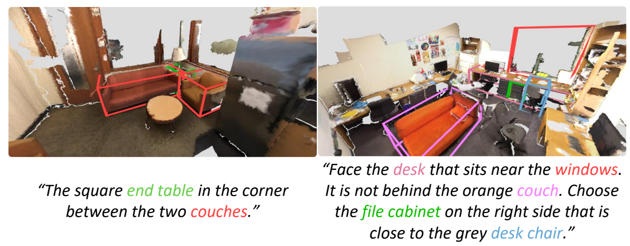
ScanEnts3D: Exploiting Phrase-to-3D-Object Correspondences for Improved Visio-Linguistic Models in 3D Scenes

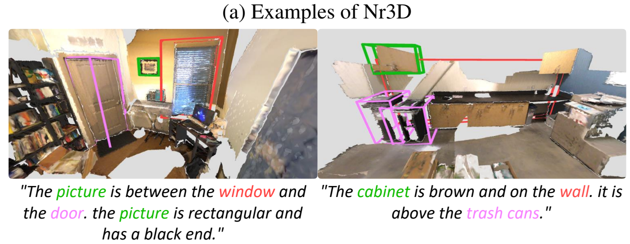
Ahmed Abdelreheem[1,2], Kyle Olszewski[2], Hsin-Ying Lee[2], Peter Wonka[1], Panos Achlioptas[2]
[1] King Abdullah University of Science and Technology (KAUST)
[2] Snap Inc.
{asamirh.95,olszewski.kyle,james371507,pwonka,pachlioptas}@gmail.com

Abstract

The two popular datasets ScanRefer [20] and ReferIt3D [5] connect natural language to real-world 3D scenes. In this paper, we curate a complementary dataset extending both the aforementioned ones. We associate all objects mentioned in a referential sentence with their underlying instances inside a 3D scene. In contrast, previous work did this only for a single object per sentence. Our ***Scan Entities*** *in* ***3D*** *(**ScanEnts3D**) dataset provides explicit correspondences between 369k objects across 84k referential sentences, covering 705 real-world scenes. We propose novel architecture modifications and losses that enable learning from this new type of data and improve the performance for both neural listening and language generation. For neural listening, we improve the SoTA in both the Nr3D and ScanRefer benchmarks by* ***4.3%*** *and* ***5.0%****, respectively. For language generation, we improve the SoTA by* ***13.2*** *CIDEr points on the Nr3D benchmark. For both of these tasks, the new type of data is only used to improve training, but no additional annotations are required at inference time. Our introduced dataset is available on the project's webpage at* `https://scanents3d.github.io/`.



(a) Examples of Nr3D



(b) Examples of ScanRefer

Figure 1. **Typical annotation examples from ScanEnts3D**. Our annotations link each noun phrase in a given referential sentence to one or more corresponding objects in a 3D ScanNet scene. The target object and its corresponding noun phrase are shown in green. The anchor objects and their corresponding noun phrases are shown in different colors. The couches on the top left and the trash cans on the bottom right are examples where one noun phrase corresponds to multiple objects in the scene.

1. Introduction

Connecting natural language to real-world 3D scenes enables us to tackle fundamental problems such as language-assisted object localization and fine-grained object identification [5, 20], object captioning [19], scene-based Q/A [9], and language-based semantic segmentation [50].

To improve upon these types of problems, we extend two recent datasets ScanRefer and Nr3D with a new type of annotation. These two datasets collected referential sentences for real-world 3D scenes. A referential sentence describes a single ("target") object in a 3D scene. The grounding annotations in these two datasets consist of labeling the target object in the scene and associating it with the referential sentence. Such a referential sentence needs to discriminate the target object from the remaining objects in the 3D scene. This can be done by emphasizing properties of the target object such as color, material, or geometry (e.g., *the tall chair*). However, we can observe that most referential sentences contain information about other objects and object relationships in order to describe the target object (e.g., *the tall chair* → *the tall chair between the table and the fireplace*). We call these other objects ("anchor objects").

In our work, we set out to utilize anchor objects in two ways. First, we propose a new dataset ScanEnts3D. We curate grounding annotations for *all* 3D objects in each ref-

	#Utterances	#Annotated Objects	Anchor Instance Annotations	Phrase-to-Object Correspondences	#Scan Entities	Avg. # of Objects per Scan Entity
Nr3D [5]	38K	38K	✗	✗	-	-
ScanRefer [20]	46K	46K	✗	✗	-	-
			ScanEnts3D			
Nr3D-ScanEnts	38K	**126K** (+88K)	✓	✓	**96k**	**1.32**
ScanRefer-ScanEnts	46K	**243K** (+197K)	✓	✓	**182k**	**1.33**

Table 1. **Comparison between the Nr3D and ScanRefer datasets and their corresponding extensions in ScanEnts3D.** Our proposed dataset contains more annotated objects and provides annotations for the anchor objects mentioned in the referential utterances. Specifically, ScanEnts3D provides explicit phrase-to-object correspondences for *all* mentioned objects. ScanRefer has more verbose utterances compared to the more parsimonious Nr3D. This distinction is also reflected in the resulting statistics from ScanEnts3D (last two columns).

erential sentence for both Nr3D and ScanRefer. Previously, grounding annotations were only available by linking a single target object to a complete referential sentence. In contrast, we provide grounding annotations by linking target and anchor objects to noun phrases within the referential sentence. We call this new type of data a *scan entity*. A scan entity consists of phrases (e.g., tables, trash cans) along with the 3D objects of the scene that correspond to them (see Figure 1). Second, we show how this new type of data can benefit language-based 3D scene understanding in two tasks: discriminative language comprehension ('neural listening') and generative language production ('neural speaking'). It is important to note that it is not possible to directly utilize our new annotations in existing architectures. We, therefore, propose several architecture modifications and training losses to recent frameworks so we can make use of anchor objects. These modifications will make use of the additional information only during training to facilitate the incorporation of auxiliary losses, but no additional data is used during inference time. The goal of our modifications is to predict the anchor objects in addition to the target object. This idea is based on our hypothesis that 3D visio-linguistic architectures *can and should* model pairwise or higher-order object-to-object relations in order to become more robust learners. Our modifications are i) *effective*, as they result in significantly improved accuracy for both tasks in well-established benchmarks; ii) *robust*, as they have a positive performance effect across many distinct architectures, and iii) their learning effect is intuitive and *interpetable* – we show that the primary cause of the quantitative gains we attain is learning more and/or better object-to-object relations expressed in the referential language. To summarize, our main contributions are the following:

- We introduce a large-scale dataset extending both Nr3D and ScanRefer by grounding all objects in a referential sentence. Our ***ScanEnts3D*** dataset (Scan Entities in 3D) includes 369,039 language-to-object correspondences, more than three times the number from the original works.
- We propose novel training losses and architecture modifications to exploit the new annotations. We improve the performance of several 3D neural listening architectures, including improving the SoTA in Nr3D and ScanRefer by **4.3%**, and **5.0%** respectively. We improve neural speaking architectures, as measured with standard captioning metrics (e.g., BLEU, METEOR, ROUGE, and CIDEr). For instance, we improve the SoTA for neural speaking with Nr3D, per CIDEr, by **+13.2**. Importantly, we note that we do *not* train our networks with more referential sentences or use ScanEnts3D's annotations during inference. Instead, we rely on additional grounding information during training only.

We acknowledge two strong concurrent works that share a similar idea [51, 72]. As an advantage of our realization, we: 1) have professional instead of crowd-sourced annotations, 2) are the only ones to tackle neural speaking, 3) tackle both ReferIt3D and ScanRefer setups in neural listening, 4) have the best overall results across widely adopted evaluation metrics, 5) propose and explore zero-shot transfer learning for neural listeners operating in novel 3D scenes [35].

2. Related Work and Background

Modern visio-linguistic tasks for objects in 3D scenes. Increasingly more tasks involving a joint understanding of computer vision and language processing are been studied thanks to the introduction of modern 3D-oriented datasets [8, 14, 22, 25, 32, 39, 48, 52, 53, 62] equipped with linguistic annotations [6, 7, 18, 30, 54]. These include captioning of 3D objects in synthetically generated contexts [6, 28] and captioning of objects embedded in real-world scenes [19, 73], 3D object identification in scenes [5, 20, 57, 73], language-based semantic segmentation [1, 2, 31, 38, 50], and 3D question answering [9, 26, 37, 41, 59, 69]. Existing visio-linguistic datasets involving objects in real-world 3D scenes [5, 20] provide limited annotations focusing only on target objects, bypassing all other mentioned context-relevant object instances. Despite that, such limited annotations naturally impede the development of more sophisticated 3D neural listeners, a flourishing line of works is being currently developed, concentrating on neural listening [5, 10, 15, 29, 34, 49, 61, 68, 74, 75], and neural speaking [12, 15, 73, 76].

3D-based visio-linguitic grounding. Visio-linguistic grounding aims at associating information expressed in language, e.g., noun-entities, to the underlying objects present in visual stimuli [45]. Such grounding for 2D images has been extensively studied [36,42,45,65,66,70,71]. On the contrary, 3D visual grounding is still in its infancy [4, 7, 13]. Recently, ScanRefer [20] and Referit3D [5] proposed datasets for language-driven neural-based comprehension in 3D, built on top of assets of ScanNet [22]. Following these establishments, several approaches explored novel designs and new formulations [3, 12, 24, 33, 49, 58, 68] for creating improved neural listeners that *implicitly* attempt to model the grounding (visual) context of each reference [29, 34, 49, 68, 74, 75]. By using the explicit annotations provided in ScanEnts3D, we take a step in reducing the gap between the richer 2D-based and less mature 3D-based learning-based comprehension paradigms. As we show, by developing appropriate adaptations that take into account ScanEnts3D, we can improve neural listeners and neural speakers across many architectural designs, including improving two state-of-the-art methods, SAT [68] and MVT [34].

3. ScanEnts3D Dataset

3.1. Curating human annotations

Curating all correspondences between each noun phrase in a referential sentence and their underlying objects within a 3D scene is generally an error-prone task. First, it requires the annotators to be familiar with (albeit simple) linguistic and syntactic rules in the given language to parse the sentence. Second, they must be able to carefully navigate inside a complicated (and, possibly, poorly reconstructed) scene, which typically contains multiple objects of the same fine-grained object class (e.g., multiple kitchen cabinets, as in the right-most example in Figure 1), so as to select *all and only* the correct referenced objects. In order to ensure the curation of high-quality correspondences with a low error rate and high coverage, we took several critical steps. First, we developed a custom web-based UI for 3D scene navigation, which was interactive, lightweight (i.e., fast), user-friendly, and which allowed for maintaining an active dialogue with the annotators. Second, we coordinated with a team of *professional* data labelers to ensure the collection of sufficiently accurate labels for ScanEnts3D.

While a common approach to large-scale data collection today is to use crowd-sourcing techniques with platforms such as Amazon Mechanical Turk (AMT) [21], we note

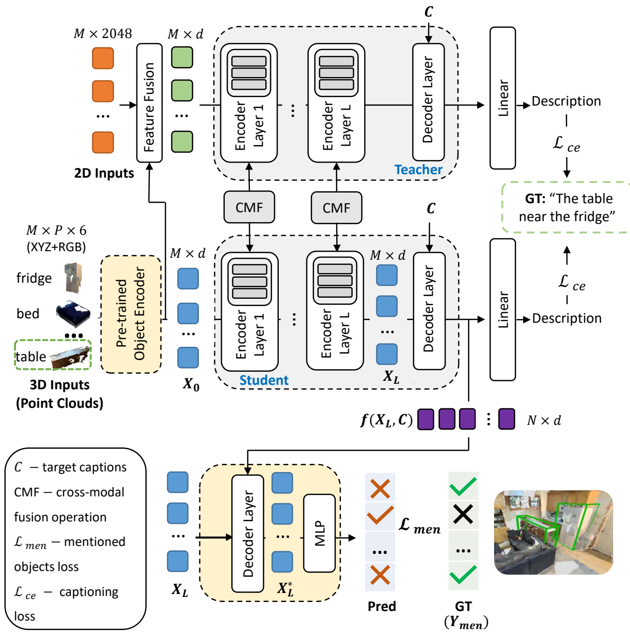


Figure 2. **Proposed M2Cap-ScanEnts model adapting $\mathcal{X}-$Trans2Cap model to operate with our proposed losses.** The model is given a set of 3D objects in a 3D scene and outputs a caption for the target object (e.g., table in green box). The $\mathcal{X}-$Trans2Cap model exploits cross-modal knowledge transfer (3D inputs together with their counterpart 2D images) and adopts a student-teacher paradigm [17, 73]. Boxes in yellow show our modifications. Here, we use a transfer learning approach by fine-tuning a pre-trained object encoder trained on the listening task to promote discriminative object feature representations. Our modular loss guides the network to predict all object instances mentioned in the ground truth caption.

