ReferIt3D: Neural Listeners for Fine-Grained 3D Object Identification in Real-World Scenes

Panos Achlioptas[1], Ahmed Abdelreheem[2], Fei Xia[1],
Mohamed Elhoseiny[1,2], Leonidas Guibas[1]

[1] Stanford University
[2] King Abdullah University of Science and Technology
[1]{panos, feixia,elhoseiny,guibas}@cs.stanford.edu,
[2]{ahmed.abdelreheem, mohamed.elhoseiny}@kaust.edu.sa

Abstract. In this work we study the problem of using referential language to identify common objects in real-world 3D scenes. We focus on a challenging setup where the referred object belongs to a *fine-grained* object class and the underlying scene contains *multiple* object instances of that class. Due to the scarcity and unsuitability of existent 3D-oriented linguistic resources for this task, we first develop two large-scale and complementary visio-linguistic datasets: i) **Sr3D**, which contains 83.5K template-based utterances leveraging *spatial relations* among fine-grained object classes to localize a referred object in a scene, and ii) **Nr3D** which contains 41.5K *natural, free-form*, utterances collected by deploying a 2-player object reference game in 3D scenes. Using utterances of either datasets, human listeners can recognize the referred object with high (>86%, 92% resp.) accuracy. By tapping on this data, we develop novel neural listeners that can comprehend object-centric natural language and identify the referred object *directly* in a 3D scene. Our key technical contribution is designing an approach for combining linguistic and geometric information (in the form of 3D point clouds) and creating multi-modal (3D) neural listeners. We also show that architectures which promote object-to-object communication via graph neural networks outperform less context-aware alternatives, and that fine-grained object classification is a bottleneck for language-assisted 3D object identification.

1 Introduction

The progress on connecting language and vision in the past decade has rekindled interest in tasks like visual question answering (e.g., [12,54]), image captioning (e.g., [28,63,68,41,6]), and sentence-to-image similarity (e.g., [28,31]). Recent works have enhanced the accessibility of visual content through language via grounding (e.g., [49,48]), showing strong results in locating linguistically described visual elements in images. However, most of these works focus on developing better models that connect vision to language in images, which express after all only a 2D view of our 3D reality. Even in embodied AI most works (e.g., embodied QA [21], or embodied visual recognition [69]), fine-grained 3D object identification is not explicitly modeled. Fine-grained 3D understanding however

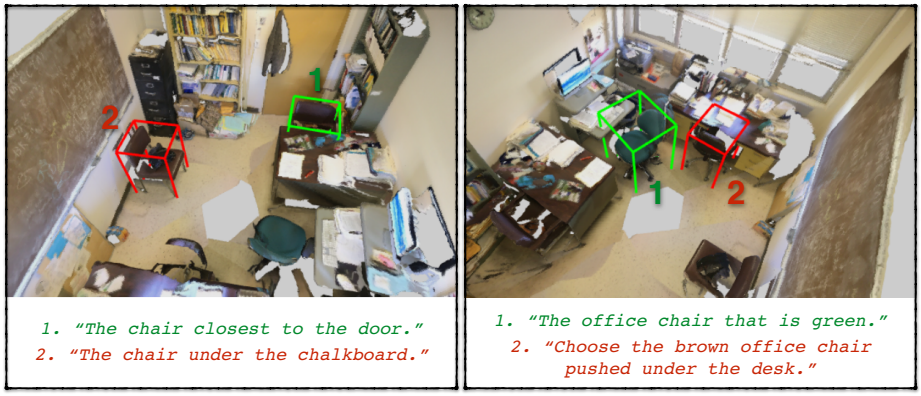

Fig. 1. Examples of natural free-form utterances. Each color-coded utterance distinguishes the corresponding object (marked with same color) against a distracting object in the underlying scene; contrasting two simple chairs (left) and two office-chairs (right). The use of a specific contrasting context inside a scene (as delineated by the bounding boxes surrounding *all and only those* objects of the same fine-grained class) fosters the production of *discriminative* language and lifts the reference problem beyond fine-grained classification.

can be essential to more 3D-oriented visually-grounded embodied tasks, such as those that need to be performed by autonomous robotic agents [66,57].

Humans possess an astonishing capacity to reason about, describe, and locate 3D objects. Over time we have developed efficient communication protocols to linguistically express such processes – e.g., given the utterance *"the laptop placed on the table next to the main door"*, one can identify the referred object in the room, as long as the reference conveys some unique aspect of that object. Solving such a reference problem *directly* in 3D space – i.e., *without* a camera view dependency – can benefit many downstream robotics applications, including embodied question answering [21], visual- and language-based navigation [9], instruction following [58], and manipulating objects in a scene [67,37]. Despite this, developing *datasets* and *methods* with characteristics that enable machine learning models to perform well on this 3D reference task is far from straightforward; in this work, we examine how to address both.

Leveraging 3D visual understanding for solving vision and language tasks has been recently explored in Visual Question Answering [33] and Visual Grounding [50]. Still, the focus has been on synthetic datasets without 3D understanding going beyond (at best) multiple 2D views. An alternative, yet more direct way to gain this understanding is by analyzing point cloud data of real-world scenes [3,52]. Point clouds carry the entire geometric and appearance characteristics of objects and provide access to a larger spatial context (within a scene) than a single 2D view [13]. This flexibility enables us also to bypass camera view dependency (e.g., having access to parts of a scene occluded by a fixed camera) when we refer linguistically to objects.

