ScanERU: Interactive 3D Visual Grounding Based on Embodied Reference Understanding

Ziyang Lu[1], Yunqiang Pei[1], Guoqing Wang[1]*, Peiwei Li[2], Yang Yang[1], Yinjie Lei[3], Heng Tao Shen[1]

[1]University of Electronic Science and Technology of China
[2]University of Science and Technology of China
[3]Sichuan University
zi.yang.lu@foxmail.com, simon1059770342@foxmail.com, gqwang0420@uestc.edu.cn,
1pw0622@mail.ustc.edu.cn, yang. yang@uestc.edu.cn, yinjie@scu.edu.cn, shenhengtao@hotmail.com

Abstract

Aiming to link natural language descriptions to specific regions in a 3D scene represented as 3D point clouds, 3D visual grounding is a very fundamental task for human-robot interaction. The recognition errors can significantly impact the overall accuracy and then degrade the operation of AI systems. Despite their effectiveness, existing methods suffer from the difficulty of low recognition accuracy in cases of multiple adjacent objects with similar appearance. To address this issue, this work intuitively introduces the human-robot interaction as a cue to facilitate the development of 3D visual grounding. Specifically, a new task termed Embodied Reference Understanding (ERU) is first designed for this concern. Then a new dataset called ScanERU is constructed to evaluate the effectiveness of this idea. Different from existing datasets, our ScanERU dataset is the first to cover semi-synthetic scene integration with textual, real-world visual, and synthetic gestural information. Additionally, this paper formulates a heuristic framework based on attention mechanisms and human body movements to enlighten the research of ERU. Experimental results demonstrate the superiority of the proposed method, especially in the recognition of multiple identical objects. Our codes and dataset are available in the ScanERU repository.

Introduction

The ability to understand and localize objects from natural expression is critical for the operation of AI systems. To this end, both 2D (Deng et al. 2018, 2021; Yang et al. 2022; Qiao, Deng, and Wu 2020; Wang et al. 2019) and 3D (Huang et al. 2022; Chen, Chang, and Nießner 2020; Yang et al. 2021; Huang et al. 2021; Yuan et al. 2021; Zhao et al. 2021) visual groundings refer to understanding how words and language can be linked to visual information in images and videos, and how this information can be used to recognize and understand objects in the environment. The former takes 2D images or video frames as input and suffers from the limitation of fully localize objects and problems, where object localization fails to capture the true 3D extent of an object (Chen, Chang, and Nießner 2020), which is caused by the restricted nature of 2D images. Compared with 2D visual grounding methods, 3D visual grounding methods capture the true physical extent and spatial relationships of objects, adapt better to the 3D context, and improve performance on complex and diverse descriptions (Chen, Chang, and Nießner 2020). Therefore, the research of 3D visual grounding based on 3D point cloud data has emerged, and this line of research aims at locating a specific object or region in a 3D scene referred by a natural language description. Because of 3D point cloud data, it can describe the objects in

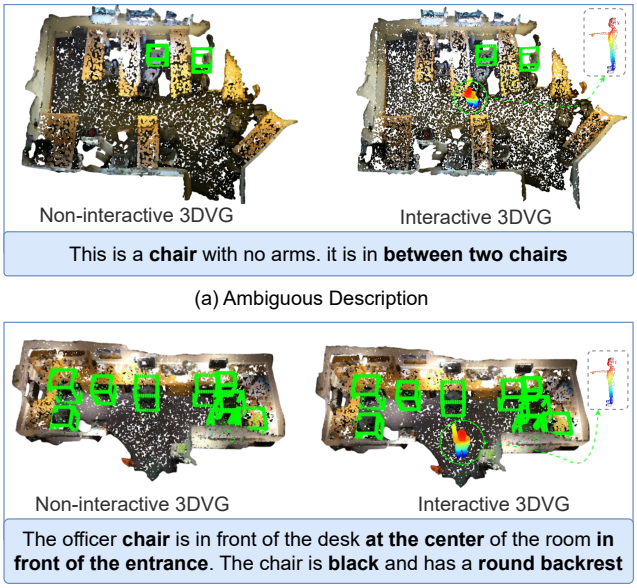

Figure 1: The comparisons between non-interactive and interactive visual grounding. The image displays two typical difficult scenarios, ambiguous description and complex spatial relationship, for non-interactive visual grounding. Nevertheless, the inclusion of gestural information conveyed by a human agent can aid in the localization of the referred object by robots and AI systems.

*Corresponding author
Our project page: https://github.com/MrLearnedToad/ScanERU

the environment in more detail, such as their shape and size. These properties guarantee its wide range of applications, including robotics, augmented reality (Pei et al. 2023), virtual reality, and human-robot interaction (Yang et al. 2021; Huang et al. 2022), where natural language can be used as an intuitive and flexible way to interact with 3D environments.

Several works have studied 3D visual grounding using various methods, including new datasets (such as ScanRefer (Chen, Chang, and Nießner 2020), ReferIt3D (Achlioptas et al. 2020)) and frameworks (Zhao et al. 2021), approaches (Liu et al. 2021) for identifying objects, and 3D visual grounding in RGB-D images without requiring scene reconstruction (Liu et al. 2021). Despite the success of existing work, identifying the referred object among multiple adjacent objects with similar appearances is still a great challenge in this task. However, the explorations on this concern are clearly insufficient. In situations where the description itself is ambiguous or the spatial relationships are complex in ScanRefer (Chen, Chang, and Nießner 2020), the accuracy of distinguishing multiple similar objects is relatively low, as demonstrated in Figure 1.

