Refer-it-in-RGBD: A Bottom-up Approach for 3D Visual Grounding in RGBD Images

Haolin Liu[1,2] Anran Lin[1] Xiaoguang Han[1,2,*] Lei Yang[4]
Yizhou Yu[3,4] Shuguang Cui[1,2]
[1]SRIBD, CUHK-Shenzhen[†] [2]FNii, CUHK-Shenzhen[‡]
[3]Deepwise AI Lab [4]The University of Hong Kong

Abstract

Grounding referring expressions in RGBD image has been an emerging field. We present a novel task of 3D visual grounding in single-view RGBD image where the referred objects are often only partially scanned due to occlusion. In contrast to previous works that directly generate object proposals for grounding in the 3D scenes, we propose a bottom-up approach to gradually aggregate content-aware information, effectively addressing the challenge posed by the partial geometry. Our approach first fuses the language and the visual features at the bottom level to generate a heatmap that coarsely localizes the relevant regions in the RGBD image. Then our approach conducts an adaptive feature learning based on the heatmap and performs the object-level matching with another visio-linguistic fusion to finally ground the referred object. We evaluate the proposed method by comparing to the state-of-the-art methods on both the RGBD images extracted from the ScanRefer dataset and our newly collected SUNRefer dataset. Experiments show that our method outperforms the previous methods by a large margin (by 11.2% and 15.6% Acc@0.5) on both datasets.

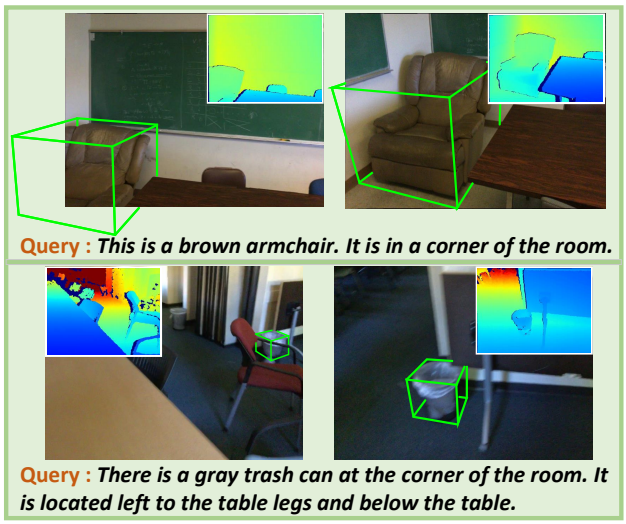

Figure 1: We present a novel task of 3D visual grounding in single-view RGBD images given a referring expression, and propose a bottom-up neural approach to address it. Our goal is to estimate the bounding box that encloses the full shape of the referred object even this object is only partially observed (top-left). Predicted bounding boxes of the referred objects are in green.

1. Introduction

Localizing objects described by referring expressions in vision signals, also known as *visual grounding*, has long been a major motive for robotics and embodied vision. So far, we have seen growing efforts devoted to visual grounding in images [17, 36, 13, 40, 24, 29, 33, 5, 41, 11, 42, 10, 9, 12, 19, 47, 18, 35, 38, 39, 20] and videos [46, 45, 43, 37, 30, 31, 44]. Suppose that a robot is going to take 'the spoon on the table in the kitchen' following your command [14, 23]; this would require a more accurate localization result, preferable the 3D coordinate of the referred object rather than a 2D bounding box. Recent works [4, 1] extend the visual grounding task to 3D scenes [7] and localize the object referred by a natural language expression. While promising results are produced, these methods can only perform 3D visual grounding in complete scenes that are reconstructed and/or segmented [4, 1] in advance. Thus, they are not readily applicable to single-view RGBD images with partial observation, RGBD streaming data, or any dynamically changing environments.

To this end, we propose a novel task for 3D visual grounding: Given a single-view RGBD image of a scene,

*Corresponding Email: hanxiaoguang@cuhk.edu.cn
†Shenzhen Research Institute of Big Data
‡The Future Network of Intelligence Institute, The Chinese University of Hong Kong, Shenzhen

we aim at estimating the 3D bounding box of the full target object described by a given referring expression. While this novel task opens up many promising possibilities, it also poses a major challenge due to the nature of single-view RGBD images that they contain only incomplete information about the scene and often partial observation of the referred object. Compared to 2D visual grounding on image and 3D visual grounding on complete scene where geometry information is complete in 2D and 3D space, various occlusion cases in our task require a holistic understanding of the object geometry to infer the 3D bounding boxes enclosing the full target object.

Image-based grounding methods may be applied to RGBD images but require an extra effort to lift the produced 2D bounding boxes to 3D. Chen et al. [4] proposed a one-stage search and match strategy for 3D grounding. However, it fails to handle single-view RGBD images where the referred objects are partially observed. The reason is two-fold: First, it is inadequate to directly match between features of these object proposals and the referring expression to achieve reliable grounding, as each proposal contains only incomplete information due to the partial observation. Moreover, this is worsened by the fact that only content-free object proposals are generated by a detection network that searches the scene globally, thus failing to accumulate useful information about the referred object for grounding.