that we conducted an AMT-based *pilot* study to determine whether such an approach is sufficient, given the aforementioned complexity and specificity of this task. We found that the error rate within the collected annotations was significantly higher than that in the annotations provided by the professional labelers (error rates of 16% vs. $< 5\%$, respectively). Rather than attempting to evaluate our approach using data with such a high percentage of erroneous labels, we ultimately decided to employ professional annotators, which significantly improved the attained quality of ScanEnts3D.

Finally, we split the curation process into two phases; the annotation phase and the verification phase. The verification phase also involved *correcting* the mistakes found so as to provide high-quality annotations. In Figure 1, we show examples from the ScanEnts3D dataset for Nr3D and ScanRefer, which demonstrate that our annotations cover different classes of anchor objects and that our annotations provide rich contexts for these utterances.

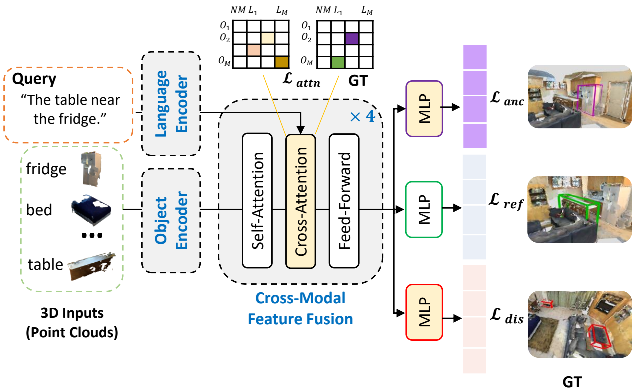


Figure 3. **Demonstration of our proposed listening losses adjusted for the MVT model.** The losses are applied independently of each other on top of object-centric and context-aware features. Crucially, the extended MVT-ScanEnts model can predict all anchor objects (shown in purple), same-class distractor objects (red), and the target (green). The default model only predicts the target.

3.2. Key Characteristics of ScanEnts3D

In this section, we briefly present key characteristics of the ScanEnts3D dataset. In Table 1, we present the number of collected annotations for 37,842 examples from the Nr3D dataset and 46,173 examples from ScanRefer. We observe that in general ScanRefer annotations provide more entities per single utterance compared to Nr3D (182,300 vs. 96,032, respectively), as ScanRefer utterances are typically longer and more verbose than Nr3D utterances (on average, there are 20.3 words per utterance in ScanRefer, vs. 11.4 in Nr3D).

We also calculate how frequently an object is used as an anchor object when it is the *only* 3D instance of its class inside a scene (e.g., the *window* in the lower left example in Figure 1). We find that 24.3% of all anchor objects are 'unique' in Nr3D. However, significantly more anchor objects are unique in ScanRefer (39.1%). Such anchors typically represent *salient* objects [5], and can be particularly useful for locating the target, esp. when many other objects are being described in context (explaining the differential between the two datasets).

Last, by using our collected annotations, we can extract *object-to-object* spatial relationships of scan entities (with $\sim$91% verified sampled accuracy), using existing spatial relation classifiers [43]. Crucially, to attain this accuracy level, we explicitly apply such a classifier on *ground-truth* referred entities found in ScanEnts3D. Out of the 13 spatial relationship *types* found, the most frequently used relation in Nr3D and ScanRefer is the "closest" and the "on top of", respectively. For a more detailed analysis of these findings, we encourage the reader to consult the Supp.

4. Method

In this section, we propose modifications to several existing state-of-the-art architectures to utilize the additional annotations provided by ScanEnts3D during training. The main idea of the modifications is to use the prediction of anchor objects as an auxiliary task during training. While the exact implementation of this idea depends on the specific architecture, it seems intuitive that an additional understanding of anchor objects will lead to better models. We explore two tasks: neural listening and neural speaking and multiple architectures per task. Our main goal is to demonstrate the inherent value of the curated annotations. All proposed modifications are simple to implement and lead to substantial improvements. We, therefore, conjecture that similar modifications are (or will be) possible to existing (and future) architectures making use of ScanEnts3D. We also encourage the reader to consult the supplementary material for more details regarding our modifications and their effect.

For neural listeners, we propose three new loss functions. We try these losses on two recent listening architectures, SAT [68] and MVT [34]. In addition, we also propose modifications to 3DJCG [12] described in the supplementary. For neural speakers, we propose corresponding modifications and appropriate losses for the Show, Attend, and Tell model [63] and $\mathcal{X}-$Trans2Cap model [73].

4.1. 3D Grounded Language Comprehension

The goal of a neural listener is to identify the target object in a 3D scene described in a referential sentence. Following [5], the input to our neural listener is a set of M 3D object proposals present in a 3D scene, where each proposal is represented as a 3D point cloud, and an input utterance describing the target object, represented as a sequence of N tokens. Most recent neural listeners are transformer-based models [34,68,75], each of which applies bi-modal attention between the features of the 3D objects and the features of the words of the input utterance. Assuming this generic setup, we now detail our three proposed loss functions.

