SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition

Kaiyu Yang
Princeton University
kaiyuy@cs.princeton.edu

Olga Russakovsky
Princeton University
olgarus@cs.princeton.edu

Jia Deng
Princeton University
jiadeng@cs.princeton.edu

Abstract

Understanding the spatial relations between objects in images is a surprisingly challenging task (Fig. 1). A chair may be "behind" a person even if it appears to the left of the person in the image (depending on which way the person is facing). Two students that appear close to each other in the image may not in fact be "next to" each other if there is a third student between them.

We introduce SpatialSense, a dataset specializing in spatial relation recognition which captures a broad spectrum of such challenges, allowing for proper benchmarking of computer vision techniques. SpatialSense is constructed through adversarial crowdsourcing, in which human annotators are tasked with finding spatial relations that are difficult to predict using simple cues such as 2D spatial configuration or language priors. Adversarial crowdsourcing significantly reduces dataset bias and samples more interesting relations in the long tail compared to existing datasets. On SpatialSense, state-of-the-art recognition models perform comparably to simple baselines, suggesting that they rely on straightforward cues instead of fully reasoning about this complex task. The SpatialSense benchmark provides a path forward to advancing the spatial reasoning capabilities of computer vision systems. The dataset and code are available at `https://github.com/princeton-vl/SpatialSense`.

1. Introduction

Visual understanding of space is essential for an intelligent agent. Such an understanding is the basis for describing scenes and referring to objects [7]; it is also the foundation required for tasks such as navigation and manipulation [32]. To understand space it is important to understand *spatial relations*, that is, how different spatial entities are configured relative to each other to compose a scene. Consider the following description: "Inside the living room, under the window next to the wall is a table, on top of which lies a vase with flowers". The sentence may be structured in many different ways, but its meaning is determined by the objects (room, window, table, vase, flowers) and their spatial relations (table in room, table under window, table next to wall, vase on table, flowers in vase).

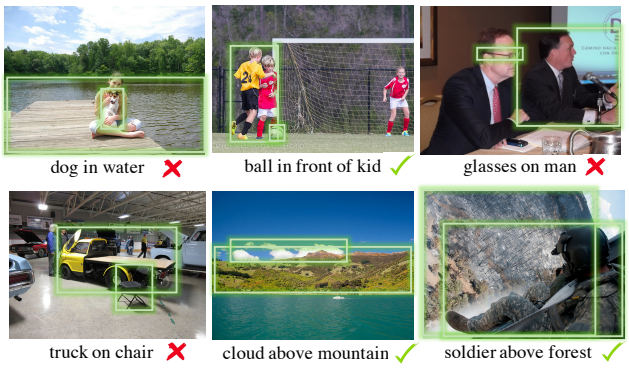

Figure 1: Spatial relation recognition in images is a challenging task which requires a deep understanding of all the objects in the image, their 3D configuration, and their interactions. Understanding that the dog is *not* in the water (top left) requires reasoning about the pier in addition to the dog and the water. Understanding that the ball is in fact in front of the kid despite being off to the right in the image space (top center) requires inferring the 3D spatial configuration. Properly evaluating computer vision abilities on this task is difficult because it requires collecting a benchmark covering the full spectrum of such challenges. We introduce SpatialSense, a novel dataset collected through *adversarial crowdsourcing*, which provides a diverse and challenging testbed for the task of spatial relation recognition.

This raises the problem of spatial relation recognition: given two objects in a scene, what is their spatial relation? This problem is important, interesting, and challenging because the semantics of spatial relations are rich and complex. The spatial semantics between objects depend not only on geometric properties such as location, pose, and shape, but also the frame of reference (*e.g.* "left of the car" can be relative to the observer or the car) and object-specific common sense knowledge (*e.g.* "hand over bed" does not

imply physical contact while "blanket over bed" does).

Benchmarking spatial relations. Despite the importance of this problem, there is no benchmark dataset specializing in spatial relations. Large datasets such as Visual Genome [11] and Visual Relationship Detection [19] provide annotations for generic visual relations and thus include a significant number of spatial relations (51.5% and 66.0% respectively). But several characteristics make them less suitable for evaluating spatial relation recognition.

One issue is that Visual Genome and the recent Open Images [12], for example, do not provide negative examples in its annotations. Each image in Visual Genome comes with a set of human annotated relations (*subject-predicate-object* triplets such as "person-riding-bike"), but the annotations are not exhaustive and many valid relations are left unannotated. While the Recall@K metric (evaluating the recall of ground truth relations given K predicted relations) has been used as a proxy to avoid exhaustive annotation [19, 33, 14, 17, 5, 34, 22, 35, 31, 37, 29, 30], it remains impossible to distinguish between a good system producing valid albeit unannotated predictions and a bad system producing false positives, or to fully evaluate the system's ability to distinguish between positive and negative relations.

Another issue in current visual relation datasets is significant language bias—the examples are dominated by relations that can be guessed without an actual spatial understanding. For example, 66% of all spatial relations in Visual Genome consist of one object "on" another. Among relations involving a table, 89.37% of them define an object "on" the table. This means that a system can take advantage of such priors to do well without even looking at the image. This is undesirable because evaluating on these examples will not provide a proper gauge of an algorithm's ability to visually understand spatial relations.