With these observations, we propose a novel, bottom-up matching approach for fine-grained grounding of the referred objects in given single-view RGBD images. To this end, our approach first matches the query expression to the input RGBD image and generates a content-aware heatmap on the voxel domain converted from the RGBD image. This bottom-level matching amounts to coarse localization of regions, which are relevant to the referred object. Then, based on the content-aware coarse localization, an adaptive search-and-match strategy is employed. This enables our network to conduct fine-grained search in the relevant regions and generate visio-linguistic features by fusing the query with more informative features from the visual modality. These fused features are used to generate and refine the 3D object proposals to the final bounding box enclosing the target object in the given referring expression. Compared to previous works, our bottom-up approach exploits the language features at different levels, and thus enables our network to be content-aware during searching and matching stages. The adaptive search guided by the content-aware heatmap also ensures the feature learning to be concentrated in the relevant regions, mitigating the challenge posed by the partial geometry.

Mauceri et al. [22] present the SUN-Spot dataset that provides spatial referring expressions to raw single-view RGBD images in the SUNRGBD dataset[32]. However, the amount of language annotations as well as their linguistic variations (only spatial references) are inadequate. Alternatively, one can extract the RGBD frames from 3D scene datasets [3, 28, 7, 4, 1] that also provide rich object-centric language descriptions. Yet, referring expressions provided in these scene datasets are constructed based on the scene context; they may contain other supporting objects that exist in the scene but are not observed in a particular frame. This artifact between annotations and the extracted frames from the 3D scenes motivates us to contribute a large-scale annotation dataset, SUNRefer, to facilitate future studies on visual grounding in single-view RGBD images. Built on the SUNRGBD dataset, our dataset contains 7,699 RGBD images with a total of 38,495 annotations of referring expressions on 7,699 objects.

We evaluated our proposed 3D visual grounding method on both SUNRefer, our newly constructed dataset, and the ScanRefer dataset using the extracted RGBD frames. We show that our approach outperforms the state-of-the-art methods by a significant margin and validate our desgin choices via extensive ablation study. Our method can also be applied to 3D visual grounding in streaming RGBD images at a processing rate of 10 frame-per-second.

Our key contributions are summarized as follows:

1) We present a novel and challenging task – 3D visual grounding in single-view RGBD images with possible incomplete information or partial occlusion

2) We propose a content-aware, bottom-up approach that significantly outperforms the state-of-the-art methods on both our newly collected dataset and the ScanRefer dataset.

3) We contribute a large-scale dataset of referring phrases and the corresponding ground-truth bounding boxes for a large amount of publicly available RGBD images.

2. Related Work

Given a referring expression along with an input image, 2D visual grounding tasks aim to estimate a 2D bounding box or instance segmentation mask that localizes in the image the object most relevant to the referring expression. Previous works [13, 40, 24, 29, 33, 5, 41, 11, 42, 10, 9, 12, 19, 47, 36] usually adopt a two-stage approach. Firstly, object proposals or segmentation masks are produced by pre-trained models. Then, these methods rank the matching scores between the referring expression and the proposals to retrieve the best-matched object proposal as the grounding result. However, performance of the two-stage approach is bounded by the pre-trained models chosen. one-stage methods are also proposed [18, 35, 38, 39, 20] which produce both object proposals and matching scores. In addition, similar approaches can be applied to object visual grounding in streaming video frames [43, 37, 30, 31, 44, 46, 45] to ground objects or referring expressions in videos.

However, visual grounding in 2D images is limited since

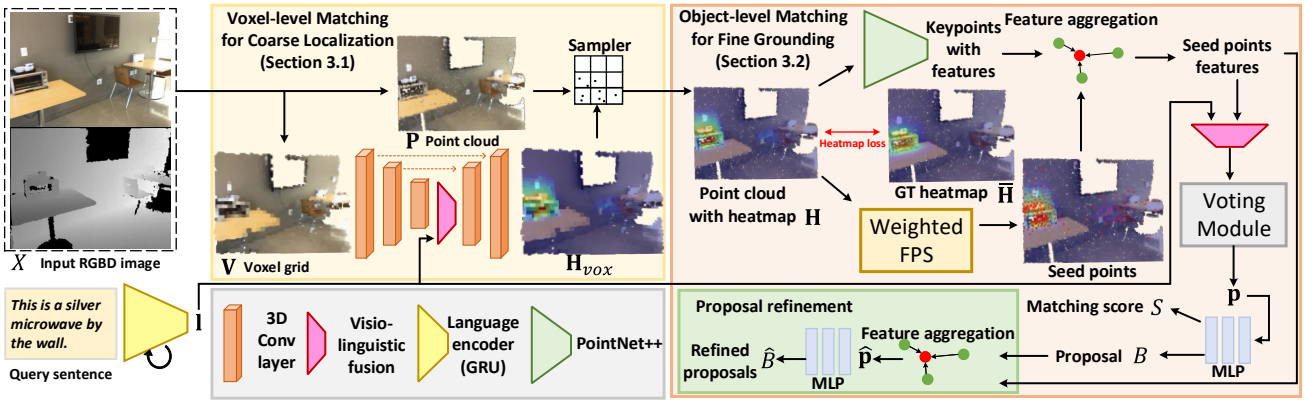

Figure 2: Overview of the proposed method. Our bottom-up approach consists of two main modules. Given an RGBD image and a referring sentence, we first match the input pair at the voxel level to coarsely localize the relevant regions. This is followed by an adaptive feature learning enabled by weighted FPS based on the coarse localization result. An object-level matching between the language feature and the content-aware visual features is performed to generate object proposals for fine grounding. Finally, the referring sentence is grounded in the object proposal with the highest score.

it is only able to localize the referred objects with 2D bounding boxes in the image domain. It is still unavailable to get the precise 3D locations of the objects which are desirable for more advanced tasks, such as embodied AI[28, 21, 2, 34, 8].

