2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) | 979-8-3503-7770-5/24/$31.00 ©2024 IEEE | DOI: 10.1109/IROS58592.2024.10802292

Boosting 3D Visual Grounding by Object-Centric Referring Network

Ruilong Ren[1], Jian Cao[1], Weichen Xu[1], Tianhao Fu[1], Yilei Dong[1], Xinxin Xu[1], Zicong Hu[1] and Xing Zhang[1,2]

***Abstract*— 3D visual grounding is tasked with locating a specific object within a 3D scene, as described by a given textual reference. This task is challenging because it requires (1) the accurate recognition of various objects in a 3D scene and (2) the understanding of spatial relations in the description. However, current studies encounter difficulties in situations where multiple similar objects are present or when the descriptions involve intricate and abstract relations. In this paper, a novel, simple, and efficient Object-Centric Referring network, namely 3D-OCR, is presented to take high-quality semantic representation and deep relation modeling into account. Specifically, an offline Fine-grained Semantic Enhancement (FSE) module is designed to reinforce the object-centric semantic awareness with fine-grained high-quality object semantic representations. To achieve superior object-centric relation awareness, we propose a Deep Relation Modeling (DRM) module with the explicit and implicit relation self-attention module, enriching object features with relational context. Moreover, we utilize a vision-language contrastive loss to further improve the matching process between point cloud and language. Comprehensive experiments conducted on the challenging ScanRefer and Nr3D datasets corroborate the exceptional performance of our method, with an increase of +1.47% on ScanRefer and +1.2% on Nr3D.**

I. INTRODUCTION

In recent years, the quest to comprehend the real world through multiple lenses - including language, images, and point clouds - has become a hot topic in research. Within this context, 3D visual grounding stands out as a key foundational task in this field. It involves pinpointing the specific object within a given 3D scene and plays an important role in various robotic applications, such as indoor navigation [1], embodied agent [2] and human-robot interactions [3]. Common methods typically adopt a two-stage strategy [4]–[10], following a detection-then-matching pipeline: they first utilize pre-trained detectors to generate proposals, which are then aligned with language cues to pinpoint the most accurate match as the predicted target, as illustrated in Fig. 1 (a).

Two key observations can be gleaned when delving into the above pipeline: First, 3D visual grounding demands that the model be adept at differentiating between a myriad of objects in a scene, especially when there are many objects similar to the target, as is often the case in real-world scenarios. This means the model must possess a keen object-centric semantic awareness, capable of discerning unique attributes to differentiate between different classes of objects and cluster those of the same class together within the feature space. Second, a deeper understanding of the relation among objects is typically the intrinsic objective for an effective fusion and alignment module. Drawing on this, the information interaction and matching between point cloud and language is enabled, thereby highlighting the object-centric relation awareness in the 3D visual grounding task.

[1]School of Software and Microelectronics, Peking University, Beijing, China {ruilongren, xuweichen1999, tianhaofu1, yileidong0101, xuxinxin, huzc}@stu.pku.edu.cn, caojian@ss.pku.edu.cn
[2]Peking University Shenzhen Graduate School, Shenzhen China zhx@pku.edu.cn

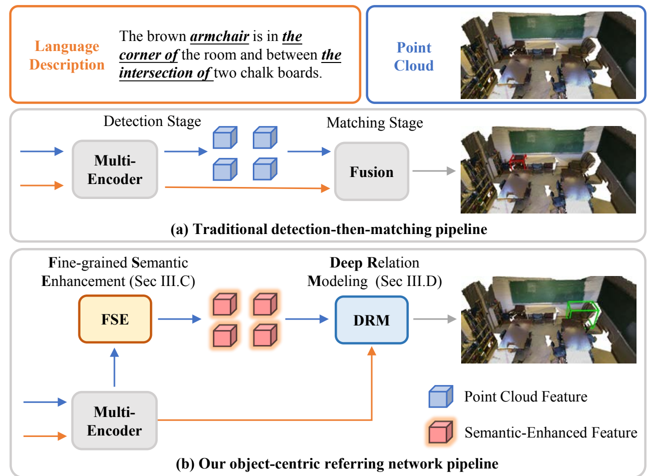

Fig. 1. The comparison between traditional detection-then-matching pipeline (a) and our method (b), where the **green** bounding box in the output denotes the correct result and the **red** means incorrect.

Prevailing methods make substantial advancements from following aspects: 1) explicitly decoupling textual attributes to alleviate ambiguity in a sentence [11]; 2) modeling shallow relational information among objects [7], [9]; 3) injecting generated view-robust scene representations into corresponding 3D scene features [10], [12]; 4) enhancing generic cross-modality representation with pre-training methods [13], [14]. Despite the effectiveness, the outcomes are still less-than-ideal. On one hand, the poor-quality of object semantic representation remains a drawback in most methods. This is because they either rely solely on point clouds [11] as visual input or use rendered 2D images as auxiliary [10]. Such limited visual representation often leads to incorrect recognition of target objects. On the other hand, most methods adopt shallow relationship modeling [7]–[9], which does a decent job in simple descriptions. While they tend to fall short when it comes to understanding more intricate or abstract relations. Therefore, is it possible to obtain high-quality of object semantic representations while achieving the deep relation understanding?