Contributions. In this paper we introduce *SpatialSense*, a dataset specializing in spatial relation recognition. A key feature of the dataset is that it is constructed through *adversarial crowdsourcing*: a human annotator is asked to come up with adversarial examples to confuse a robot. This yields two desirable properties. First, the dataset includes negative examples as some human annotators are explicitly tasked with choosing spatial relations that are false. Second, the dataset focuses on questions that require more advanced reasoning and cannot be answered by simple spatial and language priors.

SpatialSense has 17,498 relations on 11,569 images. Given two objects (names and bounding boxes), the task is to classify whether a particular spatial relation holds. We provide the object names and localizations to decouple object detection from spatial relation recognition, such that a successful relation recognition system can be directly placed on top of any object detection system. The dataset contains relations between 3,679 unique object classes, with 2,139 of these object classes appearing only once, providing a challenging long-tail distribution of concepts.

SpatialSense provides a rigorous testbed for spatial relation reasoning that is not easily amenable to simple priors. First, each predicate ("on", "under", etc.) has an equal number of positive and negative relations. Second, simple baselines using only 2D or language cues are significantly less effective on SpatialSense than on other existing spatial reasoning benchmarks. SpatialSense is complementary to large-scale datasets such as Visual Genome or Open Images, in that it enables testing spatial relation recognition models with challenging examples in the long tail.

We evaluate multiple state-of-the-art visual relationship detection models on SpatialSense. Experimental results reveal that these models rely too much on dataset bias and now perform comparably to simple baselines. This demonstrates that adversarial crowdsourcing is effective for reducing dataset bias, and showcases that SpatialSense is an important step towards improving spatial reasoning capabilities of computer vision systems.

2. Related Work

Visual relationship recognition. Recognition of visual relations has recently emerged as a frontier of high-level computer vision moving beyond object recognition. Sadeghi & Farhadi [23] studied detecting *visual phrases* from images. A visual phrase can be a spatial relation (*e.g*. "person next to bicycle"). But their dataset contains only 17 unique visual phrases, 9 of which are spatial relations. This means that each spatial predicate only occurs with a small number of object categories: *e.g*. "next to" only occurs with "person","car", and "bicycle". Thus the dataset is unsuitable for evaluating a general understanding of "next to" that is agnostic to object categories.

Lu *et al*. [19] introduced the task of *visual relationship detection* —given an image, the algorithm predicts *subject-predicate-object* triplets as well as the object bounding boxes. In contrast, our task is classification rather than detection, the object pairs are given, and there are both positive and negative relations. Our task setup leads to a proper evaluation of relation understanding.

The VRD dataset introduced by Lu *et al*. [19] includes spatial relations, but unlike our dataset, does not have negative examples. Thus evaluation using VRD has been based on Recall@K, which is ill-suited for spatial relations because numerous valid spatial relations can hold in an image, making it difficult to choose the appropriate K. In addition, as we will show in Section 4, the spatial relations in VRD are significantly easier to predict using simple priors. Visual Genome [11] is another dataset that is larger and has extensive annotations of visual relations. Similar to VRD, it covers a significant number of spatial relations but has no

negative examples, and the spatial relations are easily predictable from simple priors.

VRD and Visual Genome have spurred the development of new approaches for visual relationship detection. Successful methods typically build on top of an object detection module, and reason jointly over language and visual features [14, 33, 17, 5, 37, 31]. Multiple independent directions have been proven fruitful, including learning features that are agnostic to object categories [29], facilitating the interaction between object features and predicate features [29, 5, 14], overcoming the scarcity of labeled data through weakly supervised learning [22, 34], and detecting the relationships among multiple objects jointly as scene graphs [28, 16, 27, 15]. We benchmark some of the state-of-the-art approaches on our dataset and compare them with simple baselines based on language or 2D cues.

Peyre *et al*. introduced a dataset of unusual relations (UnRel) [22] sharing similar motivation to ours. To address the problem of missing annotations, the relations in UnRel are annotated exhaustively. Annotating every instance is made feasible by having a small predefined list of relations that are carefully designed to be unusual (such as "car under elephant"). This method is not scalable to a large number of relations. First, it is difficult to manually pick a large set of unusual relations. Second, for annotating every single instance, the amount of crowdsourcing efforts grows linearly with the number of relations. Our method samples interesting relations by encouraging crowd workers to discover them in images, and thus circumvents the scalability problem. As a result, UnRel has 76 unique relation triplets formed by 18 predicates while SpatialSense has 13,229 unique relations with 9 predicates.

Dataset bias. Instead of studying the misaligned distributions between data and the visual world [25], we focus on a specific aspect of dataset bias, which allows models to take shortcuts leading to a superficial impression of good performance. Extensive research on such bias has been conducted in the context of visual question answering (VQA) [36, 6, 3, 9, 1, 20]. Many VQA datasets suffer from language bias; the questions can be answered well simply by using language priors while ignoring images. Zhang *et al*. [36] balance the data for yes/no questions on abstract scenes. They show a question-image pair and ask the annotator to compose a new scene on which the answer for the question is different. Goyal *et al*. [6] applied the same idea to real images. Instead of asking annotators to create new scenes, they provide a few semantically similar images for the annotator to choose from.

Our adversarial crowdsourcing approach addresses the same issue (ensuring the input image is required to answer questions) in a substantially different way. Spatial relation recognition can be understood as a special case of VQA, where the questions are restricted to verifying spatial relations. In this sense, Zhang *et al*. [36] and Goyal *et al*. [6] ask the crowd to select hard images—images that defy the expected answers from language priors—with the questions fixed, whereas we ask the crowd to select hard questions with the images fixed. One potential advantage of selecting hard questions is that humans can easily compose new questions but cannot easily synthesize photorealistic images, and it can also be hard to find images that defy language priors—those images are by definition less common because language priors reflect common occurrences.