Grounding texts in the indoor scenes has been an important research field. An earlier work [16] explored the text-to-image coreference in RGBD scenes, which takes sentences describing the whole scene as input and matches each noun word with an object in the scene. This differs from ours task where input sentences are referring to a particular object in the scene. Qi et al. [28] proposed a language-guided navigation task in the indoor scenes where an agent needs to follow the language instructions to find a remote target. In this study, we focus on the task of 3D visual grounding, i.e., localizing the 3D bounding box of the referred object in the single-view RGBD images. A close related work is ScanRefer [4] where the authors propose a 3D visual grounding task to find the 3D bounding boxes of referred objects in a complete RGBD scene which requires reconstruction from raw scans beforehand. ReferIt3D [1] is another recent work on 3D visual grounding in scenes. This work assumes the availability of segmentation masks of object instances for both training and testing. We are different to both previous works in that we focus on grounding in single-view RGBD images where partial occlusions or incomplete data are our major challenges.

We also differ from ScanRefer in terms of methodology. ScanRefer employs a one-stage matching between the object proposals generated by [26] and the language features to obtain the grounding results. In contrast, we propose to conduct 3D visual grounding using a bottom-up approach. We first conduct a visual-language matching on the voxel domain to coarsely localize the relevant regions. Then, we accumulate features from these relevant regions for the following fine-grained matching on the object level.

3. Method

Overview We present a novel task of grounding a referring expression of a 3D object in a given RGBD image. The expected output of this task is the 3D bounding box of the full object referred in the language expression. This is not a trivial task as the input RGBD image may contain only part of the referred object. We propose a novel neural solution to tackle this task in a bottom-up approach. Our framework consists of two main modules as shown in Figure 2. We first produce a confidence heatmap from voxel-level matching between voxel and the query sentence, to coarsely localizes the relevant regions to the referred object in the image. Next, an adaptive search strategy is proposed to conduct fine-grained, content-aware search for the referred object based on the heatmap. Finally, we fuse the features accumulated from adaptive search and the language feature to conduct object-level matching, yielding matching scores and axis-aligned bounding box.

3.1. Voxel-level Matching for Coarse Localization

Given a referring expression and an RGBD image, this module aims to learn from the two input modalities and produces a heatmap using the global context to localize relevant regions to the referred object in the input image.

Heatmap generation. The input RGBD image $\mathbf{X}$ is converted to a voxel grid $\mathbf{V}$ with $0.05m$ cell size. We adopt a U-Net model [29], an encoder-decoder structure with skip links constructed by 3D sparse convolutions [6], to regress a voxel-wise heatmap $\mathbf{H}_{vox}$ indicating the relevancy of the

referred object in the voxel grid. For further grounding in the point cloud data, we project the produced heatmap $\mathbf{H}_{vox}$ to the point cloud $\mathbf{P}$, and eventually yield a heatmap $\mathbf{H}$ on the point cloud domain $\mathbf{P}$.

The voxel grid is chosen to produce the heatmap due to two considerations. First, they can maintain the spatial awareness of the objects compared to the 2D images. Second, they are insensitive to superficial details compared to the point clouds. This design choice is further validated in our ablation study (see Table 4 and Figure 7).

Fusing language features for coarse localization. To fuse the language input for the grounding purpose, We first pass the given referring expression to a language encoder and obtain language feature $\mathbf{l}$. In particular, We employ GloVe [25] to obtain the vector embedding for each word, and use Gated Recurrent Units (GRU) to extract feature $\mathbf{l}$ from the sequence of word embedding vectors.

We then introduce language feature $\mathbf{l}$ to modulate the heatmap generation process for content-aware coarse localization. Specifically, we concatenate the output feature maps of the voxel encoder with $\mathbf{l}$, and feed the concatenated features to a visio-linguistic encoder to obtain the fused features. It is followed by the voxel decoder that generates the content-aware heatmap $\mathbf{H}_{vox}$ that coarsely localizes the regions relevant to the referring expression.

3.2. Object-level Matching for Fine Grounding

We leverage the generated heatmap to perform fine-grained grounding by a high-level object and language matching. To this end, we first propose a content-aware sampling to generate seed points that concentrated in the regions highlighted by the heatmap. We then conduct a visio-linguistic fusion between the seeds and language feature, and employ a voting mechanism proposed in [26] for proposals generation and object-language matching.

Adaptive feature learning. Based on the heatmap, we conduct an adaptive feature learning using PointNet++[27] to aggregate information only at the relevant regions.

To this end, we employ a weighted farthest point sampling (weighted FPS) with a modified distance metric to obtain adaptive samples guided by the heatmap. In particular, we use the heatmap value at the query point $\mathbf{q}$ as the scaling factor, and modify the Euclidean distance metric used in farthest point sampling (FPS) as follow:

$$\hat{d}(\mathbf{q}, \mathbf{c}) = h(\mathbf{q}) d(\mathbf{q}, \mathbf{c}) \tag{1}$$

where $d(\mathbf{q}, \mathbf{c})$ is the Euclidean distance between points $\mathbf{q}$ and $\mathbf{c}$. $h(\mathbf{q})$ is the heat value of point $\mathbf{q}$. This weighted FPS can ensure that a point with a higher heat value will have a higher chance to be chosen, and thus result in seed points densely distributed in relevant regions identified by the heatmap. The weighted FPS also maintains the uniformity of the sampled points in the relevant region (as shown in Figure 6).

