, and object, respectively. In this paper, s is always

[2]Some combinations of objects and actions are infeasible.

Table 1. **An overview of different benchmark datasets covering HOI detection, few-shot learning, and abstract visual reasoning.** In the first row, the abbreviation *ctx* denotes context; *generalization types* indicates if a benchmark includes multiple test splits to examine different types of generalization. *We consider the concept of object counts as compositional while others such as object attributes and categories not [44]).

	concept	compositional concept	natural image	few-shot	ctx-dependent reasoning	hard negatives	generalization types	#concepts	#tasks
HAKE [24, 25]	HOI	✓	✓	✗	✗	✓	✗	600	122.6K
Omniglot [21]	shape	✗	✗	✓	✓	✗	✗	50	1.62K
miniImageNet [47]	image label	✗	✓	✗	✓	✗	✗	100	60K
Meta-Dataset [45]	image label	✗	✓	✓	✗	✗	✗	4,934	52.8M
ORBIT [31]	frame label	✗	✓	✓	✗	✗	✗	486	2.69M
RPM [2]	shape	✗	✗	✓	✓	✗	✓	50	11.36M
V-PROM [44]	attributes & counts	✓*	✓	✓	✓	✗	✓	478	235K
Bongard-LOGO [33]	shape	✗	✗	✓	✓	✗	✓	627	12K
Bongard-HOI (ours)	HOI	✓	✓	✓	✓	✓	✓	242	53K

`person`. We start with selecting a set of positive images $\mathcal{I}_c = \{I_1, \dots\}$ from HAKE that depict such a relationship. We also need negative images, where the visual concept c is not contained by them. In specific, we collect another set of images $\mathcal{I}_{\bar{c}}$ with concept $\bar{c} = \langle s, \bar{a}, o \rangle$, where $\bar{a} \neq a$, meaning that we select *hard negatives*. As a result, images from both $\mathcal{I}_c$ and $\mathcal{I}_{\bar{c}}$ contain the same categories of objects and the only differences are the action labels, *making it impossible to trivially distinguish positive images from the negatives by doing visual recognition of object categories only*. Rather, detailed visual reasoning about the interactions of human and objects are desired. Fig. 2 illustrates the difficulties introduced by the hard negatives. Finally, as an entire image may contain multiple HOI instances, we use image regions (crops) around each HOI instance instead of the original image to ensure only a single HOI instance is presented in a single image.

Next, we need to sample few-shot instances from the positive images $\mathcal{I}_c$ and the negatives $\mathcal{I}_{\bar{c}}$. We randomly sample images to form $\mathcal{P}, \mathcal{N}$, and a query image I_q. Two parameters control the sampling process: M, the number of images in $\mathcal{P}$ and $\mathcal{N}$ ($M = 6$ in Bongard-HOI), and the overlap threshold τ, indicating the maximum number of overlapped images between two few-shot instances. We want to sample as many few-shot instances as possible, but we also need to avoid significant image overlap between few-shot instances, which limits the diversity of the data. We end up setting $\tau = 3$ and $\tau = 2$ for training and test sets, respectively. More details can be found in the supplementary material.

2.2. Data Curation

Although the HAKE dataset [24, 25] has provided high-quality annotations, we found that curations are still needed to construct few-shot instances. Recall, to sample negative images, we assume a particular action is not depicted in them. In HAKE, an image region may have multiple action labels. Naively relying on the provided annotations is problematic as the action labels are either not manually exclusive or not exhaustively annotated. We hire high-quality testers on the Amazon Mechanical Turk (MTurk) platform, who maintain a good job approval record, to curate existing HOI annotations. We discuss the data curation process in detail and show visual examples in detail in the supplementary material.

After the aforementioned data curations, each image region is assigned to a single action label, describing the most salient visual relationship. With the curated annotations, action labels between a person and objects of a certain category are mutually exclusive so that we can significantly reduce the ambiguity when constructing few-shot instances. Finally, we hire high-quality testers on the MTurk platform to further remove the ambiguous few-shot instances in the test set. Every single few-shot instance is assigned to three independent testers. We compare their responses with the ground-truth labels and discard about 2.5% few-shot instances where none of the three testers correctly classifies the query images. In the end, we report the accuracy of human testers on those left unambiguous few-shot instances as a human study to examine human-level performance on our Bongard-HOI benchmark, where the average accuracy is 91.42%.

2.3. Generalization Tests

Transferring the knowledge that an agent has seen and learned is a hallmark of visual intelligence, which is a long-stand goal for the entire AI community. It is also a core focus of the Bongard-HOI benchmark. Following [2], we provide multiple test splits to investigate different types of generalization, aiming at a systematic understanding of how the tested models generalize on our benchmark. Specifically, the visual concept we consider in Bongard-HOI is an HOI triplet $\langle s, a, o \rangle$ and we have two variables of freedom: action a and object o. Therefore, by controlling whether an action or object is seen during training, we can study generalization to unseen actions, unseen objects, or a combination of two. We end up introducing four separate test sets, as shown in Fig. 3. We provide detailed statistics on our training and test

sets in the supplementary material.

Ideally, after learning from examples of `sit_on bed`, a machine learning model can quickly grasp the concept `sit_on bench`. More importantly, such a model should learn *how to learn* from just a few examples, so that they can still induce the correct concept (visual relationship) in the most challenging cases, where both actions and objects are not seen during training (*e.g.*, `shear sheep`).

3. Possible Models for Bongard-HOI

There are many possible ways of tackling Bongard-HOI, such as few-shot learning, conventional HOI detection, etc. We are particularly interested in investigating few-shot learning methods, as our benchmark requires the learner to identify the visual concept with very few samples (positive and negative images in $\mathcal{P}$ and $\mathcal{N}$, respectively). To further improve the few-shot learning methods, we consider encoding the images with Relation Network [41], aiming at better compositionality in the learned representations. Finally, we introduce an oracle model to testify whether Bongard-HOI can be trivially solved using state-of-the-art HOI detection models.

3.1. Few-shot Learning in Bongard-HOI

We start with a formal definition of the few-shot learning problem in Bongard-HOI. Specifically, each task includes multiple few-shot *instance* with $N = 2$ classes and $2M$ samples, *i.e.*, the model learns from a training set $\mathcal{S} = \mathcal{P} \cup \mathcal{N} = \{(I_1^P, 1), \ldots, (I_M^P, 1), (I_1^N, 0), \ldots, (I_M^N, 0)\}$ and is evaluated on a query image (I_q, y_q). Each example (I, y) includes an image $I \in \mathbb{R}^{H \times W \times 3}$ and a class label $y \in \{0, 1\}$, indicating whether I contains the visual concepts depicted in $\mathcal{P}$. In Bonagrd-HOI, we set $M = 6$ as our default parameter and therefore each few-shot instance is "2-way, 6-shot". Following [45], we propose to solve these few-shot prediction instances with the following two families of approaches:

Non-episodic methods. In these methods, a simple classifier is trained to map all the images in a few-shot instance (including images in $\mathcal{P}$, $\mathcal{N}$, and the query image) to the class of the query. The classifier can be parameterized as a neural network over some learned image embeddings, *i.e.* representations produced by convolutional neural networks (CNNs). In other words, we view each few-shot instance as a single training sample $(\bigcup_{i=1}^{2M+1} I_i, y_q)$ rather than a few-shot instance with multiple training samples (I, y). Our experiments cover two different ways to encode the images: CNN and Wide Relational Network (WReN) [2, 33].

Meta-learning methods. These methods adopt the episodic learning setting, *i.e.*, they learn to train a classifier using $2M$ samples from $\mathcal{S}$ and evaluate their trained classifier on the query (I_q, y_q). In general, their objective (also called *meta-objective*) is to minimize the prediction error on the query. Different meta-learning methods have their own ways to build the classifier and optimize the meta-objective. In our experiments, we consider the following state-of-the-art methods: 1) *ProtoNet* [43], a metric-based method; 2) *MetaOptNet* [23] and *ANIL* [35], two optimization-based approaches. Moreover, we also use a strong baseline meta-learning model, *Meta-Baseline* [6], which reports competitive results in many few-shot prediction tasks. We refer readers to the related papers for more details.