4.1.1 Anchor Prediction Loss

The anchor prediction loss $\mathcal{L}_{\text{anc}}$ guides the neural listener to predict the anchor objects (non-target objects that are mentioned in an input utterance). In order to identify the target object, one must typically also identify the mentioned anchor objects. The anchor prediction loss can be applied to any output token of an attention or self-attention layer. We obtain a suitable set of tokens (feature vectors) for the M input 3D object proposals denoted as $F_O = \{f_0, f_i, \ldots, f_M\}$ as follows. For the MVT model [34], F_O is obtained from a sequence of transformer decoder layers followed by aggregation over multiple views as shown in Figure 3. For the SAT model [68], F_O is obtained from a sequence of multi-modal attention layers. We derive $X_{\text{anc}} = \phi(F_O)$ with an auxiliary classification head using an MLP to encode $\phi(.)$. The MLP consists of two fully connected layers, where X_{anc} represents a vector capturing the listener's confidence of each object being an anchor object (of shape $M \times 1$ expressing the logits). We apply a binary cross entropy loss as in Equation (1), where Y_{anc} is the ground truth vector of shape $M \times 1$.

$$\mathcal{L}_{\text{anc}} = BCE(X_{\text{anc}}, Y_{\text{anc}}) \tag{1}$$

4.1.2 Cross-Attention Map Loss

The Cross-Attention Map loss encourages the network to attain high relevance values between the objects and the words corresponding to the same underlying scan entity. This loss operates on cross-attention maps A (before applying the softmax operation) between the features of the input scene 3D objects and the word tokens of the input utterance, where A is of shape $M \times N$. The target matrix Y_{attn} is a binary matrix of shape $M \times N$, where a cell $(y_{i,j})$ has a value of 1 if the ith object and the jth word correspond to one another. For each row R_i of shape $1 \times N$ and the corresponding row Y^i_{attn} in the target matrix, the cross-attention map loss $(\mathcal{L}_{\text{attn}})$ is measured as:

$$\mathcal{L}_{\text{attn}} = \frac{1}{M} \sum_{i=1}^{M} BCE(R_i, Y^i_{\text{attn}}) \tag{2}$$

4.1.3 Same-Class Distractor Prediction Loss

This loss guides the neural listener to predict the same-class distractor objects $(\mathcal{L}_{\text{dis}})$. It does not directly leverage ScanEnts3D but as we show it offers beneficial synergies with the above losses as it helps to better disentangle the target from distracting objects with the same (fine-grained) object class. Such same-class distractors are objects from the same class as the target co-existing in the scene. As with the anchor objects, we treat the same-class distractor prediction problem as a multi-label classification problem. Thus, we use an approach similar to Section 4.1.1. Specifically, we obtain the logits for predicting the same-class distractor $X_{\text{dis}} = \psi(F_O)$ of shape $M \times 1$ with an MLP $\psi(.)$. This loss is also binary cross entropy-based, like in Equation (3), where Y_{dis} is a multi-hot target vector of shape $M \times 1$. Note that a same-class distractor object may not be mentioned in the given input utterance.

$$\mathcal{L}_{\text{dis}} = BCE(X_{\text{dis}}, Y_{\text{dis}}) \tag{3}$$

4.1.4 Training Objective Function

The proposed losses can serve as auxiliary add-ons to the original loss term $(\mathcal{L}_{\text{org}})$ of existing neural listeners, such as the MVT and SAT models. We train these models in an end-to-end fashion as:

$$\mathcal{L} = \mathcal{L}_{\text{org}} + \mathcal{L}_{\text{aux}}, \quad \text{where} \quad \mathcal{L}_{\text{aux}} = \alpha\mathcal{L}_{\text{anc}} + \beta\mathcal{L}_{\text{attn}} + \gamma\mathcal{L}_{\text{dis}} \tag{4}$$

Where α, β, and γ are scalar values controlling the relevant importance of each term. In our experiments, we use $\alpha = \beta = 3.0$ and $\gamma = 0.5$.

4.2. Grounded Language Production in 3D

We describe our modifications to existing architectures for neural speaking. We call our versions of these architectures SATCap-ScanEnts and M2Cap-ScanEnts.

4.2.1 SATCap-ScanEnts

The "Show, Attend, and Tell" model is an encoder-decoder network originally designed for 2D-based image captioning. To make it amenable to purely 3D inputs, we replace the image encoder with the encoder network found in the MVT model [34], which is a point cloud PointNet++ encoder together with 3D object self-attention layers. Crucially, to improve the generalization of this speaker, we use a *pretrained* MVT-based encoder solving the neural-listening task and then fine-tune it for the speaking task. For the decoder network, we use a unidirectional LSTM cell [27]. The encoder part is given the ground-truth objects as input in a similar manner to [73]. The speaker model is trained via teacher-forcing [60]. Importantly, we also apply our proposed entity prediction loss during the decoding steps. At each decoding step, if the current word to be predicted corresponds to a scan entity, our loss pushes the object corresponding to the underlying scan entity to be the highest scoring among all objects present in the input scene.

4.2.2 M2Cap-ScanEnts

We employ a similar approach on the $\mathcal{X}-$Trans2Cap model [73], referred to as M2Cap-ScanEnts detailed in Figure 2. We introduce the following two changes to the $\mathcal{X}-$Trans2Cap architecture. First, we use a pre-trained PointNet++ encoder followed by the pre-trained 3D object self-attention layers in the MVT [34] network. Second, we add a new cross-attention layer after the captioning layer found in the student network. The layer applies a cross-attention operation between the features of the 3D objects X_L of shape $M \times d$ and the features of the predicted tokens $N \times d$, where d is the latent feature dimension, to obtain new enhanced features X^*_L of shape $M \times d$. Finally, the logit vector is obtained with an MLP $\theta(.)$, representing a confidence value for each object as to whether it is mentioned in the target caption. A binary cross-entropy loss $\mathcal{L}_{\text{men}} = BCE(\theta(X^*_L), Y_{\text{men}})$ is employed, in which the target vector Y_{men} is a multi-hot vector (y^i_{men} is 1 if the ith object is mentioned in the target