3D-OCR, a novel object-centric referring network, gives a "yes" answer. As shown in Fig. 1 (b), (1) we construct an offline Fine-grained Semantic Enhancement (FSE) module to reinforce object-centric semantic awareness. This inspiration arises from the fine-grained object segmentation capability of the celebrated 2D segmenter SAM [15], as well as the high-quality object semantic representation afforded by CLIP [16]. In detail, we begin by extracting valid class-agnostic 2D instance masks based on SAM [15] with a well-designed filtering strategy. Then our model aggregates per-mask features by fusing multi-view CLIP [16] image embeddings to form fine-grained 2D visual representations. Finally, we obtain the semantic-enhanced object features by concatenating the 3D features with its corresponding 2D counterpart. (2) A Deep Relation Modeling (DRM) module is introduced to achieve superior object-centric relation awareness. Initially, we utilize an explicit and implicit relation self-attention module that incorporates positional information of objects to enrich their features with relational context. Following this, we adopt a Co-Attention [17] mechanism to amalgamate these relation-enhanced object features with textual features. This fusion process promotes effective communication and alignment between the point cloud and language modalities.

The contributions of this paper can be summarized as:

- We propose a simple and efficient framework called 3D-OCR to improve 3D visual grounding from a new object-centric view, taking high-quality semantic representation and deep relation modeling into account.
- We design an offline fine-grained semantic enhancement module for strengthening object-centric semantic awareness and a deep relation modeling module to achieve superior object-centric relation awareness. Moreover, a vision-language contrastive loss is proposed to improve matching process between point cloud and language.
- Through extensive experiments, our method exhibits outstanding performance on challenging datasets such as ScanRefer [4] and Nr3D [5], thereby validating the effectiveness of our proposed approach.

II. Related Works

A. 2D Semantic Assist for 3D Visual Grounding

Many 3D visual grounding methods endeavor to use rich and clean 2D features to enrich semantic representations due to the inherent limitations of 3D point clouds, which are often sparse, noisy, and incomplete. There are two representative approaches: 1) utilizing 2D image semantics to enhance 3D point cloud representation [6], [18], and 2) integrating features of multi-view images rendered by the object proposal [19], [20]. However, the above approaches suffer from drawbacks such as slow extraction of image features in real-time and poor-quality semantic representation contained in rendered images. Instead, our approach proposes a novel offline fine-grained semantic enhancement module that not only effectively alleviates these issues, but also facilitates the enhancement of object-centric semantic awareness.

B. Relation Modeling with Transformer

Transformer [21] is widely applied in natural language processing, vision, vision-language and several other domains.The attention mechanism, a core component of the transformer architecture, is order-independent, and the positional information can be injected into each token. Inspired by this, some 3D visual grounding methods [7], [9] model relational information based on specially designed modules. 3DVG-Transformer [9] puts forward a coordinate-guided attention to refine the neighboring relations among clusters. TransRefer3D [7] defines the nonlinear relation representations based on entity features. However, they only adopt shallow relationship modeling and fail to represent more complex and abstract relations. In contrast, our method introduces a deep relation modeling module that effectively and fully models relations, improving object-centric relation awareness.

III. Proposed Method

A. Overview

The 3D-OCR is based on the powerful Multi3DRefer [19]. As shown in Fig. 2, given the point cloud $\boldsymbol{P} \in \mathbb{R}^{N\times(3+D)}$ (comprising N points characterized by xyz coordinates and extra D-dimensional attributes) and the textual description $\boldsymbol{T} = \{w_i\}_{i=1}^{L}$(a sentence with L words), we first adopt PointGroup [22] as our 3D encoder and CLIP [16] as the text encoder. Then we employ the fine-grained semantic enhancement module to enrich semantic representations of each object. Subsequently, we utilize the deep relation modeling module to model complicated relations among objects. At last, the grounding head uses a duo of fully connected layers to output the target object with the maximum probability.

B. Point Cloud and Language Encoder

Point Cloud Encoder. We first encode the point cloud $\boldsymbol{P}$ as object proposals features $\boldsymbol{F}_{\text{obj}}^{\text{3d}} \in \mathbb{R}^{M\times 32}$ and coordinates $\boldsymbol{C}_{\text{obj}}^{\text{3d}} \in \mathbb{R}^{M\times 6}$ by pre-trained PointGroup, where M denotes the number of proposals. These encoded object proposals will be used to regress the target object by a grounding head. Then, we fuse $\boldsymbol{F}_{\text{obj}}^{\text{3d}}$ with $\boldsymbol{F}_{\text{obj}}^{\text{2d}}$ extracted by FSE module to generate the semantic-enhanced object features $\boldsymbol{F}_{\text{obj}}^{\text{se}}$.

Language Encoder. Following [19], we encode the language input $\boldsymbol{T}$ with CLIP to obtain word-level and sentence-level feature vectors: $\boldsymbol{W} \in \mathbb{R}^{L\times 256}$, $\boldsymbol{S} \in \mathbb{R}^{1\times 256}$, where $\boldsymbol{W}$ is then used to predict the target object class and $\boldsymbol{S}$ will be calculated for the contrastive loss.

C. Fine-grained Semantic Enhancement (FSE)

To enrich visual representations of each object in 3D scenes, we devise a fine-grained semantic enhancement module based on the powerful SAM [15] and CLIP [16].