The adaptive sampled seed points are denoted as $\{\mathbf{s}_i\}_{i=1}^{M}$. To obtain the features associated with the seed points, PointNet++[27] is first employed to extract features (residing in a set of key-points) from point cloud $\mathbf{P}$. We then use the Feature Propagation module as in [27] to aggregate features from the nearby key-points to our seed points.

Visual-language fusion for proposal generation. Next, we fuse the language features and visual features sampled in a content-aware manner for the object-level grounding. Specifically, we concatenate seed point features $\mathbf{f}_i$ with the language feature $\mathbf{l}$, and process them using another visio-linguistic encoder for fusion.

$$\hat{\mathbf{f}}_i = E_{VL}(cat(\mathbf{f}_i, \mathbf{l})). \tag{2}$$

We denote the seed points with associated fused features as $\{\hat{\mathbf{s}}_i = [\mathbf{s}_i \in \mathbb{R}^3,\ \hat{\mathbf{f}}_i \in \mathbb{R}^C]\}_{i=1}^{M}$.

We adopt the voting module proposed in [26] to generate votes $\{\mathbf{v}_i\}_{i=1}^{M}$ pointing at objects' center from the obtained seed points $\{\hat{\mathbf{s}}_i\}_{i=1}^{M}$. Then, the votes are clustered in the 3D space to produce object proposal features $\{\mathbf{p}_k \in \mathbb{R}^{3+C}\}_{k=1}^{K}$, where the first three channels represent the cluster center of the proposal. Finally, we use a shared MLP-based regressor to estimate, for each of the K proposal features, the proposal box $B_k = (x, y, z, w, l, h)$ and its score S_k to match the referred object. Among K candidate proposals, we choose the highest scored proposal as our predicted bounding box to localize the referred object.

Proposal refinement: As the proposal features are aggregated from each cluster formed by the votes, they contain only local information from that cluster. Thus, we provide an optional module to refine the proposal features $\{\mathbf{p}_k\}_{k=1}^{K}$. We perform another feature aggregation among $\{\mathbf{p}_k\}_{k=1}^{K}$ and $\{\mathbf{s}_i\}_{i=1}^{M}$ to produce a set of refined proposal features $\{\hat{\mathbf{p}}_k\}_{k=1}^{K}$. Then we feed $\{\hat{\mathbf{p}}_k\}_{k=1}^{K}$ to the regressor to produce final proposal boxes $\{\hat{B}_k\}_{k=1}^{K}$.

3.3. Loss Function

We train our network with the following loss function,

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{heat} + \lambda_2 \mathcal{L}_{vote} + \lambda_3 \mathcal{L}_{match} + \lambda_4 \mathcal{L}_{resp} \tag{3}$$

Heatmap loss $\mathcal{L}_{heat}$ encourages the generation of the distinctive heatmap which is able to determine regions' relevancy to the sentence query. We firstly construct a ground-truth (GT) point cloud heatmap $\bar{\mathbf{H}}$ according to the target object's location. We calculate it using a Gaussian kernel so that further distance from the target object's center will have diminished value. $\mathcal{L}_{heat}$ is obtained by the mean square error (MSE) between $\bar{\mathbf{H}}$ and generated $\mathbf{H}$ as follow:

$$\mathcal{L}_{heat} = \left\| \bar{\mathbf{H}} - \mathbf{H} \right\|_2 \tag{4}$$

Table 1: Statistics of SUNRefer dataset comparing some public 3d referring expression datasets.

dataset	# of annotations	# of objects	Annotated on	averaged description length
ScanRefer [4]	51,583	11,046	704 scenes	20.3
Nr3D [1]	41,503	5,879	642 scenes	11.4
REVERIE [28]	21,702	4,140	90 scenes	18.0
SUN-Spot [22]	7,990	3,245	1,948 RGBD images	14.1
SUNRefer	38,495	7,699	7,699 RGBD images	16.3

Vote loss $\mathcal{L}_{vote}$is defined following [26]:

$$\mathcal{L}_{vote} = \frac{1}{M_{pos}} \sum_{i} \|v_i - \bar{v}_i\|_2 \, \mathbb{1}[s_i \text{ on objects}], \quad (5)$$

where M_{pos} is the total number of seed points on objects, v_i and $\bar{v}_i$ are the predicted vote coordinates and GT centers of the object that the seed point s_i resides on, if applicable. Thus, the vote loss enforces the network to produce a correct vote, when a seed point on any object is given.

Matching loss $\mathcal{L}_{match}$ is designed to ensure matching score S_k to approximate the Intersection of Union (IoU) between the proposal B_k and the GT bounding box of the target object $\bar{B}$. This way, it can return high responses to correct proposals while suppressing any mismatched proposals. Thus, the matching loss is computed using MSE as follows:

$$\mathcal{L}_{match} = \frac{1}{K} \sum_{i=1}^{K} \|IoU_k - S_k\|_2 \quad (6)$$

where K is the number of proposals and IoU_k denotes the IoU between B_k and $\bar{B}$.

Response loss $\mathcal{L}_{resp}$ aims to pull the proposal B_i that has the maximum IoU with GT bounding box closer to it. Given target bounding box $\bar{B}$, The loss is written as follow:

$$\mathcal{L}_{resp} = \left\|B_i - \bar{B}\right\|_2, \quad i = \arg\max_k(IoU_k). \quad (7)$$

When proposal refinement is employed, $\{B_k\}_{k=1}^{K}$ will be supervised by GT bounding box $\bar{B}^{partial}$ enclosing the partial object in the input and the refined proposals $\{\hat{B}_k\}_{k=1}^{K}$ will be supervised by GT bounding box $\bar{B}^{intact}$ enclosing the full object. Otherwise, Proposals $\{B_k\}_{k=1}^{K}$ will be directly supervised by $\bar{B}^{intact}$.