In this paper, we investigate object references when multiple instances of the same fine-grained object class are present in a 3D scene. Discriminative understanding of object classes is important at the fine-grained level and can be achieved with models combining appearance understanding and spatial reasoning skills (e.g., spatial understanding is not critical by itself if we are looking for the unique *office* chair in the presence of one or more *dining* chair(s)). ***Creating discriminative linguistic descriptions:*** We focus on designing a data collection strategy that covers *both* spatial and appearance based identification (Sec. 3). As we show in our experiments, this step is critical for progress in 3D visual object identification from free-form language descriptions. Our strategy involves both synthetic and human-based generation of utterances, and has the following characteristics: (i) in every single example there are multiple object instances of the same class referred in the language describing the target 3D object; (ii) in the case of human language utterances, we explicitly ask the human subject to describe a target object in contrast to other instances of the *same* object class. By explicitly contrasting the same fine-grained class instances *and only them,* the resulting utterances are discriminative, even if uttered by crowd-sourced annotators unfamiliar with the environment. Fig. 1 illustrates examples of this. ***Developing a 3D neural listener:*** We also design a novel visio-linguistic graph-convolution network that predicts the referred object given a language description, by enabling communication among objects in a 3D visual scene. Our contributions can be summarized below:

1. **Fine-Grained ReferIt3D** task: We introduce the task of language-based identification of specific 3D object instances, where fine-grained object-centric and multi-object understanding is necessary for its completion.
2. **Nr3D** and **Sr3D** datasets: We contribute a new dataset that contains natural and synthetic language descriptions, namely Nr3D and Sr3D respectively. For Sr3D we propose a simple but effective methodology for building template-based and *spatially-oriented* object referential language in 3D scenes. We show that training with Sr3D in addition to natural language data (Nr3D or [18]) improves neural-based pipelines.
3. **ReferIt3DNet**: We explore the task of understanding object references grounded in real-world 3D data (including both language and scenes) by designing a novel visio-linguistic graph neural network, termed *ReferIt3DNet*[3].

2 Related Work

2D High-Level Vision & **Language:** Vision & Language, also sometimes called Visual Semantic modeling, has been extensively studied in a variety of 2D tasks. Among early approaches of combining Vision & Semantics are tasks such as zero-shot learning where language/unseen descriptions of an unseen class are provided to describe it (e.g., [70,35,7,26,55,59,25,39,24,38,74,61]). Similar approaches have been developed to model image-sentence similarity for bi-directional retrieval of images given a sentence (e.g., [28,31]). More recently, the

[3] The datasets and neural listener code are available at https://referit3d.github.io

development of a large scale dataset of 2D Visual Question answering (VQA) [12] enabled new approaches on how to best represent question and images for this task. However, a huge language bias was revealed: just by looking at the language and without necessarily understanding the visual content, the predictive performance of the right answer is high [5]. The same bias was shown in tasks such as image-captioning [22]. More balanced VQA benchmarks [27] mitigated some of the biases and motivated the development of better attention mechanisms (e.g., [30]) and modular networks [71,11]. As per the example of the *2D Vision and Language* community, properly modeling 3D visio-lingual tasks requires establishing carefully designed connections between language and the 3D visual data, encoded with point clouds.

2D ReferIt Game and Grounded Vision & **Language:** Several papers explore connecting referential language to image regions for co-reference resolution (e.g., [15,53,42] – in videos, e.g., [8]), for generating referring expressions [29,44,42], and more. Recent work grounds noun phrases in image captions, such as in Flickr30KEntities [49] and ActivityNetEntities [73] in videos. In [23,43], the authors proposed the use of referring expressions for human-robot interaction and object localization in real-word environments but using primarily 2D images in contrast to our work.

Visual Relationships and Spatial Reasoning: Detecting visual relationships in images such as `<woman, carrying, umbrella>` (e.g., [40,72,1]) has been explored using datasets such as VRD [40] and more recently on the large Visual Genome dataset [32]. Spatial relations have also been studied in 3D by Rosema *et al.* [56]. However, relations in that work are not described in free form and hence are of restricted vocabulary. Also, the goal in [56] is simpler than identifying a target object in a complex 3D environment (our goal).

3D Vision & **Language:** Connecting 3D vision to natural language is relatively understudied. From a generative angle, [16] presented conditional generation of 3D models from text, which could be useful in augmented reality applications. In a concurrent work [18], Chen *et al.*, collected natural language to localize/discover referred objects in 3D scenes. In contrast, we assume that we are given the segmented object instances in a room and focus on identifying a referred object among instances of *the same* fine-grained category.

ReferIt 3D Game. Our 41,503 human language utterances were collected via a reference game between two humans, as inspired by 2D ReferItGame [29] and ShapeGlot [3]. The basic arrangement of such games can be traced back to the language games explored by Wittgenstein [65] and Lewis [36]. Recently, these approaches have also been adopted as a benchmark for discriminative and context-aware NLP [47,10,46,19,62,60,34,4]. Our paper goes beyond this prior work by grounding language behavior in a reference task containing objects in complex (real-world) 3D scenes, thereby eliciting compositional spatial and color/shape-oriented language.

3 Developing Referential 3D-Centric Data

The problem of language driven disambiguation of common objects in real world 3D scenes is new, and as such, not many datasets exist that are well suited for this task. With this in mind we introduce a two-part dataset: a high quality synthetic dataset of referential utterances (**Sr3D**) and a dataset with natural (human) referential utterances (**Nr3D**). Both Sr3D and Nr3D are built on top of ScanNet [20], a real-world 3D scene dataset with extensive semantic annotations that we utilize to create appropriate contrastive *communication contexts*. We define *communication context* as a `(scene, target, distractor(s))` tuple, where *scene* is one of the 707 unique indoor scenes of ScanNet, *target* is one of 76 fine-grained object classes (e.g., office-chairs, armchairs, etc.), and *distractors* are instances of the same fine-grained object class as the target that are contained in the same scene. We generate a total of 5878 unique tuples. We select the 76 object classes by applying the following intuitive criteria. A class is a valid class for a *target* if: (a) it is contained in at least 5 *scenes*; and (b) each *scene* contains multiple *distractors* but not more than six (to promote a problem beyond fine-grained (FG) classification without making it too hard even for human annotators). We add the constraint of having 5 such *scenes* per class, to foster generalization and make the problem less heavy-tailed (15.26% of all annotated ScanNet classes appear with multiple instances in exactly one *scene*). We also exclude the few classes that are object parts (e.g. a door of a closet) or are structural elements of the scenes (i.e., walls, floors, and ceiling) to ensure that we are working with common objects.

3.1 Creating Template Based Spatial References

We introduce the **Spatial Reference in 3D** (**Sr3D**) dataset, consisting of 83,572 utterances. Each utterance aims to uniquely refer to a *target* object in a ScanNet 3D scene by defining a relationship between the *target* and a surrounding object (*anchor*). *Anchors* are object instances that can belong to a set of 100 object classes in ScanNet, comprising of the 76 mentioned above and an additional 24 that: (a) frequently appear as singletons in a scene; and (b) are large objects (e.g., a fireplace or a TV). However, an *anchor* can never belong to the same class as the *target* and, as such, its *distractors*.

