ScanRefer: 3D Object Localization in RGB-D Scans Using Natural Language

Dave Zhenyu Chen[1](✉), Angel X. Chang[2], and Matthias Nießner[1]

[1] Technical University of Munich, Munich, Germany
zhenyu.chen@tum.de
[2] Simon Fraser University, Burnaby, Canada

Abstract. We introduce the task of 3D object localization in RGB-D scans using natural language descriptions. As input, we assume a point cloud of a scanned 3D scene along with a free-form description of a specified target object. To address this task, we propose **ScanRefer**, learning a fused descriptor from 3D object proposals and encoded sentence embeddings. This fused descriptor correlates language expressions with geometric features, enabling regression of the 3D bounding box of a target object. We also introduce the ScanRefer dataset, containing $51{,}583$ descriptions of $11{,}046$ objects from 800 ScanNet [8] scenes. ScanRefer is the first large-scale effort to perform object localization via natural language expression directly in 3D (Code: https://daveredrum.github.io/ScanRefer/).

1 Introduction

In recent years, there has been tremendous progress in both semantic understanding and localization of objects in 2D images from natural language (also known as visual grounding). Datasets such as ReferIt [27], RefCOCO [70], and Flickr30K Entities [46] have enabled the development of various methods for visual grounding in 2D [21,22,38]. However, these methods and datasets are restricted to 2D images, where object localization fails to capture the true 3D extent of an object (see Fig. 1, left). This is a limitation for applications ranging from assistive robots to AR/VR agents where understanding the global 3D context and the physical size is important, e.g., finding objects in large spaces, interacting with them, and understanding their spatial relationships. Early work by Kong et al. [30] looked at coreference in 3D, but was limited to single-view RGB-D images.

In this work, we address these shortcomings by proposing the task of object localization using natural language directly in 3D space. Specifically, we develop a neural network architecture that localizes objects in 3D point clouds given natural language descriptions referring to the underlying objects; i.e., for a given text

Fig. 1. We introduce the task of object localization in 3D scenes using natural language. Given as input a 3D scene and a natural language expression, we predict the bounding box for the target 3D object (right). The counterpart 2D task (left) does not capture the physical extent of the 3D objects.

description in a 3D scene, we predict a corresponding 3D bounding box matching the best-described object. To facilitate the task, we collect the ScanRefer dataset, which provides natural language descriptions for RGB-D scans in Scan-Net [8]. In total, we acquire $51,583$ descriptions of $11,046$ objects. To the best of our knowledge, our ScanRefer dataset is the first large-scale effort that combines 3D scene semantics and free-form descriptions. In summary, our contributions are as follows:

– We introduce the task of localizing objects in 3D environments using natural language descriptions.
– We provide the ScanRefer dataset containing $51,583$ human-written free-form descriptions of $11,046$ objects in 3D scans.
– We propose a neural network architecture for localization based on language descriptions that directly fuses features from 2D images and language expressions with 3D point cloud features.
– We show that our end-to-end method outperforms the best 2D visual grounding method that simply backprojects its 2D predictions to 3D by a significant margin (9.04 Acc@0.5IoU vs. 22.39 Acc@0.5IoU).

2   Related Work

Grounding Referring Expressions in Images. There has been much work connecting images to natural language descriptions across tasks such as image captioning [25,26,58,63], text-to-image retrieval [24,60], and visual grounding [22,38,69]. The task of visual grounding (with variants also known as referring expression comprehension or phrase localization) is to localize a region described by a given referring expression, the query. Localization can be specified by a 2D bounding box [27,38,46] or a segmentation mask [21], with the input description being short phrases [27,46] or more complex descriptions [38].

Table 1. Comparison of referring expression datasets in terms of the number of objects (#objects), number of expressions (#expressions), average lengths of the expressions, data format and the 3D context.

Dataset	#objects	#expressions	AvgLeng	Data format	3D context
ReferIt [27]	96,654	130,364	3.51	Image	–
RefCOCO [70]	50,000	142,209	3.50	Image	–
Google RefExp [38]	49,820	95,010	8.40	Image	–
SUN-Spot [40]	3,245	7,990	14.04	Image	Depth
REVERIE [51]	4,140	21,702	18.00	Image	Panoramic image
ScanRefer (ours)	**11,046**	**51,583**	**20.27**	**3D scan**	**Depth, size, location, etc.**

Recently, Acharya et al. [1] proposed visual query detection where the input is a question. The focus of our work is to lift this task to 3D, focusing on complex descriptions that can localize an unique object in a scene.

Existing methods focus on predicting 2D bounding boxes [11,22,36,45,54, 59,60,69,70] and some predict segmentation masks [5,21,32,34,39,68]. A two-stage pipeline is common, where first an object detector, either unsupervised [73] or pretrained [53], is used to propose regions of interest, and then the regions are ranked by similarity to the query, with the highest scoring region provided as the final output. Other methods address the referring expression task with a single stage end-to-end network [21,42,67]. There are also approaches that incorporate syntax [16,35], use graph attention networks [61,65,66], speaker-listener models [38,71], weakly supervised methods [10,62,72] or tackle zero-shot settings for unseen nouns [55].

However, all these methods operate on 2D image datasets [27,46,70]. A recent dataset [40] integrates RGB-D images but lacks the complete 3D context beyond a single image. Qi et al. [51] study referring expressions in an embodied setting, where semantic annotations are projected from 3D to 2D bounding boxes on images observed by an agent. Our contribution is to lift NLP tasks to 3D by introducing the first large-scale effort that couples free-form descriptions to objects in 3D scans. Table 1 summarizes the difference between our ScanRefer dataset and existing 2D datasets.