Arch.	Overall	Easy	Hard	View-dep.	View-indep.
ReferIt3DNet [5]	35.6%±0.7%	43.6%±0.8%	27.9%±0.7%	32.5%±0.7%	37.1%±0.8%
3DVG-Transformer [75]	40.8%±0.2%	48.5%±0.2%	34.8%±0.4%	34.8%±0.7%	43.7%±0.5%
TransRefer3D [29]	42.1%±0.2%	48.5%±0.2%	36.0%±0.4%	36.5%±0.6%	44.9%±0.3%
LanguageRefer [49]	43.9%	51.0%	36.6%	41.7%	45.0%
SAT [68]	49.2%±0.3%	56.3%±0.5%	42.4%±0.4%	46.9%±0.3%	50.4%±0.3%
PhraseRefer [72]	54.4%	62.1%	47.0%	51.2%	56.0%
MVT [34]	55.1%±0.3%	61.3%±0.4%	49.1%±0.4%	54.3%±0.5%	55.4%±0.3%
SAT-ScanEnts (ours)	52.5%±0.2% (+3.3%)	59.8%±0.2% (+3.6%)	45.6%±0.3% (+3.2%)	51.3%±0.5% (+4.4%)	53.2%±0.1% (+2.8%)
MVT-ScanEnts (ours)	**59.3%±0.1%** (+4.2%)	**65.4%±0.3%** (+4.1%)	**53.5%±0.2%** (+4.4%)	**57.3%±0.3%** (+3.0%)	**60.4%±0.2%** (+5.0%)

Table 2. **Listening performance on Nr3D dataset across key categories of ReferIt3D.** The neural listeners trained with the Nr3D-ScanEnts dataset and our proposed losses noticeably improve the performance of the corresponding original architectures (relative boost in green numbers) while also establishing a new SoTA.

caption). We do not train a speaker and a listener jointly, which is the key contribution of D3Net [16]. Instead, our focus is on the introduction and utilization of dense annotations.

$\mathcal{L}_{attn}$	$\mathcal{L}_{anc}$	$\mathcal{L}_{dis}$	Overall	Easy	Hard	View-dep.	View-indep.
			55.1%	61.3%	49.1%	54.3%	55.4%
✓			56.6%	63.0%	50.5%	55.4%	57.2%
		✓	56.9%	63.5%	50.6%	55.3%	57.8%
✓		✓	57.4%	64.3%	50.8%	55.6%	58.3%
	✓	✓	57.9%	63.7%	52.3%	56.0%	58.9%
	✓		58.1%	63.8%	52.6%	56.7%	58.8%
✓	✓		58.7%	64.6%	53.1%	**57.5%**	59.3%
✓	✓	✓	**59.3%**	**65.4%**	**53.5%**	57.3%	**60.4%**

Table 3. **Ablation study of loss functions.** We ablate different combinations of our proposed auxiliary losses on the MVT neural listener, trained on Nr3D using ScanEnts3D.

5. Experiments

5.1. Experimental Setup

Datasets and splits. We use the Nr3D [5] and ScanRefer [20] datasets with their original annotations as well as our additional annotations in ScanEnts3D dataset. We use the official ScanNet [22] training and validation splits.

	Unique Acc.		Multiple Acc.		Overall Acc.	
	@0.25	@0.5	@0.25	@0.5	@0.25	@0.5
3DJCG [12]	78.75	**61.30**	40.13	30.08	47.62	36.14
3DJCG-ScanEnts (ours)	**79.49**	60.74	**41.51**	**31.34**	**48.88**	**37.04**

Table 4. **Effect of ScanEnts3D for object detector-based listeners**. This ablation shows the effectiveness of using ScanEnts3D on a different listener design (ScanRefer setup). The attained performance boost further suggests the usefulness and generality of the ScanEnts3D-induced loss functions.

Metrics. For the neural listening experiments, we report the attained referential accuracy across the same data categories used by ReferIt3D. For the neural speaking experiments we evaluate the output text generations against the ground-truth annotations, based on the metrics of BLEU-4 [44], ROUGE [40], METEOR [11], and CIDEr [55].

5.2. Neural Listening

We demonstrate the effectiveness of the proposed ScanEnts3D by comparing state-of-the-art models trained with and without the additional annotations. For all experiments, we note that our dataset only leads to modifications at training time. At inference time, our trained models and their respective baseline models use the same input data.

Neural listeners trained with ScanEnts3D achieve state-of-the-art performance. As shown in Table 2 and Table 6, our MVT-ScanEnts neural listener, which is trained with our proposed dataset (Nr3D-ScanEnts) and our auxiliary losses, achieves state-of-the-art results, outperforming the current SoTA models. MVT-ScanEnts outperforms the original MVT [34] on both the Nr3D (+4.3%) and the ScanRefer (+5.0%) datasets, while the SAT-ScanEnts model similarly outperforms the original SAT [68] model on both the Nr3D (+3.3%) and ScanRefer (+2.4%) datasets.

Further analysis. Furthermore, we observe considerable improvements in each context for Nr3D, particularly in the view-independent and hard contexts (5.0% and 4.4% as in Table 2, respectively). In addition, we report the F_1 score [46], which measures the overall accuracy of a test taking into account its precision and recall, of the anchor object classification in the MVT-ScanEnts model. The F_1 score of 0.64 (out of a possible maximum of 1) suggests that the full potential value of our proposed dataset ScanEnts3D may still be attained with the development of more sophisticated losses, a promising area for future work.

Finally, in Figure 5, we present qualitative examples of how recognizing the anchor objects allows the model to identify the target object correctly. Comparing the proposed model MVT-ScanEnts and the current state-of-the-art method MVT, we demonstrate that guiding our network to understand the anchor entities mentioned in the input utterances promotes the listener to accurately identify the target object. In the third column of this Figure, we demonstrate the predicted target object and the predicted anchor objects by MVT-ScanEnts in green and purple bounding boxes, respectively.

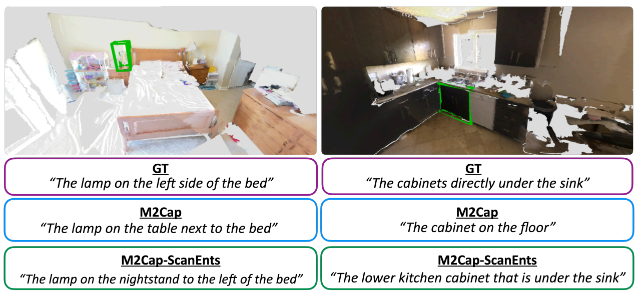


Figure 4. **Qualitative comparison of neural speaker variants.** The M2Cap-ScanEnts generations tend to be more discriminative (e.g., *left of the bed vs. next to the bed*) compared to the default M2Cap variant. In addition, M2Cap-ScanEnts makes better use of relationships between the target object and anchor objects (*cabinet, sink*).