Consider, for instance, an underlying 3D scene with a *target* object (e.g. desk) that can be completely disambiguated from its *distractors* with the help of a spatial relation (e.g., closest) to an *anchor* object (e.g., door). We synthesize discriminative **Sr3D** utterances using the following compositional template:

$$< \text{target-class} > \quad < \text{spatial-relation} > \quad < \text{anchor-class(es)} > \tag{1}$$

e.g., *"the desk that is closest to the door"*. Per (1), the **Sr3D** template consists of three placeholders. Our goal is to find combinations of them that can uniquely characterize target objects among their distractors in their scenes.

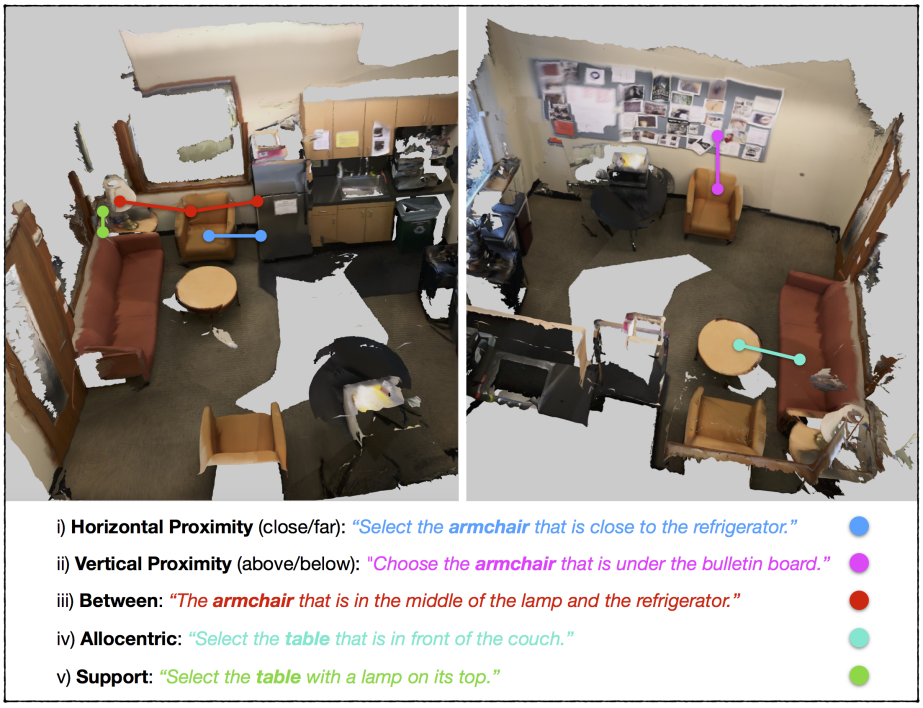

Fig. 2. Examples of spatial reference types of **Sr3D**. In the left image, there are examples of "horizontal proximity", "between", and "support" relations; the target object in the first two is an armchair and the target in the third relation is a table. In the right image, there are examples of "vertical proximity" and "allocentric" relations; the target objects are an armchair and a table respectively. The left and the right images represent a ScanNet scene where there exist two armchairs (one is beside the refrigerator and the other under the bulletin board) and three tables (a black one under the bulletin board, one in front of the couch, and one in the corner of the room).

We define the following five types of spatial object-to-object relations. For more details we refer the reader to Table 1 for a summary of statistics and to the Supplementary Material [2].

(i) **Horizontal Proximity:** This type indicates how close/far is a *target* from the *anchors* in the scene (Fig. 2, i). It applies to distance on the horizontal placement of the objects.
(ii) **Vertical Proximity:** It indicates that the *target* is either above or below the *anchor* (Fig. 2, ii).
(iii) **Between:** Between relations indicate the existence of a *target* between two *anchors* (Fig. 2, iii).
(iv) **Allocentric:** Allocentric relations encode information about the location of the *target* with respect to the intrinsic self-orientation of an *anchor* (Fig. 2, iv). To define the aforementioned orientation, we need to know; (a) the ori-

entated bounding boxes of the *anchor* and (b) whether the *anchor* has an intrinsic front (e.g., a chair with a back) or not (e.g., a stool). For (a) we utilized the Scan2CAD [14] annotations that provide 9DOF alignments between ShapeNet models and ScanNet objects, and for (b) we used a combination of PartNet's [45] and manual annotations.

(v) **Support:** Support relations indicate that the *target* is either supported by or supporting the *anchor* (Fig. 2, v).

Table 1. Statistics of Sr3D. The first row contains the number of *distinct* communication contexts yielded by each reference-type. The second row contains the number of synthetically generated utterances. *Please note that communication context in the Sr3D setup also takes into account spatial relationships and anchors.*

Relationship	Horizontal Prox.	Vertical Prox.	Support	Allocentric	Between	All
|Context|	34001	1589	747	1880	3569	41786
|Utterances|	68002	3178	1494	3760	7138	83572

Discussion Our protocol for generating Sr3D is *simple* but also **effective**:

- A user study conducted in Amazon Mechanical Turk (AMT) revealed that 86.1% of the time, humans guessed correctly the target when provided with a sampled utterance of **Sr3D** (2K samples, $p<0.001$).
- As shown in Sec. 5, **Sr3D** allows us to investigate the reference problem in a more controlled manner than **Nr3D**, by providing a homogeneous vocabulary and a specific type of reasoning. For example, it bypasses color- or shape- based reference, and other complicated factual reasoning (e.g., use of brand names or metaphors).

Sr3D+: In addition to the dataset generation described above, we augment Sr3D with more utterances choosing the target object's class among those that do not comply with the criterion of having more than one *distractors* in the scene. Given the synthetic nature of the data, we can generate a large amount of utterances in a cost-free way. We explore the contribution of Sr3D+ to the final performance of our neural listener in Sec. 5. This additional set of data will be particularly useful when comparing our method to the *Unique* setting of [18] (Table 4), since it assumes that the target object is the only instance of that class in the scene.

3.2 Natural Reference in 3D Scenes

The Natural Reference in 3D (**Nr3D**) dataset contains 41,503 human utterances collected by deploying an online reference game in AMT. The game is played between two humans: a 'speaker' who was asked to describe a designated *target* object in a ScanNet 3D scene and a 'listener' who, given the speaker's utterance,

was asked to select the referred object among its *distractors*. The game is structured such that both 'speaker' and 'listener' are rewarded when the *target* is successfully selected, hence incentivising descriptions that are most discriminative in the context of a scene and a general audience.