This work solves the challenge of identifying referred objects among visually similar adjacent objects from a new perspective of human-robot interaction, which can act as rich cues to promote the recognition of visually similar objects in 3D space. A similar idea in 2D visual grounding is formulated by Chen et al. (Chen et al. 2021) who proposes to model the coordination between language and gestural information. Specifically, they constructed the YouRefIt dataset, mainly consisting of videos, where humans jointly leverage language and gestures to refer to objects, and thus significantly improved the accuracy of 2D visual grounding by the incorporation of gesture information. Inspired by their success, this work aims to further improve the accuracy of 3D visual grounding on multiple adjacent objects with similar appearances by incorporating human gestures to disambiguate referring expressions and accurately identify the referred object. Specifically, we are the first to design a new task for 3D visual grounding termed Embodied Reference Understanding (ERU), which is built upon the embodied perspective of the agent. To better evaluate such a task, a new dataset called ScanERU is constructed based on existing datasets by incorporating textual, real-world visual, and synthetic gestural information into semi-synthetic scenes. Finally, to validate the effectiveness, we formulate a heuristic framework based on attention mechanisms (Vaswani et al. 2017), which is widely used in the community (Wang, Sun, and Sowmya 2021; Wang et al. 2021; Wang, Sun, and Sowmya 2019; Wu et al. 2023), human body movements and constructed a real-world test set. Different from prior work, this work incorporates considerations for interactions with other intelligent agents, which thus provides a more natural way of human-robot interaction and a more human-like understanding of the 3D world.

In conclusion, our contributions are as follows:

- **A novel task** called Embodied Reference Understanding (ERU) for 3D visual grounding is designed, which first jointly leverages language and gestures to refer to objects in 3D point clouds. This approach provides a more comprehensive and meaningful representation of the 3D visual information in the environment.
- **A new dataset** called ScanERU is constructed, which covers diverse and challenging semi-synthetic scenarios with synthetic gestural information. Models trained on ScanERU is tested using real-world testing scenarios. This dataset provides a valuable resource for researchers to develop and evaluate new methods for the ERU task.
- **A heuristic framework** based on attention mechanisms and human body movements is proposed to evaluate our effectiveness on the recognition of multiple identical objects or complex spatial relations. This framework offers a structured means of assessing ERU model performance, identifying strengths and weaknesses, and guiding future research.

Related Work

Non-Interactive Visual Grounding

3D visual grounding involves establishing a connection between language and 3D objects in a point cloud environment, allowing the model to capture spatial features. The objective is to align natural language descriptions with the corresponding 3D objects and their attributes in a 3D setting. Chen et al. (Chen, Chang, and Nießner 2020) presented the ScanRefer Dataset and a comprehensive end-to-end framework for the visual grounding task. The ReferIt3D (Achlioptas et al. 2020) introduced two datasets, similar to the ScanRefer dataset, but with a focus on identifying the referred object among instances of the same fine-grained category. Most approaches (Chen, Chang, and Nießner 2020; Zhao et al. 2021; Chen et al. 2022; Yuan et al. 2021; Bakr, Alsaedy, and Elhoseiny 2022; Huang et al. 2022) adopt the two-stage framework established by ScanRefer, while the 3D-SPS method (Luo et al. 2022) devises a single-stage solution to the task. Cai et al. (Cai et al. 2022) developed a unified joint framework that accommodates both the grounding task and captioning task. Inspired by the transformer, recent works such as 3DVG-Transformer (Zhao et al. 2021) and SAT (Yang et al. 2021) have integrated attention mechanisms into the framework. The most recent work, HAM (Chen et al. 2022), leverages the spatially-global and spatially-local attention to locate referred objects, achieving the best result of Acc@0.5 on the ScanRefer Challenge. However, the sparse, noisy, and limited semantic information of point clouds compared to 2D images make it difficult to accurately locate a referred object (Yang et al. 2021). Additionally, the proximity of the referent to adjacent objects in the scene can lead to localization errors (Bakr, Alsaedy, and Elhoseiny 2022; Achlioptas et al. 2020; Chen, Chang, and Nießner 2020; Zhao et al. 2021) , and view-dependent descriptions can result in poor localization performance for referent localization based on spatial terms (Huang et al. 2022; Yang et al. 2021; Zhao et al. 2021; Huang et al. 2021; He et al. 2021; Yuan et al. 2021; Feng et al. 2021) . There are also localization errors when locating a unique referent among multiple visually similar objects (Bakr, Alsaedy, and Elhoseiny 2022; Luo et al. 2022; Huang et al. 2022; Yang

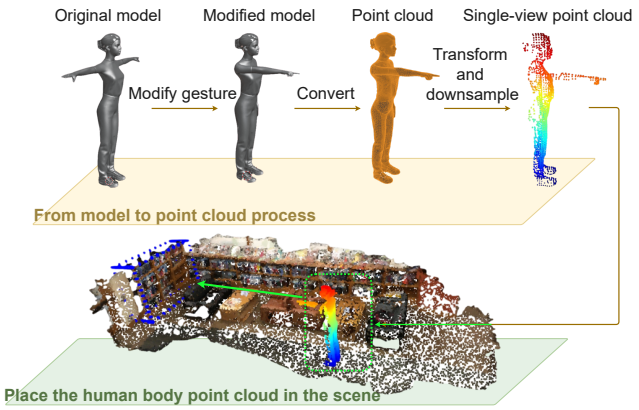

Figure 2: Generation procedure of human point cloud.

