2.5D Visual Relationship Detection

Yu-Chuan Su, Soravit Changpinyo, Xiangning Chen\textsuperscript{1,2*}, Sathish Thoppay, Cho-Jui Hsieh\textsuperscript{2}, Lior Shapira, Radu Soricut, Hartwig Adam, Matthew Brown, Ming-Hsuan Yang, Boqing Gong
\textsuperscript{1}Google Research \hspace{1cm} \textsuperscript{2}UCLA

Abstract

Visual 2.5D perception involves understanding the semantics and geometry of a scene through reasoning about object relationships with respect to the viewer in an environment. However, existing works in visual recognition primarily focus on the semantics. To bridge this gap, we study 2.5D visual relationship detection (2.5VRD), in which the goal is to jointly detect objects and predict their relative depth and occlusion relationships. Unlike general VRD, 2.5VRD is egocentric, using the camera's viewpoint as a common reference for all 2.5D relationships. Unlike depth estimation, 2.5VRD is object-centric and not only focuses on depth. To enable progress on this task, we create a new dataset consisting of 220k human-annotated 2.5D relationships among 512K objects from 11K images. We analyze this dataset and conduct extensive experiments including benchmarking multiple state-of-the-art VRD models on this task. Our results show that existing models largely rely on semantic cues and simple heuristics to solve 2.5VRD, motivating further research on models for 2.5D perception. The new dataset is available at https://github.com/google-research-datasets/2.5vrd.

1. Introduction

Visual 2.5D perception involves understanding the semantics and geometry of a scene: the relationships between objects with the viewer as the main reference point in an environment [51]. For instance, we may refer to a chair solely through semantics by its name and attributes (e.g., the wooden chair), through both semantics and geometry by its spatial relationship to other objects (e.g., the chair on the right of the table), or through the geometry only—by distance to our viewpoint (e.g., the chair that is closer). However, object recognition [10], detection [37, 29], and segmentation [37], among other hallmark computer vision tasks, primarily focus on the semantics component. As a result, most visual perception models operate in a 2D world and lack 2.5D visual understanding.

Motivated by this, we introduce 2.5D visual relationship detection (2.5VRD). The goal of 2.5VRD is to detect objects and predict their relative depth and occlusion relationships as a unified task, as illustrated in Figure 1. We study the relative depth in two settings: “within an image” and “across images” (e.g., the depth of the tree with respect to the man and to the vehicle.) Occlusion on the other hand only applies to the “within an image” setting. Clearly, to be able to perform well on this task, the geometry of a scene cannot be ignored.

Our task is primarily motivated by the scientific question “Do machines possess 2.5D visual understanding capability like humans do?” An answer to this question would benefit our understanding of machine visual perception. Furthermore, we believe an effective model for the core problem of 2.5D visual relationships can benefit a wide range of applications, e.g., helping a self-driving vehicle to understand scenes beyond its LiDAR range, assisting a robot to navigate and manipulate objects, and improving (amodal) instance detection [13, 20, 78, 25, 11] and segmentation [34, 48, 21, 52, 23, 61, 7], to name a few.

2.5VRD shares similar high-level motivation as visual relationship detection (VRD) and depth estimation, yet with
important conceptual differences. Our task differs from general VRD as it focuses on depth and occlusion. It also differs from recent work on spatial VRD [53, 41, 68, 27, 28]. For example, the spatial relationships in SpatialSense [68] are concerned with both locations and poses of objects with respect to each other, while 2.5VRD is egocentric, defining occlusion and depth orders from the viewer’s perspective. Indeed, “a chair (in SpatialSense) may be behind a person even if it appears to the left of the person (depending on where that person faces)”. The most similar work to ours is Rel3D [16] which also consists of view-dependent relationships, but, unlike ours, they are situated in synthetic environments. Finally, the depth in 2.5VRD is object-centric, unlike the pixel-wise depth studied in monocular depth estimation [54, 55, 38, 57, 30, 39, 12, 33, 18, 56, 64, 31, 65, 4].

To enable progress on our proposed 2.5VRD task, we introduce a new large-scale dataset of 219,570 2.5D relationships among 511,545 objects from 11,084 images of the Open Images [29]. Our dataset is an order-of-magnitude larger than existing VRD datasets [41, 68]. It is also the first large-scale human-annotated dataset with 2.5D visual relationships (in two settings) on natural images. Additionally, unlike existing benchmarks, the annotations on our validation and test sets are exhaustive, allowing us to use both precision and recall as the evaluation metrics.

We analyze our dataset and conduct extensive experiments that shed light on the difficulty of 2.5VRD. First, we use the rich annotations to analyze how humans and visual recognition models tackle 2.5VRD. We build a simple baseline and find that our baseline’s performance and the agreement among five raters are both correlated well with the relative depth between two objects. Second, we study the effect of various cues on the performance of our baseline model. Our results show that the object sizes and locations are important at predicting the relative depth, suggesting that high-quality object detection is key to 2.5VRD. On the other hand, the appearance cue is more important for occlusion prediction. Finally, we benchmark four state-of-the-art VRD models on our 2.5VRD dataset. We find that they do not significantly outperform our simple baseline and that they do not generalize well from the within-image setting to the across-image setting. These results suggest that existing models designed for 2D VRD are not sufficient for relative depth or occlusion reasoning.

In summary, our main contributions are as follows. We propose the 2.5VRD task, promoting the object-centric depth and occlusion reasoning as the first-class citizen. We concretize the task with an extensively labeled dataset, which is an order-of-magnitude larger than existing VRD datasets and unique in the exhaustive annotations on the validation and test sets. We propose a model to study various factors that may come into play, and we hope the findings will help design improved models in future. Finally, we evaluate four state-of-the-art VRD methods for 2.5VRD. Results show that their performance is comparable to the baseline model, highlighting the new challenges in 2.5VRD of which are not taken into account by these methods yet.