4. SUNRefer Dataset

Due to the lack of a large-scale dataset dedicated to visual grounding in single-view RGBD images, we contribute a large-scale referring expression dataset, named SUNRefer. Our dataset contains 7,699 RGBD images from SUN-RGBD [32] and annotation of 3d orientated bounding boxes enclosing the full objects. We hire annotation workers to annotate each target object with referring expressions.

Given an RGBD image with the bounding box enclosing an object, we ask the annotators to describe this object using its own attributes (e.g. color, shape, and material) and/or spatial relationship in the surrounding environment with multiple sentences. A description is expected to distinguish the target object from its neighboring objects and those with similar appearances but different locations.

For each target object, we collect five descriptions from different annotators in order to ensure linguistic diversity. Verification was conducted to obtain high-quality annotations. An example of SUNRefer dataset is shown in Figure 3. Statistics of the SUNRefer dataset and comparison with some public datasets are shown in Table 1. More details can be found in our supplementary.

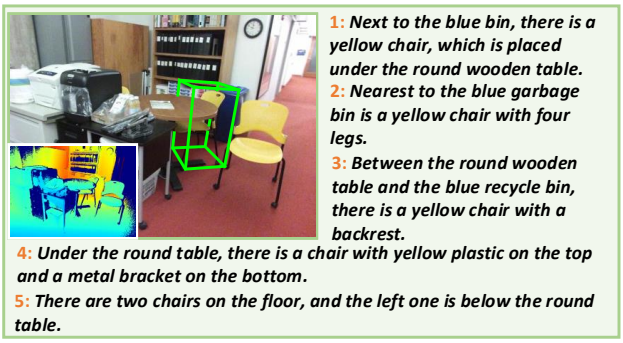

Figure 3: An example of the SUNRefer dataset with five different language descriptions referring to the chair enclosed by the green bounding box.

5. Experiments

5.1. Implementation Details

We first train the language-guided coarse localization module (and the GRU) independently with the heatmap loss for 10 epochs. Based on this pre-trained module, we then train the entire framework with the described loss function (Equation 3) for another 60 epochs. The coefficients for the loss terms are $\lambda_1 = 0.5, \lambda_2 = 0.5, \lambda_3 = 2.0, \lambda_4 = 1.0$. The Adam optimizer [15] is used for optimization. The learning rate is set to 0.0001 for the coarse localization module, and 0.001 for the rest of the model parameters. Both learning rates are decayed by 0.2 at the 50-th epoch. Data aug-

Table 2: Quantitative comparison between our methods and the state-of-the-art methods. Our method outperforms the comparing methods by a large margin on both SUNRefer and ScanRefer datasets under the single-view RGBD setting.

Dataset	SUNRefer			ScanRefer (single-RGBD)		
Methods	Acc@0.5	Acc@0.25	R@5	Acc@0.5	Acc@0.25	R@5
ReSC[38]	7.4	23.6	9.4	12.4	34.3	17.1
One Stage[39]	2.6	12.8	7.9	5.8	26.1	21.7
ScanRefer[4]	20.3	45.0	23.6	20.3	48.2	33.8
Ours	**35.9**	**49.6**	**52.9**	**31.5**	**56.5**	**49.3**