Neural listeners trained with ScanEnts3D are more context aware. To show this, first, we conduct additional experiments on both MVT and MVT-ScanEnts neural listeners (Table 7). In these experiments, we change the input to the neural listeners in multiple ways to investigate if the listener becomes better at relying on the context of the 3D scene to robustly (and more naturally) predict a target object. The changes to the input are the following: (a) an input scene *without* the 3D object proposals of the anchor objects, (b) an input scene with *only* the object proposals of the anchor objects and the same-class distractor objects, and (c) an input utterance where the words that correspond to the anchor objects are replaced with the <unk> token denoting an out-of-vocabulary word. We observe that removing the object proposals that correspond to the anchor objects from the input scene results in a massive drop in the listening performance in MVT-ScanEnts. The drop in the performance in MVT-ScanEnts is much higher than the drop found in the original MVT model (-15.3% vs. -7%). This result suggests that the neural listeners trained with ScanEnts3D similar to humans, learn to rely heavily on the anchor objects to identify the target object and are less influenced by the non-anchor/mentioned objects. At the same time, we also observe an improved performance for MVT-ScanEnts compared to MVT (70.5% vs. 67.0%) when providing as input a 3D scene consisting of only the target object, its same-class distractors (to keep the problem highly challenging), and the anchor objects. In other words, on references where humans depend on information about anchors to communicate the target object in a unique manner, we find that visual information about these anchors is both *necessary and sufficient* for the performance of our model.

5.3. Neural Speaking

With the proposed ScanEnts3D dataset, the modified speaker models, SATCap-ScanEnts and M2Cap-ScanEnts, improve significantly against their corresponding baseline, as shown in Table 5. The encoder networks in SATCap and M2Cap models use the pre-trained encoder weights of an original MVT neural listener trained without ScanEnts3D, while the encoder networks in SATCap-ScanEnts and M2Cap-ScanEnts use the pre-trained weights of an MVT-ScanEnts listener. We observe that ScanEnts3D helps our speaker models to provide better captions for Nr3D and ScanRefer across all metrics (BLEU, CIDEr, METEOR, and ROUGE). The M2Cap-ScanEnts model improves the SoTA for neural speaking with Nr3D, per CIDEr, by **+13.2**. In all experiments, we use the ground truth instances as input. Also, we do not provide an extra 2D modality during the inference phase and do not use the additional CIDEr-based loss in the final objective function as in [64]. In Figure 4, we show captions by M2Cap-ScanEnts on the Nr3D dataset; we compare these captions to those generated by the M2Cap model. We observe that the captions generated by M2Cap-ScanEnts tend to be more discriminative and make explicit use of valid anchor objects to achieve this desideratum. We refer the reader to the Supp. for ablations on M2Cap-ScanEnts.

5.4. Ablation Studies

Effectiveness of the proposed losses. We conduct an ablation study for neural listeners by applying different combinations of our proposed losses. We try each possible combination of our losses ($\mathcal{L}_{\text{anc}}$, $\mathcal{L}_{\text{attn}}$, and $\mathcal{L}_{\text{dis}}$) with the MVT [34] architecture and report their performance on the Nr3D dataset, as shown in Table 3. When applying $\mathcal{L}_{\text{attn}}$ alone, we obtain an overall boost of 1.5% over the baseline MVT model (using none of our proposed losses). We obtain an improvement of 1.8% upon applying $\mathcal{L}_{\text{dis}}$ alone. This result is unsurprising, as we find that the same-class distractors are mentioned in 17.2% of the utterances in the Nr3D and 12.4% in the ScanRefer datasets. Applying the $\mathcal{L}_{\text{anc}}$ provides the best boost in every experiment where it is applied compared to the other losses. We observe that incorporating the anchor prediction loss is useful for all the Nr3D contexts, especially the hard contexts. The aforementioned result demonstrates how useful the knowledge of the anchor objects mentioned in the input sentence is. The best-performing model applies all three losses, and the performance is better than using $\mathcal{L}_{\text{anc}}$ and $\mathcal{L}_{\text{dis}}$ by 0.6%.

Arch.	Nr3D				ScanRefer			
	C	B-4	M	R	C	B-4	M	R
Scan2Cap [19]	61.89	32.02	28.88	64.17	64.44	36.89	28.42	60.42
$\mathcal{X}-$Trans2Cap [73]	80.02	37.90	30.48	67.64	87.09	44.12	30.67	64.37
SATCap (ours)	76.57	29.12	24.97	55.62	80.98	37.47	26.91	56.98
SATCap-ScanEnts (ours)	84.37	30.73	25.90	56.57	84.81	38.85	27.18	57.62
M2Cap (ours)	86.15	37.03	30.63	67.00	85.75	44.02	30.74	64.80
M2Cap-ScanEnts (ours)	**93.25**	**39.33**	**31.55**	**68.33**	**87.20**	**44.81**	**30.93**	**65.24**

Table 5. **Speaking performance on Nr3D and ScanRefer datasets.** The results of incorporating ScanEnts3D dataset in our proposed approaches for the speaking (captioning) task. A speaking model trained with our rich annotations performs better than one trained without them for both the Nr3D and ScanRefer datasets.

Arch.	Acc.
ScanRefer [20]	44.5%
ReferIt3DNet [5]	46.9%±0.2%
SAT [68]	53.8%±0.1%
MVT [34]	54.8%±0.1%
SAT-ScanEnts (ours)	56.2%±0.2%
MVT-ScanEnts (ours)	**60.8%±0.2%**

Table 6. **Listening performance on the ScanRefer dataset.** The neural listeners are trained using the ground truth boxes as input with or without using the additional annotations from the ScanEnts3D dataset and our proposed losses.

Figure 5. **Qualitative results for our proposed model (MVT-ScanEnts) compared to the MVT model.** The rows from top to bottom show the ground truth (green box), the target object predicted by MVT (red box), the predicted target object predicted by MVT-ScanEnts (green box) along with the predicted anchor objects (purple boxes), and the input utterance. The above examples show that the model can accurately predict the target object and simultaneously also predict the underlying anchor objects mentioned in the input utterance.