2. Visual Relationships in 2.5D

This section first formalizes the 2.5VRD task, followed by a detailed strategy for data and label collection. Next, we analyze the resulting dataset and study how humans approach 2.5VRD. Finally, we compare 2.5VRD with related datasets and work.

2.1. Problem Formulation

We formalize our task as follows. Given two input images \((I_a, I_b)\), the 2.5VRD task is to predict a set of 2.5D relationships. Each 2.5D relationship consists of a triplet \((o_a, \text{predicate}, o_b)\), where \(o_a\) is an object in \(I_a\) specified by a tight bounding box and its class name and \(\text{predicate}\) is the relationship between \(o_a\) and \(o_b\). By treating \(I_a\) and \(I_b\) as separate input, this formulation is applicable to both within-image and across-image setting. For the within-image setting, i.e., \(I_a\) and \(I_b\) are identical, we consider both relative depth and occlusion relationships. For the across-image setting, only the relative depth relationship is relevant. Possible values for the \(\text{predicate}\) include \{is closer than, is farther than, is at the same depth as\} for relative depth, where the \(\text{predicate} \ “\text{is at the same depth as}” \) is on only in the within-image setting since we find it too challenging to label across images. For occlusion, \(\text{predicate} \in \{\text{occludes, does not occlude, mutual occlusion}\}\). Note that there could be no occlusion between two objects, and they could be mutually occluded. Because the relationships are defined between any two objects, there should be \(N_a \times (N_a - 1)\) and \(N_a \times N_b\) relationships in within-image and across-image setting respectively, where \(N_x\) is the number of objects in \(I_x\).
Table 1: Overview of the 2.5VRD dataset.

|                | Training | Validation | Test  |
|----------------|----------|------------|-------|
| Images         | 105,694  | 1,200      | 4,000 |
| Objects        | 493,498  | 4,063      | 13,893|
| Within pairs   | 105,694  | 6,339      | 23,724|
| Object A closer| 39.1%    | 39.9%      | 38.7% |
| Object B closer| 39.9%    | 39.6%      | 38.9% |
| Same depth     | 10.1%    | 6.3%       | 7.5%  |
| A occludes B   | 10.9%    | 10.3%      | 9.3%  |
| B occludes A   | 11.0%    | 10.0%      | 9.1%  |
| Mutual occlusion| 3.5%     | 1.9%       | 2.1%  |
| Across pairs   | 52,484   | 600        | 2,000 |
| Object A closer| 43.5%    | 40.9%      | 43.8% |
| Object B closer| 45.2%    | 48.0%      | 44.5% |

One may alternatively formulate the problem for relative depth relationship as ranking of objects, but we find that some difficult pairs of objects often invalidate the ranking lists. It becomes especially troublesome when we merge the ranking lists from different raters. Hence, we instead use the $\langle o_a, \text{predicate}, o_b \rangle$ triplets considering their flexibility and the annotation cost.

2.2. Data and Label Collection

We construct the 2.5VRD dataset on top of the Open Image Dataset (OID) [28]. OID mostly consists of scenery images from Flickr, where each image may contain multiple objects and/or people. We maintain the original train/validation/test split of the images. We use the annotated bounding boxes and 600 class names of objects in the OID images. As the annotations in OID are over-complete (i.e., multiple bounding boxes for an object), we filter the boxes before collecting 2.5VRD labels. We also ignore extremely small or large boxes (occupying less than 2% or more than 70% area of the image) to avoid ill-defined cases. Finally, we remove boxes containing group of objects. See supp. for details.

Labeling 2.5VRD within an image. For each training image, we have a rater to label one randomly formed pair of objects. The label consists of both relative depth and occlusion relationships (see the top panel in Figure 2, which is a screenshot of the annotation UI). We also include an “unsure” option for ambiguous cases in light of the difficulty of the problem. For each validation or test image, we collect five 2.5VRD labels for every pair of objects from five raters respectively. We then use majority voting to determine the final labels. This strategy ensures that the labels for the validation and test set are of high quality. It also results in comprehensive annotations for all pairs of objects in a validation or test image, allowing us to evaluate model performance in terms of precision and recall. Note that we annotate only one pair of objects in the training set to maximize the number of training samples under the budget constraint, based on the hypothesis that higher diversity in the training data is important for model performance.

Labeling 2.5VRD across images. We split the training images into two groups and then construct pairs by selecting one image from either group. Given a pair of images, we randomly choose an object from each of them. A rater ranks the two objects by their depths using the annotation UI illustrated by the bottom panel in Figure 2. We pair up validation and test images in the same way, but provide dense labels for all across-image object pairs. In addition, we assign each of them to five raters to secure high-quality labels for the validation and test sets.

2.3. Dataset Statistics and Analyses

Table 1 shows the statistics for the proposed 2.5VRD dataset. Note that the within-image and across-image setting share the common sets of images and objects within each split. Out of the within-image object pairs, about 80% have apparent depth disparities (rows of “A is closer” and “B is closer”), and approximately 10% are at about the same depth. Furthermore, one object occludes the other (but not vice versa) in about 20% of the pairs (rows of “A occludes B” and “B occludes A”), and about 3% are mutually occluded. For across-image 2.5VRD, the raters managed to tell the difference between two objects’ depths for approximately 88% of all the pairs (see the last two rows in the table). The dataset preserves the object pairs for which the raters selected the unsure option, as learning models may make sense of them in the future.