1) Valid Masks Extraction Based on SAM: Given K multi-view images within a scene, we first extract $\boldsymbol{N}_i, i \in [1, K]$ masks of the *i*-th image based on SAM and obtain initial masks $\boldsymbol{Q}$. These masks may contain many invalid results taking the ambiguity and fuzzy boundaries into account, such as excessive overlap masks and those encompassing multiple

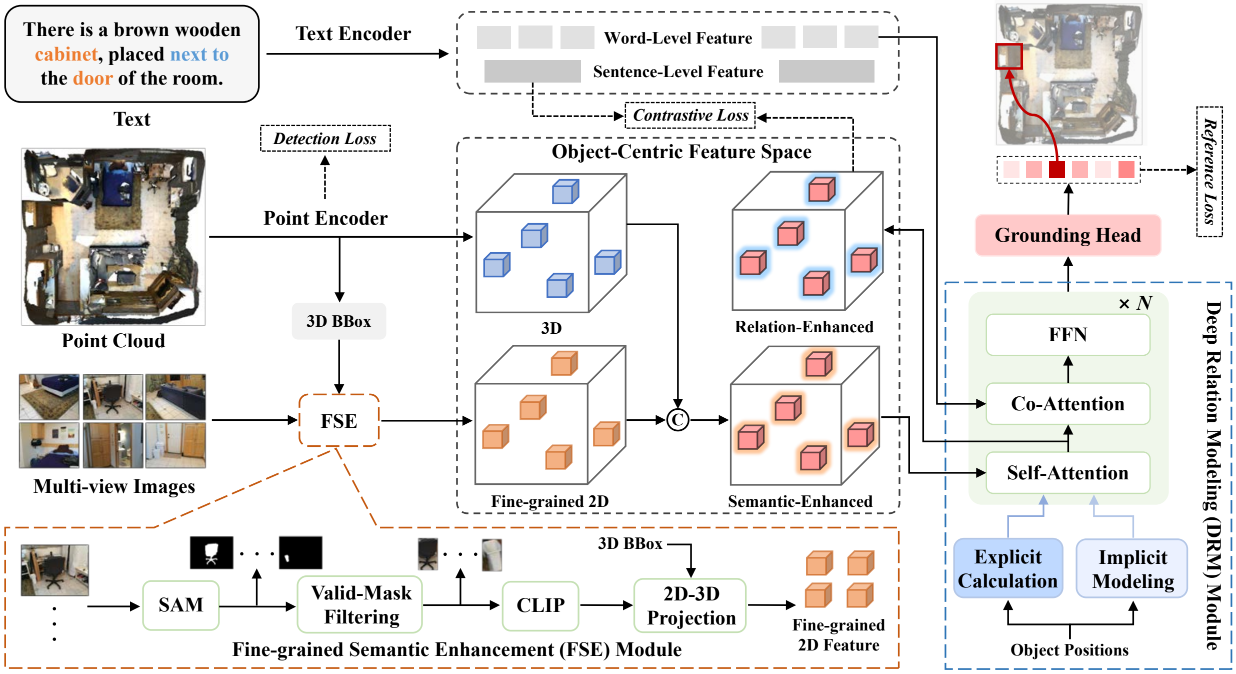

Fig. 2. The overview of our 3D-OCR framework, which takes a 3D point cloud and a text description as inputs and outputs the bounding box of the object that is most relevant to the input text. The Fine-grained Semantic Enhancement (FSE) module utilizes multi-view images corresponding to a 3D scene to enhance object-centric semantic awareness. Furthermore, the Deep Relation Modeling (DRM) module fuses relation-enhanced object features with text features to achieve superior object-centric relation awareness.

objects. Thus we introduce a valid-mask filtering strategy to filter out invalid masks, as shown in Algorithm 1: 1) We first set two ratio thresholds t_{min} and t_{max} to limit the area of each mask within a reasonable range. 2) Then, we calculate the similarities and IoU matrix among remaining masks and set a threshold T to retain masks that are highly representative. Thus we obtain the final valid masks $\boldsymbol{Q}_{\text{final}} = [\boldsymbol{q}_1, \dots, \boldsymbol{q}_K]$, where $\boldsymbol{q}_i, i \in [1, K]$ represents masks in the i-th image.

2) Fine-grained 2D-3D Feature Aggregation: The valid masks are then maped back to the original image to gain the corresponding image blocks. Then, we follow [19] and use the CLIP image encoder to extract the block features which will be filled into the corresponding area of the original image. Thus, we acquire the fine-grained 2D features for the i-th multi-view image as $\boldsymbol{F}_i^{\text{img}} = \sum_{j=1}^{N_i} \boldsymbol{f}_{i,j}^{\text{img}} \in \mathbb{R}^{256}$, where $\boldsymbol{f}_{i,j}^{\text{img}}$ denotes features of the j-th image block in the i-th image. We devise a transformation matrix $\boldsymbol{W}_{\text{t}}$ that projects 2D features to 3D space, utilizing the camera intrinsic parameters. Therefore, we obtain fine-grained 2D representations of the input 3D scene as $\boldsymbol{F}^{\text{2d}} = \sum_{i=1}^{K} \boldsymbol{W}_{\text{t}} \boldsymbol{F}_i^{\text{img}} \in \mathbb{R}^{256}$.

The above steps are executed offline for efficiency as we save the fine-grained 2D features to local storage and retrieve them directly while training. We apply PointGroup [22] to generate object proposals with 3D features $\boldsymbol{F}_{\text{obj}}^{\text{3d}}$ and coordinates $\boldsymbol{P}_{\text{obj}}^{\text{3d}}$. Then, these coordinates enable direct retrieval of the corresponding 2D features $\boldsymbol{F}_{\text{obj}}^{\text{2d}}$ from the 3D space. In the end, the 2D and 3D features are concatenated into a 288-dimensional set, which is then compacted to 256 dimensions via a 1D convolution to yield the semantic-enhanced object features: $\boldsymbol{F}_{\text{obj}}^{\text{se}} = \text{Conv}_{\text{1d}}([\boldsymbol{F}_{\text{obj}}^{\text{3d}}; \boldsymbol{F}_{\text{obj}}^{\text{2d}}]) \in \mathbb{R}^{M \times 256}$.