Arch.	Anchor Objects Lesioned (↓)	Anchor Words Lesioned (↓)	Anchor Info Present (↑)
MVT	48.1% (-7%)	45.5% (-9.6%)	67.0%
MVT-ScanEnts (ours)	**44.0%** (-15.3%)	**44.3%** (-15.0%)	**70.5%**

Table 7. **Evaluating the anchor-object-*awareness* for neural listeners trained w/ and w/o ScanEnts3D**. A listener trained with ScanEnts3D (MVT-ScanEnts) learns to depend heavily on the mentioned anchor objects, similar to humans. As seen here, its performance accuracy drops significantly ($\sim$15%) when the anchors are lesioned from the underlying input, *and* at the same time, its performance gets boosted when only the anchor(s) and the objects of the same class as the target are provided as input.

Can ScanEnts3D improve 3D object detector-based methods? As a last experiment, we investigate the extent to which our proposed dataset can improve the performance of different types of neural listeners. In particular, a widely used design proposed by ScanRefer [20] requires a listener to first *predict* 3D object proposals and then identify the target object (i.e., 3D object localization). To that end, we adapt the anchor prediction loss to work with the recent 3DJCG network [12]. In Table 4, we see attained improvements in the 3D object localization performance when using our ScanEnts3D. Most importantly, we can observe an improvement in the more complex and harder cases (Multiple).

6. Conclusion

This work takes substantial initial steps to bring object-to-object interactions, *grounded in language*, to the frontline of relevant learning-based methods. First, we curate and share a set of rich correspondences covering all referential entities mentioned in Nr3D and ScanRefer. Second, we use these annotations to train neural networks with better generalization and understanding of 3D objects w.r.t. their language-based grounding. By adapting existing methods and integrating our proposed loss functions, we attain *SoTA* results in both neural listening and speaking tasks for real-world scenes. We expect the derived insights to open new opportunities to advance related multimodal 3D object-centric tasks.

References

[1] Ahmed Abdelreheem, Abdelrahman Eldesokey, Maks Ovsjanikov, and Peter Wonka. Zero-shot 3d shape correspondence, 2023. 2

[2] Ahmed Abdelreheem, Ivan Skorokhodov, Maks Ovsjanikov, and Peter Wonka. Satr: Zero-shot semantic segmentation of 3d shapes, 2023. 2

[3] Ahmed Abdelreheem, Ujjwal Upadhyay, Ivan Skorokhodov, Rawan Al Yahya, Jun Chen, and Mohamed Elhoseiny. 3DRef-Transformer: Fine-grained object identification in real-world scenes using natural language. *WACV*, 2022. 3

[4] Panos Achlioptas. *Learning to generate and differentiate 3D objects using geometry & language*. PhD thesis, Stanford University, 2021. 3

[5] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*, 2020. 1, 2, 3, 4, 6, 8, 12, 17

[6] Panos Achlioptas, Judy Fan, Robert XD Hawkins, Noah D Goodman, and Leonidas J. Guibas. ShapeGlot: Learning language for shape differentiation. In *ICCV*, 2019. 2

[7] Panos Achlioptas, Ian Huang, Minhyuk Sung, Sergey Tulyakov, and Leonidas Guibas. ShapeTalk: A language dataset and framework for 3d shape edits and deformations. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 3

[8] Iro Armeni, Ozan Sener, Amir R Zamir, Helen Jiang, Ioannis Brilakis, Martin Fischer, and Silvio Savarese. 3D semantic parsing of large-scale indoor spaces. In *CVPR*, 2016. 2

[9] Daich Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoki Kawanabe. ScanQA: 3d question answering for spatial scene understanding. *ArXiv*, abs/2112.10482, 2021. 1, 2

[10] Eslam Mohamed Bakr, Yasmeen Alsaedy, and Mohamed Elhoseiny. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. *ArXiv*, abs/2211.14241, 2022. 3

[11] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*, 2005. 6

[12] Daigang Cai, Lichen Zhao, Jing Zhang, Lu Sheng, and Dong Xu. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16464–16473, 2022. 3, 4, 6, 8, 15

[13] Angel X. Chang. *Text to 3D Scene Generation*. PhD thesis, Stanford University, 2015. 3

[14] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An information-rich 3D model repository. *Computing Research Repository (CoRR)*, abs/1512.03012, 2015. 2

[15] Dave Zhenyu Chen, Ronghang Hu, Xinlei Chen, Matthias Nießner, and Angel X. Chang. Unit3d: A unified transformer for 3d dense captioning and visual grounding, 2022. 3

[16] Dave Zhenyu Chen, Qirui Wu, Matthias Nießner, and Angel X. Chang. D3net: A speaker-listener architecture for semi-supervised dense captioning and visual grounding in rgb-d scans, 2021. 6

[17] Guobin Chen, Wongun Choi, Xiang Yu, Tony X. Han, and Manmohan Chandraker. Learning efficient object detection models with knowledge distillation. In *NIPS*, 2017. 3

[18] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. *Computing Research Repository (CoRR)*, abs/1803.08495, 2018. 2

[19] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2Cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3193–3203, 2021. 1, 2, 8

[20] Z. Dave Chen, Angel X. Chang, and Matthias Nießner. ScanRefer: 3D object localization in RGB-D scans using natural language. *Computing Research Repository (CoRR)*, abs/1912.08830, 2019. 1, 2, 3, 6, 8, 15

[21] Kevin Crowston. Amazon mechanical turk: A research tool for organizations and information systems scholars. In Anol Bhattacherjee and Brian Fitzgerald, editors, *Shaping the Future of ICT Research. Methods and Approaches*, pages 210–221, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 3

[22] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *CVPR*, 2017. 2, 3, 6

[23] Helisa Dhamo, Fabian Manhardt, Nassir Navab, and Federico Tombari. Graph-to-3d: End-to-end generation and manipulation of 3d scenes using scene graphs. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 12

[24] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, Xiangdong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal S. Mian. Free-form description guided 3d visual graph network for object grounding in point cloud. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3702–3711, 2021. 3