Table 2: Distributions of difficulty scales for within-image and across-image object depth ordering, respectively.

|                | Easy | Moderate | Difficult | Infeasible | Ambiguous |
|----------------|------|----------|-----------|------------|-----------|
| Within-image   | 55.8%| 16.4%    | 13.1%     | 10.5%      | 4.3%      |
| Across-image   | 50.0%| 21.6%    | 16.9%     | 6.7%       | 4.8%      |

Human perception of 2.5VRD. The annotations on each validation or test example by five raters allow us to analyze how humans approach the 2.5VRD task. We define five difficulty scales for depth ordering:

Easy: Five raters all agreed on a relative depth label and did not choose the unsure option.

Moderate: Four out of five raters agreed with each other.

Difficult: Three out of five agreed on a relative depth label.

Infeasible: A majority of the raters chose “unsure”.

Ambiguous: There is no majority agreement on any label.
Table 2 shows the distribution of validation and test examples over the five difficulty scales. For the proposed dataset, more than 50% of the object pairs belong to the “easy” scale. Moreover, more within-image object pairs fall into the “easy” scale than the across-image object pairs, likely because the latter requires the raters to estimate metric depths to some extent. In contrast, relative depths are sufficient to rank two objects in the same scenery image. There are 7% and 10% infeasible object pairs for across-image and within-image 2.5VRD, respectively, meaning that a majority of the raters were “unsure” how to rank them by depth. Finally, less than 5% of the object pairs received no label because there was no majority winner. Overall, the 2.5VRD task is more difficult for humans than we expected, considering that a notable proportion of examples are “infeasible” or “ambiguous” for human raters to reach a consensus.

Table 3 contrasts our 2.5VRD dataset with the related, representative datasets. VRD and SpatialSense are arguably the most widely-used benchmarks for VRD. Our dataset is an order of magnitude larger than them in the numbers of images, objects, and relations. While 2.5VRD makes depth and occlusion the first-class citizen, almost all relationships in the existing VRD datasets are 2D. Only 5,132 relations in SpatialSense have “behind” or “front” in their predicates, and yet a cell phone could be “in front of” a person as long as the person faces to the phone even if the

2.4. Related Datasets and Work

VRD. Sadeghi and Farhadi studied VRD using 17 unique relationships [53]. Lu et al. scaled up the study by a new benchmark with 37,993 relations over 5,000 images [41]. They showed that language prior was effective for detecting the visual relations with few to no training examples. Peyre et al. collected 76 unusual relationships to evaluate model generalization for VRD [46]. SpatialSense curated 17,498 spatial relationships over 11,569 images [68]. The Visual Genome (VG) [27] and OID [28] provide VRD labels albeit sparse per image.
Table 3: 2.5VRD vs. existing datasets (*2.5D relations for existing datasets have “behind” or “front” in their predicates)

|                  | Images | Objects | Classes | Predicates | Relations | 2.5D relations* | Occlusion | Across-img relations |
|------------------|--------|---------|---------|------------|-----------|-----------------|-----------|---------------------|
| VRD [41]         | 5,000  | 32,901  | 100     | 70         | 37,993    | 4,780           | 0         | 0                   |
| SpatialSense [68]| 11,569 | 33,861  | 3,679   | 9          | 17,498    | 5,132           | 0         | 0                   |
| VG (relationship) [27] | 108,077 | 2,254,357 | 65,405  | 4,016      | 2,316,104 | 66,660          | 0         | 0                   |
| OID (relationship) [28] | 596,308 | 1,478,971 | 303     | 31         | 3,284,282 | 0               | 0         | 0                   |
| 2.5VRD (ours)    | 110,894 | 511,454 | 600     | 5          | 219,570   | 219,570         | 29,105    | 83,813              |

Besides the existing VRD datasets, our work is also closely related to the rich line of VRD methods and models [36, 75, 76, 81, 35, 47, 74, 8, 70, 67, 77, 63, 72, 46, 71], which will speed up tackling 2.5VRD. We leave the exploration into them to future work. Instead, our experiments aim to help readers gain more insights into 2.5VRD, especially about how different visual cues interplay in the 2.5D visual relationships. Our work is also related to human-object interactions [17, 69, 9, 26, 15, 3, 49, 2, 82], which may be viewed as human-centric VRD. In contrast, 2.5VRD is egocentric, using the viewpoint as the reference for the relationships between two arbitrary objects.

Figure 4: Examples of across-image 2.5VRD. The difficulty is easy, moderate, difficult, infeasible, and ambiguous from top to bottom.

2.5D perception. Monocular depth estimation [54, 55, 38, 57, 30, 39, 12, 33, 18, 56, 4, 64, 31, 65] infers a dense, pixel-wise depth map from an image, and thus not object-centric. Our empirical results show that, while dense depth maps provide informative cues to 2.5VRD, it is far from solving depth ordering for objects and does not account for occlusion. Unlike depth, existing works mostly do not consider occlusion as an independent task, but a latent factor to improve object detection [13, 20, 78], semantic and instance segmentation [52, 23, 61, 7], and other applications [24, 73]. Instead, 2.5VRD directly deals with occlusion.

Some works study occlusion and depth ordering between image regions instead of objects [59, 19, 79, 23, 42, 80, 48]. Because they define occlusion along the object or scene boundaries, the relationship is always binary. Also, most of them couple the two relationships and define depth order based on occlusion [23, 42, 80, 48], so the relative depth is only defined within a connected component where the regions overlap. While some works define depth order globally, they rely on the existence of the ground [19] or 3D object bounding boxes [79] and consider only objects on the ground or cars. In contrast, we consider the 2.5D relationships between arbitrary objects, leading to a more general relationships definition and a larger dataset.