et al. 2021; Zhao et al. 2021; Huang et al. 2021; He et al. 2021; Yuan et al. 2021; Feng et al. 2021; Liu et al. 2021; Achlioptas et al. 2020; Chen, Chang, and Nießner 2020) . Our approach introduces a new task of 3D visual grounding in a human-in-the-loop-based scenario, where body gestures are integrated into the scene to mitigate localization errors resulting from sparse, noisy, and semantically limited point clouds, object proximity, difficulty in distinguishing a unique referent among visually similar objects, and view-dependent descriptions.

Interactive Visual Grounding

Interactive visual grounding is a task that involves using natural language and gestures to refer to objects or regions in an image or a 3D scene. To improve the accuracy and facilitate a more natural method of communication between humans and agents, Chen et al. (Chen et al. 2021) introduced the ERU (Embodied Reference Understanding) task, in which an agent uses both language and gestures to refer to an object to another agent in a shared physical environment. This is accompanied by the introduction of the YouRefIt dataset, which is a 2D multi-modal dataset encompassing textual, visual, and gestural information. Building on psychological studies of human pointing gestures, Li et al. (Li et al. 2023) proposed a new architecture using the virtual touch line (a line connecting the eye and the fingertip) and a transformer, leading to a significant improvement in the ERU task. Our work expands the ERU task to point cloud environment and studies the disambiguation effect of human gesture. Figure 1 shows the comparisons between non-interactive and interactive visual grounding.

Dataset

To study ERU task in 3D environment, we propose the ScanERU dataset, a semi-synthetic dataset for ERU task. Our dataset is based on the ScanRefer (Chen, Chang, and Nießner 2020) and ScanNet (Dai et al. 2017) datasets and includes 706 unique indoor scenes, 9,929 referred objects, and 46,173 descriptions. We synthesize human point cloud data and pointing gestures for each referred object. We also propose the real-world ScanERU test set for testing the validity of our method in real-world scenarios to evaluate its effectiveness and generalization capabilities. For the further information of ScanERU, please refer to the supplement.

Data Annotation

The procedure of annotating the semi-synthetic dataset is conducted through a visualization UI and LabelImg (Tzutalin 2015), which presents the workers with both the point cloud and a top view of the scene with the non-referred objects (e.g. ceiling) faded out. To ensure the quality and accuracy of our annotations, our workers are instructed to annotate five possible positions of the synthetic agent on the top view. The annotations are then subjected to automatic checks to verify that there is no object obstructing the line of sight between the synthetic agent and the referred object. Upon successful completion of the checks, the verified annotations are assigned to each referred object in the dataset. In our dataset generation method, we adjusted the angle fluctuation range of the arm movement by ±3° to ensure that the ray emitted by the gesture can pass through the referred object, rather than pointing directly at the object's center. The output of the annotation process are the index between point cloud of synthetic human agents and referred objects and the coordinate of the synthetic agent in the scene point cloud.

Generation Procedure

To ensure the diversity and variability of the synthetic gestural information, we utilize 10 different human models both male and female. The 3D models of the human body used in our model were obtained from publicly available and open-source models from Sketchfab and Mixamo. The meshes are batch-processed using Blender. The detailed process is shown as Figure 2:

Rotate the skeleton to generate a character pool. To generate a character pool, we apply a rotation transformation to the skeleton of each character. Specifically, we create a 'pointing' gesture for each character by rotating their hand and arm towards the referred object. The 'pointing' gesture consists of two parts: the choice of left or right hand, and the rotation angle of the arm. We vary the rotation angle from -90 to +90 degrees with an interval of 0.5°. Moreover, we introduce random angle perturbation to ensure the diversity and realism of each synthetic agent. The randomly rotated skeleton includes the left or right arm (both upper and lower parts), the left or right hand, and the head. The perturbation angle is drawn from a Gaussian distribution with a range of -3° to +3°. This is to make the gesture more realistic and natural, and to avoid overfitting to a specific gesture.

Convert the mesh files of human models into point clouds. We efficiently convert the mesh files of human models into 3D point clouds and perform voxel down-sampling to align them with the ScanNet dataset. The voxel size is consistently set to 0.25cm.

Transform point cloud to a "single-view" format. In real-world scenarios, it is impractical for the human agent to remain idle until the scanning process is complete. Consequently, the point cloud representing the human is typically incomplete and most likely manifests in a "single-view"

form because of the machine agent's inability to thoroughly scan the human agent in the same manner as the entire scene. To address this, we initially select a random angle from which the machine agent scans the human agent. Subsequently, we compute the point cloud to eliminate points not visible to the machine agent. Following this, random down-sampling is employed to decrease the point count to 3,000. In cases where the number of points falls below 3,000, zero-padding is utilized to occupy the remaining points.