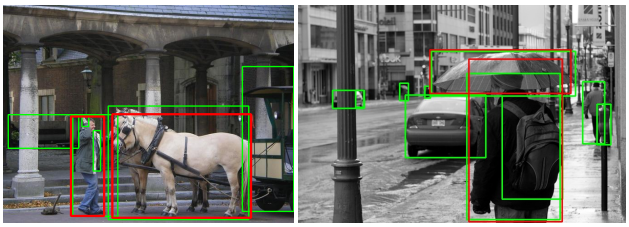


Figure 4. **Class-agnostic (objectness) detections.** We show the detections from our class-agonostic detector (in green) and ground-truth human and object boxes (in red).

3.1.1 Image Encoding with Relational Network

As mentioned above, representation learning of the input images can be crucial to the success of few-shot learning methods on Bongard-HOI. As our benchmark demands learning compostional concepts (HOIs), simply feeding an image into a Convolutional Neural Network (CNN) is not optimal. To this end, we propose to use the Relational Network [41], which shows promising compositional reasoning accuracy on a Visual Question Answering (VQA) benchmark [16], to explicitly encode the compositionality of visual relationships. In specific, the feature representations of the image I is computed as

$$\mathrm{RN}(I) = f_\phi \circ \sum_{i,j} g_\psi \left(\texttt{concat}(h_\theta(o_i, I), h_\theta(o_j, I))\right),$$

where o_i and o_j are two detected objects of the image I, provided by ground truth object annotations or a pre-trained object detector like Faster R-CNN [38]. h_θ denotes the RoI Pooled features of o_i from a ResNet backbone [12] followed by a MLP (multi-layer perceptron) [38], which is parameterized by θ. g_ψ and f_ϕ are two additional MLPs.

A challenge we are facing is the unseen object categories in the test sets. Since the object detector has to be pre-trained on a dataset without the unseen object categories, it is likely to fail on our test set where images could contain objects belonging to these categories. To tackle this issue, we train a binary class-agnostic (objectness) detection model instead to get o_i and o_j. Class-agnostic object detections are shown in Fig. 4. As we can see, all objects of interest have been successfully detected. But at the same time, there are a lot of other distracting ones, such as the bench and the wagon in the left image of Fig. 4. This is a unique challenge of

dealing with visual reasoning over real-world images. We devote discussions to it in the experiment section.

3.2. Oracle

One may wonder if our Bongard-HOI benchmark could be trivially solved using the state-of-the-art HOI detection model. To address this concern, we develop an oracle model resorting to the HOITrans [55], which is based on the Transformer model [46] and reports state-of-the-art accuracy on the HICO [4] and V-COCO [10] benchmarks. In specific, let's denote the HOI detections in the $\mathcal{P}$ and $\mathcal{N}$ as $\mathcal{D}^P$ and $\mathcal{D}^N$, respectively. $\mathcal{D}^P$ contains the detections from all of the images in the $\mathcal{P}$, defined as $\mathcal{D}^P = \{c_i^P\}_{i=1}^{N_P}$, where c_i^P is a HOI triplet introduced in Section 2.1. N_P is the total number of detections. Note that there may be multiple or no detections for a single image. Similarly, $\mathcal{D}^N$ is defined as $\mathcal{D}^N = \{c_i^N\}_{i=1}^{N_N}$. According to the property of Bongard-HOI, the visual concept c_P should only appear in the $\mathcal{P}$, not in the $\mathcal{N}$. We, therefore, compute c_P as

$$c_P = \texttt{majority\_vote}(\mathcal{D}^P - \mathcal{D}^N),$$

where $-$ is the set operator for set subtraction. Here we first exclude the HOIs detected in $\mathcal{N}$ from $\mathcal{D}^P$, then the majority of the remaining HOIs will be viewed as the visual concept c_P. Given the detections $\mathcal{D}^q = \{c_i^q\}_{i=1}^{N_q}$ for the query image I_q, our prediction y becomes

$$y = \begin{cases} 1, & \text{if } c_P \in \mathcal{D}^q, \\ 0, & \text{otherwise.} \end{cases}$$

Discussions of how to deal with the corner cases, *e.g.*, `majority_vote` returns more than 1 concept, $\mathcal{D}^q$ is empty, etc, are provided in the supplementary material. We illustrate how this model works in Fig. 5, where we show HOI detections in each image.

We call it our oracle model as it has privileged information, *i.e.*, the entire HOI action & object vocabulary, including those held-out ones in the test set. As we shall we in Section 4, such an oracle model still struggles on our Bongard-HOI benchmark, achieving only 62.46% accuracy on average, which is far below the human-level performance of 91.42%. It suggests that our Bongard-HOI benchmark is not trivial to solve.

4. Experiments

4.1. Implementation Details

We benchmark the models introduced in Section 3 on Bongard-HOI to test their performance on human-level few-shot visual reasoning. We use a ResNet50 [12] as an encoder for the input images. We consider different pre-training strategies: 1) no pre-training at all (scratch), 2) pre-trained on the ImageNet dataset with manual labels [8], and 3) latest self-supervised approach [5] pre-trained on ImageNet but without manual labels. We train an Faster R-CNN [38]

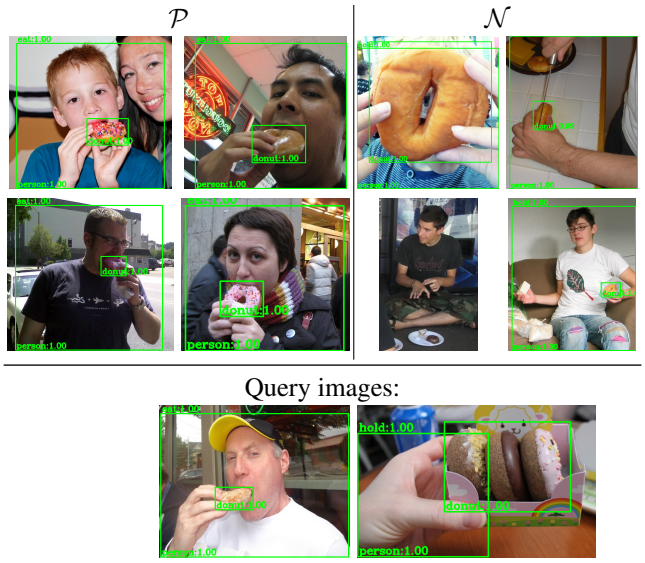


Figure 5. **Illustration of our oracle model.** We first generate some detections for all the images using HOITrans [55]. Note that some images may not have any detection at all. According to the detections in the $\mathcal{P}$ and $\mathcal{N}$, the common concept is `eat donut`. As a result, in the bottom row, the first query image is considered to be positive as its HOI detections contain `eat donat`. The second query image is negative. Zoom in for the best view.

class-agnostic objectness detection model on the COCO dataset [34] using a ResNet101 [12] pre-trained on ImageNet [8] as the backbone. We use the RoIPool operation [38] to get feature representations for each bounding box. We also use ground-truth bounding boxes provided in HAKE [24] as input to diagnose the effectiveness of the visual perception. In addition to RoIPooled region features, we also concatenate each bounding box's normalized coordinates (center and spatial dimensions) as spatial information to the Relational Network encoder introduced in Section 3.1.1.