Object Detection in 3D. Recent work on 3D object detection on volumetric grids [12,18,19,31,41] has been applied to several 3D RGB-D datasets [3,8,57]. As an alternative to regular grids, point-based methods, such as PointNet [49] or PointNet++ [50], have been used as backbones for 3D detection and/or object instance segmentation [13,64]. Recently, Qi et al. [48] introduced VoteNet, a 3D object detection method for point clouds based on Hough Voting [20]. Our
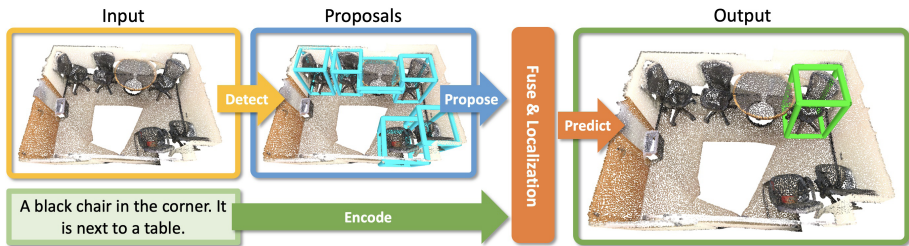
Fig. 2. Our task: ScanRefer takes as input a 3D scene point cloud and a description of an object in the scene, and predicts the object bounding box.
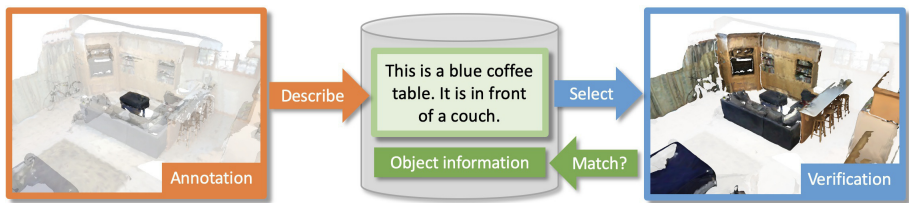

Fig. 3. Our data collection pipeline. The annotator writes a description for the focused object in the scene. Then, a verifier selects the objects that match the description. The selected object is compared with the target object to check that it can be uniquely identified by the description.

approach extracts geometric features in a similar fashion, but backprojects 2D feature information since the color signal is useful for describing 3D objects with natural language.

3D Vision and Language. Vision and language research is gaining popularity in image domains (e.g., image captioning [25,37,58,63], image-text matching [14, 15,23,29,33], and text-to-image generation [15,52,56]), but there is little work on vision and language in 3D. Chen et al. [6] learn a joint embedding of 3D shapes from ShapeNet [4] and corresponding natural language descriptions. Achlioptas et al. [2] disambiguate between different objects using language. Recent work has started to investigate grounding of language to 3D by identifying 3D bounding boxes of target objects for simple arrangements of primitive shapes of different colors [47]. Instead of focusing on isolated objects, we consider large 3D RGB-D reconstructions that are typical in semantic 3D scene understanding. A closely related work by Kong et al. [30] studied the problem of coreference in text description of single-view RGB-D images of scenes, where they aimed to connect noun phrases in a scene description to 3D bounding boxes of objects.

3   Task

We introduce the task of object localization in 3D scenes using natural language (Fig. 2). The input is a 3D scene and free-form text describing an object in the
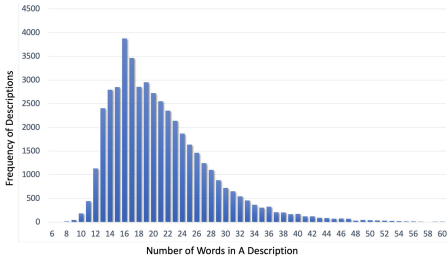
Fig. 4. Description lengths

Table 2. ScanRefer dataset statistics.

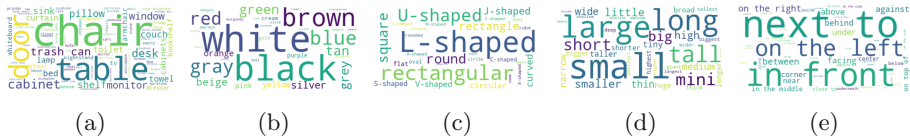|	
Number of descriptions	51,583
Number of scenes	800
Number of objects	11,046
Number of objects per scene	13.81
Number of descriptions per scene	64.48
Number of descriptions per object	4.67
Size of vocabulary	4,197
Average length of descriptions	20.27

(a)   (b)   (c)   (d)   (e)

Fig. 5. Word clouds of terms for (a) object names (b) colors (c) shapes (d) sizes, and (e) spatial relations for the ScanRefer dataset. Bigger fonts indicate more frequent terms in the descriptions.

scene. The scene is represented as a point cloud with additional features such as colors and normals for each point. The goal is to predict the 3D bounding box of the object that matches the input description.

4   Dataset

The ScanRefer dataset is based on ScanNet [8] which is composed of 1,613 RGB-D scans taken in 806 unique indoor environments. We provide 5 descriptions for each object in each scene, focusing on complete coverage of all objects that are present in the reconstruction. Here, we summarize the annotation process and statistics of our dataset (see supplement for more details).

4.1   Data Collection

We deploy a web-based annotation interface on Amazon Mechanical Turk (AMT) to collect object descriptions in the ScanNet scenes. The annotation pipeline consists of two stages: i) description collection, and ii) verification (Fig. 3). From each scene, we select objects to annotate by restricting to indoor furniture categories and excluding structural objects such as "Floor" and "Wall". We manually check the selected objects are recognizable and filter out objects with reconstructions that are too incomplete or hard to identify.

Annotation. The 3D web-based UI shows each object in context. The workers see all objects other than the target object faded out and a set of captured image frames to compensate for incomplete details in the reconstructions. The

initial viewpoint is random but includes the target object. Camera controls allow for adjusting the camera view to better examine the target object. We ask the annotator to describe the appearance of the target and its spatial location relative to other objects. To ensure the descriptions are informative, we require the annotator to provide at least two full sentences. We batch and randomize the tasks so that each object is described by five different workers.