Load the matching point cloud from the character pool into the scene. Due to the large size of the point cloud file, we do not load it into the scene point cloud until training or evaluation time. The dataset is loaded along with label, vertex, and normal information based on the annotated position and the index between point cloud of synthetic human agents and referred objects. And each referred object is associated with 3-5 different synthetic agents pointing at it from different positions.

ScanERU Test Set

To evaluate the effectiveness of the ScanERU method and dataset, we conducted a real-world test set for 3D ERU task. This section outlines the test set creation process.

In our study, we employ the Azure Kinect DK, a Time-of-Flight (ToF) RGB-D camera with an inertial measurement unit (IMU), as our 3D sensor for reconstructing the scene. The RGB sensor operate at a resolution of 1280*540, while the depth sensor function at a resolution of 512*512. We utilize the official SDK of the Azure Kinect DK and the Open3D (Zhou, Park, and Koltun 2018) reconstruction system as our software. Following data acquisition, we manually process the 3D mesh data in MeshLab (Cignoni et al. 2008), by deleting distorted areas in the point cloud, especially in the edge regions, and by applying down-sampling to align the original ScanNet (Dai et al. 2017) point cloud. Since these data are only used for test set, we do not perform a complete semantic segmentation annotation on the point cloud. We develop a Blender script to annotate the referred object and the human agent. For the descriptions, we apply the similar tool as ScanERU data annotation to help our workers to describe the referred object. The whole task is assigned to four workers with the request of each description (e.g., the length of description, the way to describe the referred object etc.). In addition, each referred object is described by all four workers. Notably, to evaluate the disambiguation ability and comprehension of complex spatial information, all the referred objects have at least one similar object in the scene.

Methodology

This section describes our work in detail. Sec 4.1 gives an overview of the ScanERU method. Sec 4.2 explains how we generate proposals, encode gestures, and encode language. Sec 4.3 presents how we fuse multi-modal features. Sec 4.4 defines the loss function.

Overview

Shown as Figure 3, the ScanERU method comprises three inputs: the point cloud of the entire 3D scene, the description of the referred object, and the point cloud of the human agent. The scene point cloud $P_p \in \mathbb{R}^{N\times(3+K)}$ contains N points' coordinate and K-dimensional features such as RGB and normal vectors. The description is tokenized and transformed into word embeddings using the GloVE (Pennington, Socher, and Manning 2014) model. The human agent point cloud is similar to the scene point cloud, except that its features only include normal vectors. The objective of the task is to locate the referred object and output its bounding-box in world coordinates.

The ScanERU method consists of four modules: proposal generation, gestural encoding, language-aware, and multi-modal fusion. To better leverage the features among language, gesture, and the scene point cloud, an attention mechanism (Vaswani et al. 2017) is employed in our work. The proposal generation module is the same as that used in 3DVG-Transformer (Zhao et al. 2021) and it generates a bounding-box from object proposal while extracting context-aware features. The proposal features are represented by $F_p \in \mathbb{R}^{M\times H}$, for M proposals with H-dimensional features. In the gestural encoding module, the Point2Skeleton technique (Lin et al. 2021) is employed to abstract a human's skeletal structure, which is subsequently encoded into gestural features F_g via a convolution layer. Similar to ScanRefer (Chen, Chang, and Nießner 2020) and 3DVG-Transformer (Zhao et al. 2021), the language-aware module aggregates the word embeddings into the language features $F_l \in \mathbb{R}^{L\times H}$ and global language features using a GRU (Chung et al. 2014) cell and a self-attention module. The multi-modal fusion module leverages attention mechanism (Vaswani et al. 2017) to fuse proposal features F_p, gestural features F_g, and word features F_l, thereby generating the confidence score of each bounding-box. Specifically, this study centers on the combination of gestural information with the proposal and word information, aiming to disambiguate referring expressions and accurately identify the referred object.

Feature Encoding Modules

Proposal generation module. Our proposed method utilizes a PointNet++ (Qi et al. 2017) backbone and a voting and grouping module (Qi et al. 2019), similar to ScanRefer (Chen, Chang, and Nießner 2020) and 3DVG-Transformer (Zhao et al. 2021), to process the point cloud of the scene and group them into individual clusters. Subsequently, we employ the coordinate-guided contextual aggregation (CCA) module, as utilized in 3DVG-Transformer (Zhao et al. 2021), to generate refined proposal features as F_{p0} and bounding-boxes. To further refine the proposal features, a self-attention module is applied, which takes the refined proposal features F_{p0} as input and studies the contextual relationships within the refined proposal features F_{p0}. In addition, we employ a copy&paste module, akin to the method used in 3DVG-Transformer (Zhao et al. 2021), to leverage the over-fitting issue, producing the output as F_p.