4.2. Quantitative Results

The quantitative results of different models on our Bongard-HOI benchmark can be found in Table 2. We make the following observations: First of all, despite the overall difficulties brought by our benchmarks, most models perform worse on the challenging test splits, where actions and/or object categories are completely unseen during training. This observation aligns well with our hypothesis, *i.e.* existing machine learning approaches can be limited in terms of generalizing beyond training concepts. It also echos the findings in Bongard-LOGO [33], a dataset studying a similar problem with synthetic images. Second, meta-learning approaches generally tend to perform better than non-episodic counterparts, which can be on par with or even worse than random guesses (50% chance). We hypothesize the reason to be the focus on *learning to learn* in these methods, which is essentially required to solve the few-shot instances in

Table 2. **Quantitative results on the Bongard-HOI benchmark.** All the models use a ResNet50 as the image encoder. For the input of bounding boxes (bbox), we consider two options: from an object detection model (det) and ground-truth annotations (gt). For the ResNet50 encoder, we experiment with different pre-training strategies: no pre-training at all (scratch), pre-trained on the ImageNet dataset with manual labels (IN), and state-of-the-art self-supervised approach MoCoV2 [5]. (* denotes that we are unable to get meaningful results; # indicates that the trained model has a run-time error during the inference stage since the condition of the QP solver can not be satisfied).

	bbox	pre-train	test set				avg.
			seen act., seen obj.	seen act., unseen obj.	unseen act., seen obj.	unseen act., unseen obj.	
CNN-Baseline [33]	-	scratch	50.03	49.89	49.77	50.01	49.92
WReN-BP [2, 33]	-	IN	50.31	49.72	49.97	49.01	49.75
ProtoNet* [43]	det	IN	-	-	-	-	-
ProtoNet [43]	gt	IN	58.90	58.77	57.11	58.34	58.28
MetaOptNet# [23]	det	IN	-	-	-	-	-
MetaOptNet [23]	gt	IN	58.60	58.28	58.39	56.59	57.97
ANIL [35]	det	IN	50.18	50.13	49.81	48.83	49.74
ANIL [35]	gt	IN	52.73	50.11	49.55	48.19	50.15
Meta-Baseline [6]	det	scratch	54.61	53.79	54.58	53.94	54.23
Meta-Baseline [6]	det	MoCoV2	55.23	54.54	54.32	53.11	54.30
Meta-Baseline [6]	det	IN	56.45	56.02	55.60	55.21	55.82
Meta-Baseline [6]	gt	IN	58.82	58.75	58.56	57.04	58.30
HOITrans [55] (oracle)	-	-	59.50	64.38	63.10	62.87	62.46
Human (Amateur)	-	-	87.21	90.01	93.61	94.85	91.42

the Bongard-HOI benchmark, especially for the challenging test splits with novel categories. Similar observations have also been made in Bongard-LOGO. Moreover, some meta-learning models are distracted by bounding boxes provided by an object detection model. We will discuss this issue in the next section.

Surprisingly, the oracle model (*HOITrans*) also struggles on our tests with an averaged accuracy of 62.46%, albeit being trained with direct HOI supervision and all action&object categories. It suggests a clear gap between the existing HOI detection datasets, *e.g.* HAKE [24] and Bongard-HOI, where the latter one requires capabilities beyond perception, *e.g.* HOI recognition. Rather, a model might also need context-dependent reasoning, learning-to-learn from very few examples, etc., to perform well on our benchmarks.

Finally, machine learning models still largely fall behind amateur human testers (*e.g.*, 55.82% of Meta-Baseline vs 91.42%). While we only give human testers a couple of examples about visual relationships before they start working on solving Bongard-HOI, they can quickly learn how to induce visual relationships from just a few examples, reporting an average 91.42% accuracy on our Bongard-HOI benchmark. Particularly, there are no significant differences for the different subsets of the test set. We hope our findings will foster more research efforts on closing this gap.

4.3. Discussions

We need holistic perception and reasoning. Our work suggests that the significant challenges in current visual reasoning systems lie in both the reliability of perception and the intricacy of the reasoning task itself. Models that have only good pattern recognition performances are likely to fail on our benchmarks. Rather, an ideal learner needs to integrate visual perception in natural scenes and detailed cognitive reasoning as a whole. This marks our key motivation to propose Bongard-HOI as the first step towards studying these two problems holistically.

Pre-training improves performances. Intuitively, models for Bongard-HOI might need additional representation learning, *e.g.* pre-training, since currently we only train on binary labels of few-shot instances. We can see from Table 2 that *pre-training is very helpful*. Compared to no pre-training, using either manual labels or self-supervision leads to a performance boost. In particular, the self-supervised pre-training [5] does not use any manual labels for supervision. Yet it can produce better results than learning from scratch.

Visual perception matters in Bongard-HOI. Finally, an imperfect perception could still be a major obstacle here. Different from Bongard-LOGO [33] which uses synthetic shapes instead, Bongard-HOI studies visual reasoning on natural scenes, which often contain rich visual stimuli, issuing such as large intra-class variance and cluttered background also present challenges to reliable visual perception

on which reasoning is based. In our case, bounding boxes produced by an object detection model can be inevitably noisy. Some meta-learning models, including ProtoNet [43], have difficulties inducing the true visual relationships. For MetaOptNet [23], although we can finish training, we constantly encounter run-time errors where the condition of the QP solver is not satisfied during the inference stage. Instead, when taking clean ground-truth bounding boxes as input, all of these approaches produce better accuracy. Note that using ground-truth bounding boxes only serves as an oracle, which does not indicate the models' authentic performance.

5. Related Work

Visual relationship detection benchmarks. Various benchmarks are also dedicated for visual relationship recognition and detection, particularly for human-centric relationships (*i.e.*, HOI). In the seminal work of Visual Genome [19], scene graph annotations, including relationships of different objects, are provided. A subset of the annotations is used in VRD [29] to focus on visual relationship detection. In a recent effort, large-scale visual relationships are provided in the Open Images dataset [20]. HOI, is of particular interest to understand the interactions of humans and other objects. A lot of HOI benchmarks, such as HICO [4], COCO-a [39], vCOCO [10], and HOI-COCO [14], are built on top of the object categories provided in the COCO dataset [27]. The MECCANO [36] dataset focuses on human-object interactions in egocentric settings and industrial scenarios. Ambiguous-HOI [25] is part of the HAKE project [24], where the focus is human activity understanding with a large-scale knowledge base and visual reasoning.

Although our Bongard-HOI benchmark is built on top of the dataset HAKE [24], it differs from the existing visual relationship and HOI benchmarks, since we focus on human-level cognitive reasoning instead of recognition. To solve Bongard-HOI, one might not need to explicitly name the underlying visual relationship but does need to induce the HOI from a few images and perform context-dependent reasoning. Our results also suggest that Bongard-HOI cannot be trivially solved by the state-of-the-art models on these datasets, *e.g.* HOITrans [55].

Few-shot and meta learning models. Few-shot learning aims at learning from a limited number of training samples [9, 18]. With the goal of extracting the generic knowledge across tasks and generalizing to a new task using task-specific information, meta-learning (or learning-to-learn) [13] becomes one of the leading approaches to deal with the few-shot learning problems. In general, meta-learning methods are divided into three categories: 1) memory-based methods, such as MANN [40] and SNAIL [32], 2) metric-based methods, such as Matching Networks [47] and ProtoNet [43], and 3) optimization-based methods, such as MetaOptNet [23] and ANIL [35].

These meta-learning methods have been evaluated on several commonly used few-shot learning benchmarks, including miniImageNet [47] and tieredImageNet [37]. Although state-of-the-art meta-learning algorithms have achieved excellent performance on these standard few-shot image classification benchmarks, whether these approaches can generalize to tasks where the concepts to learn (in a few-shot manner) are compositional, *e.g.* visual relationships rather than simple object categories is unknown [15, 17]. In other words, existing benchmarks fail to account for the challenging problem of generalizing to new compositional concepts in few-shot learning. Therefore, with a focus on the more challenging visual concepts of visual relationships, we propose Bongard-HOI to serve as a new benchmark for the few-shot learning methods. We believe that our benchmark can foster the development of new few-shot learning, especially meta-learning algorithms to achieve better performances on learning and generalizing with compositional concepts.

Abstract visual reasoning benchmarks. Inspired by cognitive studies, several benchmarks have been built for abstract reasoning, highlighting cognitive abstract reasoning. Notable examples include compositional question answering [16, 30], physical reasoning [1, 51], math problems [42], and general artificial intelligence [7, 50]. The most relevant to our benchmark are RPM [2, 52], its variant with natural images [44], and Bongard problems with synthetic shapes [33] and physical problems [49]. While most of them consider synthetic images [2, 33, 49], our Bongard-HOI benchmark studies cognitive reasoning on natural images, which impose unique challenges due to the difficulty of visual perception. Moreover, we use human-object interaction as the underlying concepts to construct few-shot instances, which require explicit compositional concept learning in a few-shot manner, compared to the object categories and shapes [44]. Moreover, the existence of hard negatives in the few-shot instances makes our benchmark more challenging.