Our work is also broadly related to amodal instance segmentation [34, 80, 48, 21] and amodal object detection [25, 11]. We envision that detecting depth and occlusion relationships between objects can facilitate amodal tasks and vice versa. Finally, single-view 3D object detection [62, 14, 66, 58, 6, 45, 44, 40, 32, 22, 1, 5] is related but requires labor-intensive data collection, limiting existing work to mainly indoor and self-driving environments.
3. Experiments and Analyses

In this section, we evaluate the performance of visual recognition models on 2.5VRD. The goal of our experiments is to understand the effect of different visual signals and models on the 2.5VRD performance and establish the baseline results for future work. To this end, we develop two baselines and benchmark them along with four state-of-the-art VRD methods on our proposed 2.5VRD task. Code and data will be made publicly available.

3.1. Approaches to 2.5VRD

In our experiments, we explore a two-stage approach for this task. The first stage leverages an oracle/ off-the-shelf object detectors to provide/infer multiple \((a_0, b_0)\) pairs. Then, given \((I_a, o_a, I_b, o_b)\) as input, we infer 2.5D relationships between \(o_a\) and \(o_b\) by predicting the predicate for each relationship. Our baseline and state-of-the-art models will operate in this second stage. Because 2.5D relationships are directional, we treat \((o_a, \text{occludes}, o_b)\) and \((o_b, \text{occludes}, o_a)\) as two different labels for \((o_a, o_b)\). We also include the “unsure” label for the relative depth relationship. This leads to four possible values for each 2.5D relationship.

Overview of visual cues All baselines and state-of-the-art models explored in this paper employ a subset of the following four types of visual cues. The first one is direct semantics in the form of object class labels. For example, a person is often closer to the viewpoint than trees and buildings; more examples are provided in Figure 5. The second cue is the geometric cue in the form of box size and location. For example, an overlap implies a probable occlusion relation. The third cue is appearance, both in term of object and its context. The forth cue is depth, both in terms of object and its context.

3.1.1 Rule-Based Baselines

We explore the following rule-based baselines, each of which relies on a specific visual cue.

- **Object class** predicts the most frequent predicate for the pair of object classes in the training set.
- **Size** predicts \(o_a\) is closer to the camera than \(o_b\) if \(o_a\)’s box size is larger by a margin \(\Delta_s\) (based on the fact that an object’s size in an image is inversely proportional to its depth.) For occlusion prediction, if the size of the overlap area is larger than a threshold, the object that is closer occludes the other; otherwise, no occlusion.
- **Location** predicts \(o_a\) is closer to the camera than \(o_b\) if \(o_a\)’s Y-coordinate is larger by a margin \(\Delta_l\). For occlusion prediction, we couple the rule with relative depth prediction as in Size.
- **Depth** For depth prediction, we assume a depth map produced by a monocular depth estimator MiDaS [31] and compute a depth estimate \(D_o\) for each object by averaging the depth values inside its bounding box. \(o_a\) is closer to the camera than \(o_b\) if \(D_{o_a}\) is smaller than \(D_{o_b}\) by a margin \(\Delta_d\). For occlusion prediction, we again couple the rule with relative depth prediction as in Size and Location.

The margins are set to \(\Delta_s=0.0, \Delta_l=0.02,\) and \(\Delta_d=0.02\), respectively, by a grid search on the validation set.

3.1.2 Simple MLP Baselines

We explore a two-layer multi-layer perceptron (MLP) followed by two heads, treating depth and occlusion predictions as two multi-class classification problems. This model takes in up to four types of visual signals, as detailed below.

---

We explored different methods for combining the inferred depth values but did not observe significant differences.
• **Object class feature** represents an object’s class using a one-hot vector.

• **Bounding box feature** uses the bounding boxes’ coordinates, concatenated with the overlap region’s height, width, and area.

• **Appearance feature** extracts the appearance feature for an object from the bounding box locally and from the image globally using a Faster-RCNN [50] pre-trained on OID [28] with an Inception-ResNet [60] backbone. Given an image, we first obtain a feature map from Faster-RCNN’s last convolutional layer. We then perform average pooling over the feature map to obtain the image appearance feature and ROI pooling over a bounding box to reap the corresponding object’s appearance feature. Concatenating the image feature and the object feature provides information about the object’s surroundings and the object itself.

• **Depth feature** extracts depth information using MiDaS [31]. Given the depth map (without per-image normalization), we compute the mean, standard deviation, minimum, and maximum of the depth values within the bounding box of an object as the depth feature. Further, we also compute the depth feature for the entire image and concatenate it with the object’s depth feature.

**Implementation details** We use concatenation to combine features from a pair of objects and to combine features of different types. We use a hidden layer of size 1024. We use the sum of cross-entropy losses over the two classification heads. We augment each training input with \((I_a, o_a, I_b, o_b)\) with \((I_b, o_b, I_a, o_a)\), and also randomly perturb the center, width, and height of bounding boxes by 10% in addition to other standard perturbation to the images’ saturation, contrast, brightness, and hue during training. The model is trained using Adam for 60,000 steps with a base learning rate of \(2 \times 10^{-4}\) and a batch size of 32. We add an \(L_2\) regularization with weight \(1 \times 10^{-4}\) and use dropout with ratio 0.5.

3.1.3 State-of-the-art VRD Methods

We explore the following state-of-the-art VRD models.

• **ViP-CNN** [35] predicts predicates using visual features from three bounding boxes, including the two object bounding boxes and a tight bounding box covering the union of the two objects.

• **PPR-FCN** [81] is similar to ViP-CNN but adopts a different architecture to combine the information.