Gesture encoding module. The point cloud corresponding to the human agent undergoes processing utilizing the Point2Skeleton method (Lin et al. 2021). This results in the acquisition of S skeletal points, denoted as P_s, pertaining

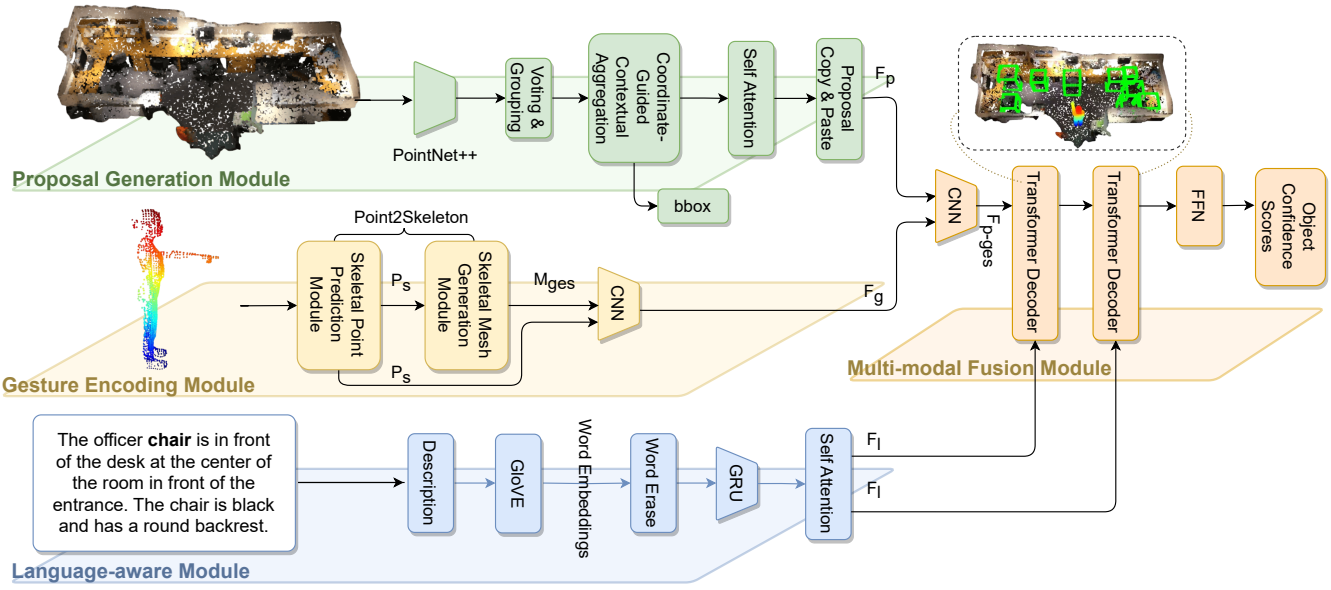

Figure 3: The architecture of the proposed ScanERU. It takes a point cloud of the scene, a point cloud of a human agent, and a description of the referred object as input, through the processing of the proposal generation module, the gesture encoding module, and the language-aware module, comprehensive semantic features of different modalities are extracted. A multi-modal fusion module is designed and followed to integrate multi-branch features to finally produce the confidence scores of the bounding boxes. The highest confidence score is the final prediction. Best viewed in color.

to the human agent, as well as the link matrix $M_{ges} \in \mathbb{R}^{S\times S}$, which describes the relationships among these skeletal points. Subsequently, the P_s points and $M_{ges} \in \mathbb{R}^{S\times S}$ matrix are concatenated and encoded via convolutional layers, ultimately producing the gestural features F_g.

Language-aware module. The textual description input is encoded using the GloVe (Pennington, Socher, and Manning 2014) and GRU (Chung et al. 2014) module, which is the same module used in the ScanRefer (Chen, Chang, and Nießner 2020) framework. Moreover, our training process is augmented with the word erase training strategy, which has been shown to be beneficial in 3DVG-Transformer (Zhao et al. 2021). Furthermore, to refine the language features, we employ a self-attention module to generate F_l from the GRU output.

Multi-Modal Fusion Module

As illustrated in Figure 3, the proposal features F_p and gestural features F_g are concatenated and fused using a convolution block, with the resulting features being denoted as F_{p-ges}. Subsequently, we use a 2-layer stacked transformer decoder to exploit the relationship of proposal-gestural features F_{p-ges} and language features F_l, where the proposal-gestural features F_{p-ges} serves as query while the language features F_l serves as key and value. Finally, the output of the stacked transformer decoder is fed into a feed-forward network (FFN) layer and a softmax activation layer to generate the confidence score of each bounding-box.

Loss Function

In our approach, we employ a loss function similar to that used in 3DVG-Transformer (Zhao et al. 2021) and Point2Skeleton (Lin et al. 2021), which is represented as $L = 0.3L_{loc} + 10L_{det} + 0.1L_{cls} + 0.3L_{skel}$. Here, L_{loc} denotes the localization loss, L_{det} represents the object detection loss, and L_{cls} indicates the language-to-object classification loss. Furthermore, we can decompose L_{det} as $L_{det} = L_{vote-reg} + 0.1L_{objn-cls} + 0.1L_{sem-cls} + L_{box}$ where $L_{vote-reg}$ is the vote regression loss, $L_{objn-cls}$ and $L_{sem-cls}$ are the objectness and semantic classification losses, respectively, and L_{box} denotes the bounding-box loss. The bounding-box loss can be further decomposed as $L_{box} = L_{center-reg} + 0.1L_{size-cls} + L_{size-reg}$ where $L_{center-reg}$ and $L_{size-reg}$ are the center and size regression losses, respectively, and $L_{size-cls}$ denotes the size classification loss. L_{skel} refers to the skeleton detection loss. We can decompose L_{skel} as $L_{skel} = L_{point} + L_{link}$ where L_{point} is skeletal point prediction loss, L_{link} is skeletal mesh generation loss.