Verification. We recruit trained workers (students) to verify that the descriptions are discriminative and correct. Verifiers are shown the 3D scene and a description, and are asked to select the objects (potentially multiple) in the scene that match the description. Descriptions that result in the wrong object or multiple objects are filtered out. Verifiers also correct spelling and wording issues in the description when necessary. We filter out 2,823 invalid descriptions that do not match the target objects and fix writing issues for 2,129 descriptions.

4.2   Dataset Statistics

We collected 51,583 descriptions for 800 ScanNet scenes[1]. On average, there are 13.81 objects, 64.48 descriptions per scene, and 4.67 descriptions per object after filtering (see Table 2 for basic statistics, Table 3 for sample descriptions, and Fig. 4 for the distribution of the description lengths). The descriptions are complex and diverse, covering over 250 types of common indoor objects, and exhibiting interesting linguistic phenomena. Due to the complexity of the descriptions, one of the key challenges of our task is to determine what parts of the description describe the target object, and what parts describe neighboring objects. Among those descriptions, 41,034 mention object attributes such as color, shape, size, etc. We find that many people use spatial language (98.7%), color (74.7%), and shape terms (64.9%). In contrast, only 14.2% of the descriptions convey size information. Figure 5 shows commonly used object names and attributes. Table 3 shows interesting expressions, including comparatives ("taller") and superlatives ("the biggest one"), as well as phrases involving ordinals such as "third from the wall". Overall, there are 672 and 2,734 descriptions with comparative and superlative phrases. We provide more detailed statistics in the supplement.

5   Method

Our architecture consists of two main modules: 1) detection & encoding; 2) fusion & localization (Fig. 6). The detection & encoding module encodes the input point cloud and description, and outputs the object proposals and the language embedding, which are fed into the fusion module to mask out invalid object proposals and produce the fused features. Finally, the object proposal with the highest confidence predicted by the localization module is chosen as the final output.

---

[1] 6 scenes are excluded since they do not contain any objects to describe.

Table 3. Examples from our dataset illustrating different types of phrases such as attributes (1–8) and parts (5), comparatives (4), superlatives (5), intra-class spatial relations (6), inter-class spatial relations (7) and ordinal numbers (8).

|	
1.	There is a brown wooden chair placed right against the wall
2.	This is a triangular shape table. The table is near the armchair
3.	The little nightstand. The nightstand is on the right of the bed
4.	This is a short trash can. It is in front of a taller trash can
5.	The couch is the biggest one below the picture. The couch has three seats and is brown
6.	This is a gray desk chair. This chair is the last one on the side closest to the open door
7.	The kitchen counter is covering the lower cabinets. The kitchen counter is under the upper cabinets that are mounted above
8.	This is a round bar stool. It is third from the wall
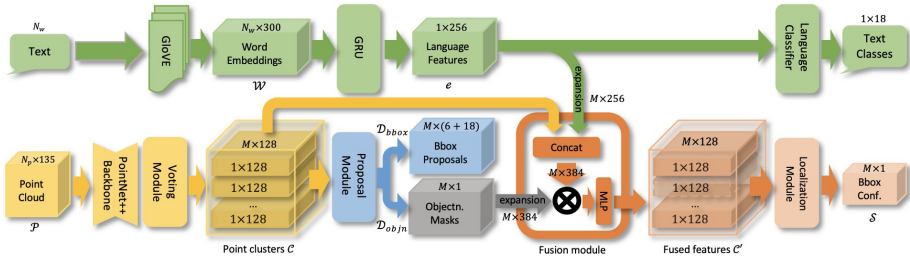
Fig. 6. ScanRefer architecture: The PointNet++ [50] backbone takes as input a point cloud and aggregates it to high-level point feature maps, which are then clustered and fused as object proposals by a voting module similar to Qi et al. [48]. Object proposals are masked by the objectness predictions, and then fused with the sentence embedding of the input descriptions, which is obtained by a GloVE [44] + GRU [7] embedding. In addition, an extra language-to-object classifier serves as a proxy loss. We apply a softmax function in the localization module to output the confidence scores for the object proposals.

5.1   Data Representations

Point Clouds. We randomly sample N_P vertices of one scan from ScanNet as the input point cloud $\mathcal{P} = \{(p_i, f_i)\}$, where $p_i \in \mathcal{R}^3$ represents the point coordinates in 3D space and f_i stands for additional point features such as colors and normals. Note that the point coordinates p_i provides only geometrical information and does not contain other visual information such as color and texture. Since descriptions of objects do refer to attributes such as color and

texture, we incorporate visual appearance by adapting the feature projection scheme in Dai et al. [9] to project multi-view image features $v_i \in \mathcal{R}^{128}$ to the point cloud. The image features are extracted using a pre-trained ENet [43]. Following Qi et al. [48], we also append the height of the point from the ground and normals to the new point features $f_i' \in \mathcal{R}^{135}$. The final point cloud data is prepared offline as $\mathcal{P}' = \{(p_i, f_i')\} \in \mathcal{R}^{N_P \times 135}$. We set N_P to $40,000$ in our experiments.

Descriptions. We tokenize the input description with SpaCy [17] and the N_W tokens to 300-dimensional word embedding vectors $\mathcal{W} = \{w_j\} \in \mathcal{R}^{N_W \times 300}$ using pretrained GloVE word embeddings [44].

5.2   Network Architecture

Our method takes as input the preprocessed point cloud $\mathcal{P}'$ and the word embedding sequence $\mathcal{W}$ representing the input description and outputs the 3D bounding box for the proposal which is most likely referred to by the input description. Conceptually, our localization pipeline consists of the following four stages: detection, encoding, fusion and localization.