Adversarial crowdsourcing. Our approach for adversarial crowdsourcing is inspired by the "Beat the Machine" framework by Attenberg *et al*. [2], in which a crowd worker is challenged to find cases that will cause an AI system to fail. Adversarial crowdsourcing is related to active learning (*e.g*. [10, 26]) in that in both cases we seek difficult examples to improve learning. The key difference, however, is that in active learning it is the machine's task to identify hard examples whereas in adversarial crowdsourcing it is on the human annotator.

3. Dataset Collection through Adversarial Crowdsourcing

Datasets are meant to evaluate the performance of algorithms under challenging and varied conditions. However, one weakness observed in many datasets is a strong language bias, allowing algorithms to perform well by exploiting language priors even while ignoring the visual input [36, 6, 3, 9, 1]. Further, in the context of spatial reasoning, algorithms may exploit simple 2D cues, circumventing a true 3D understanding of spacing [8]. We address both issues in our adversarial crowdsourcing framework.

Adversarial crowdsourcing protocol. In our data collection pipeline (Fig. 2), we ask annotators to propose spatial relations to make a robot fail. Given an image and a request for a positive or negative spatial relationship, the annotator comes up with a proposed example by clicking on two objects, entering their names and selecting a spatial predicate corresponding to the relationship between them (a true relationship if the request was for a positive one, and a false relationship otherwise). The robot then tries to guess whether the relationship is positive or negative using only the object names and the 2D coordinates given by the clicks. The task is completed if the robot is wrong, i.e., it predicts that the relationship is positive but in fact it was negative, or vice versa. Otherwise, if the robot is able to guess correctly, the robot provides feedback to the annotator about how the correct guess was made, and the annotator tries again. Additional crowdsourcing is used to verify the collected relations and annotate the object bounding boxes.

To reduce language bias and promote true 3D spatial understanding, we need the annotators to pick relations

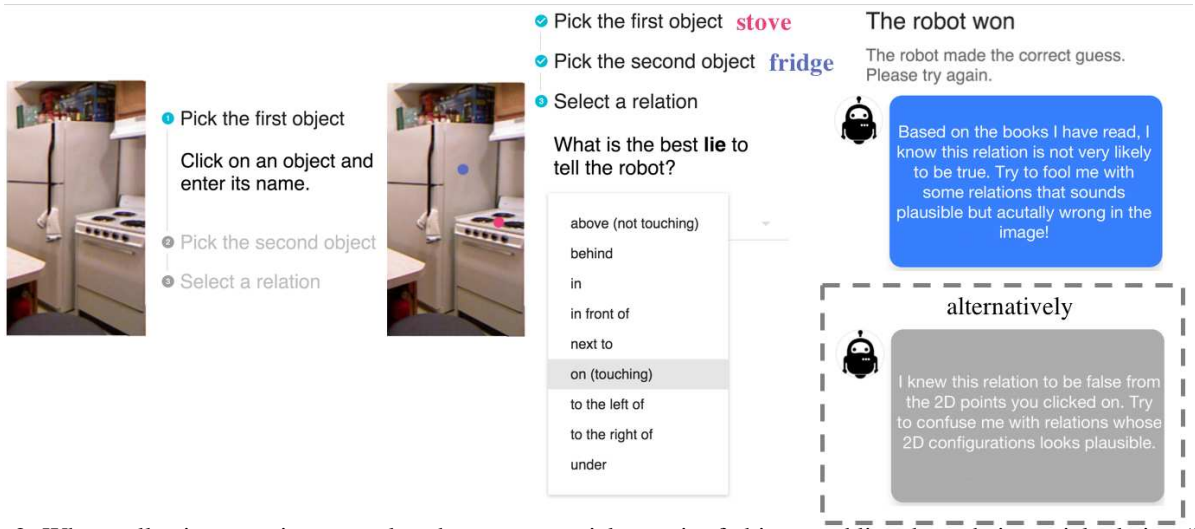

Figure 2: When collecting negative examples, the annotator picks a pair of objects and lies about their spatial relation ("stove on fridge") with the goal of making it believable enough for the robot. However, here the robot catches the annotator in the lie and explains how the correct guess was made (from language or 2D cues, or both). The UI for positive examples is similar, in which the annotator tells the robot a spatial relation that is valid but unbelievable.

that are difficult to predict given object names and 2D cues. The robot is therefore an ensemble of two models: a language-only model and a 2D-only model. The language-only model takes two object names along with the predicate, and outputs the probability that the relation holds. The object names are converted to word embeddings using Word2Vec [21], which are then encoded into a fixed-length feature vector by a gated recurrent unit (GRU) [4]. The one hot encoding of the predicate is mapped to a vector of the same size by a linear layer. The three feature vectors are fused by element-wise addition, on top of which a 2-layer fully connected network outputs the probability. For the 2D-only model, linear layers map the object coordinates to feature vectors, and the prediction is made following the same procedure of the language-only model. The final output of the robot is the average of these two models. Initially, we trained the robot on a dataset of 7,850 relations collected without adversarial crowdsourcing; during the collection of SpatialSense, we occasionally re-trained the robot using all currently available data in order to prevent the annotators from exploiting the failing modes of a particular robot.