6. Conclusion

In this paper, we introduced the Bongard-HOI benchmark focusing on the few-shot learning and the generalization with compositional concepts in real-world visual relationship reasoning. Drawing inspirations from the classic Bongard problems [3], we constructed few-shot instances using the visual relationships between humans and objects as the underlying concepts. Our benchmark is built on top of an existing HOI dataset, HAKE [24], where we carefully curated the provided annotations to construct the few-shot instances. We benchmarked state-of-the-art few-shot learning methods, including both non-episodic and meta-learning approaches. Our findings suggested that current machine learning models still struggle to generalize beyond concepts that they have seen during the training process. Moreover, natural images in our benchmark contain rich stimuli, impos-

ing great challenges to the machine learning models in the real-world visual relationship reasoning tasks. By building the Bongard-HOI benchmark, we hope to foster research efforts in real-world visual relationship reasoning, especially in holistic perception-reasoning systems and better representation learning.

A. Limitation Statement

We re-use the images collected by the HAKE [4] creators, including the ones for HICO [4], V-COCO [10], OpenImages [20], HCVRD [53], and PIC [26], which were crawled from the web. Except the images, in this paper, no identity related information were collected nor used when constructing the dataset and benchmarking other approaches. It is possible, however, that some person may be identified via facial recognition techniques. We will provide contact information of the benchmark maintainer and commit to processing request of removing some certain images from the dataset. In addition, similar to other human-centric dataset, the images we use are from just a small portion of the population, which may contain biases toward some certain races, gender, ethnic groups, etc. We are unable to measure the bias as we do not have any identity-related data. We encourage researchers to investigate such issues.

B. More details on the Bongard-HOI Benchmark

B.1. Constructing Bongard Problems

Given positive images $\mathcal{I}_c$ that depict a certain relationship $c = \langle s, a, o \rangle$ and and negative images $\mathcal{I}_{\bar{c}}$ that does not, we need to sample few-shot instances from them. We randomly sample images to form $\mathcal{P}$, $\mathcal{N}$, and a query image I_q. Two parameters control the sampling process: M, the number of images in $\mathcal{P}$ and $\mathcal{N}$ ($M = 6$ in Bongard-HOI), and the overlap threshold τ, indicating the maximum number of overlapped images between two few-shot instances. We want to sample as many few-shot instances as possible, but we also need to avoid significant image overlap between few-shot instances, which limits the diversity of the data. The sampling process in summarized in Algorithm 1. We set $\tau = 3$ and $\tau = 2$ for training and test sets, respectively.

Algorithm 1: Sample few-shot instances for a visual concept c

Input: Positive images $\mathcal{I}_c$, negative images $\mathcal{I}_{\bar{c}}$, number of images in a few-shot instance M, overlap threshold τ.
Output: Sampled few-shot instances $\mathcal{Q}$.

```
Q = ∅;
while True do
    P^i, N^i, I_q^i = sample_instance(I_c, I_c̄, M);
    if sample fails then
        break;
    t = overlap(P^i, N^i, I_q^i, Q);
    if t < τ then
        Q = Q ∪ (P^i, N^i, I_q^i);
```

B.2. Data Curation

Although the HAKE dataset [24] has provided high-quality annotations, we found that curations are still needed to construct the Bongard problems (few-shot instances) for our Bongard-HOI benchmark. Recall, to sample negative images, we assume a particular action is not depicted in them. In HAKE, an image region may have multiple action labels. Naively relying on the provided annotations is problematic as the action labels are either not manually exclusive or not exhaustively annotated. We show different cases of data curations in Fig. 6 and discuss them in details as follows.

Similar actions. Although some action labels may convey different semantic meanings, for some certain object categories, they look visually similar and indistinguishable. As shown in Fig. 6(a), `scratch cat` and `pet cat` are hard to differentiate visually. If we simply use images of `scratch cat` as negatives to construct few-shot instances for `pet cat`, such few-shot instances are ambiguous, as it violates the basic assumption that the visual concept depicted in the Set $\mathcal{A}$ is not available in the Set $\mathcal{B}$. We therefore simply merge such similar action labels to reduce the visual ambiguity.

Hierarchical actions. Action labels are inherently hierarchical. For example, as shown in Fig. 6(b), `eat carrot` very likely also means `hold carrot` visually. There are two problems to construct few-shot instances with multiple hierarchical action labels associated with the same image region. First of all, as we previously explained, using images of `eat carrot` as negatives for `hold carrot` may cause ambiguity. More importantly, there is the *visual specificity* issue. People tend to focus on capturing the most salient actions in an image, which are usually the parent actions (`eat carrot` in this case). In our preliminary experiments, images of `eat carrot` were used as positives for `hold carrot` to construct few-shot instances. We found that it caused a lot of confusion for human testers. To this end, we merge such hierarchical action labels for the same region, keeping the parent action labels only.

Hard-to-see objects. In some cases, the person or the objects in image regions are hard to see. For example, in Fig. 6(c), the person with the action label `stand_on boat` is hard to see clearly. On the one hand, it causes significant challenges for a visual perception system (*e.g.*, [11]) to accurately localize the meaningful objects. At the same time, it also imposes difficulty for annotators to accurately annotate the image region. We simply discard all image regions with hard-to-see objects.

Extrapolating actions. Actions are continuous. As a result, annotators tend to *extrapolate* the action label given a single image, instead of describing the current state the action. For example, as we can see in the top row of Fig. 6(d), the `eat` action is about to happen. Yet, the action is different from a normal `hold banana`

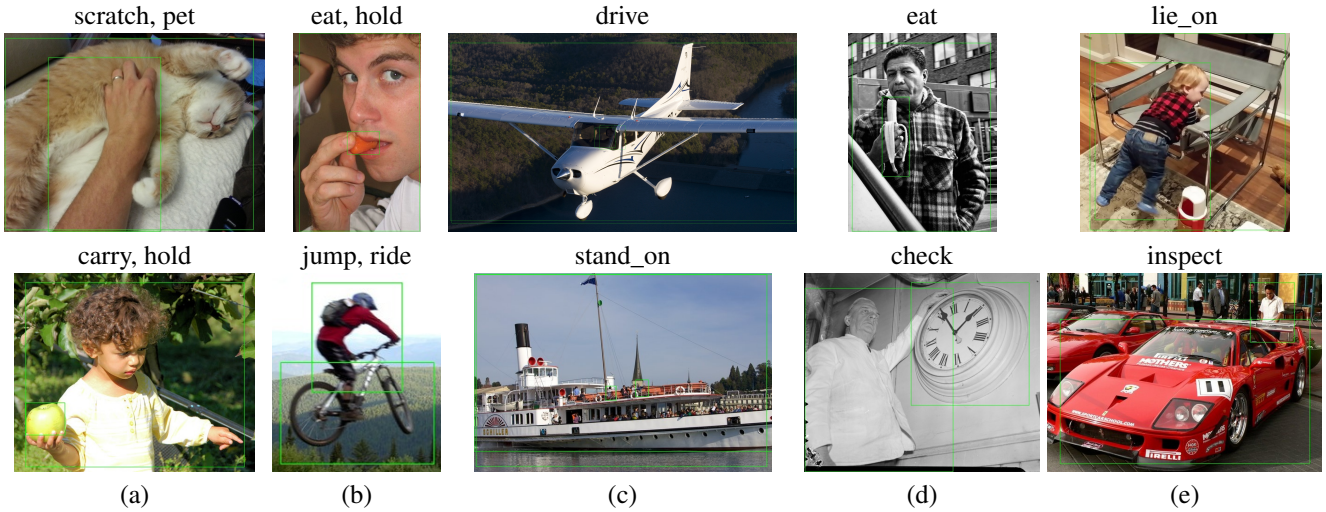


Figure 6. **Samples of annotations where curations are needed.** For each image region, its annotated action labels are show on its top and bounding boxes corresponding to the person and object are shown for visualization purpose. From left to right: (a) similar actions, (b) hierarchical annotations, (c) hard-to-see objects, (d) extrapolating annotations, and (e) inaccurate or confusing annotations.

without any indication of `eat`. To distinguish different scenarios, we introduce `hold_not_about_to_eat banana`, `hold_and_about_to_eat banana`, and `eat banana`. In this way, all the actions are mutually exclusive. We can sample image regions for form few-shot instances without worrying about causing ambiguity.