Detection. As the first step in our network, we detect all probable objects in the given point cloud. To construct our detection module, we adapt the Point-Net++ [50] backbone and the voting module in Qi et al. [48] to process the point cloud input and aggregate all object candidates to individual clusters. The output from the voting module is a set of point clusters $\mathcal{C} \in \mathcal{R}^{M \times 128}$ representing all object proposals with enriched point features, where M is the upper bound of the number of proposals. Next, the proposal module takes in the point clusters and processes those clusters to predict the objectness mask $\mathcal{D}_{\text{objn}} \in \mathcal{R}^{M \times 1}$ and the axis-aligned bounding boxes $\mathcal{D}_{\text{bbox}} \in \mathcal{R}^{M \times (6+18)}$ for all M proposals, where each $\mathcal{D}_{\text{bbox}}^i = (c_x, c_y, c_z, r_x, r_y, r_z, l)$ consists of the box center c, the box lengths r and a vector $l \in \mathcal{R}^{18}$ representing the semantic predictions.

Encoding. The sequences of word embedding vectors of the input description are fed into a GRU cell [7] to aggregate the textual information. We take the final hidden state $e \in \mathcal{R}^{256}$ of the GRU cell as the final language embedding.

Fusion. The outputs from the previous detection and encoding modules are fed into the fusion module (orange block in Fig. 6, see supplemental for details) to integrate the point features together with the language embeddings. Specifically, each feature vector $c_i \in \mathcal{R}^{128}$ in the point cluster $\mathcal{C}$ is concatenated with the language embedding $e \in \mathcal{R}^{256}$ as the extended feature vector, which is then masked by the predicted objectness mask $\mathcal{D}_{\text{objn}}^i \in \{0, 1\}$ and fused by a multilayer perceptron as the final fused cluster features $C' = \{c_i'\} \in \mathcal{R}^{M \times 128}$.

Localization. The localization module aims to predict which of the proposed bounding boxes corresponds to the description. Point clusters with fused cluster features $\mathcal{C}' = \{c_i'\}$ are processed by a single layer perceptron to produce the raw scores of how likely each box is the target box. We use a softmax function to

squash all the raw scores into the interval of $[0, 1]$ as the localization confidences $S = \{s_i\} \in \mathcal{R}^{M \times 1}$ for the proposed M bounding boxes.

5.3   Loss Function

Localization Loss. For the predicted localization confidence $s_i \in [0, 1]$ for object proposal $\mathcal{D}^i_{\mathrm{bbox}}$, the target label is represented as $t_i \in \{0, 1\}$. Following the strategy of Yang et al. [67], we set the label t_j for the jth box that has the highest IoU score with the ground truth box as 1 and others as 0. We then use a cross-entropy loss as the localization loss $\mathcal{L}_{\mathrm{loc}} = -\sum_{i=1}^{M} t_i \log(s_i)$.

Object Detection Loss. We use the same detection loss $\mathcal{L}_{det}$ as introduced in Qi et al. [48] for object proposals $\mathcal{D}^i_{\mathrm{bbox}}$ and $\mathcal{D}^i_{\mathrm{objn}}$: $\mathcal{L}_{\mathrm{det}} = \mathcal{L}_{\mathrm{vote\text{-}reg}} + 0.5\mathcal{L}_{\mathrm{objn\text{-}cls}} + \mathcal{L}_{\mathrm{box}} + 0.1\mathcal{L}_{\mathrm{sem\text{-}cls}}$, where $\mathcal{L}_{\mathrm{vote\text{-}reg}}$, $\mathcal{L}_{\mathrm{objn\text{-}cls}}$, $\mathcal{L}_{\mathrm{box}}$ and $\mathcal{L}_{\mathrm{sem\text{-}cls}}$ represent the vote regression loss (defined in Qi et al. [48]), the objectness binary classification loss, box regression loss and the semantic classification loss for the 18 ScanNet benchmark classes, respectively. We ignore the bounding box orientations in our task and simplify $\mathcal{L}_{\mathrm{box}}$ as $\mathcal{L}_{\mathrm{box}} = \mathcal{L}_{\mathrm{center\text{-}reg}} + 0.1\mathcal{L}_{\mathrm{size\text{-}cls}} + \mathcal{L}_{\mathrm{size\text{-}reg}}$, where $\mathcal{L}_{\mathrm{center\text{-}reg}}$, $\mathcal{L}_{\mathrm{size\text{-}cls}}$ and $\mathcal{L}_{\mathrm{size\text{-}reg}}$ are used for regressing the box center, classifying the box size and regressing the box size, respectively. We refer readers to Qi et al. [48] for more details.

Language to Object Classification Loss. To further supervise the training, we include an object classification loss based on the input description. We consider the 18 ScanNet benchmark classes (excluding the label "Floor" and "Wall"). The language to object classification loss $\mathcal{L}_{\mathrm{cls}}$ is a multi-class cross-entropy loss.

Final Loss. The final loss is a linear combination of the localization loss, object detection loss and the language to object classification loss: $\mathcal{L} = \alpha\mathcal{L}_{\mathrm{loc}} + \beta\mathcal{L}_{\mathrm{det}} + \gamma\mathcal{L}_{\mathrm{cls}}$, where α, β and γ are the weights for the individual loss terms. After fine-tuning on the validation split, we set those weights to 1, 10, and 10 in our experiments to ensure the loss terms are roughly of the same magnitude.

5.4   Training and Inference

Training. During training, the detection and encoding modules propose object candidates as point clusters, which are then fed into the fusion and localization modules to fuse the features from the previous module and predict the final bounding boxes. We train the detection backbone end-to-end with the detection loss. In the localization module, we use a softmax function to compress the raw scores to $[0, 1]$. The higher the predicted confidence is, the more likely the proposal will be chosen as output. To filter out invalid object proposals, we use the predicted objectness mask to ensure that only positive proposals are taken into account. We set the maximum number of proposals M to 256 in practice.

Inference. Since there can be overlapping detections, we apply a non-maximum suppression module to suppress those overlapping proposals in the inference step. The remaining object proposals are fed into the localization module to predict the final score for each proposal. The number of object proposals is less than the upper bound M in the training step.

Implementation Details. We implement our architecture using PyTorch and train the model end-to-end using ADAM [28] with a learning rate of 1e−3. We train the model for roughly 130,000 iterations until convergence. To avoid overfitting, we set the weight decay factor to 1e−5 and apply data augmentations to our training data. For point clouds, we apply rotation about all three axes by a random angle in $[-5°, 5°]$ and randomly translate the point cloud within 0.5 meters in all directions. We rotate around all axes (not just up), since the ground alignment in ScanNet is imperfect.