Concept vocabulary. We restrict the spatial predicate to a predefined list (`above`, `behind`, `in`, `in front of`, `next to`, `on`, `to the left of`, `to the right of`, `under`) instead of letting human annotators enter free-form text. Spatial relations can be encoded using a surprisingly small set of prepositions [13]. Our list of 9 predicates covers the coarse-grained semantics of most spatial relations. Although it is possible to extend the vocabulary to represent more fine-grained spatial semantics (such as `lean on` and `sit on`), fewer predicates ensures sufficient training samples for each predicate. Nevertheless, our adversarial crowdsourcing method is scalable to more predicates as the effort grows linearly w.r.t. the number of relations, but is independent of the number of predicates.

In contrast, for object names we allow the human annotator to enter free-form text. The vocabulary of objects is vast and it would be cumbersome for the human annotator to choose from a long list of objects. In addition, a limited vocabulary of objects would restrict the annotator's ability to pick the objects that form interesting spatial relations. In general, an image is rarely completely predictable, so there will be an unusual or surprising spatial relation that beats the robot's simple intuition. This setup thus provides an efficient way of obtaining spatial relations in the long tail.

Image collection. We annotated $11,569$ images in total. Of these, 10,180 are RGB images from Flickr and $1,389$ are RGB-D images from NYU Depth [24] which we include to make it possible to test the utility of depth information for spatial understanding. When querying Flickr images, we use combinations of two keywords rather than a single keyword, following the approach adopted by COCO [18] to obtain images with diverse objects. Additionally, annotators can pick an image to annotate from a set of 8 images, so as to avoid images that do not have enough objects (*e.g.* a close up shot of a single foreground object). These techniques ensure that the images are complex scenes containing multiple objects necessary for relation reasoning.

We annotated $13,156$ relations on Flickr images and $4,342$ on NYU images. Each relation consists of a spatial predicate, the names of two objects, and their bounding boxes. Importantly, there is an equal number of positive *and negative* relations for each of the 9 predicate. 20% of the relations are reserved for testing and 15% for validation.

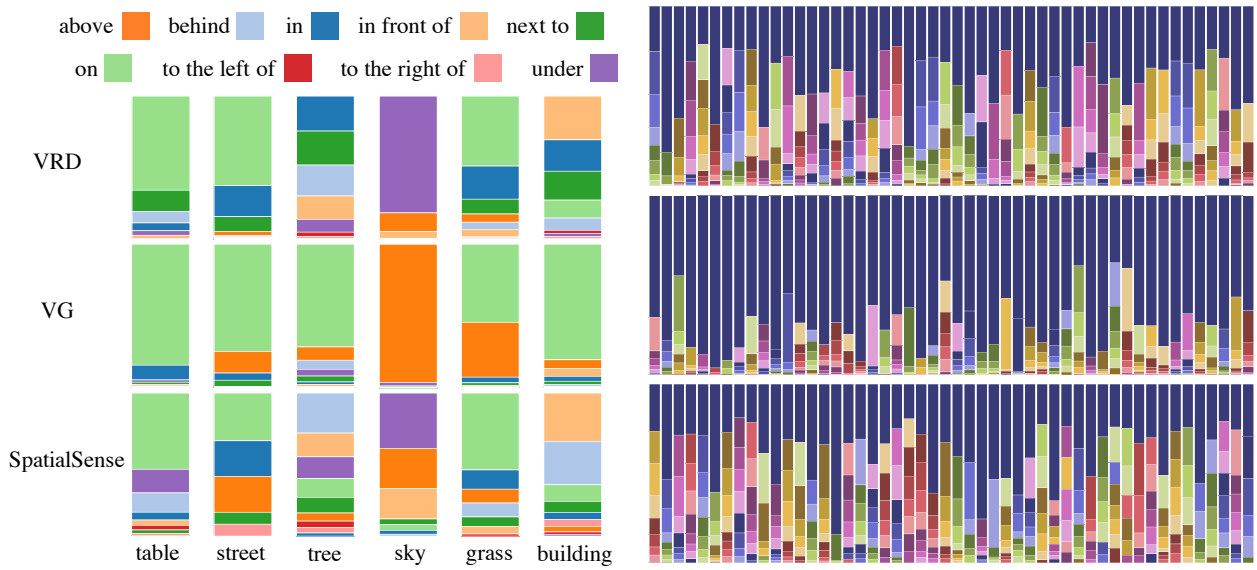

Figure 3: (Left) The predicate distributions of frequent objects in VRD-Spatial, VG-Spatial and SpatialSense-Positive. For example, the bottom-left bar shows the frequency distribution of predicates `on`, `under`, `behind`, etc. for the object "table" in SpatialSense. SpatialSense contains less language bias than other datasets since the distribution is more balanced. (Right) The predicate distributions of the top-50 objects in the three datasets, further showing the wider distribution in SpatialSense.

4. Analysis of the Dataset

The SpatialSense dataset has two key advantages compared to existing benchmarks. First, it contains positive as well as negative relations. Second, it is constructed to be challenging; due to adversarial crowdsourcing, simple language and 2D priors are not enough to do well on this dataset. We now perform an in-depth analysis, comparing SpatialSense to VRD [19], Visual Genome [11] and a version of itself without adversarial crowdsourcing.