Inaccurate or confusing actions. In some rare cases, the annotations in HAKE are inaccurate or confusing, as shown in Fig. 6(e). We modify the action labels if such a image region depicts a clear action label. Otherwise we discard such regions to avoid introducing ambiguity to sampled few-shot instances.

MTurk data curation. After performing the aforementioned data curations, each image region is assigned to a single action label, describing the most salient content. Such action labels are mutually exclusive so that we can significantly reduce the ambiguity when constructing few-shot instances. Finally, we hire high-quality testers on the Amazon Mechanical Turk (MTurk) platform, who maintain a good job approval record, to curate the testing set to further remove the ambiguous few-shot instances. Every single BP is assigned to three independent testers. We compare their responses with the ground-truth labels and disard about 2.5% few-shot instances where none of the three testers correctly classifies the query images. We provide more details of the MTurk curations in Section D.

B.3. Dataset statistics

Our Bongard-HOI benchmark provides disjoint training, validation, and testing sets. In specific, there are 118 concepts (visual relationships) and 21,956 few-shot instances in the training set. There are 17,184 and 13,941 few-shot instances in the validation and testing set, respectively, corresponding to 167 and 166 visual concepts. Detailed distribution of concepts and few-shot instances among different generalization types are provided in Table 3.

	seen object	unseen object
seen action	99 / 5008	36 / 5002
unseen action	20 / 3402	12 / 3775

(a) validation set

	seen object	unseen object
seen action	102 / 4476	27 / 4562
unseen action	21 / 3291	16 / 1612

(b) test set

Table 3. **Number of concepts and few-shot instances in the validation and test sets.** Depending on whether an action and object is seen during the training, we divide the validation and test sets into four categories, where we can study the systematic generalization of machine learning models. For each category, we show number of concepts (combinations of action and object) and number of few-shot instances.

B.4. Illustration about the Context-Dependent Reasoning Property

Two Bongard problems (few-shot instancess) are shown in Fig. 7. For the same query image, among different context (*i.e.*, positive and negative examples), it receives different classification labels. This context-dependent reasoning property distinguishes our Bongard-HOI benchmark from other few-shot learning ones, where an image always has a fixed label.

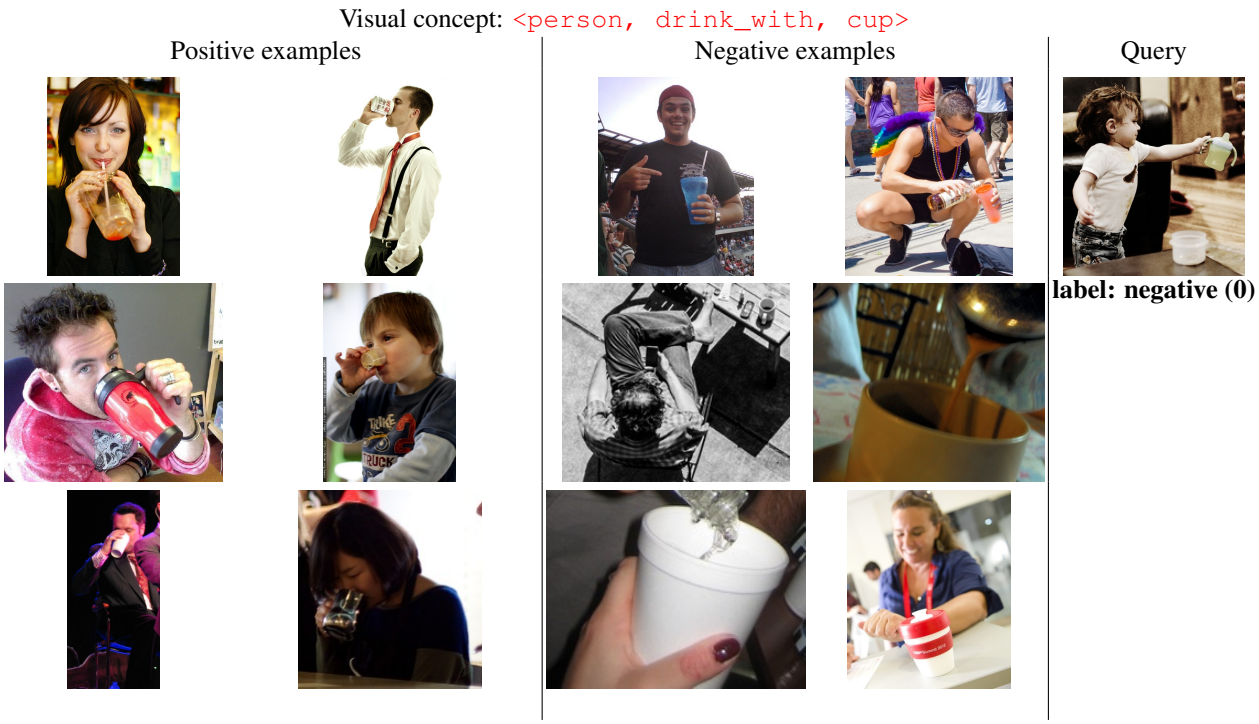
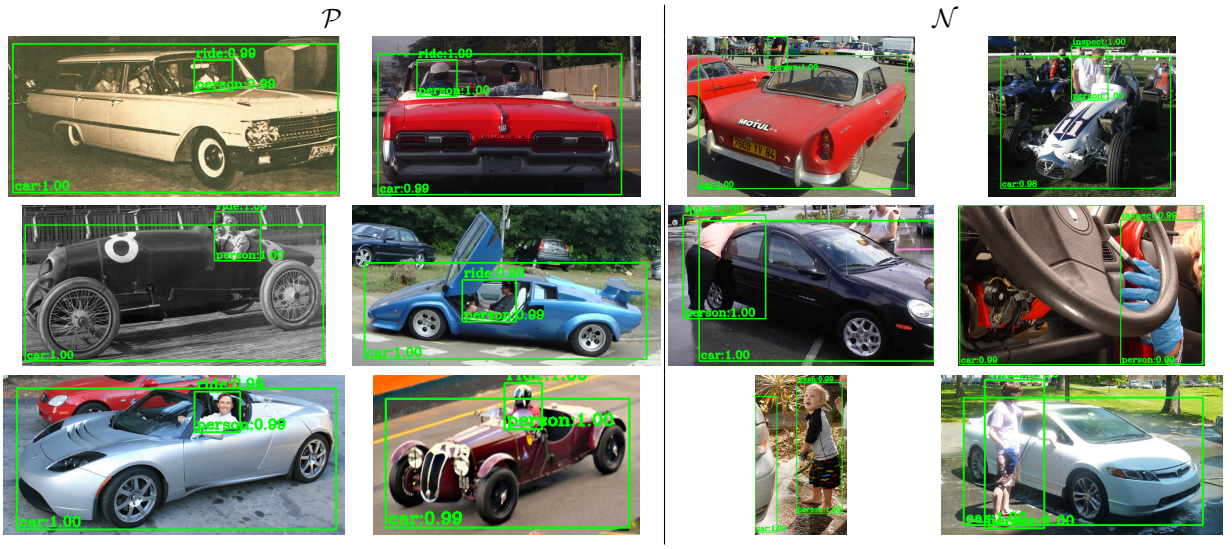


Figure 7. **Illustration of the context-dependent reasoning property of the Bongard problems (few-shot instances) in our Bongard-HOI benchmark.** Two instances are shown here with their underlying visual concepts (relationships) displayed on top with red color. The same query image receives two different labels (negative in the top and positive in the bottom) among different context (*i.e.*, positive and negative examples).

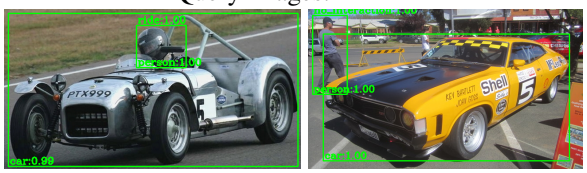
Query images:

Predictions: **positive** **negative**

Figure 8. **Illustration of our oracle model.** The concept in $\mathcal{P}$ is `wash car`.