D. Deep Relation Modeling (DRM)

Understanding spatial relations among objects is essential to accurately identify the target within the scene. Distinguishing itself from other methods, 3D-OCR uses a deep relation modeling module to provide the effective and comprehensive object-centric relation awareness.

It is necessary to provide a concise review of the original attention mechanism in [21] since our DRM module is built upon it. Given an input sequence $\boldsymbol{X} \in \mathbb{R}^{n \times d_x}$ of n elements and the single-head attention first computes the query, key and value embeddings by $\boldsymbol{Q} = \boldsymbol{W}_Q \boldsymbol{X}$, $\boldsymbol{K} = \boldsymbol{W}_K \boldsymbol{X}$, $\boldsymbol{V} = \boldsymbol{W}_V \boldsymbol{X}$, where $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{d_x \times d_z}$ are matrix parameters. Then we calculate attention matrix using the query and key embeddings and aggregate the value embeddings to output the results as follows:

$$\boldsymbol{A} = \frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_z}}, \quad \boldsymbol{Z} = \text{Softmax}(\boldsymbol{A})\boldsymbol{V} \tag{1}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is the attention matrix, whose elements $a_{i,j}$ is the attention weight between the i-th and j-th object proposal. To capture more diverse relations, multi-head self-attention is employed where each head computes an independent and the outputs from all heads are concatenated.

1) Explicit and Implicit Relation Self-Attention: To model intricate relations among objects, the Explicit and Implicit Relation Self-Attention (EIR-SA) module is proposed, as shown in Fig. 3. Given the object positions, we explicitly calculate the relative distances under x, y, z coordinates, and the Euclidean distance in 3D space, denoted as $\boldsymbol{D}_{\text{x}}$, $\boldsymbol{D}_{\text{y}}$, $\boldsymbol{D}_{\text{z}}$, and $\boldsymbol{D}_{\text{e}}$, respectively. They constitute the first part of

Algorithm 1 Valid-mask Filtering Strategy in FSE.

Input: K multi-view images
Parameter: Threshold ratios t_{min}, t_{max}, IoU threshold T
Output: Final valid masks $\boldsymbol{Q}_{\text{final}}$

for $i = 1$ to K **do**
  $\{m_1, m_2, ..., m_{n_i}\} \leftarrow$ obtain initial masks based on SAM for the i-th image
  $\{r_1, r_2, ..., r_{n_i}\} \leftarrow$ calculate the area ratio of each mask to the entire image, and remain masks with area ratio between t_{min} and t_{max}
  compute the similarities $\boldsymbol{S}$ and IoU matrix $\boldsymbol{I}_{\text{mat}}$ among remaining masks with the entire image
  filter out the masks where $I_{\text{mat}} \geqslant T$ with lower $\boldsymbol{S}$, and select final valid masks $\boldsymbol{Q}_{\text{final}}$
end for
return $\boldsymbol{Q}_{\text{final}}$

relation-attention weights in our EIR-SA module:

$$\boldsymbol{W}_{\text{ex}} = \text{Concat}\left(\boldsymbol{D}_{\text{e}}, \boldsymbol{D}_{\text{x}}, \boldsymbol{D}_{\text{y}}, \boldsymbol{D}_{\text{z}}\right) \in \mathbb{R}^{M \times M \times 4} \tag{2}$$

Although these four relative distances encapsulate the majority of object relations in linguistic descriptions (e.g., $\boldsymbol{D}_{\text{e}}$ for "near" and "far", $\boldsymbol{D}_{\text{x}}$ for "right" and "left", $\boldsymbol{D}_{\text{y}}$ for "behind" and "front", $\boldsymbol{D}_{\text{z}}$ for "low" and "top"), they fall short when it comes to grasping the more complex spatial relations, such as *it is a black chair at the corner of the table*, or *there is a cabinet against the side of the room*. Inspired by [8], we introduce an implicit relation modeling method by encoding the direction angles among objects:

$$\boldsymbol{p}_{ij} = \left[\sin\left(\boldsymbol{\theta}_{\text{h}}\right), \cos\left(\boldsymbol{\theta}_{\text{h}}\right), \sin\left(\boldsymbol{\theta}_{\text{v}}\right), \cos\left(\boldsymbol{\theta}_{\text{v}}\right)\right] \tag{3}$$

where $\boldsymbol{\theta}_{\text{h}}, \boldsymbol{\theta}_{\text{v}}$ are the horizontal and vertical angles of the line connecting centers of the i-th object and j-th object in 3D space. The implicit pairwise relation embeddings $\boldsymbol{P}_{\text{im}} = \{\boldsymbol{p}_{ij}\} \in \mathbb{R}^{M \times M \times 4}$ are used to generate the second part of relation-attention weights in our EIR-SA module:

$$\boldsymbol{W}_{\text{im}} = \frac{\text{Linear}(\boldsymbol{X})\,\text{Linear}(\boldsymbol{P}_{\text{im}})^{\text{T}}}{\sqrt{d_z}} \in \mathbb{R}^{M \times M \times 4} \tag{4}$$