6   Experiments

Train/Val/Test Split. Following the official ScanNet [8] split, we split our data into train/val/test sets with 36,665, 9,508 and 5,410 samples respectively, ensuring disjoint scenes for each split. Results and analysis are conducted on the val split (except for results in Table 4 bottom). The test set is hidden and will be reserved for the ScanRefer benchmark.

Metric. To evaluate the performance of our method, we measure the thresholded accuracy where the positive predictions have higher intersection over union (IoU) with the ground truths than the thresholds. Similar to work with 2D images, we use Acc@kIoU as our metric, where the threshold value k for IoU is set to 0.25 and 0.5 in our experiments.
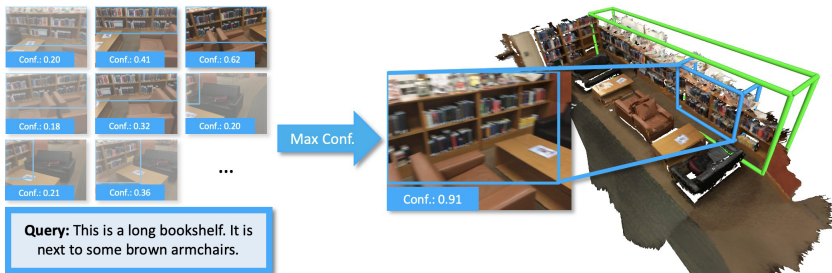

Fig. 7. Object localization in an image using a 2D grounding method and back-projecting the result to the 3D scene (blue box) vs. directly localizing in the 3D scene (green box). Grounding in 2D images suffers from the limited view of a single frame, which results in inaccurate 3D bounding boxes. (Color figure online)

Baselines. We design several baselines by 1) evaluating our language localization module on ground truth bounding boxes, 2) adapting 3D object detectors, and 3) adapting 2D referring methods to 3D using back-projection.

OracleCatRand & OracleRefer: To examine the difficulty of our task, we use an oracle with ground truth bounding boxes of objects, and predict the box by simply selecting a random box that matches the object category (OracleCatRand) or our trained fusion and localization modules (OracleRefer).

VoteNetRand & VoteNetBest: From the predicted object proposals of the VoteNet backbone [48], we select one of the bounding box proposals, either by selecting a box randomly with the correct semantic class label (VoteNetRand) or the best matching box given the ground truth (VoteNetBest). VoteNetBest provides an upper bound on how well the object detection component works for our task, while VoteNetRand provides a measure of whether additional information beyond the semantic label is required.

SCRC & One-Stage: 2D image baselines for referring expression comprehension by extending SCRC [22] and One-stage [67] to 3D using back-projection. Since 2D referring expression methods operate on a single image frame, we construct a 2D training set by using the recorded camera pose associated with each annotation to retrieve the frame from the scan video with the closest camera pose. At inference time, we sample frames from the scans (using every 20th frame) and predict the target 2D bounding boxes in each frame. We then select the 2D bounding box with the highest confidence score from the bounding box candidates and project it to 3D using the depth map for that frame (see Fig. 7).

Ours: We compare our full end-to-end model against using a pretrained VoteNet backbone with a trained GRU [7] for selecting a matching bounding box.

6.1   Task Difficulty

To understand how informative the input description is beyond capturing the object category, we analyze the performance of the methods on "unique" and "multiple" subsets with 1,875 and 7,663 samples from val split, respectively. The "unique" subset contains samples where only one unique object from a certain category matches the description, while the "multiple" subset contains ambiguous cases where there are multiple objects of the same category. For instance, if there is only one refrigerator in a scene, it is sufficient to identify that the sentence refers to a refrigerator. In contrast, if there are multiple objects of the same category in a scene (e.g., chair), the full description must be taken into account. From the OracleCatRand baseline, we see that information from the description, other than the object category, is necessary to disambiguate between multiple objects (see Table 4 Acc@0.5IoU multiple). From the OracleRefer baseline, we see that using our fused language module, we are able to improve beyond over selecting a random object of the same category (multiple Acc@0.5IoU increases from 17.84% to 32.00%), but we often fail to identify the correct object category (unique Acc@0.5IoU drops from 100.0% to 73.55%).

Table 4. Comparison of localization results obtained by our ScanRefer and baseline models. We measure percentage of predictions whose IoU with the ground truth boxes are greater than 0.25 and 0.5. We also report scores on "unique" and "multiple" subsets; unique means that there is only a single object of its class in the scene. We outperform all baselines by a significant margin.

	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
OracleCatRand (GT boxes + RandCat)	100.00	100.00	18.09	17.84	29.99	29.76
OracleRefer (GT boxes + GRU)	74.09	73.55	32.57	32.00	40.63	40.06
VoteNetRand (VoteNet [48] + RandCat)	34.34	19.35	5.73	2.81	10.00	5.28
VoteNetBest (VoteNet [48] + Best)	88.85	85.50	46.63	46.42	55.10	54.33
SCRC [22] + backproj	24.03	9.22	17.77	5.97	18.70	6.45
One-stage [67] + backproj	29.32	22.82	18.72	6.49	20.38	9.04
Ours (VoteNet [48] + GRU)	55.09	37.66	26.37	16.03	32.49	20.53
Ours (end-to-end)	**63.04**	**39.95**	**28.91**	**18.17**	**35.53**	**22.39**
Test results (ScanRefer benchmark)						
OracleRefer (GT boxes + GRU)	72.37	71.84	31.81	31.26	39.69	39.13
VoteNetBest (VoteNet [48] + Best)	86.78	83.85	45.54	45.33	53.82	53.07
Ours (VoteNet [48] + GRU)	57.67	36.96	28.31	15.16	34.90	20.05
Ours (end-to-end)	**62.90**	**40.31**	**30.88**	**16.54**	**38.06**	**21.87**