Query images:



Predictions: **positive** **negative**

Figure 9. **Illustration of our oracle model.** The concept in $\mathcal{P}$ is `ride car`.

Figure 10. **Illustration of our oracle model.** The concept in $\mathcal{P}$ is `teach person`.

C. More details on the oracle model

We first review how our oracle model works. Denoting the HOI detections in the $\mathcal{P}$ and $\mathcal{N}$ as $\mathcal{D}^P$ and $\mathcal{D}^N$, respectively. $\mathcal{D}^P$ contains the detections from all of the images in the $\mathcal{P}$, defined as $\mathcal{D}^P = \{c_i^P\}_{i=1}^{N_P}$, where c_i^P is a HOI triplet. N_P is the total number of detections. Note that there may be multiple or no detections for a single image. Similarly, $\mathcal{D}^P$ is defined as $\mathcal{D}^P = \{c_i^P\}_{i=1}^{N_P}$. According to the property of Bongard-HOI, the visual concept c_P should only appear in the $\mathcal{P}$, not in the $\mathcal{N}$. We, therefore, compute c_P as

$$c_P = \texttt{majority\_vote}(\mathcal{D}^P - \mathcal{D}^N),$$

where $-$ is the set operator for set subtraction. Given the detections $\mathcal{D}^q = \{c_i^q\}_{i=1}^{N_q}$ for the query image I_q, our prediction y becomes

$$y = \begin{cases} 1, & \text{if } c_P \in \mathcal{D}^q, \\ 0, & \text{otherwise.} \end{cases}$$

We now discuss some possible corner cases where the main paper does not cover.

What if `majority_vote` return multiple concepts? In this case, we simply enumerate each of them when making predictions for y. The predicted y will be 1 as long as at least one returned concepts present in $\mathcal{D}^q$; otherwise it will be 0.

What if $\mathcal{D}^P$, $\mathcal{D}^N$ or $\mathcal{D}^q$ is empty? In case when $\mathcal{D}^P$ is empty, we view this example as an failure case for our oracle model, as it does not induce the right concept as expected. On the contrary, it's totally fine that $\mathcal{D}^N$, meaning that no detection need to be removed from $\mathcal{D}^P$. Finally, how we handle the case when $\mathcal{D}_q$ is empty depends on the true label $y^\star$. If $y^\star$ is 1, then we view this example as an failure case. But we will make the prediction an automatic success if $y^\star$ is 0, since our oracle model finds there is no ground truth concept presenting in the query, which should be the right prediction.

We show successful cases of our oracle model in Fig. 8, Fig. 9, Fig. 10, Fig. 11. A failure case is shown in Fig. 12.

D. More Details on MTurk Data Curation

User interface. The user interface of data curation on the Amazon Mechanical Turkp (MTurk) platform is shown in Fig. 13. In the top part, we show images depicting a common visual relationship between human and objects in the left

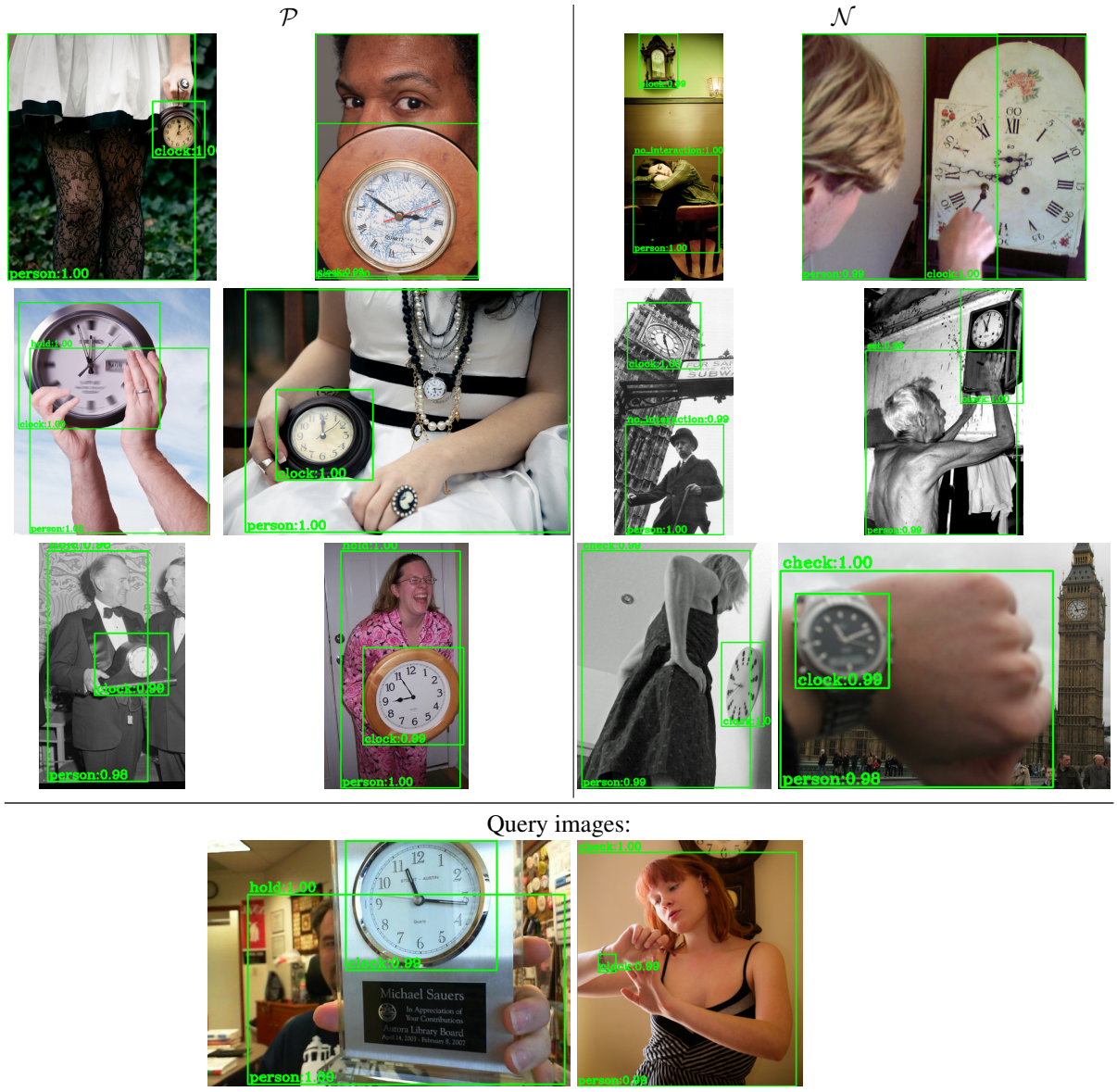


Figure 11. **Illustration of our oracle model.** The concept in $\mathcal{P}$ is `hold clock`.

(*i.e.*, positive examples $\mathcal{P}$ in our Bongard problem). In the right, images that do not contain the visual relationship are shown (*i.e.*, negative examples $\mathcal{N}$). In the bottom part, given a query image, a tester needs to decide whether it depicts the particular visual relationship or not. Each MTurk job contains two few-shot instances, where a tester can freely switch between two pages. They can only submit the job once both two tasks are finished.

We do not tell the testers what objects to focus on to induce the common visual relationship. It is intended to be similar to what a machine learning model does, which needs to do object detection first.

Simple examples given to testers. To ensure testers who see the form of few-shot instances for the first time can successfully finish the job, we provide some examples of different visual relationships and encourage them to take a look at these examples before starting working on a job. Such examples are shown in Fig. 14.

MTurk job setting. We provided more details about the job setting below.