In the end, we concatenate the above explicit and implicit relation-attention weights and construct the relation-enhanced object features $\boldsymbol{F}_{\text{obj}}^{\text{re}}$ with object-pairs attention weights as the follows:

$$\boldsymbol{F}_{\text{obj}}^{\text{re}} = \text{Softmax}\left(\boldsymbol{A} + \boldsymbol{A}_{\text{EIR}}\right)\boldsymbol{V} \tag{5}$$

$$\boldsymbol{A}_{\text{EIR}} = \text{Concat}\left[\boldsymbol{W}_{\text{ex}},\ \log\sigma(\boldsymbol{W}_{\text{im}})\right] \tag{6}$$

2) Cross-Modal Interaction and Matching: In order to facilitate the information interaction and matching between point cloud and language, we leverage a Co-Attention module [17] to fuse relation-enhanced object features and text features. Specifically, we use $\boldsymbol{F}_{\text{obj}}^{\text{re}}$ as queries and $\boldsymbol{W}$ as keys and values to conduct cross-modal interaction: $\boldsymbol{F}_{\text{obj}}^{\text{co}} = \text{CoAttn}(\boldsymbol{F}_{\text{obj}}^{\text{re}}, \boldsymbol{W}, \boldsymbol{W}) \in \mathbb{R}^{M \times d_o}$, where d_o denotes the dimension of output proposals and the FNN is used to transform the cross-modal features $\boldsymbol{F}_{\text{obj}}^{\text{co}}$ for the final matching.

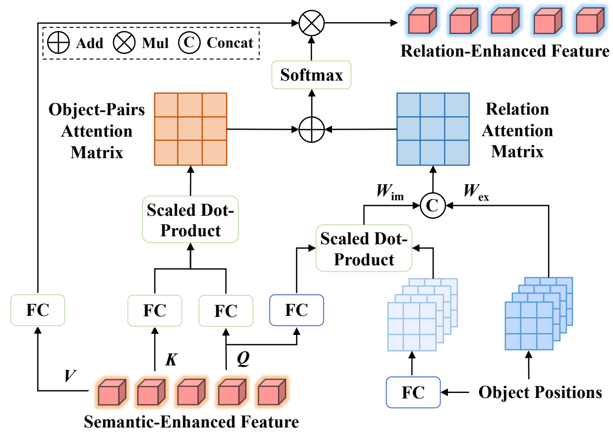

Fig. 3. The Explicit and Implicit Relation Self-Attention (EIR-SA) module, where the self-attention matrix is combined by object-pairs and relation attention matrix.

E. Training and Loss Function

We employ multiple losses $\mathcal{L} = \mathcal{L}_{\text{det}} + \mathcal{L}_{\text{ref}} + \mathcal{L}_{\text{con}}$ following the previous works [8], [19] to train our 3D-OCR model end-to-end, including the 3D detection loss $\mathcal{L}_{\text{det}}$, the reference loss $\mathcal{L}_{\text{ref}}$ and the contrastive loss $\mathcal{L}_{\text{con}}$. The detection loss $\mathcal{L}_{\text{det}}$ is introduced in PointGroup [22] for supervising per-point semantic class, offset vectors towards object center and objectness confidence score. We use the multi-class cross-entropy loss as $\mathcal{L}_{\text{ref}}$ for supervising the DRM module and grounding head to locate target object with the maximum probability. Finally, to facilitate the learning of better matching process between point cloud and language vis cross-modal alignment, we design a symmetric contrastive loss $\mathcal{L}_{\text{con}}$ which is composed of two cross-entropy loss:

$$\mathcal{L}_{\text{con}}^{\text{O}\rightarrow\text{S}} = -\log\frac{\exp(\cos(\bar{\boldsymbol{F}}_i^{\text{re}}, \boldsymbol{S}_i)/\tau)}{\sum_{j=1}^{B}\exp(\cos(\bar{\boldsymbol{F}}_i^{\text{re}}, \boldsymbol{S}_j)/\tau)} \tag{7}$$

$$\mathcal{L}_{\text{con}}^{\text{S}\rightarrow\text{O}} = -\log\frac{\exp(\cos(\boldsymbol{S}_i, \bar{\boldsymbol{F}}_i^{\text{re}}/\tau)}{\sum_{j=1}^{B}\exp(\cos(\boldsymbol{S}_i, \bar{\boldsymbol{F}}_j^{\text{re}})/\tau)} \tag{8}$$

$$\mathcal{L}_{\text{con}} = (\mathcal{L}_{\text{con}}^{\text{O}\rightarrow\text{S}} + \mathcal{L}_{\text{con}}^{\text{S}\rightarrow\text{O}})/2 \tag{9}$$

where $\left(\bar{\boldsymbol{F}}_i^{\text{re}}, \boldsymbol{S}_i\right)$ is the mean relation-enhanced features of all objects paired with a sentence for the i-th batch, B is the batch size and τ is the temperature parameter.

IV. Experiments

A. Datasets

ScanRefer [4] dataset is built on 800 3D scenes from ScanNet dataset [24]. We follow the official split and use 36,665 human-written language annotations describing 7,875 objects from 562 3D scenes for training, and evaluates on 9,508 sentences for 2,068 objects from 141 3D scenes. According to whether the target object is a unique object class in the scene, the dataset can be divided into a "unique" and a "multiple" subset in evaluation. The metric of the ScanRefer is the Acc@mIoU with $m \in \{0.25, 0.5\}$, which means the fraction of descriptions whose predicted box overlaps the Ground Truth (GT) with IoU $> m$.

TABLE I
COMPARISON RESULTS ON THE SCANREFER [4] AND NR3D [5] DATASETS. WE HIGHLIGHT THE BEST PERFORMANCE IN **BOLD**.