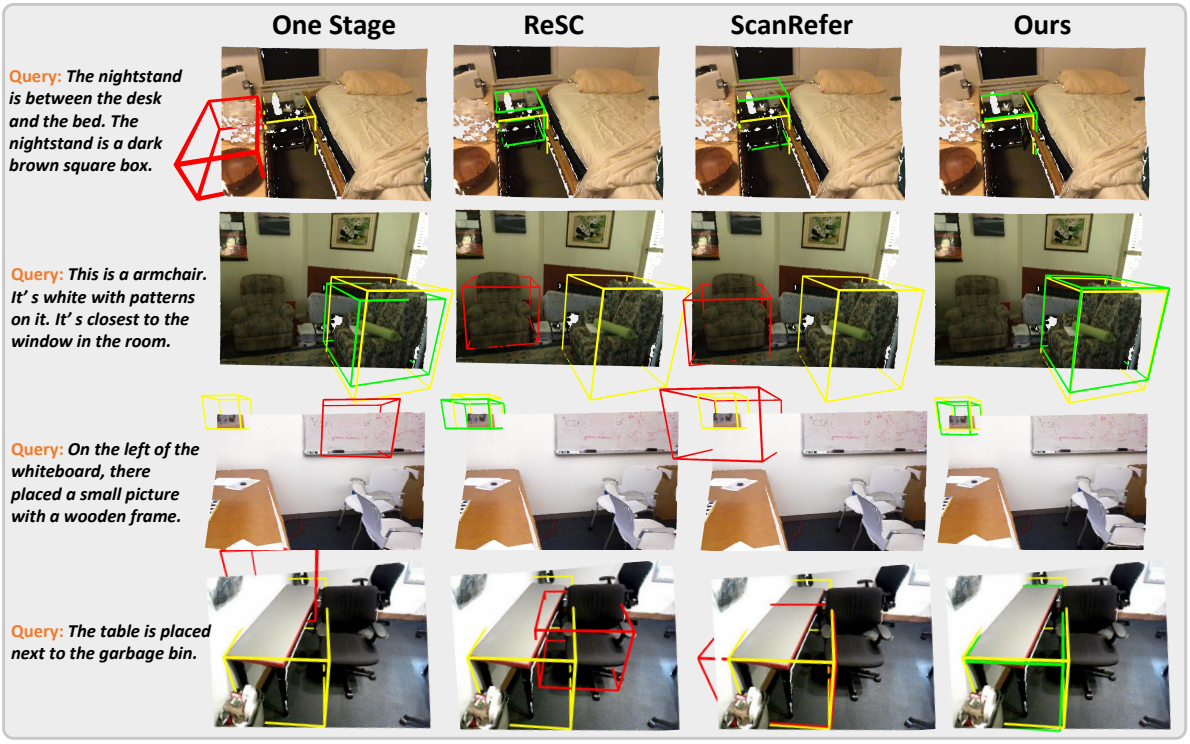

Figure 4: Comparison between our method and the state-of-the-art. A bounding box is considered as a successful prediction (green) if it has an IoU larger than 0.5 with the ground-truth box (yellow) of the entire object even if it is partially observed; otherwise, it is considered as a failure one (red). More visual results are presented in the supplementary.

Table 3: Quantitative comparison against the state-of-the-art methods on the ScanRefer dataset. Our method achieves comparable, if not better, results comparing to two variants of ScanRefer even under the whole-scene setting.

Dataset	ScanRefer (whole scene)		
Methods	Acc@0.5	Acc@0.25	R@5
ScanRefer-full	27.1	**41.0**	43.5
ScanRefer-xyz+rgb	23.1	35.9	39.7
FPS baseline-xyz+rgb	24.2	35.1	44.7
Ours-xyz+rgb	**29.0**	40.2	**47.1**

mentation is applied to both voxel and point cloud models by perturbing them with small translations and rotations in all three axes. The batch size is set to 14 for training. The training time is approximately 30 hours on a single NVIDIA RTX-2080-TI GPU. More implementation details are supplied in the supplemenetary materials

5.2. Evaluation Metric

We use the highest scored proposal from our network as the predicted bounding box of the referred object. We follow [4] to use Acc@0.5 and Acc@0.25 as our evaluation metrics. We also introduce another evaluation metric, R@5, which is widely used in 2D visual grounding to reflect the retrieval accuracy, where the retrieval is considered correct if at least one of the proposals of top-5 scores has an IoU > 0.5 with the GT box. Thus, Acc@$\{0.5,\ 0.25\}$ focus on how close the highest scored proposal to the GT box, and R@5 evaluates the quality of the proposal candidates.

5.3. Comparison with SOTA

Comparison in the single-view RGBD setting: We compare the proposed method with ScanRefer [4] and the other two methods extended from 2D grounding, i.e., One

Table 4: Quantitative results of ablation study. We compare our full model using the voxel-based heatmap and the weighted FPS with several alternative configurations. The comparisons validate our design choices.

Dataset	SUNRefer			ScanRefer (single-RGBD)		
Methods	Acc@0.5	Acc@0.25	R@5	Acc@0.5	Acc@0.25	R@5
Ours (voxel + weighted FPS)	**35.9**	**49.6**	**52.9**	**31.5**	**56.5**	**49.3**
Voxel-level matching only	12.2	31.0	18.4	16.8	39.8	28.4
FPS baseline w/o voxel-level matching	28.6	45.7	43.5	25.4	52.7	39.2
Weigthed random sampling	34.1	48.2	51.9	29.4	56.4	46.8
image modality	33.6	48.3	48.3	29.2	54.9	45.1
point cloud modality	34.0	47.8	50.4	29.8	55.2	48.3
w/o proposal refinement	35.8	49.5	52.9	30.7	55.6	47.7

Stage [39] and ReSC [38], under the single-view RGBD setting on both ScanRefer and SUNRefer datasets.

As the ScanRefer dataset provides only the complete scene annotations, we extract the raw RGBD images from ScanNet [7]. For each extracted images, we identify the objects that are visible in the image and use these objects and their corresponding referring sentences as paired input to our network for training and testing. We select a subset of the images in which the objects have different visibility ranging from high to nearly invisible. This ensures our model can cope with target objects with severe incompleteness. We follow the train/val split of the ScanRefer dataset.

For the two methods extended from 2D grounding, we train a VoteNet to generate 3D object proposals and match them with the 2D grounding results to produce the 3D grounding output. See Supplementary for more details.

Quantitative comparison between these methods is shown in Table 2. Our method significantly outperforms the other methods with a large margin on both datasets. This shows that ScanRefer may not be suitable for grounding in single-view RGBD images. Performance of 2D visual grounding methods indicates the inefficiency of obtaining 3D precise localization results from a 2D bounding box. Figure 4 depicts some qualitative results. The proposed method has both higher retrieval accuracy and localization accuracy than previous methods.

Comparison in the whole-scene setting: We then discuss the performance of our method in the whole-scene setting, the same setting as in [4]. We compare our method to two ScanRefer variants: 1) *ScanRefer-xyz+rgb* using the spatial coordinates and colors; and 2) *ScanRefer-full* using its full configuration (including pre-trained multi-view image features, point cloud normals, and a language classifier). Table 3 shows our method outperforms the two variants of ScanRefer in Acc@0.5 (29.0 against 27.1) and R@5 (47.1 against 43.5), and attains a comparable performance in Acc@0.25. Our method using xyz+rgb outperforms ScanRefer with the same input and achieves comparable results compared to the full configuration of ScanRefer which uses extra inputs. This shows our bottom-up grounding method is applicable to the 3D scenes as well.