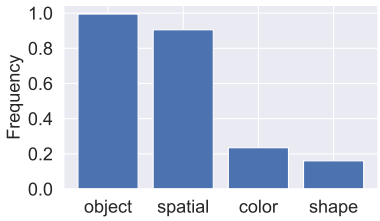

Fig. 3. Vocabulary histogram.

Both players are shown the same 3D scene in the form of a decimated mesh model and can interact with it through a 3D interface. In order to remove any camera view bias, we initialize the 'speaker' and 'listener' 3D interfaces with different randomized camera parameters. Given the specifics of the task and the difficulty of understanding the depicted 3D real-world visual content by the non-expert players, we highlight with bounding boxes (oriented when available) the *target* and *distractors*. We distinguish them for the 'speaker' with red and green color respectively, whereas for the 'listener' there is no distinction among them. To encourage players to explore the scene and familiarize themselves with all highlighted objects, we also provide them with a total count of bounding boxes they should expect in the scene. For an example of the speaker's interface we refer the reader to Fig. 1 in Supplementary Material [2].

We collect at least 7 utterances from different player pairs per target object. During the collection process, we iterate over all object instances with the same fine-grained object class in a scene (e.g., all 6 sofa chairs Fig. 1 in Supplementary Material [2]), providing to the dataset a symmetric property. Among the collected utterances, some originate from games with unsuccessful results; these are not used for training/learning purposes.

Discussion Before presenting our neural agents, we identify several important properties of Nr3D:

- Performance in the gamified data collection process was high (92.2%), but 'listeners' made significantly more errors in the more challenging "hard" contexts (90.0% vs. 94.7%, z = 17.5, p<0.001). We define "hard" contexts as those 3D scenes that contain *more than* 2 *distractors* (Fig. 5 illustrates examples of "hard" vs. "easy").
- Speakers naturally produced longer utterances on average to describe targets in hard contexts (approximately 12.5 words vs 10.2, t=-35, p<0.001). The average number of words across all utterances (ignoring punctuation) is 11.4 and the median is 10.
- Regardless of the context difficulty, we identified two attributes in the descriptive power of the utterances (Fig. 5): (a) the *target* is **scene-discoverable** when it is uniquely distinguishable among objects in the entire scene and not only its *distractors*. The majority of the utterances mention the fine-grained class type or a close synonym of the *target* (91.6%). This *naturally emerging*

property of Nr3D allows us to identify the target among *all* objects in the room; and (b) the identification of the *target* is **view-independent** thus not requiring the observer to place themselves into the scene facing certain objects. Although this attribute is not as prominent as the previous one (63%), even in the case that there is view dependency, speakers were instructed to guide the listeners on how to place themselves in the scene.
– The use of spatial prepositions is ubiquitous (90.5%), which exemplifies why Sr3D is relevant. Reference to color and shape properties is drastically less used in distinguishing instances of the same fine-grained object class. Fig. 3 is a frequency-histogram of the different types of language use.

4 Developing 3D Neural Listeners

Given a 3D scene $\mathcal{S}$ represented as an RGB-colored point-cloud of N points $\mathcal{S} \in \mathbb{R}^{N\times 6}$ and a word-tokenized utterance $U = \{u_1, \ldots, u_t\}$ we want to build a neural listener that can identify the referred target object $\mathcal{T} \subset \mathcal{S}$. To this end, we assume access to a partition of $\mathcal{S} = \{O_1, \ldots, O_M\}$ that represents the objects (O_i) present in $\mathcal{S}$. While it is feasible to attempt identification (or more precisely, in this case, object localization) by operating directly on the unstructured $\mathcal{S}$ ([51], [18]), the problem of instance localization (*especially* for FG classes) remains vastly unsolved. To overcome this and decouple the 3D instance-segmentation problem from our referential setting, we assume access to the instance-level object segmentations of the underlying scene. This choice allows us to cast the 3D reference problem into a classification problem that aims to predict the referred "target" among M segmented 3D instances.

While the above assumption eliminates the need to define each object in $\mathcal{S}$, it still leaves open the problems of: (i) FG object classification; (ii) recognition of the referred object class (per the utterance); and (iii) the original problem of selecting the referred object among the m options. For the first two tasks, we experiment with a neural listener that utilizes two auxiliary cross-entropy losses ($\mathcal{L}_{fg}$, $\mathcal{L}_{text}$) aimed to decouple these intrinsic aspects of the original task. Specifically, the two losses are added to the cross-entropy loss of the main task in hand ($\mathcal{L}_{ref}$) making a final loss that is a weighted sum of these terms:

$$\mathcal{L}_{total} = \alpha_1 \mathcal{L}_{fg} + \alpha_2 \mathcal{L}_{text} + \mathcal{L}_{ref}$$

Contextual scene understanding The above design is object-aware, but our underlying task is also scene-oriented and heavily relies on the *configuration* of the objects present in a scene. Because of this reason it is important to provide a neural listener with a signal that contains explicit information about the scene it operates. A baseline that we explored to this end, is to create a PointNet++ hierarchical scene-feature (based on a large number of points of $\mathcal{S}$) which we fused with every visual representation extracted independently for each object O_i. While the resulting representation is simultaneously object-centric and scene-aware it is not taking into account explicit object-to-object interactions. A more sophisticated approach – which is part of the **ReferIt3DNet** – uses a

structured and explicit way of capturing object-to-object interactions to provide information about the scene. Specifically, we use a dynamic graph-convolutional network (DGCN) [64] that operates on the visual features of the objects present in a scene (the object are nodes of a graph). The edges of this graph are computed dynamically at each layer of the DGCN according to the Euclidean similarity among the updated (per-node) visual features. In our experiments we use the k-nearest neighbor-graph among the nodes ($k = 7$, chosen per validation). We note that $k = 7$ creates a relatively sparse graph (the 90^{th} percentile of the number of objects in the training scenes is 52). For further details we refer the reader to the Supplementary Material [2].