• **DRNet** [8] takes as input the appearance feature, location, and word vector embedding of two objects. The model architecture is designed by unrolling a conditional random field model.

• **VTransE** [74] predicts predicates from the feature vector difference between two objects, where the features involve appearance, location, and word vector embedding.

See supp. for details. Note that ViP-CNN and PPR-FCN take the union of two boxes as input and are therefore not applicable for the cross-image task.

3.2 Evaluation Metrics

We can evaluate a model’s performance by precision and recall since we exhaustively label all pairs of objects in an image or between images of the validation and test sets. We report F1-scores in the main text and all metrics in the supplementary materials.

A 2.5VRD model detects objects and predicts *predicate* between any pair of them. We first use the same filtering procedure in data collection (see Section 2.2) to discard extremely small or big box and ill-defined relations. Assuming that \(N\) objects in an image survive this procedure, there will be \(N \times (N - 1)\) 2.5D visual relationships—for each pair of objects, we predict two *predicates* for both directions, respectively, because the depth and occlusion relationships are directional. To compute precision and recall, we find true positive 2.5D visual relations as follows. A detected relationship, \((o_a, \text{predicate}, o_b)\), is considered correct if it satisfies two conditions. 1) Both objects \(o_a\) and \(o_b\) are detected correctly. We consider \(o_a\) as correct detection if it has greater than 0.5 intersection-over-union with the groundtruth box. 2) The predicted *predicate* is correct. Similarly, we use the F1-score to evaluate across-image 2.5VRD, where the object pairs are between images.

We shift the evaluation metrics toward the quality of the *predicate*, not the object detection, by making no require-
A closer
B closer
Same
No occl.
A occl. B
B occl. A
Mutual
A closer
B closer

Within Image
Occlusion
Across Images

Figure 6: Model performance w.r.t. object size and location for the \( \langle o_a, \text{is closer than}, o_b \rangle \) relationship.

Figure 7: Class-wise results of the baseline model.

The results of our MLP baseline model are in the middle part of Table 4. We explore using multiple combinations of visual cues, starting from the bounding box (B) feature only and then adding object class (C), depth (D), and appearance (A) features. We again observe that the estimated depth is a strong cue but others are also useful, with the appearance feature being most complementary, especially in the occlusion and across-image depth sub-tasks, implying the potential of the appearance cue for future work.

Since the bounding box (B) feature couples the object location and size cues, we use Figure 6 to dive deep into the results. We discretize the objects’ (normalized) sizes and vertical positions and focus on the \( \langle o_a, \text{is closer than}, o_b \rangle \) relations. There are high F1-scores in panels (a) and (c) when \( o_a \) is larger than \( o_b \) in size or the vertical position, implying that the model heeds the geometric prior for within-image 2.5VRD. For the same reason, the F1-scores are low when \( o_a \) is closer than \( o_b \) and yet \( o_a \) is smaller than \( o_b \) in size (or vertical position). There is no obvious pattern for the cross-image depth relations (see (b) and (d)).

3.3. Results and Analyses

We use the Faster-RCNN detector pre-trained on OID to detect objects for all the experiments except in Table 5, where we employ the groundtruth bounding boxes. We present more experiments and analysis in the supp.

Overall results. The top part of Table 4 shows the results of different rule-based methods. The estimated depth performs best on all the three 2.5VRD sub-tasks (within-image depth, occlusion, and across-image depth). The object location-based rule is also strong, and especially useful for the sub-task of within-image depth. In contrast, the object size matters more in the across-image depth sub-task. Finally, we see relatively low performance using the object classes. This suggests that 2.5VRD is more dependent on geometry than the semantic class prior.

The results of our MLP baseline model are in the middle part of Table 4. We explore using multiple combinations of visual cues, starting from the bounding box (B) feature only and then adding object class (C), depth (D), and appearance (A) features. We again observe that the estimated depth is a strong cue but others are also useful, with the appearance feature being most complementary, especially in
Table 5: Sources of error in the baseline model to 2.5VRD.

| 2.5VRD | Predicate Prediction | Object Detection |
|--------|----------------------|------------------|
| Average| 0.335                | 0.782            | 0.492            |

lates the symmetric property in 8.9% of the test cases. Similarly, we also check the model’s transitive property, i.e., if it predicts $o_a \leq o_b$ and $o_b \leq o_c$, then it is supposed to predict $o_a \geq o_c$. The model fails the transitive property test in 1.7% of cases. For comparison, the groundtruth labels aggregated from five raters break the transitive property in only 0.5% of all cases. It would be interesting to design some inductive bias into the model architecture to make its prediction symmetric and transitive in future work.

Sources of error. Finally, we provide two “upper bounds” for the baseline model, through which we hope to understand the sources of error in 2.5VRD. Our approach takes two stages to tackle 2.5VRD, first detecting objects and then predicting predicates for all pairs of the detected objects. We investigate from which stage the final error mainly comes from by using the following approach variations:

- **Predicate** prediction, which supplies the model with groundtruth bounding boxes and classes to study how the predicate prediction performs.
- **Object detection**, singling out the object detection module by assuming a perfect predicate predictor, and
- **2.5VRD**, which performs both object detection and predicate prediction by the full model.

Table 5 reports the results of the three variations. Assuming perfect object detection, 2.5VRD degenerates to the task of predicate prediction, which boosts our method’s F1-score from 0.335 to 0.782. This drastic change indicates there is a big room for the object detection module to improve for tackling 2.5VRD. When we use an ideal predicate predictor, we only need object detection for 2.5VRD and observe a performance increase from 0.335 to 0.492. It is clear that the object detection module is the primary source of our model’s error, but both “upper bounds” are virtually high. Tackling 2.5VRD requires advancing not only object detection but also 2.5D predicate prediction.