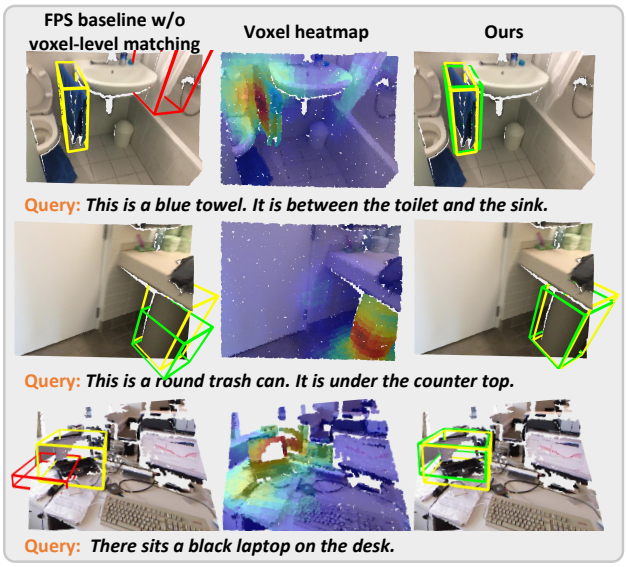

Figure 5: Comparison between the FPS baseline with the proposed bottom-up approach. Successful predictions are in green; failures in red; and ground-truth in yellow.

5.4. Ablation Study

In order to show the effectiveness of different design choices in our proposed approach, we conducted extensive ablation studies.

FPS baseline w/o voxel-level matching: We first compare our method to the FPS baseline where no voxel-level matching is performed. Thus, the FPS baseline amounts to use a standard VoteNet [26] with a single visual-language matching at the object level. Quantitative comparison demonstrates that our method with voxel-level matching is able to achieve far better results; see the first row block of Table 4. Some qualitative results are provided in Figure 5. In the first and third row, our method successfully localizes the correct object, while FPS baseline fails. In the second row, both methods are successful, yet our method obtain more accurate results.

Voxel-level matching only: We also consider the configuration where object-level matching is removed. We conducted an experiment of using only voxel-level matching

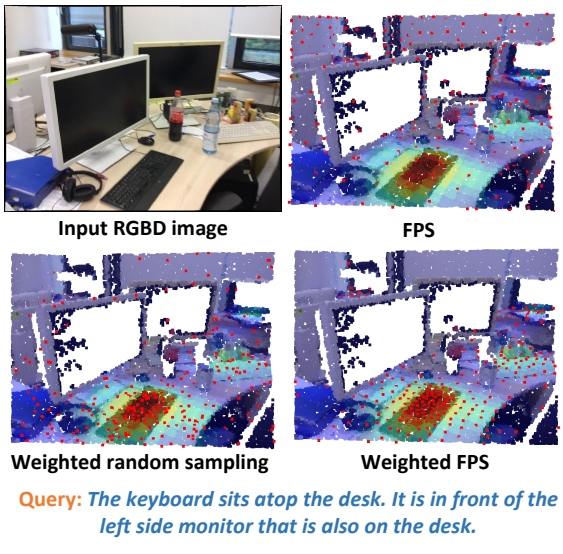

Figure 6: Comparison between different sub-sampling methods. Seed points are visualized in red.

for grounding. Specifically, we trained a VoteNet [26] to generate object proposals and ground the text in the proposal with the highest confidence value based on the point cloud heatmap. Results are shown in the first row block in Table 4. Using only voxel-level matching severely deteriorates the performance, further demonstrating the necessity of combining voxel-level and object-level matching.

Different sampling strategies: We show the advantage of weighted FPS over the weighted random sampling technique which randomly samples the point set using their heatmap value as the sampling probability. As can be seen from the first row block of Table 4, weighted FPS (Ours) can attain slightly better results compared to weighted random sampling. Both methods outperform the FPS baseline by a large margin. The performances show that weighted FPS facilitates object proposal generation. Intermediate results are shown in Figure 6 where weighted FPS can yield well-patterned seed points that are densely concentrated in the relevant regions. The weighted random sampling only scatters the majority of seed points around relevant regions, while the FPS spreads the seed points uniformly in the point cloud with few seed points in the relevant regions.

Different modalities for heatmap generation: We validate the design choice of using voxel modality for coarse localization. We compare our model using 3D voxels with its variants using 2D images and 3D point clouds. See supplementary for more details of these variants.

We show the performances of the three models using different modalities in the second row block of Table 4. Comparison shows our method using the voxel-based heatmap achieves the best result among the three optional configurations. In terms of efficiency, the voxel-based model can run at 10.0 fps, slightly lower than the frame-rate of the image-based model (12.5 fps) but much higher than that of the point cloud model (5.3 fps). Qualitative examples are given in Figure 7. Our voxel-based model can learn a more concentrated heatmap than models using the other two modalities. With such a heatmap, the network can extract content-aware features from the regions relevant to the target object, and hence precisely infer the bounding box of the target object.

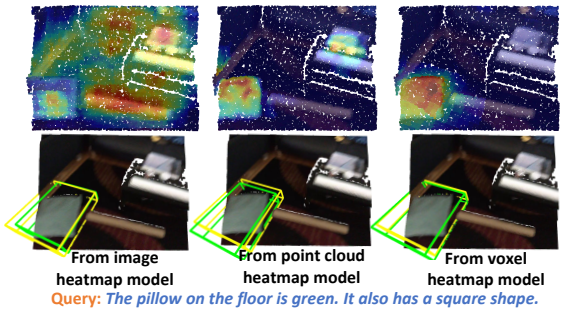

Figure 7: Comparison between different modalities used by the heatmap generation model. Successful predictions are in green; failures in red; and ground-truth in yellow.