Incorporation of language An important decision regards how one should "fuse" the linguistic signal in a pipeline like the above. Despite a chair being visually different from a door, our graph-network should inspect the relation among these objects, especially when the reference requires it (e.g., "the chair close to the door"). To promote this action we fuse the visual (object) features with the

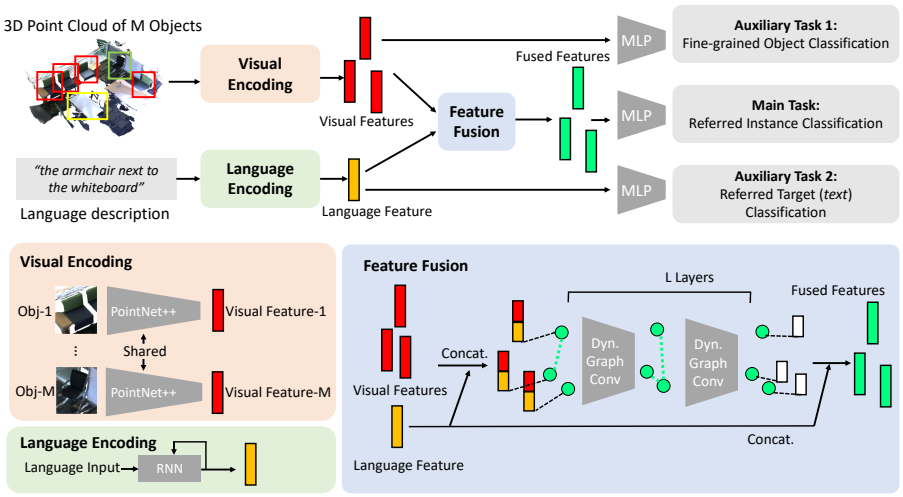

Fig. 4. The *ReferIt3DNet* neural listener. A visual encoder processes (via a shared PointNet++) each 3D object of a given scene that is represented by a 6D point cloud containing its xyz coordinates and RGB color. Simultaneously, the utterance describing the referred object (e.g., "*the armchair next to the whiteboard*") is processed by a Recurrent Neural Network (RNN). The resulting visio-linguistic representations are fused together and processed by a Dynamic Graph Convolution Network (DGCN) which creates an object-centric *and* scene- (context-) aware representation per object. The output of the DGCN is processed by an MLP classifier that estimates for every object its likelihood to be the referred one. Two auxiliary losses modulate the visio-linguistic representations before they are processed by the DGCN via an FG object-class classifier and a referential-text classifier ($\mathcal{L}_{fg}$ and $\mathcal{L}_{text}$ – see text for details).

linguistic ones (derived by an RNN) *before* we pass them to the DGCN. We also explore the effect of adding the linguistic features *after* the DGCN and in both places – which is the best performing option. An overview of our pipeline is illustrated in Fig. 4.

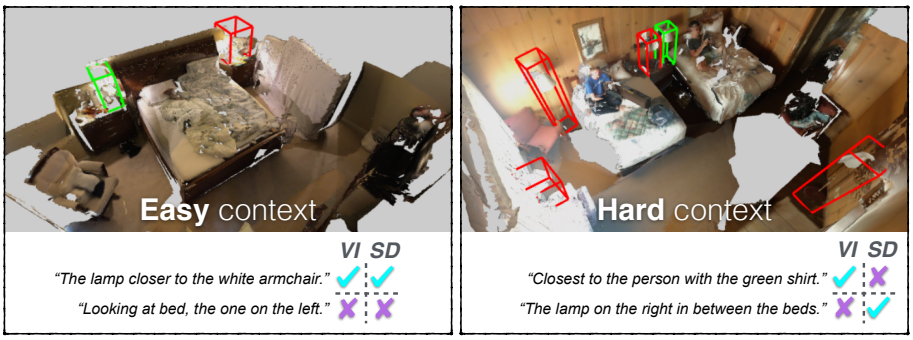

Fig. 5. Easy vs. Hard communication contexts and examples of natural utterances with attributes that affect a navigating/listening agent. **Scene-Discoverable (SD)**: does the utterance explicitly refer to the target's object class (or a synonym), hence permitting object-identification among *all* objects of the scene? **View-Independent (VI)**: Is the description in the utterance view-independent?

5 Experiments and Analysis

We explore different listening architectures [4] and report the listening accuracy; each test utterance receives a binary score (1 if the correct object is predicted as target and 0 otherwise). For all experiments we use the official-ScanNet splits.

1. **Decoupled approach**: This is a baseline listener consisting of a text classifier and an (FG) object classifier that are trained separately. Given an utterance we use the text-clf to predict the referred object-class. Then we select uniformly i.i.d. (and output) an object from $O_i \in \mathcal{S}$ for which the object-clf matches the text-based prediction. We note that in Nr3D (Sr3D) test accuracies for the two classifiers are 93.0% (100.0%) and 64.7% (67.4%), indicating a noticeable asymmetry in the difficulty of solving the two tasks.
2. **Vision + Language, no Context (V + L)**: Inspired by context-free listening architectures like those in [4,46], we ground an RNN with the visual feature of each object of $O_i \in \mathcal{S}$, *independently*, and use a shallow classifier to predict the likelihood of each O_i for being the referred target. This baseline can encode visual properties of an object beyond its FG class enabling rich (context-free) distinctions (e.g.,"very small, or yellow colored chair").

[4] Architecture details and hyper-parameters for all the experiments, are provided in the Supplementary Material [2].

Table 2. ReferIt3DNet performance on Nr3D with/out Sr3D. The first row contains the achieved accuracy on the Nr3D testing data for a listener trained solely with the Nr3D training set; the other rows showcase the effect of training simultaneously with the Sr3D/Sr3D+, respectively.

	Overall	Easy	Hard	View-dep.	View-indep.
Nr3D	35.6%±0.7%	43.6%±0.8%	27.9%±0.7%	32.5%±0.7%	37.1%±0.8%
w/ **Sr3D**	37.2%±0.3%	44.0%±0.6%	**30.6%±0.3%**	**33.3%±0.6%**	39.1%±0.2%
w/ **Sr3D+**	**37.6%±0.4%**	**45.4%±0.6%**	30.0%±0.4%	33.1%±0.5%	**39.8%±0.4%**

Table 3. Listening performance of various ablated models. The first two columns contain the obtained accuracy when no auxiliary losses are used, and the last two the accuracy when these losses are included.