6.2   Quantitative Analysis

We evaluate the performance of our model against baselines on the val and the hidden test split of ScanRefer which serves as the ScanRefer benchmark (see Table 4). Note that for all results using Ours and VoteNet for object proposal, we take the average of 5 differently seeded subsamplings (of seed points and vote points) during inference (see supplemental for more details on experimental variance). Training the detection backbone jointly with the localization module (end-to-end) leads to a better performance when compared to the model trained separately (VoteNet [48] + GRU). However, as the accuracy gap between VoteNetBest and ours (end-to-end) indicates, there is still room for improving the match between language inputs and the visual signals. For the val split, we also include additional experiments on the 2D baselines and a comparison with VoteNetRand. With just category information, VoteNetRand is able to perform relatively well on the "unique" subset, but has trouble identifying the correct object in the "multiple" case. However, the gap between the VoteNetRand and OracleCatRand for the "unique" case shows that 3D object detection still need to be improved. Our method is able to improve over the bounding box predic-
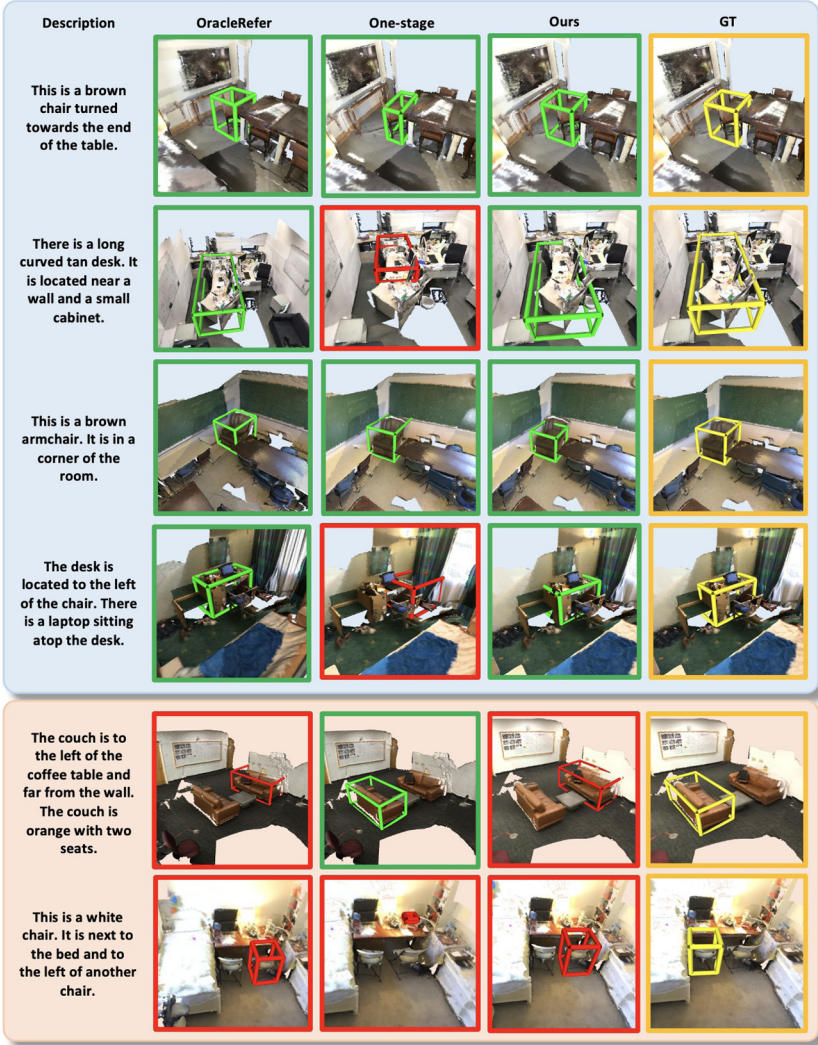
Fig. 8. Qualitative results from baseline methods and ScanRefer. Predicted boxes are marked green if they have an IoU score higher than 0.5, otherwise they are marked red. We show examples where our method produced good predictions (blue block) as well as failure cases (orange block). Image best viewed in color. (Color figure online)

tions from VoteNetRand, and leverages additional information in the description to differentiate between ambiguous objects. It adapts better to the 3D context compared to the 2D methods (SCRC and One-stage) which is limited by the view of a single frame (see Fig. 7 and Fig. 8).

Table 5. Ablation study with different features. We measure the percentages of predictions whose IoU with the ground truth boxes are greater than 0.25 and 0.5. Unique means that there is only a single object of its class in the scene.

	Unique		Multiple		Overall	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Ours (xyz)	50.83	31.81	24.38	13.98	29.51	17.43
Ours (xyz+rgb)	51.22	32.09	24.50	14.51	29.68	17.92
Ours (xyz+rgb+normals)	54.24	33.71	25.44	15.53	31.05	19.05
Ours (xyz+multiview)	56.69	35.32	25.83	14.26	31.63	19.75
Ours (xyz+multiview+normals)	55.27	35.51	25.95	16.29	31.64	20.02
Ours (xyz+lobjcls)	58.92	35.01	28.27	16.99	34.21	20.49
Ours (xyz+rgb+lobjcls)	60.11	37.89	27.21	16.49	33.59	20.65
Ours (xyz+rgb+normals+lobjcls)	60.54	39.19	26.95	16.69	33.47	21.06
Ours (xyz+multiview+lobjcls)	61.16	39.02	26.49	16.69	34.71	21.87
Ours (xyz+multiview+normals+lobjcls)	**63.04**	**39.95**	**28.91**	**18.17**	**35.53**	**22.39**

6.3 Qualitative Analysis

Figure 8 shows results produced by OracleRefer, One-stage, and our method. The successful localization cases in the green boxes show our architecture can handle the semantic correlation between the scene contexts and the textual descriptions. In contrast, even provided with a pool of ground truth proposals, OracleRefer sometimes still fails to predict correct bounding boxes, while One-stage is limited by the single view and hence cannot produce accurate bounding boxes in 3D space. The failure case of OracleRefer suggests that our fusion & localization module can still be improved. Some failure cases of our method are displayed in the orange block in Fig. 8, indicating that our architecture cannot handle all spatial relations to distinguish between ambiguous objects.