- **Region.** We restrict the regions of testers to be in the US and Germany.
- **Approval rate.** Each MTurker tester maintains a job approval rate based on their performance on previous jobs. We invite only MTurk testers whose job approval rate is

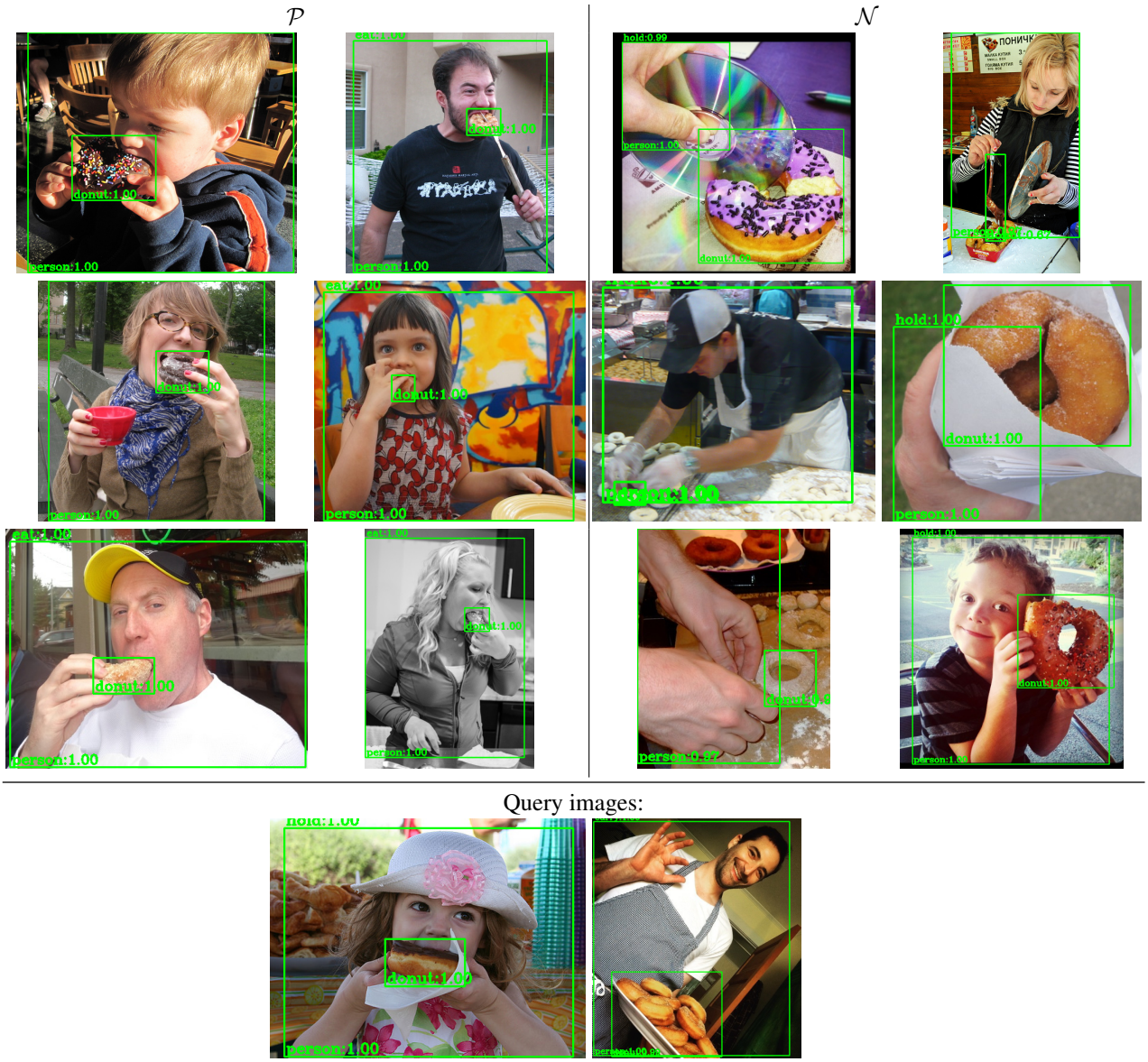


Figure 12. **A failure of our oracle model.** The concept in $\mathcal{P}$ is `eat cake`. The *HOITrans* model [54] incorrectly recognizes the first query image as `hold cake` (which should be `eat cake`). As a result, it makes a wrong prediction for the first query image.

equal to or greater than 98%.

- **Number of approved jobs.** Setting a qualification for the job approval rate only is not sufficient to hire high-quality testers since newly registered novel testers have a job approval rate of 100%. Therefore, we also set a qualification such that only testers who have more than 500 jobs approved previously are invited.
- **Invited annotators.** Through a couple of small-scale preliminary studies, we identified 35 reliable annotators on MTurk. For the large-scale data curation, we invited them to participate only.
- **Reward setting.** We provide \$0.15 for each job with an additional \$0.15 bonus if consistently high-quality annotations are made. According to our experiences of finishing the job, it roughly corresponds to about \$30 per hour.
- **Number of testers for each job.** We hire three independent testers for each job and aggregate their annotations. In specific, we only keep the few-shot instances where at least one of the three testers correctly classified the query image according to the ground-truth annotations. Otherwise, it suggests that a BP is either ambiguous or too difficult. We discard 2.5% of the few-shot instances that

we submitted to MTurk.

- **Job life time.** A job will not be available after 7 days if it is not claimed by any tester. But we found that all of the jobs were finished within such a limit.

Figure 13. **The user interface (UI) of MTurk data curation.**

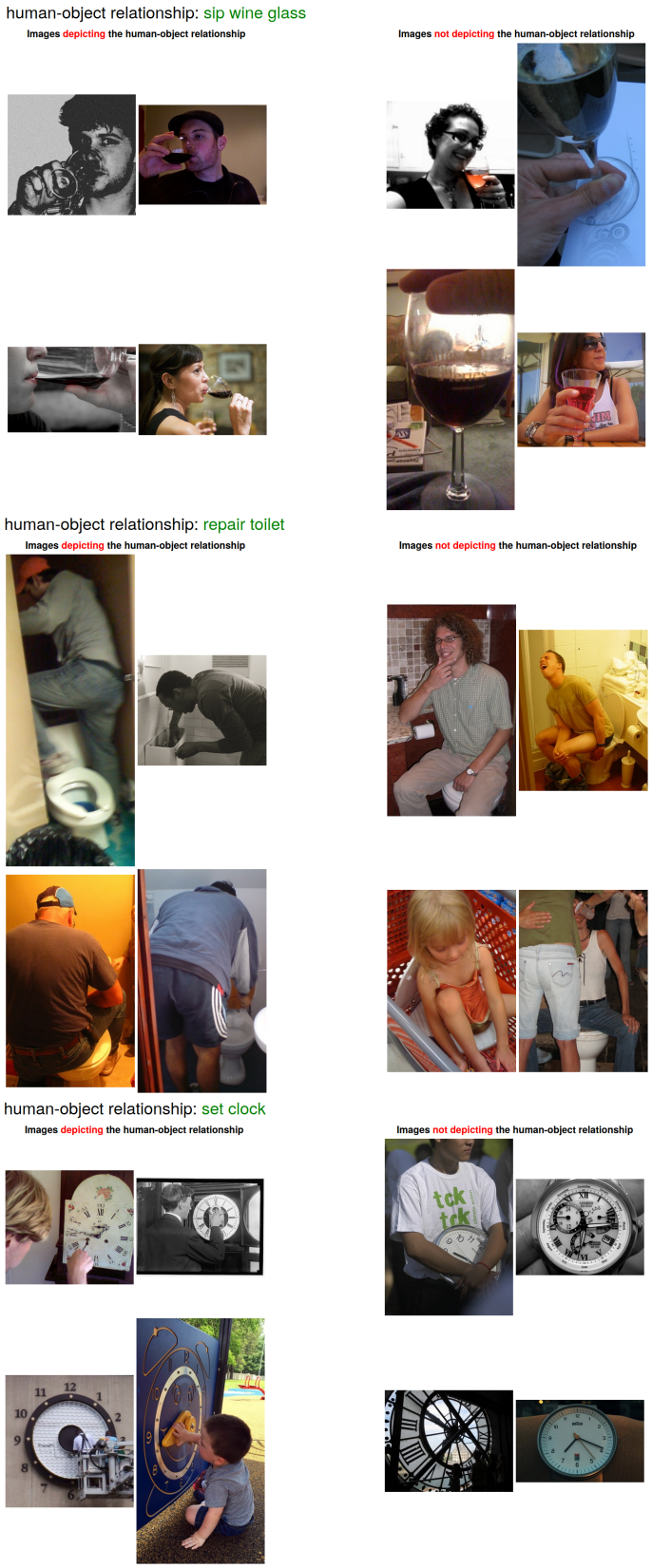


Figure 14. **Examples of different visual relationships given to MTurk testers.** For each example, we tell what the visual relationship is so that the testers can better understand the scope of the job.