	Nr3D	**Sr3D**	**Nr3D**	**Sr3D**
Aux. classification loss	**No**		**Yes**	
Decoupled	25.45%	31.73%	-	-
V + L	26.12%±0.5%	32.61%±0.4%	26.62%±0.5%	32.98%±0.4%
V + L + C	27.45%±0.6%	34.7%±0.4%	28.51%±0.6%	37.2%±0.4%
ReferIt3DNet-*A*	32.3%±0.3%	**39.7%±0.3%**	33.4%±0.3%	**41.0%±0.3%**
ReferIt3DNet-*B*	31.8%±0.3%	38.1%±0.2%	33.0%±0.3%	40.5%±0.2%
ReferIt3DNet	**32.4%±0.5%**	38.4%±0.2%	**35.6%±0.7%**	39.8%±0.2%

3. **Vision + Language + Holistic Context (V + L + C)**: Similar to the above, but also fuses a PointNet++ scene-feature with each object's visual feature to ground the RNN. This enables the inspection of non-structured context when solving the reference task (PointNet++ is applied on a non-segmented scene point cloud).
4. **Vision + Language + Graph (structured) Context (*ReferIt3DNet*)**: This is our proposed listener and comes in three variants that differ w.r.t. *where* we fuse the linguistic with the visual information.

Neural Listeners. Comparisons for the above models are presented in Table 3. We observe the following main trends[5]: i) using the visual and linguistic auxiliary classification losses improves performance; ii) Simplified language (Sr3D) makes identification easier; iii) scene context matters a lot, but most importantly how we incorporate the context (e.g., via DGCN, or direct fusion of PointNet++) makes an important difference in performance. As expected, a more structured versus a rudimentary representation favors better results; iv) where we fuse language matters as well: ReferIt3DNet-*A* fuses after the DGCN, ReferIt3DNet-*B* before, and the best performing (for Nr3D) model fuses in both places.

The results shown in Fig. 6 show the neural listener's capability to understand and locate objects in challenging 3D scenarios. For example, the top-right example was successful despite the utterance being long. Referring to this particular trashcan among other similar ones requires both spatial reasoning and

[5] In all results mean accuracies and standard errors across 5 random seeds are reported, to control for the point cloud scene sampling.

Table 4. ScanRefer performance with/out Sr3D. MeanIoU improvements when combining Sr3D data with ScanRefer's data during training.

Dataset	Unique		Multiple		Overall	
	P@0.25	**P@0.5**	**P@0.25**	**P@0.5**	**P@0.25**	**P@0.5**
ScanRefer	53.75%	37.47%	21.03%	12.83%	26.44%	16.90%
w/ **Sr3D**	59.99%	39.06%	21.69%	14.33%	28.02%	18.42%
w/ **Sr3D+**	**63.55%**	**42.18%**	**24.12%**	**15.75%**	**30.64%**	**20.12%**

visual 3D understanding. Similar capability can be demonstrated in the two examples in the second row about the door and the cabinets. Finally, the last row shows two challenging failure cases of our model. In the bottom-left example, the utterance has wrongly placed the listener in the room (the chair is at the 3 o'clock position instead of 9). The bottom-right example is particularly hard to solve; the scene is almost symmetric and stripped of visual features, making it hard to discriminate among the chairs.

Combining Nr3D & Sr3D. In Table 2, we observe how combining the two datasets provides a consistent boost in performance. This demonstrates the contribution of adding a synthetically generated dataset to a human one. We get a similar outcome when combining Sr3D to the ScanRefer [18] data (see Table 4). We performed this experiment following the implementation in [17]. Going back to Table 2, the results showcase that our neural listener performs better by a margin in "easy" versus "hard" cases. This is expected and solidifies the understanding that more work needs to be done in discriminatively distinguishing objects when there are multiple distractors in the scene. Another important finding is that view-independent utterances are easier to solve than dependent ones. This does not come as a surprise, since the network has naturally more work to do to comprehend nuances related to viewing the scene w.r.t. another object.

6 Conclusion

Language assisted object disambiguation done directly for 3D objects in 3D environments is a novel but very challenging task. This is especially true when one tries to distinguish among multiple instances of the same fine-grained object category. In addition to the intrinsic difficulty of the problem, there is a scarcity of appropriate datasets. Creating relevant visio-linguistic data that allow us to study this problem is important for advancing 3D deep-learning that, similar to 2D visual learning, is a data hungry methodology. While our neural listeners are a promising first step, more research has to be done before human-level performance and generalization is attained. In summary, this paper has (a) introduced the problem of fine-grained multi-instance 3D object identification in real-world scenes; (b) contributed two relevant public datasets; and (c) explored an array of sensible neural architectures for solving the referential task.