Method	Publication	Data	ScanRefer						Nr3D				
			Unique		Multiple		Overall		Easy	Hard	View Dep	View Indep	Overall
			@0.25	@0.5	@0.25	@0.5	@0.25	@0.5					
ScanRefer [4]	ECCV2020	3D Only	67.64	46.19	32.06	21.26	38.97	26.10	-	-	-	-	-
ReferIt3D [5]	ECCV2020	3D Only	53.80	37.50	21.00	12.80	26.40	16.90	43.6	27.9	32.5	37.1	35.6
TransRefer3D [7]	MM2021	3D Only	-	-	-	-	-	-	56.7	39.6	42.5	50.7	48.0
SAT [6]	ICCV2021	3D+2D	73.21	50.83	37.64	25.16	44.54	30.14	56.3	42.4	46.9	50.4	49.2
3DVG-Transformer [9]	ICCV2021	3D Only	77.16	58.47	38.38	28.70	45.90	34.47	48.5	34.8	34.8	43.7	40.8
MVT [10]	CVPR2022	3D+2D	77.67	66.45	31.92	25.26	40.80	33.26	61.3	49.1	54.3	55.4	55.1
3D-SPS [23]	CVPR2022	3D+2D	84.12	66.72	40.32	29.82	48.82	36.98	58.1	45.1	48.0	53.2	51.5
ViL3DRel [8]	NeurIPS2022	3D Only	81.58	68.62	40.30	30.71	47.94	37.73	70.2	57.4	**62.0**	64.5	64.4
EDA [11]	CVPR2023	3D Only	85.76	68.57	**49.13**	37.64	**54.59**	42.26	-	-	-	-	52.1
Multi3DRefer [19]	ICCV2023	3D+2D	85.30	77.20	43.80	36.80	51.90	44.70	55.6	43.4	42.3	52.9	49.4
Ours	-	3D+2D	**86.67**	**77.56**	45.35	**37.65**	53.37	**45.39**	**72.3**	**57.9**	61.8	**65.9**	**65.6**

TABLE II
ABLATION STUDY ABOUT TRAINING 3D-OCR WITH DIFFERENT TYPES OF PROPOSAL FEATURES. "3D" FOR POINT CLOUD FEATURES, "RENDERED" FOR RENDERED IMAGE FEATURES IN MULTI3DREFER [19], AND "FINE-GRAINED" FOR FINE-GRAINED IMAGE FEATURES IN OUR FSE MODULE.

Features	Unique @0.5	Multiple @0.5	Overall @0.5
3D	75.11	35.36	43.78
Rendered	73.09	32.03	42.10
Fine-grained	74.46	33.98	42.65
3D + Rendered	76.37	36.29	44.76
3D + Fine-grained	**77.56**	**37.65**	**45.39**

TABLE III
ABLATION STUDY OF THE EXPLICIT AND IMPLICIT RELATION SELF-ATTENTION (EIR-SA) MODULE. "ERM" FOR EXPLICIT RELATION MODELING, "IRM" FOR IMPLICIT RELATION MODELING.

ERM	IRM	Unique @0.5	Multiple @0.5	Overall @0.5
		74.51	34.67	43.43
✓		75.31	35.36	44.38
	✓	75.16	36.24	44.21
✓	✓	**77.56**	**37.65**	**45.39**

TABLE IV
ABLATION STUDY OF OUR METHOD 3D-OCR WITH AND WITHOUT CONTRASTIVE LOSS.

Contrastive Loss	Unique @0.5	Multiple @0.5	Overall @0.5
	77.34	35.59	44.71
✓	**77.56**	**37.65**	**45.39**

Nr3D [5] dataset is also proposed on ScanNet [24] with 41,503 free-form sentences similar to ScanRefer's text annotation. The sentences are split into "easy" and "hard" subsets in evaluation, where the target object in "easy" subset only contains one same-class distractor in the scene while it contains multiple ones in the "hard" subset. According to whether the sentence requires a specific viewpoint to ground the referred object, the dataset can also be partitioned into "view depedent" and "view independent" subsets. GT boxes for all candidate objects in the scene are provided in Nr3D dataset. The metric is the accuracy of selecting the target bounding box among proposals.

B. Implementation Details

Following [19], we adopt a re-implementation of PointGroup [22] with the Minkowski Engine as point cloud encoder. We freeze SAM with VIT-L as mask extractor and CLIP with VIT-B/32 as image encoder in FSE module. Our text encoder is also initialized from CLIP. We set $t_{\text{min}} = 0.05$, $t_{\text{max}} = 0.75$ and $T = 1.0$ during the valid-mask-filtering phase. We set the dimension $d = 256$ and use 8 heads for 3 transformer layers. Our 3D-OCR is optimized by the AdamW algorithm with the momentum of 0.9, the weight decay of 1e-4 and the initial learning rate of 5e-4. We train 3D-OCR on ScanRefer and Nr3D both for 60 epochs with a batch size of 4. For data augmentation, we randomly apply coordinate jitter, x-axis flipping and rotation around the z-axis. All experiments are conducted on a single NVIDIA RTX 4090 GPU using Pytorch.

C. Comparison Results

We compare our 3D-OCR against recent state-of-the-art (SOTA) methods on two 3D visual grounding datasets.