4.1. Comparison to Existing Datasets

Setup. Since VRD and Visual Genome contain generic relations and no negative examples, we preprocess the data to allow for fair comparison: (1) Only the positive examples in SpatialSense are considered; the resulting dataset is referred to as *SpatialSense-Positive*; (2) We filter out non-spatial relations in VRD and Visual Genome; the resulting datasets are referred to as *VRD-Spatial* and *VG-Spatial*. In addition to discarding non-spatial relations, we also map the predicates in VRD and Visual Genome to their equivalents in our list of 9 spatial predicates. For example, `rest on`, `park on` and `lying on` in VRD are all mapped to `on`. For Visual Genome, since there is no closed vocabulary, we examined the top-100 most frequent predicates to figure out the mapping (supplementary material).

Predicate distribution. Compared to VRD and VG, the predicate distribution in SpatialSense is less biased. Fig. 3 visualizes the distribution of predicates corresponding to different objects in the three datasets. For VG-Spatial, objects are frequently dominated by a single predicate, such as something `on` table or something `on` street. VRD-Spatial, which was annotated in-house rather than via crowdsourcing, looks more balanced; there are nevertheless a large number of objects `on` street and or `under` sky. This confirms that many spatial relations in VRD and VG can be predicted without even looking at the image. In contrast, SpatialSense-Positive has a more balanced distributions, which reduces language bias, making it more difficult to guess the relation from the object names alone. There are plenty of unexpected or difficult to predict relations in any scene, and the key of our adversarial crowdsourcing approach is to encourage the annotators to reveal these.

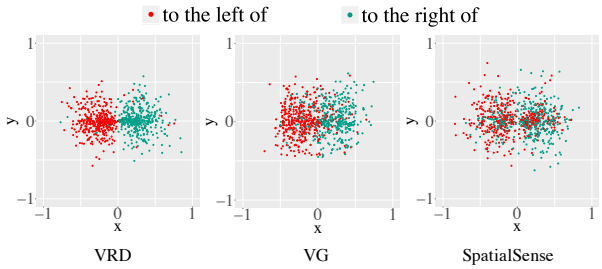

Figure 4: The 2D locations of subjects relative to objects for the predicates `to the left/right of`, normalized by image size. SpatialSense is less biased in 2D cues since the points are less separable. Each figure contains 400 randomly sampled relations for each predicate.

2D Spatial distribution. SpatialSense is also less biased in 2D cues. As an example, we evaluate whether the

Train \ Test	VRD	VG	SpatialSense	average drop
VRD	**66.9 / 59.6**	45.2 / 53.2	30.6 / 39.9	29.0 / 13.1
VG	50.8 / 38.2	**76.0 / 65.3**	36.1 / 33.8	32.6 / 29.3
SpatialSense	40.3 / 44.6	42.8 / 52.5	**39.8 / 43.4**	**-1.8 / -5.2**

Table 1: Accuracies of the language-only model / 2D-only model for predicting the spatial predicate from an object pair. Note two observations: (1) These simple baselines achieve low accuracy on SpatialSense compared to other datasets, demonstrating that SpatialSense is less susceptible to simple cues and requires more advanced reasoning. (2) SpatialSense is less biased than other datasets as evidenced by better cross-dataset generalization [25].

predicates `to the left of` and `to the right of` can be predicted from spatial cues alone, without relying on pixel-level information. Concretely, for each subject-predicate-object relationship, let (x_s, y_s) be the center of the subject bounding box, (x_o, y_o) be the center of the object bounding box, and w, h be the image width and height. We compute the normalized relative location between the object and subject as $((x_s - x_o)/w, (y_s - y_o)/h)$. These are 2D points within $[-1, 1] \times [-1, 1]$ representing the 2D location of the subject relative to object. For VRD-Spatial and VG-Spatial, an algorithm can easily distinguish between the predicates `to the left of` and `to the right of` based on these 2D locations alone, as shown in Fig. 4. For SpatialSense-Positive, however, these two predicates have similar distribution, making it difficult to tell them apart from 2D cues alone. Under our adversarial crowdsourcing framework, the annotators make extensive use of relative frame of references; a person standing to the left of a car can actually be on the right side of an image (when the car is facing towards the camera). Among other things, this relative frame of references makes SpatialSense less predictable from 2D cues and necessitates deeper spatial understanding.

Language and 2D baselines. To quantitatively show that SpatialSense is less biased in language and 2D locations, we examine the extent to which predicates can be determined from these cues alone, without pixel-level image information. For each relation, given the two object names (e.g., bed, floor) and their bounding boxes, a model has to predict the correct predicate (e.g., `on`). We train and evaluate on the three datasets and compare the accuracies. For the comparison to be fair, VRD-Spatial and VG-Spatial are randomly sampled to have the same size as SpatialSense-Positive. We adopt the official train/test split for VRD-Spatial and SpatialSense-Positive; for VG-Spatial, 20% of the images are reserved for testing.

The model architectures are similar to those used in data collection: For the language-only model, object names are encoded to fixed-length vectors using Word2Vec followed by a GRU. The two vectors are fused into one by element-wise addition, which is then classified by a 2-layer fully connected network. For the 2D-only model, bounding box coordinates are encoded by linear layers, and then fused and classified following the same procedure. The models are trained with cross-entropy loss.