6.4 Ablation Studies

We conduct an ablation study on our model to examine what components and point cloud features contribute to the performance (see Table 5).

Does a Language-Based Object Classifier Help? To show the effectiveness of the extra supervision on input descriptions, we conduct an experiment with the language to object classifier (+lobjcls) and without. Architectures with a language to object classifier outperform ones without it. This indicates that it is helpful to predict the category of the target object based on the input description.

Do Colors Help? We compare our method trained with the geometry and multi-view image features (xyz+multiview+lobjcls) with a model trained with only geometry (xyz+lobjcls) and one trained with RGB values from the reconstructed meshes (xyz+rgb+lobjcls). ScanRefer trained with geometry and preprocessed multi-view image features outperforms the other two models. The

performance of models with color information are higher than those that use only geometry.

Do Other Features Help? We include normals from the ScanNet meshes to the input point cloud features and compare performance against networks trained without them. The additional 3D information improves performance. Our architecture trained with geometry, multi-view features, and normals (xyz+multiview+ normals+lobjcls) achieves the best performance among all ablations.

7   Conclusion

In this work, we introduce the task of localizing a target object in a 3D point cloud using natural language descriptions. We collect the ScanReferdataset which contains 51,583 unique descriptions for 11,046 objects from 800 ScanNet [8] scenes. We propose an end-to-end method for localizing an object with a free-formed description as reference, which first proposes point clusters of interest and then matches them to the embeddings of the input sentence. Our architecture is capable of learning the semantic similarities of the given contexts and regressing the bounding boxes for the target objects. Overall, we hope that our new dataset and method will enable future research in the 3D visual language field.

References

1. Acharya, M., Jariwala, K., Kanan, C.: VQD: visual query detection in natural scenes. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL) (2019)
2. Achlioptas, P., Fan, J., Hawkins, R.X., Goodman, N.D., Guibas, L.J.: ShapeGlot: learning language for shape differentiation. In: Proceedings of the International Conference on Computer Vision (ICCV) (2019)
3. Chang, A., et al.: Matterport3D: learning from RGB-D data in indoor environments. In: Proceedings of the International Conference on 3D Vision (3DV) (2017)
4. Chang, A.X., et al.: ShapeNet: an information-rich 3D model repository. arXiv preprint arXiv:1512.03012 (2015)

5. Chen, D.J., Jia, S., Lo, Y.C., Chen, H.T., Liu, T.L.: See-through-text grouping for referring image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7454–7463 (2019)
6. Chen, K., Choy, C.B., Savva, M., Chang, A.X., Funkhouser, T., Savarese, S.: Text2Shape: generating shapes from natural language by learning joint embeddings. In: Jawahar, C.V., Li, H., Mori, G., Schindler, K. (eds.) ACCV 2018. LNCS, vol. 11363, pp. 100–116. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-20893-6_7
7. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
8. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the Computer Vision and Pattern Recognition (CVPR) (2017)
9. Dai, A., Nießner, M.: 3DMV: joint 3D-multi-view prediction for 3D semantic scene segmentation. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11214, pp. 458–474. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01249-6_28
10. Datta, S., Sikka, K., Roy, A., Ahuja, K., Parikh, D., Divakaran, A.: Align2Ground: weakly supervised phrase grounding guided by image-caption alignment. In: Proceedings of the IEEE International Conference on Computer Vision (2019)
11. Dogan, P., Sigal, L., Gross, M.: Neural sequential phrase grounding (SeqGROUND). In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4175–4184 (2019)
12. Elich, C., Engelmann, F., Schult, J., Kontogianni, T., Leibe, B.: 3D-BEVIS: birds-eye-view instance segmentation. arXiv preprint arXiv:1904.02199 (2019)
13. Engelmann, F., Kontogianni, T., Leibe, B.: Dilated point convolutions: on the receptive field of point convolutions. arXiv preprint arXiv:1907.12046 (2019)
14. Feng, F., Wang, X., Li, R.: Cross-modal retrieval with correspondence autoencoder. In: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 7–16. ACM (2014)
15. Gu, J., Cai, J., Joty, S.R., Niu, L., Wang, G.: Look, imagine and match: improving textual-visual cross-modal retrieval with generative models. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7181–7189 (2018)
16. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. IEEE Trans. Pattern Anal. Mach. Intell. (2019)
17. Honnibal, M., Montani, I.: spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017, to appear)
18. Hou, J., Dai, A., Nießner, M.: 3D-SIC: 3D semantic instance completion for RGB-D scans. arXiv preprint arXiv:1904.12012 (2019)
19. Hou, J., Dai, A., Nießner, M.: 3D-SIS: 3D semantic instance segmentation of RGB-D scans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4421–4430 (2019)
20. Hough, P.V.: Machine analysis of bubble chamber pictures. In: Conference Proceedings, vol. 590914, pp. 554–558 (1959)
21. Hu, R., Rohrbach, M., Darrell, T.: Segmentation from natural language expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 108–124. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_7