ScanRefer. The left half of Table I shows the comparison results on ScanRefer. 3D-OCR outperforms most SOTA method comprehensively, especially by +1.47 at Acc@0.25 and +0.69 at Acc@0.5 compared to the powerful method Multi3DRefer [19]. Delving into the details, for the "Unique" subset—a challenging subset that requires robust object-centric semantic discernment for the precise identification of unique target object—the accuracy rates achieved by most pre-existing solutions fail to surpass the 85% at Acc@0.25 and 70% at Acc@0.5. In contrast, we realize the remarkable 86.67% and 77.56% thanks to the semantic-enhanced object features generated by FSE module.

Nr3D. The comparison results on Nr3D are reported at the right half of Table I. While descriptions in Nr3D are detailed and complex, our 3D-OCR still surpasses SOTA methods across most metrics and reaches the best performance of

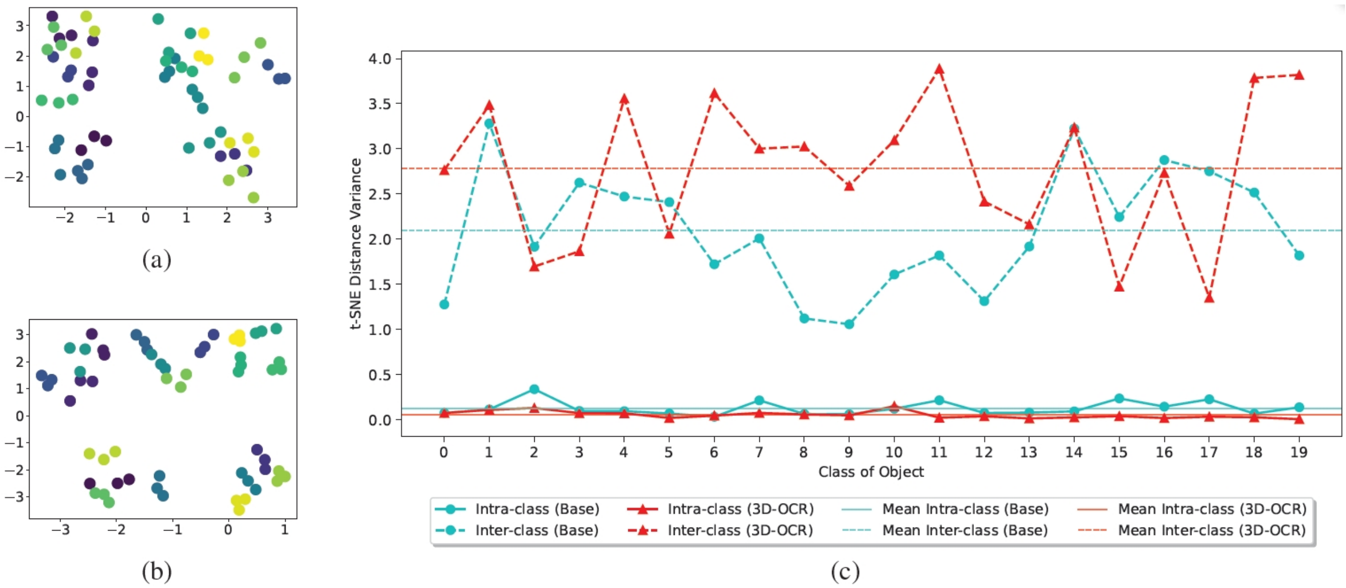

Fig. 4. t-SNE [25] visualization and distance variances of proposal features. Base is the variant of 3D-OCR without FSE and DRM modules. (a) and (b) represent the t-SNE visualization of Base and 3D-OCR, while (c) denotes the t-SNE distance variances for inter-class and intra-class objects in ScanRefer.

65.6% in the "Overall" setting. Additionally, the "Hard" subset within the Nr3D, characterized by the presence of multiple distracting objects from identical categories, serves as a rigorous test for any object recognition system. Remarkably, our 3D-OCR method demonstrates a significant enhancement of 0.5% in accuracy compared to the SOTA method Vil3DRel in the "Hard" subset, thereby underscoring the proficiency of our method in effectively differentiating between objects of the same class situated within 3D scenes.

D. Ablation Study

We conduct ablation studies on the ScanRefer validation set to investigate the contribution of each component.

1) In Table II, we present the ablation results for different types of proposal features, which clearly showcase the superiority of our fine-grained 2D features over the rendered image features utilized in the baseline Multi3DRefer [19]. Moreover, the integration of fine-grained 2D features with 3D features results in the highest performance enhancement, demonstrating better object-centric semantic awareness obtained by our method. Additionally, we observe that both types of 2D features are less effective than 3D features alone, underscoring the value of 3D information.

2) To showcase the capabilities of our 3D-OCR in relation modeling, we perform ablation analysis on EIR-SA module, with results presented in Table III. Rows 2 and 3 illustrate that both explicit and implicit relation modeling significantly enhance spatial reasoning. Row 4 confirms that using a multi-perspective approach to model relative positions results in a more complete understanding of object relations, implying the advanced object-centric relation awareness.

3) Table IV reports the performance of our 3D-OCR with and without contrastive learning on ScanRefer validation set. We note a consistent enhancement across all subsets due to the incorporation of contrastive loss, with a particularly notable improvement of +2.06% in the 'hard' subsets. This finding substantiates that the contrastive loss contributes to better fusion of visual and language features, thereby leading to a more precise differentiation of similar objects.