The results verify our intuition that SpatialSense is much more difficult to tackle using simple language and 2D cues than prior datasets (Table 1). Concretely, the language-only model achieves strong predicate prediction accuracy of 66.9% on VRD-Spatial and 76.0% on VG-Spatial, significantly higher than 39.8% on SpatialSense-Positive. Similarly, the 2D-only baseline is able to achieve impressive accuracies of 59.6% on VRD-Spatial and 65.3% on VG-Spatial without ever seeing pixel information; in contrast, this simple model again struggles on SpatialSense-Positive and yields only 43.4% accuracy.

Cross-dataset generalization. Finally, we evaluate the generality of our collected dataset using the method proposed by Torralba and Efros [25]. Models trained on one dataset are evaluated on other datasets, and the resulting drop in accuracy is used as a metric for dataset bias (more drop corresponds to more bias). Table 1 shows the results of cross-dataset predicate classification using the language-only and 2D-only model described above. Models trained on SpatialSense generalize impressively well to other datasets. The language-only model trained on SpatialSense but evaluated on VRD or VG (instead of on SpatialSense) achieves a 1.8% average *increase* in accuracy; similarly, the 2D model achieves a 5.2% average *increase* as well. In contrast, models trained on VRD or VG are not able to generalize well and experience a average 26.0% *drop* in accuracy when evaluated on a different dataset.

4.2. The Effect of Adversarial Crowdsourcing

The bias reduction demonstrated in Section 4.1 is a result of adversarial crowdsourcing. To verify, we compare with a dataset constructed without adversarial crowdsourcing. The dataset is collected by annotators who propose positive spatial relations freely (without the need to beat a robot), and the negative relations are randomly generated and verified by humans. The resulting dataset is SpatialNaive and contains 3,015 images with 3,925 positive and 3,925 negative relations. Just like in SpatialSense, each predicate has an equal number of positive and negative relations.

We quantify the amount of bias in SpatialSense and SpatialNaive by comparing the performance on the task of spatial relation recognition. Given two object names, their bounding boxes and a predicate, a model classifies whether or not the relation holds. Since SpatialNaive is smaller, we randomly sample a subset of SpatialSense, enforcing the two datasets to have exactly the same number of relations for each predicate. The model architectures are the same as those used for collecting data (described in section 3),

Dataset	Language	2D Locations
SpatialNaive	69.2	71.3
SpatialSense	**56.4**	**65.2**

Table 2: SpatialSense, constructed with adversarial crowdsourcing, is significantly more challenging (lower accuracy of baselines) than the ablation dataset SpatialNaive.

except that now the 2D locations are represented by object bounding boxes rather than coordinates. 20% of the images in SpatialNaive are used for testing and another 15% of the images are for hyperparameters tuning.

The results are in Table 2. The models performs much worse on SpatialSense than SpatialNaive: 12.8% accuracy drop for the language-only model, 6.1% for the 2D-only, confirming the effectiveness of adversarial crowdsourcing for reducing dataset bias, especially the language bias.

5. Baselines for Spatial Relation Recognition

Having verified that SpatialSense is an effective benchmark for spatial relation recognition, we evaluate multiple methods on SpatialSense, including simple baselines based on language and 2D cues as well as state-of-the-art models for visual relationship detection. Experimental results reveal the difficulty for state-of-the-art models to go beyond simple priors and learn to reason about visual content; a simple baseline based on 2D cues performs competitively with state-of-the-art models. We also conduct a human evaluation quantifying the level of ambiguity in the task.

Model architectures. The task is spatial relation recognition: given the image, two objects (their names and bounding boxes) and a spatial predicate, the model classifies whether the relation holds. Two simple baselines are evaluated: a language-only model and a 2D-only model. Their architectures are the same as in section 4.2. We also report the performance of combining their predictions by a weighted average. We evaluate five state-of-the-art models: Vip-CNN [14], Peyre *et al*. [22], PPR-FCN [37], DRNet [5] and VTransE [33]. They were created for visual relationship detection but can be adapted to our task straightforwardly: First, object detectors are replaced by ground truth objects. Second, object names are encoded using word embeddings rather than one hot encoding, since SpatialSense has unconstrained object categories. Third, for each relation *subject-predicate-object*, the model takes *subject* and *object* as input, and generates scores for all predicates; the score for that particular *predicate* is the final binary classification score. Details are in the supplementary material.

Implementation details. Object names are encoded to fixed-length vectors using Word2Vec followed by a GRU. When combining the language and 2D baselines by a weighted average, we find 80% from 2D and 20% from language to perform well (measured by validation accuracy).

For state-of-the-art models, we crop the union bounding box of the two objects, resize it to 280×280 and normalize the pixel values by the mean and standard deviation of all training images. During training, we then crop it randomly, resize to 224×224 and apply color jittering; during testing, we simply take a 224×224 crop at the center.

Analyzing the results. Table 3 summarizes the testing accuracies. Due to the challenging nature of SpatialSense, the best models perform around 70%, which is quite low for a binary classification task. DRNet is the best model without ensemble, closely followed by other state-of-the-art models, which confirms that models for visual relationship detection can be well adapted to our task. Notably, the 2D baseline performs closely to state-of-the-art models and is on a par with DRNet when combined with the language baseline. This suggests state-of-the-art models might rely too much on 2D cues and fail to develop deeper visual reasoning.

To validate this conjecture, we examine the correlations between the errors. For each model, consider an error vector of length N where N is the size of the testing data. The values in the vector are 1 when the model predicts incorrectly and 0 otherwise. Table 4 shows the correlation matrix between these error vectors. The language baseline does not correlate well with the 2D baseline, suggesting that they make rather different kinds of errors. However, all state-of-the-art models have high correlation with the 2D-only baseline, which implies they make similar predictions to a 2D-only model and supports our conjecture.