4. Conclusion

We introduce 2.5VRD, a new task for studying the relationships between objects via depth and occlusion. We collect a large-scale dataset with rich human annotations, through which we conduct extensive analyses to gain insights into 2.5VRD. Experiments reveal that 2.5VRD desires progress on both object detection and predicate prediction, and the latter may benefit from a model’s inductive bias that satisfies symmetric and transitive properties.

References

[1] Garrick Brazil, Gerard Pons-Moll, Xiaoming Liu, and Bernt Schiele. Kinematic 3d object detection in monocular video. arXiv preprint arXiv:2007.09548, 2020.
[2] Yu-Wei Chao, Yunfan Liu, Xieyang Liu, Huaiy Zeng, and Jia Deng. Learning to detect human-object interactions. In WACV, 2018.
[3] 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.
[4] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild. In NeurIPS, 2016.
[5] Weifeng Chen, Shengyi Qian, David Fan, Noriyuki Kojima, Max Hamilton, and Jia Deng. Oasis: A large-scale dataset for single image 3d in the wild. In CVPR, 2020.
[6] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In CVPR, 2016.
[7] Yi-Ting Chen, Xiaokai Liu, and Ming-Hsuan Yang. Multi-instance object segmentation with occlusion handling. In CVPR, 2015.
[8] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In CVPR, 2017.
[9] Vincent Delaitre, Josef Sivic, and Ivan Laptev. Learning person-object interactions for action recognition in still images. In NeurIPS, 2011.
[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In CVPR. Ieee, 2009.
[11] Zhuo Deng and Longin Jan Latecki. Amodal detection of 3d objects: Inferring 3d bounding boxes from 2d ones in rgb-depth images. In CVPR, 2017.
[12] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In ICCV, 2015.
[13] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. In CVPR, 2011.
[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In CVPR, 2012.
[15] Georgia Gkioxari, Ross Girshick, Piotr Dollár, and Kaiming He. Detecting and recognizing human-object interactions. In CVPR, 2018.
[16] Ankit Goyal, Kaiyu Yang, Dawei Yang, and Jia Deng. Rel3d: A minimally contrastive benchmark for grounding spatial relations in 3d. Advances in Neural Information Processing Systems 33 pre-proceedings (NeurIPS), 2020.
[17] Abhinav Gupta, Aniruddha Kembhavi, and Larry S Davis. Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 31(10):1775–1789, 2009.
[18] Christian Hane, Lubor Ladicky, and Marc Pollefeys. Direction matters: Depth estimation with a surface normal classifier. In CVPR, 2015. 2, 5

[19] Derek Hoiem, Alexei A Efros, and Martial Hebert. Recovering occlusion boundaries from an image. International Journal of Computer Vision, 91(3):328–346, 2011. 5

[20] Edward Hsiao and Martial Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. IEEE Transactions on Pattern Analysis and Machine Intelligence, 36(9):1803–1815, 2014. 1, 5

[21] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In CVPR, 2019. 1, 5

[22] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single rgb image via perspective points. In NeurIPS, 2019. 5

[23] Zhaoyin Jia, Andrew Gallagher, Yao-Jen Chang, and Tsuhan Chen. A learning-based framework for depth ordering. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 294–301. IEEE, 2012. 1, 5

[24] Ziyu Jiang, Buyu Liu, Samuel Schulter, Zhangyang Wang, and Manmohan Chandraker. Peek-a-boo: Occlusion reasoning in indoor scenes with plane representations. In CVPR, 2020. 5

[25] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jiten-Dra Malik. Amodal completion and size constancy in natural scenes. In ICCV, 2015. 1, 5

[26] Keizo Kato, Yin Li, and Abhinav Gupta. Compositional learning for human object interaction. In ECCV, 2018. 5

[27] Ranjay Krishna, Yue Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision, 123(1):32–73, 2017. 2, 4, 5

[28] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. arXiv preprint arXiv:1811.00982, 2018. 2, 3, 4, 5, 7

[29] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Mallocci, Alexander Kolesnikov, et al. The open images dataset v4. IJCV, pages 1–26, 2020. 1, 2

[30] Lubor Ladicky, Jianbo Shi, and Marc Pollefeys. Pulling things out of perspective. In CVPR, 2014. 2, 5

[31] Katrin Lasinger, René Ranftl, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341, 2019. 2, 5, 6, 7

[32] Buyu Li, Wanli Ouyang, Lu Sheng, Xingyu Zeng, and Xiaogang Wang. G3d: An efficient 3d object detection framework for autonomous driving. In CVPR, 2019. 5

[33] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In CVPR, 2015. 2, 5

[34] Ke Li and Jitendra Malik. Amodal instance segmentation. In ECCV, 2016. 1, 5

[35] Yikang Li, Wanli Ouyang, and Xiaogang Wang. Vip-cnn: A visual phrase reasoning convolutional neural network for visual relationship detection. arXiv preprint arXiv:1702.07191, 2.17. 2017. 5, 7, 12

[36] Xiaoan Liang, Lisa Lee, and Eric P Xing. Deep variation-structured reinforcement learning for visual relationship and attribute detection. In CVPR, 2017. 5

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

[38] Beyang Liu, Stephen Gould, and Daphne Koller. Single image depth estimation from predicted semantic labels. In CVPR, 2010. 2, 5

[39] Fayao Liu, Chunhua Shen, and Guosheng Lin. Deep convolutional neural fields for depth estimation from a single image. In CVPR, 2015. 2, 5