References

[1] Anton Bakhtin, Laurens van der Maaten, Justin Johnson, Laura Gustafson, and Ross Girshick. Phyre: A new benchmark for physical reasoning. In *Advances in Neural Information Processing Systems*, pages 5083–5094, 2019. 8

[2] David Barrett, Felix Hill, Adam Santoro, Ari Morcos, and Timothy Lillicrap. Measuring abstract reasoning in neural networks. In *ICML*, pages 511–520, 2018. 2, 4, 5, 7, 8

[3] Mikhail Moiseevich Bongard. The recognition problem. Technical report, Foreign Technology Div Wright-Patterson AFB Ohio, 1968. 2, 8

[4] Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. HICO: A benchmark for recognizing human-object interactions in images. In *ICCV*, 2015. 3, 6, 8, 10

[5] Xinlei Chen, Haoqi Fan, Ross B. Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *CoRR*, abs/2003.04297, 2020. 6, 7

[6] Yinbo Chen, Xiaolong Wang, Zhuang Liu, Huijuan Xu, and Trevor Darrell. A new meta-baseline for few-shot learning. *arXiv preprint arXiv:2003.04390*, 2020. 3, 5, 7

[7] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019. 8

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009. 1, 6

[9] Li Fe-Fei et al. A bayesian approach to unsupervised one-shot learning of object categories. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 1134–1141. IEEE, 2003. 8

[10] Saurabh Gupta and Jitendra Malik. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474*, 2015. 3, 6, 8, 10

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *IEEE TPAMI*, 42(2):386–397, 2020. 10

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 5, 6

[13] Sepp Hochreiter, A Steven Younger, and Peter R Conwell. Learning to learn using gradient descent. In *International Conference on Artificial Neural Networks*, pages 87–94. Springer, 2001. 8

[14] Zhi Hou, Yu Baosheng, Yu Qiao, Xiaojiang Peng, and Dacheng Tao. Detecting human-object interaction via fabricated compositional learning. In *CVPR*, 2021. 8

[15] Zhi Hou, Xiaojiang Peng, Yu Qiao, and Dacheng Tao. Visual compositional learning for human-object interaction detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 584–600. Springer, 2020. 8

[16] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, 2017. 5, 8

[17] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–251. Springer, 2018. 8

[18] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015. 8

[19] Ranjay Krishna, Yuke Zhu andd Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017. 3, 8

[20] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020. 3, 8, 10

[21] Brenden M. Lake, Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. One shot learning of simple visual concepts. In Laura A. Carlson, Christoph Hölscher, and Thomas F. Shipley, editors, *CogSci*, 2011. 4

[22] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015. 1

[23] Kwonjoon Lee, Subhransu Maji, Avinash Ravichandran, and Stefano Soatto. Meta-learning with differentiable convex optimization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10657–10665, 2019. 5, 7, 8

[24] Yonglu Li, Liang Xu, Xijie Huang, Xinpeng Liu, Ze Ma, Mingyang Chen, Shiyi Wang, Haoshu Fang, and Cewu Lu. HAKE: human activity knowledge engine. *CoRR*, abs/1904.06539, 2019. 2, 3, 4, 6, 7, 8, 10

[25] Yong-Lu Li, Xinpeng Liu, Han Lu, Shiyi Wang, Junqi Liu, Jiefeng Li, and Cewu Lu. Detailed 2d-3d joint representation for human-object interaction. In *CVPR*, 2020. 2, 3, 4, 8

[26] Yue Liao, Si Liu, Fei Wang, Yanjie Chen, Chen Qian, and Jiashi Feng. PPDM: parallel point detection and matching for real-time human-object interaction detection. In *CVPR*, 2020. 3, 10

[27] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 3, 8

[28] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, December 2015. 3

[29] Cewu Lu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li. Visual relationship detection with language priors. In *ECCV*, 2016. 8

[30] Xiaojian Ma, Weili Nie, Zhiding Yu, Huaizu Jiang, Chaowei Xiao, Yuke Zhu, Song-Chun Zhu, and Anima Anandkumar. Relvit: Concept-guided vision transformer for visual relational reasoning. In *International Conference on Learning Representations*, 2022. 8

[31] Daniela Massiceti, Luisa Zintgraf, John Bronskill, Lida Theodorou, Matthew Tobias Harris, Ed Cutrell, Cecily Morrison, Katja Hofmann, and Simone Stumpf. Orbit: A real-world

few-shot dataset for teachable object recognition. In *ICCV*, 2021. 4

[32] Nikhil Mishra, Mostafa Rohaninejad, Xi Chen, and Pieter Abbeel. A simple neural attentive meta-learner. *ICLR*, 2018. 8

[33] Weili Nie, Zhiding Yu, Lei Mao, Ankit B. Patel, Yuke Zhu, and Anima Anandkumar. Bongard-logo: A new benchmark for human-level concept learning and reasoning. In *NeurIPS*, 2020. 2, 4, 5, 6, 7, 8

[34] Genevieve Patterson and James Hays. Coco attributes: Attributes for people, animals, and objects. *European Conference on Computer Vision*, 2016. 1, 6

[35] Aniruddh Raghu, Maithra Raghu, Samy Bengio, and Oriol Vinyals. Rapid learning or feature reuse? towards understanding the effectiveness of maml. *ICLR*, 2020. 5, 7, 8

[36] Francesco Ragusa, Antonino Furnari, Salvatore Livatino, and Giovanni Maria Farinella. The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In *WACV*, pages 1569–1578, 2021. 8

[37] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *ICLR*, 2018. 8

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2017. 5, 6

[39] Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. In *BMVC*, 2015. 8

[40] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. Meta-learning with memory-augmented neural networks. In *International conference on machine learning*, pages 1842–1850, 2016. 8

[41] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *NeurIPS*, 2017. 5

[42] David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*, 2019. 8

[43] Jake Snell, Kevin Swersky, and Richard S. Zemel. Prototypical Networks for Few-shot Learning. *Advances in Neural Information Processing Systems*, 2017. 5, 7, 8

[44] Damien Teney, Peng Wang, Jiewei Cao, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel. V-prom: A benchmark for visual reasoning using visual progressive matrices. In *AAAI*, 2020. 2, 4, 8

[45] Eleni Triantafillou, Tyler Zhu, Vincent Dumoulin, Pascal Lamblin, Utku Evci, Kelvin Xu, Ross Goroshin, Carles Gelada, Kevin Swersky, Pierre-Antoine Manzagol, et al. Meta-dataset: A dataset of datasets for learning to learn from few examples. In *ICLR*, 2020. 1, 4, 5

[46] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 6

[47] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. In *NeurIPS*, pages 3630–3638, 2016. 1, 4, 8

[48] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 3

[49] Erik Weitnauer and Helge Ritter. Physical bongard problems. In *Ifip international conference on artificial intelligence applications and innovations*, pages 157–163. Springer, 2012. 8

[50] Sirui Xie, Xiaojian Ma, Peiyu Yu, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Halma: Humanlike abstraction learning meets affordance in rapid problem solving. *arXiv preprint arXiv:2102.11344*, 2021. 8

[51] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. CLEVRER: collision events for video representation and reasoning. In *ICLR*, 2020. 8

[52] Chi Zhang, Feng Gao, Baoxiong Jia, Yixin Zhu, and Song-Chun Zhu. Raven: A dataset for relational and analogical visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5317–5327, 2019. 8

[53] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian D. Reid, and Anton van den Hengel. HCVRD: A benchmark for large-scale human-centered visual relationship detection. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *AAAI*, 2018. 3, 10

[54] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, et al. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 16

[55] Cheng Zou, Bohan Wang, Yue Hu, Junqi Liu, Qian Wu, Yu Zhao, Boxun Li, Chenguang Zhang, Chi Zhang, Yichen Wei, and Jian Sun. End-to-end human object interaction detection with hoi transformer. In *CVPR*, 2021. 2, 6, 7, 8