**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\(^3,4\), 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, anima@caltech.edu

github.com/nvlabs/Bongard-HOI

**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.

**1. Introduction**

In recent years, great strides have been made on visual recognition benchmarks, such as ImageNet [8] and COCO [33]. 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 [21]. 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 [46], Meta-Dataset [44], and ORBIT [44] have been dedicated to studying few-shot visual learning, they focus on recog-
We construct the few-shot instances in Bongard-HOI on top of the HAKE dataset [23, 24]. 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, 32, 43]. Comparisons with different benchmarks can be found in Table 1.

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.

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 [52] 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 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 Bongard-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 Bongard-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 [47], facial attributes [27]. 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 [18], 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 [23, 24] 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], OpenImages [19], HCVRD [51], and PIC [25]. HAKE has 80 object categories, which are consistent with the vocabulary defined in the standard COCO dataset [26]. It also has 117 action labels, leading to 600 human-object relationships.\footnote{Some combinations of objects and actions are infeasible.}

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 person. We start with selecting a set of positive images...
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 [43]).

<table>
<thead>
<tr>
<th>concept</th>
<th>compositional concept</th>
<th>natural image</th>
<th>few-shot</th>
<th>ctx-dependent reasoning</th>
<th>hard negatives</th>
<th>generalization types</th>
<th>#concepts</th>
<th>#tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>HAKE [23, 24]</td>
<td>HOI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>600</td>
</tr>
<tr>
<td>Omniglot [20]</td>
<td>shape</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>50</td>
</tr>
<tr>
<td>miniImageNet [46]</td>
<td>image label</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>100</td>
</tr>
<tr>
<td>Meta-Dataset [44]</td>
<td>image label</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>4,934</td>
</tr>
<tr>
<td>ORBIT [30]</td>
<td>frame label</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>486</td>
</tr>
<tr>
<td>RPM [2]</td>
<td>shape</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>50</td>
</tr>
<tr>
<td>V-PROM [43]</td>
<td>attributes &amp; counts</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>478</td>
</tr>
<tr>
<td>Bongard-LOGO [32]</td>
<td>shape</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>627</td>
</tr>
<tr>
<td>Bongard-HOI (ours)</td>
<td>HOI</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>242</td>
</tr>
</tbody>
</table>

\( \mathcal{I}_c = \{ I_1, \ldots \} \) 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 \( \mathcal{I}_c \) with concept \( \bar{c} = (s, a, o) \), 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 \( \mathcal{P}, \mathcal{N}_a \), and a query image \( I_q \). Two parameters control the sampling process: \( M \), the number of images in \( \mathcal{P} \) and \( \mathcal{N} \) \((M = \bar{M} \in \text{Bongard-HOI}) \), and the overlap threshold \( \tau \), 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 [23, 24] 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-standing 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 \((s, a, o)\) 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 [40], 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 \( \mathcal{S} = \mathcal{P} \cup \mathcal{N} = \{(I^p_1, 1), \ldots, (I^p_M, 1), (I^n_1, 0), \ldots, (I^n_M, 0)\} \) 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 Bongard-HOI, we set \( M = 6 \) as our default parameter and therefore each few-shot instance is "2-way, 6-shot". Following [44], 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, 32].

**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 [42], a metric-based method; 2) MetaOpt-Net [22] and ANIL [34], two optimization-based approaches. Moreover, we also use a strong baseline meta-learning model, Meta-Baseline [5], which reports competitive results in many few-shot prediction tasks. We refer readers to the related papers for more details.

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 compositional 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 [40], which shows promising compositional reasoning accuracy on a Visual Question Answering (VQA) benchmark [15], to explicitly encode the compositionality of visual relationships. In specific, the feature representations of the image \( I \) is computed as

\[
\text{RN}(I) = f_{\theta} \circ \sum_{i,j} g_{\psi} \left( \text{concat}(h_{\theta}(o_i, I), h_{\theta}(o_j, I)) \right),
\]

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 [37]. \( h_{\theta} \) denotes the RoI Pooled features of \( o_i \) from a ResNet backbone [11] followed by a MLP (multi-layer perceptron) [37], which is parameterized by \( \theta \). \( g_{\psi} \) and \( f_{\theta} \) 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 [52], which is based on the Transformer model [45] 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 $$c_P = \text{majority\_vote}(\mathcal{D}^P - \mathcal{D}^N),$$ where $\text{\{-\}}$ 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., $\text{majority\_vote}$ 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 [11] 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 [37] class-agnostic objectness detection model on the COCO dataset [33] using a ResNet101 [11] pre-trained on ImageNet [8] as the backbone. We use the RoIPool operation [37] to get feature representations for each bounding box. We also use ground-truth bounding boxes provided in HAKE [23] as input to diagnose the effectiveness of the visual perception. In addition to RoIPooled region features, we also concatenate each bounding box’s normalized co-ordinates (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 echoes the findings in Bongard-LOGO [32], 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 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.
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.