[40] Lijie Liu, Jiwen Lu, Chunjing Xu,Qi Tian, and Jie Zhou. Deep fitting degree scoring network for monocular 3d object detection. In CVPR, 2019. 5

[41] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In ECCV, 2016. 2, 4, 5

[42] Rui Lu, Feng Xue, Menghan Zhou, Anlong Ming, and Yu Zhou. Occlusion-shared and feature-separated network for occlusion relationship reasoning. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10343–10352, 2019. 5

[43] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and HansPeter Graf. Attend and interact: Higher-order object interactions for video understanding. In CVPR, 2018. 5

[44] Xinzhu Ma, Zhihui Wang, Haojie Li, Pengbo Zhang, Wanli Ouyang, and Xin Fan. Accurate monocular 3d object detection via color-embedded 3d reconstruction for autonomous driving. In ICCV, 2019. 5

[45] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Alexander G Schwing. Sail-vos: Semantic amodal instance segmentation with kins dataset. In CVPR, 2019. 1, 5

[46] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Open images dataset v4: Unified image classification, object detection and geometry. In CVPR, 2018. 5

[47] Julia Peyre, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Weakly-supervised learning of visual relations. In ICCV, 2017. 4, 5

[48] Bryan A Plummer, Arun Mallya, Christopher M Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In ICCV, 2017. 5

[49] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In CVPR, 2019. 1, 5

[50] Siyuan Qi, Wenguan Wang, Baoxiong Jia, Jianbing Shen, and Song-Chun Zhu. Learning human-object interactions by graph parsing neural networks. In ECCV, 2018. 5
Appendices

We supplement the main text by the following materials.

Appendix A provides dataset construction details.

Appendix B provides the distributions of object classes in our dataset.
Appendix C analyzes the dataset’s potential bias in terms of the occlusion relationships.

Appendix D describes the implementation details of state-of-the-art VRD methods.

Appendix E presents more results evaluated by precision, recall, and F1-score.

Appendix F studies the models’ transferability between the within-image 2.5VRD and across-image 2.5VRD.

Appendix G analyzes model performance against different difficulty scales.

Appendix H analyzes model performance against the objects’ locations in an image.

Appendix I qualitatively compares the model’s predictions with the groundtruth labels.

If not mentioned specifically, we use the MLP baseline with all features in the analysis.

A. Dataset construction

This section provides the extended description of the dataset construction process presented in Section 2.2. In particular, we describe the data filtering process in more detail.

We randomly sample 110,894 images from the Open Image Dataset (OID) V4. These images all have a Creative Commons Attribution license. Most of them are scenery images from Flickr, each containing multiple objects and/or people. We maintain the original train/validation/test split of the images and use the annotated bounding boxes and 600 class names in the OID images.

As the annotations in OID are over-complete (i.e., multiple bounding boxes for an object), we filter the boxes before collecting 2.5VRD labels. We remove the boxes of human body parts and clothing if the person box is available. Next, we discard the object pairs whose two boxes are highly overlapped (intersection-over-union is greater than 0.7). In addition, we exclude the pairs of one object being part of the other (e.g., auto part and vehicle). To avoid ill-defined cases, we ignore extremely small or big boxes (occupying less than 2% or more than 70% area of the image). Finally, we remove any box that contains not one object, but a group of objects using the original OID label.

B. Object class distribution

Figure 8 and Figure 9 show the sizes of top-100 single object classes and the distribution of top-100 pairs of object classes, respectively. See Section 2.2 on how we process bounding boxes of object classes to arrive at these distributions. The most frequent object classes and pairs are human-centric — they are about people or the objects with which people interact the most, indicating that the dataset is a fair representation of our daily scenes.

C. Potential bias of occlusion relationships

In Section 2.3, we discuss the dataset’s potential bias of the depth relationships between objects. Similarly, we focus on the within-image scenario and investigate occlusion labels. Figure 10 shows the top six object classes (and object pairs) with the highest percentage for each label. We also observe a bias. For instance, objects that “interact” with human body parts, such as musical instruments (guitar, cello), vehicles (motorcycle, bike), rifle, or camera, tend to be part of “Mutual occlusion.” We also observe that salient, often small objects (keyboard, laptop, camera, coffee cup, butterfly, bee, elephant, sculpture, elephant, train) tend to be occluded by less salient, bigger objects or stuff (couch, bed, tree, boat, airplane, building), or human/body parts (face, man). Finally, objects that are likely to be part of “No occlusion” are window, door, picture frame, face, and tie. We keep them as is, rather than correct them, as they are a result of the natural prior of the visual world and daily scenes.

D. State-of-the-art VRD methods

We implement state-of-the-art VRD methods based on the implementation released with the SpatialSense dataset[68]. For fair comparison, we use the same appearance feature as the MLP baselines, i.e., a Faster-RCNN pre-trained on OID. For location features, we encode the object bounding boxes using binary masks following DRNet[8]. For word vector embedding, we learn the word embedding for OID classes end-to-end without pre-training. The appearance feature is used in all four methods, i.e., ViP-CNN[35], PPR-FCN[81], DRNet[8], and VTransE[74], and the location and word vector embedding features are used in DRNet and VTransE. All methods are trained using the same setup as the MLP baselines. We verify our implementations on the SpatialSense dataset and achieve comparable accuracy as that reported in the SpatialSense paper (e.g., 71.3% vs. 71.0% for DRNet).

E. Overall results

In this section, we show additional results expanded from Table 4. We first show the corresponding precision and recall in Table 6. The results show that combining multiple features improves both precision and recall consistently. Next, we show the F1-score for each predicate in Table 7. The results show that the models are not fully symmetric, i.e., $o_a$ is closer than $o_b$ does not always imply $o_b$ is further than $o_a$. Also, we can see that the appearance feature is important for occlusion prediction, especially for mutually occluded cases.