22. Hu, R., Xu, H., Rohrbach, M., Feng, J., Saenko, K., Darrell, T.: Natural language object retrieval. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4555–4564 (2016)
23. Huang, Y., Wang, W., Wang, L.: Instance-aware image and sentence matching with selective multimodal LSTM. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2310–2318 (2017)
24. Huang, Y., Wu, Q., Song, C., Wang, L.: Learning semantic concepts and order for image and sentence matching. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6163–6171 (2018)
25. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3128–3137 (2015)
26. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in Neural Information Processing Systems, pp. 1889–1897 (2014)
27. Kazemzadeh, S., Ordonez, V., Matten, M., Berg, T.: ReferItGame: referring to objects in photographs of natural scenes. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 787–798 (2014)
28. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
29. Kiros, R., Salakhutdinov, R., Zemel, R.S.: Unifying visual-semantic embeddings with multimodal neural language models. arXiv preprint arXiv:1411.2539 (2014)
30. Kong, C., Lin, D., Bansal, M., Urtasun, R., Fidler, S.: What are you talking about? Text-to-image coreference. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3558–3565 (2014)
31. Lahoud, J., Ghanem, B., Pollefeys, M., Oswald, M.R.: 3D instance segmentation via multi-task metric learning. arXiv preprint arXiv:1906.08650 (2019)
32. Li, R., et al.: Referring image segmentation via recurrent refinement networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5745–5753 (2018)
33. Li, S., Xiao, T., Li, H., Yang, W., Wang, X.: Identity-aware textual-visual matching with latent co-attention. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1890–1899 (2017)
34. Liu, C., Lin, Z., Shen, X., Yang, J., Lu, X., Yuille, A.: Recurrent multimodal interaction for referring image segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1271–1280 (2017)
35. Liu, D., Zhang, H., Wu, F., Zha, Z.J.: Learning to assemble neural module tree networks for visual grounding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4673–4682 (2019)
36. Liu, X., Wang, Z., Shao, J., Wang, X., Li, H.: Improving referring expression grounding with cross-modal attention-guided erasing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1950–1959 (2019)
37. Lu, J., Xiong, C., Parikh, D., Socher, R.: Knowing when to look: adaptive attention via a visual sentinel for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 375–383 (2017)
38. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A.L., Murphy, K.: Generation and comprehension of unambiguous object descriptions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 11–20 (2016)

39. Margffoy-Tuay, E., Pérez, J.C., Botero, E., Arbeláez, P.: Dynamic multimodal instance segmentation guided by natural language queries. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11215, pp. 656–672. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01252-6_39

40. Mauceri, C., Palmer, M., Heckman, C.: SUN-Spot: an RGB-D dataset with spatial referring expressions. In: Proceedings of the IEEE International Conference on Computer Vision Workshops (2019)

41. Narita, G., Seno, T., Ishikawa, T., Kaji, Y.: PanopticFusion: online volumetric semantic mapping at the level of stuff and things. arXiv preprint arXiv:1903.01177 (2019)

42. Nguyen, A., Do, T.T., Reid, I., Caldwell, D.G., Tsagarakis, N.G.: Object captioning and retrieval with natural language. arXiv preprint arXiv:1803.06152 (2018)

43. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: ENet: a deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147 (2016)

44. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

45. Plummer, B.A., Kordas, P., Kiapour, M.H., Zheng, S., Piramuthu, R., Lazebnik, S.: Conditional image-text embedding networks. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11216, pp. 258–274. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01258-8_16

46. 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: Proceedings of the IEEE International Conference on Computer Vision, pp. 2641–2649 (2015)

47. 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. arXiv preprint arXiv:1910.01210 (2019)

48. Qi, C.R., Litany, O., He, K., Guibas, L.J.: Deep hough voting for 3D object detection in point clouds. In: Proceedings of the IEEE International Conference on Computer Vision (2019)

49. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: PointNet: deep learning on point sets for 3D classification and segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 652–660 (2017)

50. 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, pp. 5099–5108 (2017)

51. Qi, Y., Wu, Q., Anderson, P., Liu, M., Shen, C., van den Hengel, A.: REVERIE: remote embodied visual referring expression in real indoor environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2020)

52. Reed, S., Akata, Z., Yan, X., Logeswaran, L., Schiele, B., Lee, H.: Generative adversarial text to image synthesis. arXiv preprint arXiv:1605.05396 (2016)

53. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems, pp. 91–99 (2015)

54. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9905, pp. 817–834. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46448-0_49

55. Sadhu, A., Chen, K., Nevatia, R.: Zero-shot grounding of objects from natural language queries. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4694–4703 (2019)
56. Sharma, S., Suhubdy, D., Michalski, V., Kahou, S.E., Bengio, Y.: ChatPainter: improving text to image generation using dialogue. arXiv preprint arXiv:1802.08216 (2018)
57. Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: a RGB-D scene understanding benchmark suite. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 567–576 (2015)
58. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2015)
59. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Trans. Pattern Anal. Mach. Intell. **41**(2), 394–407 (2018)
60. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
61. Wang, P., Wu, Q., Cao, J., Shen, C., Gao, L., van den Hengel, A.: Neighbourhood watch: referring expression comprehension via language-guided graph attention networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1960–1968 (2019)
62. Xiao, F., Sigal, L., Jae Lee, Y.: Weakly-supervised visual grounding of phrases with linguistic structures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5945–5954 (2017)
63. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
64. Yang, B., et al.: Learning object bounding boxes for 3D instance segmentation on point clouds. arXiv preprint arXiv:1906.01140 (2019)
65. Yang, S., Li, G., Yu, Y.: Cross-modal relationship inference for grounding referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4145–4154 (2019)
66. Yang, S., Li, G., Yu, Y.: Dynamic graph attention for referring expression comprehension. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4644–4653 (2019)
67. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4683–4693 (2019)
68. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10502–10511 (2019)
69. Yu, L., et al.: MAttNet: modular attention network for referring expression comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1307–1315 (2018)
70. Yu, L., Poirson, P., Yang, S., Berg, A.C., Berg, T.L.: Modeling context in referring expressions. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9906, pp. 69–85. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_5
71. Yu, L., Tan, H., Bansal, M., Berg, T.L.: A joint speaker-listener-reinforcer model for referring expressions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7282–7290 (2017)

72. Zhao, F., Li, J., Zhao, J., Feng, J.: Weakly supervised phrase localization with multi-scale anchored transformer network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5696–5705 (2018)
73. Zitnick, C.L., Dollár, P.: Edge boxes: locating object proposals from edges. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 391–405. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_26