<table>
<thead>
<tr>
<th></th>
<th>bbox</th>
<th>pre-train</th>
<th>test set</th>
<th>avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>seen act., seen obj.</td>
<td>seen act., unseen obj.</td>
<td>unseen act., seen obj.</td>
<td>unseen act., unseen obj.</td>
</tr>
<tr>
<td>CNN-Baseline [32]</td>
<td>-</td>
<td>scratch</td>
<td>50.03</td>
<td>49.89</td>
</tr>
<tr>
<td>WReN-BP [2, 32]</td>
<td>-</td>
<td>IN</td>
<td>50.31</td>
<td>49.72</td>
</tr>
<tr>
<td>ProtoNet* [42]</td>
<td>det</td>
<td>IN</td>
<td>58.90</td>
<td>58.77</td>
</tr>
<tr>
<td>MetaOptNet# [22]</td>
<td>det</td>
<td>IN</td>
<td>58.60</td>
<td>58.28</td>
</tr>
<tr>
<td>ANIL [34]</td>
<td>det</td>
<td>IN</td>
<td>50.18</td>
<td>50.13</td>
</tr>
<tr>
<td>Meta-Baseline [6]</td>
<td>det</td>
<td>scratch</td>
<td>54.61</td>
<td>53.79</td>
</tr>
<tr>
<td>Meta-Baseline [6]</td>
<td>det</td>
<td>MoCoV2</td>
<td>55.23</td>
<td>54.54</td>
</tr>
<tr>
<td>Meta-Baseline [6]</td>
<td>det</td>
<td>IN</td>
<td>56.45</td>
<td>56.02</td>
</tr>
<tr>
<td>Meta-Baseline [6]</td>
<td>gt</td>
<td>IN</td>
<td>58.82</td>
<td>58.75</td>
</tr>
<tr>
<td>HOITrans [52] (oracle)</td>
<td>-</td>
<td>-</td>
<td>59.50</td>
<td>64.38</td>
</tr>
<tr>
<td>Human (Amateur)</td>
<td>-</td>
<td>-</td>
<td>87.21</td>
<td>90.01</td>
</tr>
</tbody>
</table>

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 [23] 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 [32] 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 [42], have difficulties inducing the true visual relationships. For MetaOptNet [22], 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 [18], scene graph annotations, including relationships of different objects, are provided. A subset of the annotations is used in VRD [28] to focus on visual relationship detection. In a recent effort, large-scale visual relationships are provided in the Open Images dataset [19]. HOI, is of particular interest to understand the interactions of humans and other objects. A lot of HOI benchmarks, such as HICO [4], COCO [38], vCOCO [10], and HOI-COCO [13], are built on top of the object categories provided in the COCO dataset [26]. The MECCANO [35] dataset focuses on human-object interactions in egocentric settings and industrial scenarios. Ambiguous-HOI [24] is part of the HAKE project [23], 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 [23], 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 [52].

Few-shot and meta learning models. Few-shot learning aims at learning from a limited number of training samples [9, 17]. 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) [12] 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 [39] and SNAIL [31], 2) metric-based methods, such as Matching Networks [46] and ProtoNet [42], and 3) optimization-based methods, such as MetaOptNet [22] and ANIL [34].

These meta-learning methods have been evaluated on several commonly used few-shot learning benchmarks, including miniImageNet [46] and tieredImageNet [36]. 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 [14, 16]. 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 [15, 29], physical reasoning [1, 50], math problems [41], and general artificial intelligence [7, 49]. The most relevant to our benchmark are RPM [2], its variant with natural images [43], and Bongard problems with synthetic shapes [32] and physical problems [48]. While most of them consider synthetic images [2, 32, 48], 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 [43]. 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 [23], 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, imposing 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.
References


[23] 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