Some random failing examples for the language and 2D baselines are shown in Fig. 5. Models based solely on 2D cues struggle in scenarios that involves the relative frame of reference or require depth-based reasoning. Language cues fall short when a common spatial relation does not appear or, in contrast, an unusual spatial relation is present. These observations indicate that neither language nor 2D cues are sufficient for spatial relation recognition. Beyond these simple cues, it is crucial to learn visual reasoning that elude the current state-of-the-art. Our benchmark takes a step towards that goal by providing a more accurate gauge of a model's visual reasoning ability.

Human evaluation. Finally, recognizing spatial relations is inherently noisy; it is not always clear whether a relation holds. We conduct a human evaluation, in which annotators are asked to make predictions on the testing data. Multiple human responses on the same relation is merged by majority vote. For quality control purpose, the annotators who answer "yes" more than 80% of the time are considered outliers, and their responses are excluded. We collect 10,205 predictions and the accuracies are in the last row of Table 3. Although not perfect, humans perform very well on this task, reaching an accuracy of 94.6%. The gap between humans and algorithms provides a large room for future improvement on this benchmark.

Model	Overall	above	behind	in	in front of	next to	on	to the left of	to the right of	under
Language-only	60.1	60.4	62.0	54.4	55.1	56.8	63.2	51.7	54.1	70.3
2D-only	68.8	58.0	66.9	**70.7**	63.1	62.0	76.0	66.3	**74.7**	67.9
Language + 2D	71.1	61.1	67.5	69.2	66.2	64.8	77.9	**69.7**	**74.7**	**77.2**
Vip-CNN [14]	67.2	55.6	68.1	66.0	62.7	62.3	72.5	**69.7**	73.3	66.6
Peyre *et al.* [22]	67.5	59.0	67.1	69.8	57.8	**65.7**	75.6	56.7	69.2	66.2
PPR-FCN [37]	66.3	61.5	65.2	70.4	64.2	53.4	72.0	69.1	71.9	59.3
DRNet [5]	**71.3**	**62.8**	**72.2**	69.8	**66.9**	59.9	**79.4**	63.5	66.4	75.9
VTransE [33]	69.4	61.5	69.7	67.8	64.9	57.7	76.2	64.6	68.5	76.9
Human	94.6	90.0	96.3	95.0	95.8	94.5	95.7	88.8	93.2	94.1

Table 3: The testing accuracies of baseline methods on spatial relation recognition. The 2D baseline performs closely to state-of-the-art models, which suggests state-of-the-art models might learn to exploit simple priors and fail to develop deeper visual reasoning capabilities.

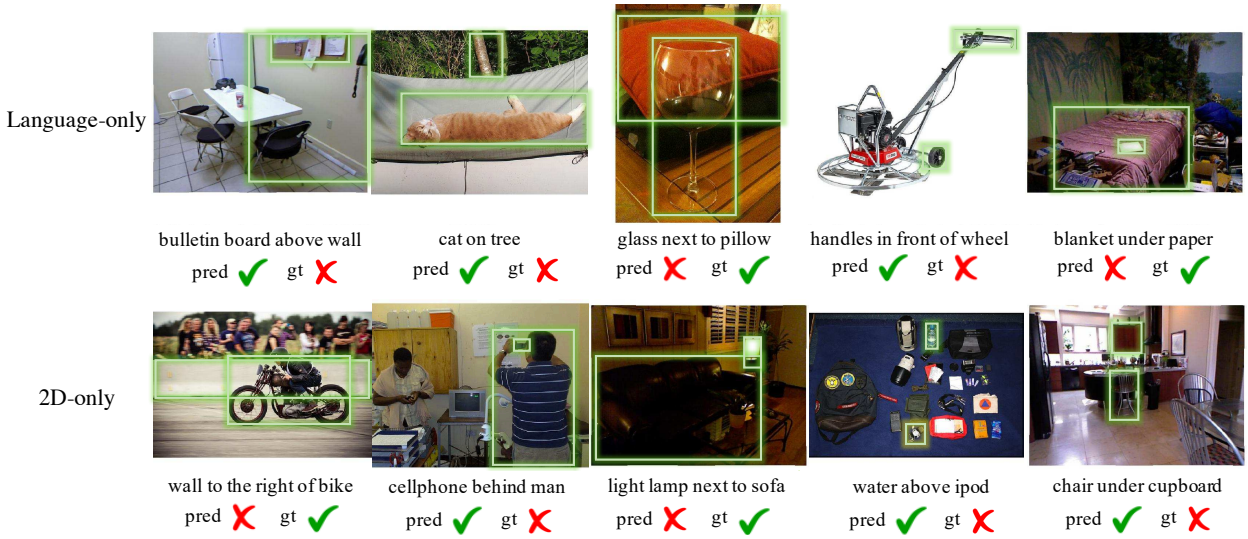

Figure 5: Failing examples of the language and 2D baselines (*pred*: prediction, *gt*: ground truth). The language baseline fails when a frequent spatial relation does not occur in a particular image (e.g., "cat on tree"), or a technically valid spatial relation is expressed in an unusual way (e.g., "blanket under paper"). The 2D baseline fails to consider the relative frame of reference (e.g., "wall to the right of bike") and depth information (e.g., "light lamp next to sofa" and "chair under cupboard").