Experiment

Dataset Split. In our experimental evaluation, we conduct tests on the ScanERU dataset. Following the same protocol as the ScanRefer dataset, we split it into train and validation sets with 36,665 and 9,508 samples, respectively.

Baseline. We meticulously devise the baselines by comparing our innovative method with state-of-the-art methods on 3D visual grounding task.

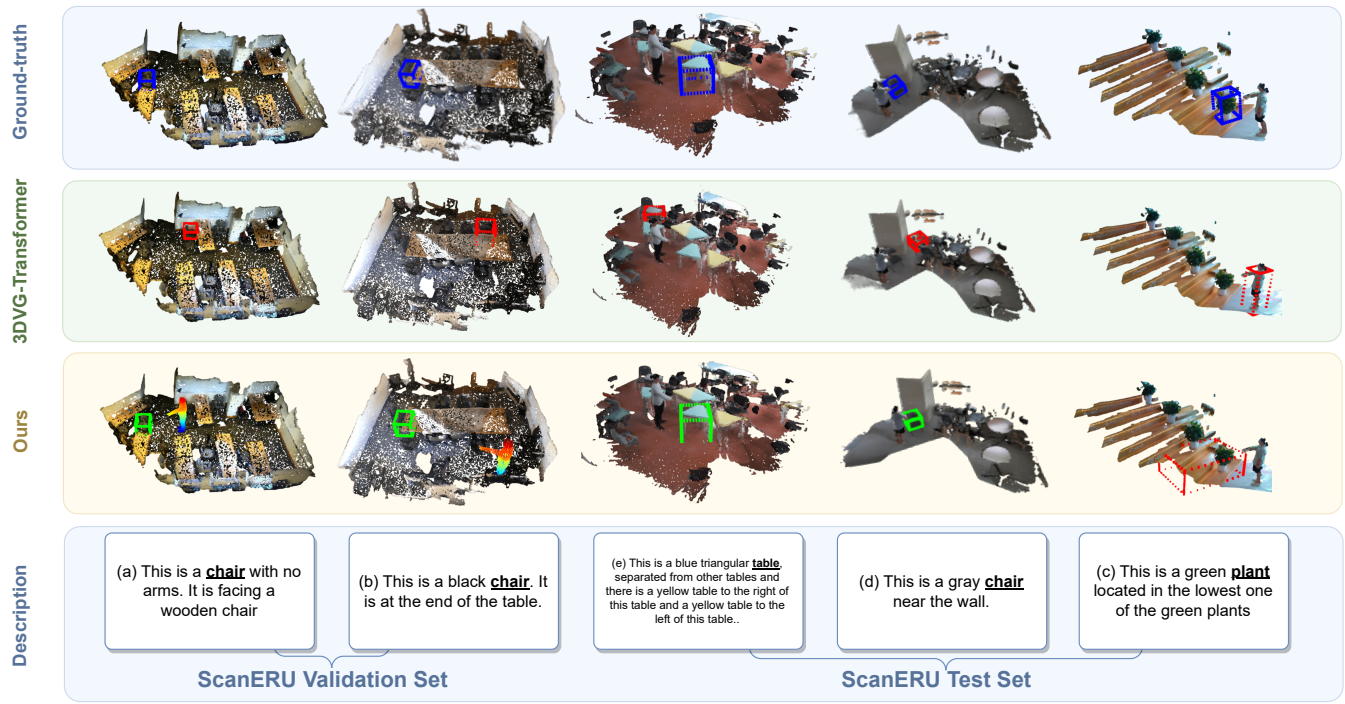

Figure 4: Qualitative results from 3DVG-Transformer(Zhao et al. 2021) and our ScanERU. The GT boxes are marked in blue. If one predicted box has an IoU score higher than 0.5, this box is marked in green, otherwise it is marked in red.

Metric. Following the standard evaluation metric for 3D visual grounding tasks, we employ two commonly used metrics, namely Acc@0.25IoU and Acc@0.5IoU, to measure the performance of our method. Additionally, we also report the "unique," "multiple," and "overall" scores, as defined in the ScanRefer dataset (Chen, Chang, and Nießner 2020). The "unique" score measures the performance when there is only a single object of its class in the scene, whereas the "multiple" score measures the performance when there are more than one similar object of its class in the scene. The "overall" score is the weighted average of the "unique" and "multiple" scores.

Quantitative Study

In Table 1, We compare the performance of our ScanERU method with several existing 3D visual grounding methods.

To compare our work with other 3DVG methods based on ScanRefer, we adopted the same experimental setup as them. This was to ensure consistency and fairness in the evaluation. Since our ScanERU dataset and ScanRefer dataset are the same in terms of 3D scenes and text descriptions, the results obtained by other methods through ScanRefer training or ScanERU training are consistent.

Experiment on the ScanERU validation set. In the "multiple" subset, our innovatively proposed method exhibits superior performance compared to the state-of-the-art (SOTA) method, with a significant improvement of 24.88% for Acc@0.25 and 21.23% for Acc@0.5. Moreover, our method's overall accuracy remarkably surpasses the SOTA method by 20.34% for Acc@0.25 and 17.11% for Acc@0.5 due to its enhanced disambiguation ability and deep comprehension of complex spatial information, which underscores the effectiveness of incorporating human gestures in localizing and distinguishing multiple similar objects. These results strongly validate our proposed method's efficacy in significantly improving localization performance in complex indoor environments.