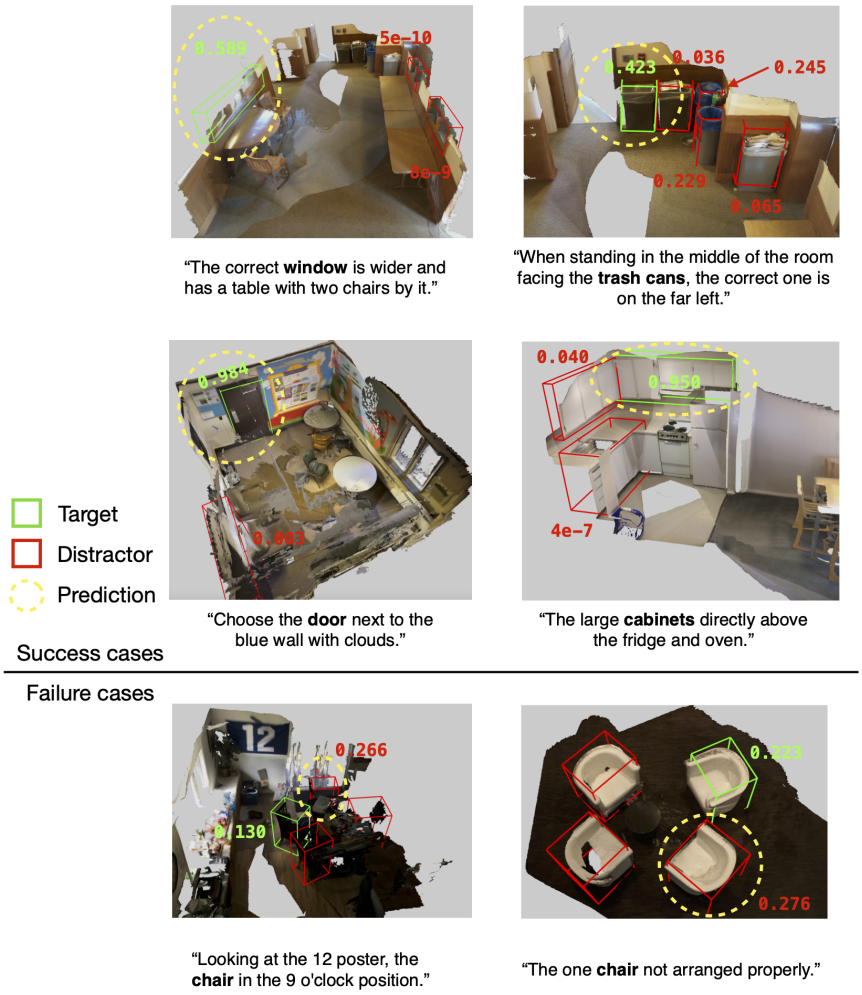

Fig. 6. Qualitative results. Success cases are in the top four images and Failure in the bottom two. Targets are shown in green boxes and distractors in red. The network predictions are shown in a dashed yellow circle, along with the predicted probabilities. Please note that probabilities of inter class distractors are not illustrated here.

Acknowledgments

The authors wish to acknowledge the support of a Vannevar Bush Faculty Fellowship, a grant from the Samsung GRO program and the Stanford SAIL Toyota Research Center, NSF grant IIS-1763268, KAUST grant BAS/1/1685-01-01, and a research gift from Amazon Web Services. Also, they wish to thank Prof. Angel X. Chang for the inspiring discussions regarding the creation of synthetic 3D spatial data, and Iro Armeni and Antonia Saravanou for their help in writing.

References

1. Abdelkarim, S., Achlioptas, P., Huang, J., Li, B., Church, K., Elhoseiny, M.: Long-tail visual relationship recognition with a visiolinguistic hubless loss. CoRR abs/2004.00436 (2020)
2. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Supplementary material for: ReferIt3D: Neural listeners for fine-grained 3d object identification in real world 3d scenes (2020)
3. Achlioptas, P., Diamanti, O., Mitliagkas, I., Guibas, L.: Learning representations and generative models for 3d point clouds. In: International Conference on Machine Learning (ICML) (2018)
4. Achlioptas, P., Fan, J., Hawkins, R.X., Goodman, N.D., Guibas, L.J.: ShapeGlot: Learning language for shape differentiation. In: International Conference on Computer Vision (ICCV) (2019)
5. Agrawal, A., Batra, D., Parikh, D.: Analyzing the behavior of visual question answering models. In: Empirical Methods in Natural Language Processing (EMNLP) (2016)
6. Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., Anderson, P.: nocaps: novel object captioning at scale. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
7. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR (2015)
8. Anayurt, H., Ozyegin, S.A., Cetin, U., Aktas, U., Kalkan, S.: Searching for ambiguous objects in videos using relational referring expressions. CoRR abs/1908.01189 (2019)
9. Anderson, P., Wu, Q., Teney, D., Bruce, J., Johnson, M., Sünderhauf, N., Reid, I., Gould, S., van den Hengel, A.: Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
10. Andreas, J., Klein, D.: Reasoning about pragmatics with neural listeners and speakers. CoRR abs/1604.00562 (2016)
11. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
12. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: International Conference on Computer Vision (ICCV) (2015)
13. Armeni, I., Sener, O., Zamir, A.R., Jiang, H., Brilakis, I., Fischer, M., Savarese, S.: 3d semantic parsing of large-scale indoor spaces. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
14. Avetisyan, A., Dahnert, M., Dai, A., Savva, M., Chang, A.X., Nießner, M.: Scan2cad: Learning cad model alignment in rgb-d scans. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
15. C. Kong, D. Lin, M.B.R.U., Fidler, S.: What are you talking about? text-to-image coreference. In: CVPR) (2014)
16. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2shape: Generating shapes from natural language by learning joint embeddings. CoRR abs/1803.08495 (2018)
17. Chen, Z.D., Chang, A.X., Nießner, M.: `https://github.com/daveredrum/ScanRefer`, accessed: 2020-07-17

18. Chen, Z.D., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. CoRR abs/1912.08830 (2019)
19. Cohn-Gordon, R., Goodman, N., Potts, C.: Pragmatically informative image captioning with character-level inference. CoRR abs/1804.05417 (2018)
20. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
21. Das, A., Datta, S., Gkioxari, G., Lee, S., Parikh, D., Batra, D.: Embodied question answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
22. Devlin, J., Gupta, S., Girshick, R., Mitchell, M., Zitnick, C.L.: Exploring nearest neighbor approaches for image captioning. arXiv preprint arXiv:1505.04467 (2015)
23. Doğan, F.I., Kalkan, S., Leite, I.: Learning to generate unambiguous spatial referring expressions for real-world environments. CoRR (2019)
24. Elhoseiny, M., Elfeki, M.: Creativity inspired zero-shot learning. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5784–5793 (2019)
25. Elhoseiny, M., Saleh, B., Elgammal, A.: Write a classifier: Zero-shot learning using purely textual descriptions. In: ICCV (2013)
26. Elhoseiny, M., Zhu, Y., Zhang, H., Elgammal, A.: Link the head to the "beak": Zero shot learning from noisy text description at part precision. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
27. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
28. Karpathy, A., Joulin, A., Li, F.F.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems (NeurIPS) (2014)
29. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: Referitgame: Referring to objects in photographs of natural scenes. In: Empirical Methods in Natural Language Processing (EMNLP) (2014)
30. Kim, J.H., Jun, J., Zhang, B.T.: Bilinear attention networks. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)
31. Kiros, R., Salakhutdinov, R., Zemel, R.S., et al: Unifying visual-semantic embeddings with multimodal neural language models. Transactions of the Association for Computational Linguistics (TACL) (2015)
32. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (2017)
33. Kulkarni, N., Misra, I., Tulsiani, S., Gupta, A.: 3d-relnet: Joint object and relational network for 3d prediction. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
34. Lazaridou, A., Hermann, K.M., Tuyls, K., Clark, S.: Emergence of linguistic communication from referential games with symbolic and pixel input. arXiv preprint arXiv:1804.03984 (2018)
35. Lei Ba, J., Swersky, K., Fidler, S., et al.: Predicting deep zero-shot convolutional neural networks using textual descriptions. In: ICCV (2015)
36. Lewis, D.: Convention: A philosophical study. John Wiley & Sons (2008)
37. Li, C., Xia, F., Martín-Martín, R., Savarese, S.: Hrl4in: Hierarchical reinforcement learning for interactive navigation with mobile manipulators. In: Conference on Robot Learning (2020)