	L	2D	Vi	Pe	PP	D	VT
Language-only	–	**0.04**	**0.05**	**0.09**	**0.05**	0.22	0.33
2D-only	**0.04**	–	0.60	0.35	0.46	0.43	0.31
Vip-CNN [14]	**0.05**	0.60	–	0.30	0.46	0.25	0.20
Peyre *et al.* [22]	**0.09**	0.35	0.30	–	0.30	0.27	0.20
PPR-FCN [37]	**0.05**	0.46	0.46	0.30	–	0.24	0.21
DRNet [5]	0.22	0.43	0.25	0.27	0.24	–	0.44
VTransE [33]	0.33	0.31	0.20	0.20	0.21	0.44	–

Table 4: The correlation matrix between errors of the models, with low correlation in **green** and high correlation in red. The 2D and language baselines make independent errors (low correlation). DRNet [5] and VTransE [33] correlate strongly with both baselines, suggesting that their predictions are very similar to a combination of 2D and language cues. The other models are also highly correlated with the 2D baseline (they do not utilize language cues).

6. Conclusion

We introduced a novel dataset SpatialSense for the challenging task of spatial relation recognition. SpatialSense was constructed through adversarial crowdsourcing, which significantly reduced its dataset bias compared to alternative datasets. We evaluated multiple baselines on SpatialSense, demonstrating e.g., that a simple 2D baseline performs competitively with state-of-the-art models. This reveals that state-of-the-art models rely too much on dataset bias in existing benchmarks, validating the need for SpatialSense as a new challenging testbed for spatial relation recognition.

Acknowledgements. This work is partially supported by National Science Foundation under grant 1633157, DARPA under grant FA8750-18-2-0019, and IARPA under grant D17PC00343.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Dont just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[2] Josh Attenberg, Panagiotis G Ipeirotis, and Foster J Provost. Beat the machine: Challenging workers to find the unknown unknowns. *Human Computation*, 11(11), 2011.

[3] Wei-Lun Chao, Hexiang Hu, and Fei Sha. Being negative but constructively: Lessons learnt from creating better visual question answering datasets. *arXiv preprint arXiv:1704.07121*, 2017.

[4] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[5] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[7] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, Trevor Darrell, et al. Grounding spatial relations for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1640–1647. IEEE, 2013.

[8] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[9] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 1983–1991. IEEE, 2017.

[10] Ashish Kapoor, Kristen Grauman, Raquel Urtasun, and Trevor Darrell. Active learning with gaussian processes for object categorization. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[11] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 2017.

[12] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[13] Barbara Landau and Ray Jackendoff. Whence and whither in spatial language and spatial cognition? *Behavioral and brain sciences*, 16(2):255–265, 1993.

[14] Yikang Li, Wanli Ouyang, Xiaogang Wang, and Xiao'ou Tang. Vip-cnn: Visual phrase guided convolutional neural network. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[15] Yikang Li, Wanli Ouyang, Bolei Zhou, Jianping Shi, Chao Zhang, and Xiaogang Wang. Factorizable net: an efficient subgraph-based framework for scene graph generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 335–351, 2018.

[16] Yikang Li, Wanli Ouyang, Bolei Zhou, Kun Wang, and Xiaogang Wang. Scene graph generation from objects, phrases and region captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1261–1270, 2017.

[17] Xiaodan Liang, Lisa Lee, and Eric P. Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[18] 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.

[19] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pages 852–869. Springer, 2016.

[20] Varun Manjunatha, Nirat Saini, and Larry S Davis. Explicit bias discovery in visual question answering models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9562–9571, 2019.

[21] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[22] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[23] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE, 2011.

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision*, pages 746–760. Springer, 2012.

[25] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *ICCV*, 2011.

[26] Sudheendra Vijayanarasimhan and Kristen Grauman. Large-scale live active learning: Training object detectors with crawled data and crowds. *International Journal of Computer Vision*, 108(1-2):97–114, 2014.

[27] Sanghyun Woo, Dahun Kim, Donghyeon Cho, and In So Kweon. Linknet: Relational embedding for scene graph. In *Advances in Neural Information Processing Systems*, pages 558–568, 2018.

[28] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[29] Xu Yang, Hanwang Zhang, and Jianfei Cai. Shuffle-then-assemble: Learning object-agnostic visual relationship features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 36–52, 2018.

[30] Guojun Yin, Lu Sheng, Bin Liu, Nenghai Yu, Xiaogang Wang, Jing Shao, and Chen Change Loy. Zoom-net: Mining deep feature interactions for visual relationship recognition. In *European Conference on Computer Vision*, pages 330–347. Springer, 2018.

[31] Ruichi Yu, Ang Li, Vlad I. Morariu, and Larry S. Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[32] Zhen Zeng, Zheming Zhou, Zhiqiang Sui, and Odest Chadwicke Jenkins. Semantic robot programming for goal-directed manipulation in cluttered scenes. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7462–7469. IEEE, 2018.

[33] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[34] Hanwang Zhang, Zawlin Kyaw, Jinyang Yu, and Shih-Fu Chang. Ppr-fcn: Weakly supervised visual relation detection via parallel pairwise r-fcn. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[35] Ji Zhang, Mohamed Elhoseiny, Scott Cohen, Walter Chang, and Ahmed Elgammal. Relationship proposal networks. In *CVPR*, volume 1, page 2, 2017.

[36] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.

[37] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.