Experiment on the ScanERU test set shows the performance comparison of ScanRefer (Chen, Chang, and Nießner 2020), 3DVG-Transformer (Zhao et al. 2021), and ScanERU on the task of 3D object localization. Because all the referred objects are "multiple", there are only "overall" results reported in Table 1. Our method achieves significantly higher accuracy than other methods. The results indicate that our dataset and method is generalizable to real-world environments and the combination of textual and gestural information is crucial for accurate object localization in the complex real-world situation.

Qualitative Study

Figure 4 depicts a visualization of the performance of our proposed method and the baseline 3DVG-Transformer (Zhao et al. 2021). The ground-truth bounding boxes are denoted in blue, whereas the predicted boxes are highlighted in green if their IoU score with the ground truth is above 50%, and in red otherwise.

Methods	Venue	Unique		Multiple		Overall	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
Results on the ScanERU validation set							
ScanRefer (Chen, Chang, and Nießner 2020)	ECCV2020	76.33	53.51	32.73	21.11	41.19	27.40
TGNN (Huang et al. 2021)	ICCV2021	68.61	56.80	29.84	23.18	37.37	29.70
InstanceRefer (Yuan et al. 2021)	ICCV2022	77.45	**66.83**	31.27	24.77	40.23	32.93
3DVG-Transformer (Zhao et al. 2021)	ICCV2021	81.93	60.64	39.30	28.42	47.57	34.67
3DJCG (Cai et al. 2022)	CVPR2022	83.47	64.34	41.39	30.82	49.56	37.33
3D-SPS (Luo et al. 2022)	CVPR2022	84.12	66.72	40.32	29.82	48.82	36.98
ScanERU		**85.01**	64.37	**66.27**	**52.05**	**69.90**	**54.44**
Results on the ScanERU test set							
ScanRefer (Chen, Chang, and Nießner 2020)	ECCV2020	-	-	18.79	16.05	18.79	16.05
3DVG-Transformer (Zhao et al. 2021)	ICCV2021	-	-	28.56	23.80	28.56	23.80
ScanERU		-	-	**49.51**	**42.13**	**49.51**	**42.13**
Results of ScanERU ablation study							
ScanERU$_{lang-only}$		81.40	59.37	41.42	29.90	49.18	35.62
ScanERU$_{ges-only}$		53.98	41.42	55.15	44.21	54.92	43.67
ScanERU$_{full}$		**85.01**	**64.37**	**66.27**	**52.05**	**69.90**	**54.44**

Table 1: Comparasion of visual grounding performances on ScanERU dataset.

Experiment on the ScanERU validation set demonstrates that our method is capable of successfully localizing the referred object in complex environments with multiple similar objects, while the baseline method exhibits failure cases. We identify two main causes of failure for 3DVG-Transformer (Zhao et al. 2021). The first cause is the difficulty of distinguishing fine-grained features from point cloud data. As illustrated in Figure 4 (a), the terms "no arms" and "wooden chair" are not sufficiently descriptive to differentiate the object from others, even for human observers. The second cause is the ambiguity or complexity of the description. In Figure 4 (b), the phrase "at the end of the table" is unclear, meaning that there are multiple possible objects being referred. These results indicate that the language-only modality has limitations in disambiguating the correct object, particularly in challenging environments.

Experiment on the ScanERU test set illustrates the superiority of our methodology when dealing with intricate situations in real-world environments, compared to the 3DVG-Transformer (Zhao et al. 2021). Figures 4 (c) and 4 (d) highlight the primary causes of failure in 3D visual grounding cases - excessive complexity in spatial information and vague descriptions. Incorporating gestural information can mitigate these factors, particularly in real-world settings. Nonetheless, in the proposal generation phase, gestural information is not provided, potentially leading to imprecise proposal generation. For instance, Figure 4(d) reveals that the ScanERU method fails to accurately generate a proposal for the plant. Consequently, the impact of gestural information remains restricted in such circumstances.

Ablation Study

In this subsection, we aim to conduct a detailed analysis of the contributions of textual and gestural modalities in our proposed approach. Table 1 results reveal that ScanERU$_{ges-only}$ demonstrates considerably lower performance compared to both ScanERU$_{lang-only}$ and ScanERU$_{full}$ within the "Unique" subset for Acc@0.25. This is attributed to ScanERU$_{ges-only}$ solely incorporating gestural details while disregarding descriptions with details about the referred object itself and spatial information among all objects in the scene. Consequently, when several objects (e.g., sofa and table) share the same direction indicated by the human agent, gestural information becomes highly ambiguous, leading to imprecise localization outcomes. This underlines the importance of textual information for resolving potential ambiguities pertaining to the referenced object. Additionally, ScanERU$_{lang-only}$ illustrates that incorporating gestural cues, as evidenced by the superior performance of ScanERU$_{full}$ in Table 1, is crucial for the ERU task. The integration of gestural information thus markedly bolsters the model's capacity to distinguish similar objects and enhances localization performance.