38. Long, Y., Shao, L.: Describing unseen classes by exemplars: Zero-shot learning using grouped simile ensemble. In: Winter Conference on Applications of Computer Vision (WACV) (2017)
39. Long, Y., Shao, L.: Learning to recognise unseen classes by a few similes. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 636–644. ACM (2017)
40. Lu, C., Krishna, R., Bernstein, M., Fei-Fei, L.: Visual relationship detection with language priors. In: European conference on computer vision. pp. 852–869. Springer (2016)
41. Mao, J., Xu, W., Yang, Y., Wang, J., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). In: International Conference on Learning Representations (ICLR) (2015)
42. Mauceri, C., Palmer, M., Heckman, C.: SUN-Spot: An RGB-D Dataset With Spatial Referring Expressions. In: International Conference on Computer Vision Workshop on Closing the Loop Between Vision and Language (2019)
43. Mauceri, C., Palmer, M., Heckman, C.: SUN-Spot: An RGB-D Dataset With Spatial Referring Expressions. In: International Conference on Computer Vision Workshop on Closing the Loop Between Vision and Language (2019)
44. Mitchell, M., van Deemter, K., Reiter, E.: Generating expressions that refer to visible objects. In: North American Chapter of the Association for Computational Linguistics (NAACL)) (2013)
45. Mo, K., Zhu, S., Chang, A.X., Yi, L., Tripathi, S., Guibas, L.J., Su, H.: PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3D object understanding. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
46. Monroe, W., Hawkins, R.X., Goodman, N.D., Potts, C.: Colors in context: A pragmatic neural model for grounded language understanding. Transactions of the Association for Computational Linguistics (TACL) (2017)
47. Paetzel, M., Racca, D.N., DeVault, D.: A multimodal corpus of rapid dialogue games. In: LREC. pp. 4189–4195 (2014)
48. Plummer, B.A., Shih, K.J., Li, Y., Xu, K., Lazebnik, S., Sclaroff, S., Saenko, K.: Revisiting image-language networks for open-ended phrase detection. CoRR abs/1811.07212 (2018)
49. Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
50. Prabhudesai, M., Tung, H.Y.F., Javed, S.A., Sieb, M., Harley, A.W., Fragkiadaki, K.: Embodied language grounding with implicit 3D visual feature representations. CoRR abs/1910.01210 (2019)
51. Qi, C.R., Litany, O., He, K., Guibas, L.: Deep hough voting for 3d object detection in point clouds. In: International Conference on Computer Vision (ICCV) (2019)
52. Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in Neural Information Processing Systems (NeurIPS) (2017)
53. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people with "their" names using coreference resolution. In: European Conference on Computer Vision (ECCV) (2014)
54. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems (NeurIPS) (2015)

55. Romera-Paredes, B., Torr, P.: An embarrassingly simple approach to zero-shot learning. In: ICML. pp. 2152–2161 (2015)
56. Rosman, B., Ramamoorthy, S.: Learning spatial relationships between objects. The International Journal of Robotics Research (2011)
57. Savva, M., Kadian, A., Maksymets, O., Zhao, Y., Wijmans, E., Jain, B., Straub, J., Liu, J., Koltun, V., Malik, J., et al.: Habitat: A platform for embodied ai research. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
58. Shridhar, M., Thomason, J., Gordon, D., Bisk, Y., Han, W., Mottaghi, R., Zettlemoyer, L., Fox, D.: Alfred: A benchmark for interpreting grounded instructions for everyday tasks. CoRR abs/1912.01734 (2019)
59. Socher, R., Ganjoo, M., Manning, C.D., Ng, A.: Zero-shot learning through cross-modal transfer. In: NeurIPS. pp. 935–943 (2013)
60. Su, J.C., Wu, C., Jiang, H., Maji, S.: Reasoning about fine-grained attribute phrases using reference games. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 418–427 (2017)
61. Tsai, Y.H.H., Huang, L.K., Salakhutdinov, R.: Learning robust visual-semantic embeddings. In: ICCV (2017)
62. Vedantam, R., Bengio, S., Murphy, K., Parikh, D., Chechik, G.: Context-aware captions from context-agnostic supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 251–260 (2017)
63. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
64. Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M.: Dynamic graph cnn for learning on point clouds. ACM Transactions on Graphics (TOG) (2019)
65. Wittgenstein., L.: Philosophical investigations: The english text of the third edition (1953)
66. Xia, F., Zamir, A.R., He, Z., Sax, A., Malik, J., Savarese, S.: Gibson env: Real-world perception for embodied agents. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
67. Xiang, F., Qin, Y., Mo, K., Xia, Y., Zhu, H., Liu, F., Liu, M., Jiang, H., Yuan, Y., Wang, H., et al.: Sapien: A simulated part-based interactive environment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11097–11107 (2020)
68. Xu, K., Ba, J., Kiros, R., Courville, A., Salakhutdinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning (ICML) (2015)
69. Yang, J., Ren, Z., Xu, M., Chen, X., Crandall, D., Parikh, D., Batra, D.: Embodied visual recognition. CoRR abs/1904.04404 (2019)
70. Yang, Y., Hospedales, T.M.: A unified perspective on multi-domain and multi-task learning. In: ICLR (2015)
71. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
72. Zhang, J., Kalantidis, Y., Rohrbach, M., Paluri, M., Elgammal, A., Elhoseiny, M.: Large-scale visual relationship understanding. In: AAAI Conference on Artificial Intelligence (2019)
73. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2019)

74. Zhu, Y., Elhoseiny, M., Liu, B., Peng, X., Elgammal, A.: A generative adversarial approach for zero-shot learning from noisy texts. In: CVPR (2018)