[25] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10933–10942, 2021. 2

[26] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4089–4098, 2018. 2

[27] Klaus Greff, Rupesh Kumar Srivastava, Jan Koutník, Bas R. Steunebrink, and Jürgen Schmidhuber. Lstm: A search space odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28:2222–2232, 2017. 5, 15

[28] Zhizhong Han, Chao Chen, Yu-Shen Liu, and Matthias Zwicker. Shapecaptioner: Generative caption network for 3d shapes by learning a mapping from parts detected in multiple

views to sentences. In *ACM International Conference on Multimedia (MM)*, 2020. 2

[29] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. TransRefer3D: Entity-and-relation aware transformer for fine-grained 3D visual grounding. *Computing Research Repository (CoRR)*, abs/2108.02388, 2021. 3, 6

[30] Yining Hong, Qing Li, Song-Chun Zhu, and Siyuan Huang. VLGrammar: Grounded grammar induction of vision and language. *ICCV*, 2021. 2

[31] Ji Hou, Benjamin Graham, Matthias Nießner, and Saining Xie. Exploring data-efficient 3d scene understanding with contrastive scene contexts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15587–15597, 2021. 2

[32] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *2016 fourth international conference on 3D vision (3DV)*, pages 92–101. Ieee, 2016. 2

[33] Ian Huang, Panos Achlioptas, Tianyi Zhang, Sergey Tulyakov, Minhyuk Sung, and Guibas Leonidas. LADIS: Language disentanglement for 3D shape editing. In *Findings of Empirical Methods in Natural Language Processing*, 2022. 3

[34] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, 2022. 3, 4, 5, 6, 7, 8, 12, 15

[35] Wald Johanna, Avetisyan Armen, Navab Nassir, Tombari Federico, and Niessner Matthias. Rio: 3d object instance re-localization in changing indoor environments. *Proceedings IEEE International Conference on Computer Vision (ICCV)*, 2019. 2, 12

[36] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 3

[37] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2

[38] Juil Koo, Ian Huang, Panos Achlioptas, Leonidas J. Guibas, and Minhyuk Sung. PartGlot: Learning shape part segmentation from language reference games. In *CVPR*, 2022. 2

[39] Yuchen Li, Ujjwal Upadhyay, Habib Slim, Ahmed Abdelreheem, Arpita Prajapati, Suhail Pothigara, Peter Wonka, and Mohamed Elhoseiny. 3d compat: Composition of materials on parts of 3d things. In *European Conference on Computer Vision*, 2022. 2

[40] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*, 2004. 6

[41] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes, 2023. 2

[42] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3

[43] Eric Nichols and Fadi Botros. SpRL-CWW: Spatial relation classification with independent multi-class models. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*. Association for Computational Linguistics, 2015. 4, 12

[44] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002. 6

[45] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *CVPR*, 2015. 3

[46] David M. W. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *CoRR*, abs/2010.16061, 2020. 6

[47] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *NeurIPS*, 2017. 17

[48] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 2

[49] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-language model for 3D visual grounding. *Computing Research Repository (CoRR)*, abs/2107.03438, 2021. 3, 6

[50] Dávid Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. *ArXiv*, abs/2204.07761, 2022. 1, 2

[51] Akshit Sharma. Denserefer3d: A language and 3d dataset for coreference resolution and referring expression comprehension, 2022. 2

[52] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 2

[53] Habib Slim, Xiang Li, Yuchen Li, Mahmoud Ahmed, Mohamed Ayman, Ujjwal Upadhyay, Ahmed Abdelreheem, Arpit Prajapati, Suhail Pothigara, Peter Wonka, and Mohamed Elhoseiny. 3dcompat^{++}: An improved large-scale 3d vision dataset for compositional recognition, 2023. 2

[54] Jesse Thomason, Mohit Shridhar, Yonatan Bisk, Chris Paxton, and Luke Zettlemoyer. Language grounding with 3d objects. *Computing Research Repository (CoRR)*, abs/2107.12514, 2021. 2

[55] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 6

[56] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 15

[57] Guangzhi Wang, Hehe Fan, and Mohan Kankanhalli. Text to point cloud localization with relation-enhanced transformer. *arXiv preprint arXiv:2301.05372*, 2023. 2

[58] Heng Wang, Chaoyi Zhang, Jianhui Yu, and Weidong (Tom) Cai. Spatiality-guided transformer for 3d dense captioning on point clouds. In *IJCAI*, 2022. 3

[59] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. Embodied question answering in photorealistic environments with point cloud perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019. 2

[60] Ronald J. Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Computation*, 1989. 5, 15

[61] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding, 2022. 3

[62] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *Proceedings of the IEEE international conference on computer vision*, pages 1625–1632, 2013. 2

[63] Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, 2015. 4

[64] Muli Yang, Cheng Deng, Junchi Yan, Xianglong Liu, and Dacheng Tao. Learning unseen concepts via hierarchical decomposition and composition. In *CVPR*, 2020. 7, 15

[65] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *European Conference on Computer Vision*, pages 387–404. Springer, 2020. 3

[66] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4683–4693, 2019. 3

[67] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4682–4692, 2019. 15

[68] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. SAT: 2d semantics assisted training for 3D visual grounding. *Computing Research Repository (CoRR)*, abs/2105.11450, 2021. 3, 4, 6, 8, 13, 15

[69] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L Berg, and Dhruv Batra. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6309–6318, 2019. 2

[70] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018. 3

[71] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 3

[72] Zhihao Yuan, Xu Yan, Zhuo Li, Xuhao Li, Yao Guo, Shuguang Cui, and Zhen Li. Toward explainable and fine-grained 3d grounding through referring textual phrases. *arXiv preprint arXiv:2207.01821*, 2022. 2, 6

[73] Zhihao Yuan, Xu Yan, Yinghong Liao, Yao Guo, Guanbin Li, Shuguang Cui, and Zhen Li. X-trans2cap: Cross-modal knowledge transfer using transformer for 3d dense captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8563–8573, June 2022. 2, 3, 4, 5, 8

[74] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Zhen Li, and Shuguang Cui. InstanceRefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. *ICCV*, 2021. 3

[75] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, 2021. 3, 4, 6

[76] Yufeng Zhong, Long Xu, Jiebo Luo, and Lin Ma. Contextual modeling for 3d dense captioning on point clouds, 2022. 3