Without proposal refinement: The grounding performance can be increased by a moderate margin on both datasets with the additional module designed for refining the final proposals. Comparison against results without this refinement module is shown in the bottom block in Table 4.

6. Conclusions

This paper presents a novel task of 3D visual grounding in single-view RGBD images. A bottom-up method is proposed that matches the query sentence to visual features at both voxel level and object level. We also contribute a large-scale referring expression dataset, SUNRefer, on more than 7,000 single-view RGBD images corresponding to 38,495 descriptions in total. Extensive experiments on both the SUNRefer and ScanRefer datasets show that the proposed method significantly outperforms previous methods.

7. Acknowledgements

This work was partially supported by National Key Research and Development Program of China (No.2020YFC2003902). It was supported in part by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the National Key R&D Program of China with grant No. 2018YFB1800800, by Shenzhen Outstanding Talents Training Fund, by Guangdong Research Project No. 2017ZT07X152, and by National Natural Science Foundation of China 61902334, Shenzhen Fundamental Research (General Project) JCYJ20190814112007258.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020.

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3674–3683, 2018.

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3D: Learning from RGB-D data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[4] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. *16th European Conference on Computer Vision (ECCV)*, 2020.

[5] Kan Chen, Rama Kovvuri, and Ram Nevatia. Query-guided regression network with context policy for phrase grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 824–832, 2017.

[6] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019.

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[8] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2054–2063, 2018.

[9] Samyak Datta, Karan Sikka, Anirban Roy, Karuna Ahuja, Devi Parikh, and Ajay Divakaran. Align2ground: Weakly supervised phrase grounding guided by image-caption alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[10] Chaorui Deng, Qi Wu, Qingyao Wu, Fuyuan Hu, Fan Lyu, and Mingkui Tan. Visual grounding via accumulated attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7746–7755, 2018.

[11] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019.

[12] Zhiyuan Fang, Shu Kong, Charless Fowlkes, and Yezhou Yang. Modularized textual grounding for counterfactual resilience. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6378–6388, 2019.

[13] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564, 2016.

[14] Lucas Janson, Brian Ichter, and Marco Pavone. Deterministic sampling-based motion planning: Optimality, complexity, and performance. *The International Journal of Robotics Research*, 37(1):46–61, 2018.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[16] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler. What are you talking about? text-to-image coreference. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3558–3565, 2014.

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[18] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10880–10889, 2020.

[19] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. Improving referring expression grounding with cross-modal attention-guided erasing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1950–1959, 2019.

[20] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-task collaborative network for joint referring expression comprehension and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10034–10043, 2020.

[21] Arjun Majumdar, Ayush Shrivastava, Stefan Lee, Peter Anderson, Devi Parikh, and Dhruv Batra. Improving vision-and-language navigation with image-text pairs from the web. *arXiv preprint arXiv:2004.14973*, 2020.

[22] Cecilia Mauceri, Martha Palmer, and Christoffer Heckman. Sun-spot: An rgb-d dataset with spatial referring expressions. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[23] Dipendra Misra, Kejia Tao, Percy Liang, and Ashutosh Saxena. Environment-driven lexicon induction for high-level instructions. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 992–1002, 2015.

[24] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016.

[25] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[26] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9277–9286, 2019.

[27] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in neural information processing systems*, pages 5099–5108, 2017.

[28] Yuankai Qi, Qi Wu, Peter Anderson, Xin Wang, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9982–9991, 2020.

[29] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer, 2016.

[30] Arka Sadhu, Kan Chen, and Ram Nevatia. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10417–10427, 2020.

[31] Jing Shi, Jia Xu, Boqing Gong, and Chenliang Xu. Not all frames are equal: Weakly-supervised video grounding with contextual similarity and visual clustering losses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10444–10452, 2019.

[32] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.

[33] Liwei Wang, Yin Li, and Svetlana Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5005–5013, 2016.

[34] 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 Conference on Computer Vision and Pattern Recognition*, pages 6659–6668, 2019.

[35] Sibei Yang, Guanbin Li, and Yizhou Yu. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4145–4154, 2019.

[36] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4644–4653, 2019.

[37] Xun Yang, Xueliang Liu, Meng Jian, Xinjian Gao, and Meng Wang. Weakly-supervised video object grounding by exploring spatio-temporal contexts. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1939–1947, 2020.

[38] Zhengyuan Yang, Tianlang Chen, Liwei Wang, and Jiebo Luo. Improving one-stage visual grounding by recursive sub-query construction. In *ECCV*, 2020.

[39] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *ICCV*, 2019.

[40] 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.

[41] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7282–7290, 2017.

[42] Hanwang Zhang, Yulei Niu, and Shih-Fu Chang. Grounding referring expressions in images by variational context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4158–4166, 2018.

[43] Zhu Zhang, Zhou Zhao, Zhijie Lin, Baoxing Huai, and Nicholas Jing Yuan. Object-aware multi-branch relation networks for spatio-temporal video grounding. *arXiv preprint arXiv:2008.06941*, 2020.

[44] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020.

[45] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6578–6587, 2019.

[46] Luowei Zhou, Nathan Louis, and Jason J Corso. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834*, 2018.

[47] Bohan Zhuang, Qi Wu, Chunhua Shen, Ian Reid, and Anton Van Den Hengel. Parallel attention: A unified framework for visual object discovery through dialogs and queries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4252–4261, 2018.