---

2[https://github.com/princeton-vl/SpatialSense](https://github.com/princeton-vl/SpatialSense)
Figure 8: The distribution of top-100 object class labels, sorted by the log of frequency.
Figure 9: The distribution of top-100 pairs of object class labels, sorted by the log of frequency.
Figure 10: Distributions of occlusion labels with respect to object classes (Y-axis: Object A) and object pairs (Y-axis: (Object A, Object B)), respectively.

Table 6: Precision and recall of rule-based models (top part of the table), our approach with different visual cues (middle of the table; B: bounding box feature, C: object class feature, D: depth feature, A: appearance feature), and existing methods (bottom part of the table).

F. Model transferability

In our previous experiments, we train separate models for within-image and across-image 2.5VRD, though the models share exactly the same architecture. However, we can unify them into one. To evaluate how well a model generalizes across the two settings, we test the models’ transferability across within-image and across-image relationships.

Table 8 shows the results. Not surprisingly, the model performance degrades when it transfers from the within-image sub-task to the across-image sub-task, and vice versa. The unified model, which is trained by pooling the in-image and across-image training examples, is in between of the other models. The results show that 2.5VRD models do not fulfill the desired property of transferability, which raises the need for further development of models and/or learning algorithms.

Figure 11: Model performance w.r.t. objects’ horizontal locations for the \((o_a, \text{ is closer than, } o_b)\) relationship.
Table 7: Predicate-wise 2.5VRD results of rule-based models (top part of the table), our approach with different visual cues (middle; B: bounding box feature, C: object class feature, D: depth feature, A: appearance feature), and existing method (bottom).

| Rule: Object class | Within Image | Occlusion | Across Images |
|--------------------|--------------|-----------|--------------|
| A closer           | 0.011        | 0.178     | 0.023        |
| B closer           | 0.010        | 0.100     | 0.027        |
| Same distance      | 0.015        | 0.064     | 0.100        |
| No occl.           | 0.005        | 0.000     | 0.225        |
| A occludes B       | 0.039        | 0.171     | 0.218        |
| B occludes A       | 0.171        | 0.000     | 0.256        |
| Mutual             | 0.000        | 0.000     | 0.243        |

| Rule: Location     | Within Image | Occlusion | Across Images |
|--------------------|--------------|-----------|--------------|
| Rule: Size         | 0.324        | 0.371     | 0.320        |
| Rule: Depth        | 0.325        | 0.371     | 0.317        |

MLP: B           | 0.217        | 0.343     | 0.235        |
MLP: B+C         | 0.261        | 0.344     | 0.319        |
MLP: B+D         | 0.000        | 0.349     | 0.328        |
MLP: B+A         | 0.000        | 0.203     | 0.371        |
MLP: B+C+D+A     | 0.298        | 0.211     | 0.372        |
ViP-CNN           | 0.343        | 0.360     | 0.343        |
PPR-FCN           | 0.344        | 0.239     | 0.245        |
DRNet             | 0.344        | 0.228     | 0.224        |
VTransE           | 0.342        | 0.259     | 0.274        |

Table 8: Model transferability across 2.5VRD sub-tasks.

|          | Within-Image | Across-Image |
|----------|--------------|--------------|
| Within-Image | 0.322        | 0.308        |
| Across-Image  | 0.304        | 0.370        |

Table 9: 2.5VRD results of various difficulty scales.

|                | Within Image | Occlusion | Across Images | Average |
|----------------|--------------|-----------|---------------|---------|
| Easy           | 0.886        | 0.821     | 0.950         | 0.886   |
| Moderate       | 0.644        | 0.781     | 0.833         | 0.753   |
| Difficult      | 0.483        | 0.781     | 0.668         | 0.644   |

G. Results of various difficulty scales

This section shows the model performance at different difficulty scales. The results are shown in Table 9. Note that the difficulties are defined only on annotated objects, so we use the groundtruth objects in this experiment (i.e., predicate prediction in Table 5). The model’s performance aligns very well with the human raters’ assessments about the examples’ difficulty scales.

H. Object location distribution

Figure 6 shows that the model’s accuracy correlates with the objects’ vertical positions in an image. In contrast, Figure 11 shows that the model performance is not sensitive to the objects’ horizontal position. The results are consistent with our observation that an object’s depth is highly correlated with the Y-coordinate of the object center.

I. Qualitative results

In this section, we present qualitative results of the MLP baseline. Figure 12 shows examples for within-image depth relationships with different difficulty scales, and Figure 13 and Figure 14 are about examples for within-image occlusion and across-image depth, respectively. We can clearly see the increasing ambiguity in different difficulty levels. The models manage to differentiate the objects’ relative depths for these examples. We also show failure examples in Figure 15, Figure 16, Figure 17, and Figure 18. We can see that the labels may depend on minor differences in objects’ depths, and the models’ predictions are reasonable despite that they do not match the groundtruth labels, e.g., the last two examples in Figure 15.
Figure 12: Qualitative examples for within-image depth prediction with different difficulties (groundtruth = predicted labels).
Figure 13: Qualitative examples for within-image occlusion prediction (groundtruth = predicted labels).
Figure 14: Qualitative examples for across-image depth prediction (groundtruth = predicted labels).
Figure 15: Failure cases for within-image depth prediction. This figure show examples where the model correctly detects the objects but predicts wrong predicates.

Figure 16: Failure cases for within-image depth prediction. This figure show examples where the model fails to detect the objects.
Figure 17: Failure examples for occlusion prediction.

Figure 18: Failure examples for across-image depth prediction.