Conclusion

This paper introduces embodied reference understanding (ERU) within 3D point cloud environments, where agents utilize language and gestures to refer to objects in a shared physical space. To support research in this domain, we present ScanERU, a semi-synthetic dataset evaluated for effectiveness on our real-world test set. Next, we propose a novel framework for ERU in 3D environments, integrating multi-modal features and attention mechanisms. Our approach surpasses existing 3D visual grounding methods, particularly in identifying multiple identical objects. This research significantly advances 3D visual grounding by integrating human gestures as an additional modality to disambiguate referring expressions and accurately pinpoint objects. Moreover, it underscores the significance of embodied perspective, human-centered AI, and human-agent interaction for a more natural and human-like understanding of the 3D world. Our future work aims to diversify our dataset with varied scenes and gestures while exploring additional modalities to further augment ERU performance.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under grant U23B2011, 62102069, U20B2063 and 62220106008, the Sichuan Science and Technology Program under grant 2022YFG0032, and the China Academy of Space Technology (CAST) Innovation Program.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, 422–440. Springer.

Bakr, E. M.; Alsaedy, Y.; and Elhoseiny, M. 2022. Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding. *arXiv preprint arXiv:2211.14241*.

Cai, D.; Zhao, L.; Zhang, J.; Sheng, L.; and Xu, D. 2022. 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16464–16473.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, 202–221. Springer.

Chen, J.; Luo, W.; Wei, X.; Ma, L.; and Zhang, W. 2022. Ham: Hierarchical attention model with high performance for 3d visual grounding. *arXiv preprint arXiv:2210.12513*.

Chen, Y.; Li, Q.; Kong, D.; Kei, Y. L.; Zhu, S.-C.; Gao, T.; Zhu, Y.; and Huang, S. 2021. Yourefit: Embodied reference understanding with language and gesture. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1385–1395.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Cignoni, P.; Callieri, M.; Corsini, M.; Dellepiane, M.; Ganovelli, F.; Ranzuglia, G.; et al. 2008. Meshlab: an open-source mesh processing tool. In *Eurographics Italian chapter conference*, 129–136.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5828–5839.

Deng, C.; Wu, Q.; Wu, Q.; Hu, F.; Lyu, F.; and Tan, M. 2018. Visual grounding via accumulated attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7746–7755.

Deng, J.; Yang, Z.; Chen, T.; Zhou, W.; and Li, H. 2021. Transvg: End-to-end visual grounding with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1769–1779.

Feng, M.; Li, Z.; Li, Q.; Zhang, L.; Zhang, X.; Zhu, G.; Zhang, H.; Wang, Y.; and Mian, A. 2021. Free-form description guided 3d visual graph network for object grounding in point cloud. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3722–3731.

He, D.; Zhao, Y.; Luo, J.; Hui, T.; Huang, S.; Zhang, A.; and Liu, S. 2021. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2344–2352.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2021. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1610–1618.

Huang, S.; Chen, Y.; Jia, J.; and Wang, L. 2022. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15524–15533.

Li, Y.; Chen, X.; Zhao, H.; Gong, J.; Zhou, G.; Rossano, F.; and Zhu, Y. 2023. Understanding Embodied Reference with Touch-Line Transformer. In *ICLR*.

Lin, C.; Li, C.; Liu, Y.; Chen, N.; Choi, Y.-K.; and Wang, W. 2021. Point2skeleton: Learning skeletal representations from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4277–4286.

Liu, H.; Lin, A.; Han, X.; Yang, L.; Yu, Y.; and Cui, S. 2021. Refer-it-in-rgbd: A bottom-up approach for 3d visual grounding in rgbd images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6032–6041.

Luo, J.; Fu, J.; Kong, X.; Gao, C.; Ren, H.; Shen, H.; Xia, H.; and Liu, S. 2022. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16454–16463.

Pei, Y.; Huang, R.; Wang, G.; Yang, Y.; Xie, N.; and Shen, H. T. 2023. Multimodal Apology: Using WebXR to Repair Trust with Virtual Companion. In *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, 727–728. IEEE.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 1532–1543.

Qi, C. R.; Litany, O.; He, K.; and Guibas, L. J. 2019. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, 9277–9286.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30.

Qiao, Y.; Deng, C.; and Wu, Q. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 4426–4440.

Tzutalin. 2015. LabelImg. https://github.com/tzutalin/labelImg. Accessed: 2023-2-11.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, G.; Sun, C.; and Sowmya, A. 2019. Erl-net: Entangled representation learning for single image de-raining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5644–5652.

Wang, G.; Sun, C.; and Sowmya, A. 2021. Context-enhanced representation learning for single image deraining. *International Journal of Computer Vision*, 1650–1674.

Wang, G.; Yang, Y.; Xu, X.; Li, J.; and Shen, H. 2021. Enhanced context encoding for single image raindrop removal. *Science China Technological Sciences*, 2640–2650.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1960–1968.

Wu, Y.; Pan, C.; Wang, G.; Yang, Y.; Wei, J.; Li, C.; and Shen, H. T. 2023. Learning Semantic-Aware Knowledge Guidance for Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1662–1671.

Yang, L.; Xu, Y.; Yuan, C.; Liu, W.; Li, B.; and Hu, W. 2022. Improving visual grounding with visual-linguistic verification and iterative reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9499–9508.

Yang, Z.; Zhang, S.; Wang, L.; and Luo, J. 2021. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1856–1866.

Yuan, Z.; Yan, X.; Liao, Y.; Zhang, R.; Wang, S.; Li, Z.; and Cui, S. 2021. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 1791–1800.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2928–2937.

Zhou, Q.-Y.; Park, J.; and Koltun, V. 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847*.