4) In order to assess the capacity of our 3D-OCR to differentiate various objects within 3D scenes, we present a t-Distributed Stochastic Neighbor Embedding (t-SNE) [25] visualization of proposal features, as depicted in Fig. 4. We assign 20 labels in ScanRefer [4] to the proposals with the nearest GT object [19], [26]. We compare the performance of 3D-OCR with its variant that does not include FSE and DRM modules, namely Base. We set each point represents the features of one object proposal and different colors are used to distinguish different classes of proposals in Fig. 4 (a) and (b). The delineation is markedly noticeable, with our 3D-OCR demonstrating a significant consolidation of features for intra-class objects and a clear dispersion for those from inter-class. Furthermore, We obtain the same assertion more intuitively by calculating the feature distance variances among these objects, which is presented in Fig. 4 (c). The inter-class mean distance variances from 3D-OCR surpass those from Base, while the intra-class variances demonstrate the converse trend. Consequently, we conduct that our 3D-OCR is better at distinguishing objects in the scene, which facilitates the optimization of downstream tasks.

E. Qualitative Results

We perform a qualitative comparison of 3D-OCR with the baseline model Multi3DRefer [19] on ScanRefer [4] dataset and present representative examples in Fig. 5. These examples indicate that 3D-OCR has a better understanding of the intricate relations between scene and language.

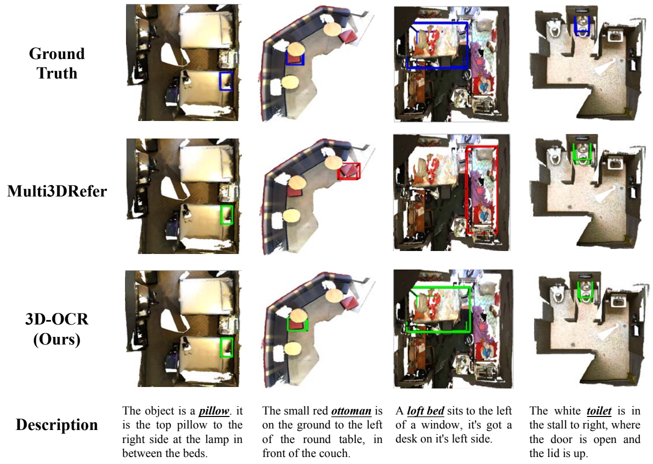

Fig. 5. The qualitative results with ScanRefer [4]. The blue/red/green colors indicate the ground truth/incorrect/correct bounding boxes.

V. Conclusion

In this paper, we introduce 3D-OCR, a simple and efficient framework designed to boost 3D visual grounding from a novel object-centric view. In detail, we propose an offline Fine-grained Semantic Enhancement (FSE) module to reinforce object-centric semantic awareness with fine-grained high-quality object semantic representation. A Deep Relation Modeling (DRM) module is proposed by explicitly and implicitly embedding positional information and modeling contextual relations among objects, achieving superior object-centric relation awareness. Extensive experiments underscores the outstanding performance and effectiveness of our method. In the future, we will delve deeper into the interaction between object and scene to further advance 3D visual grounding and related 3D vision-language tasks.

References

[1] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, "3d-llm: Injecting the 3d world into large language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 36, 2023, pp. 20 482–20 494.

[2] T. Wang, X. Mao, C. Zhu, R. Xu, R. Lyu, P. Li, X. Chen, W. Zhang, K. Chen, T. Xue *et al.*, "Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai," *arXiv preprint arXiv:2312.16170*, 2023.

[3] J. Kim, G.-C. Kang, J. Kim, S. Shin, and B.-T. Zhang, "Gvcci: Lifelong learning of visual grounding for language-guided robotic manipulation," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 952–959.

[4] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *In European Conference on Computer Vision (ECCV)*, 2020, pp. 202–221.

[5] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *In European Conference on Computer Vision (ECCV)*, 2020, pp. 422–440.

[6] Z. Yang, S. Zhang, L. Wang, and J. Luo, "Sat: 2d semantics assisted training for 3d visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2021, pp. 1856–1866.

[7] D. He, Y. Zhao, J. Luo, T. Hui, and S. Liu, "Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 2344–2352.

[8] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Language conditioned spatial relation reasoning for 3d object grounding," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 20 522–20 535.

[9] L. Zhao, D. Cai, L. Sheng, and D. Xu, "3dvg-transformer: Relation modeling for visual grounding on point clouds," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 2928–2937.

[10] S. Huang, Y. Chen, J. Jia, and L. Wang, "Multi-view transformer for 3d visual grounding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 15 524–15 533.

[11] Y. Wu, X. Cheng, R. Zhang, Z. Cheng, and J. Zhang, "Eda: Explicit text-decoupling and dense alignment for 3d visual grounding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 19 231–19 242.

[12] A. Delitzas, M. Parelli, N. Hars, G. Vlassis, S. Anagnostidis, G. Bachmann, and T. Hofmann, "Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes," in *34th British Machine Vision Conference (BMVC)*, 2023.

[13] Z. Zhu, X. Ma, Y. Chen, Z. Deng, S. Huang, and Q. Li, "3D-VisTA: Pre-trained transformer for 3D vision and text alignment," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 2911–2921.

[14] Z. Jin, M. Hayat, Y. Yang, Y. Guo, and Y. Lei, "Context-aware alignment and mutual masking for 3d-language pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 10 984–10 994.

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, and R. Girshick, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 4015–4026.

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine lLearning (ICML)*, 2021, pp. 8748–8763.

[17] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Advances in neural information processing systems*, vol. 29, 2016.

[18] Z. Guo, Y. Tang, R. Zhang, D. Wang, Z. Wang, B. Zhao, and X. Li, "Viewrefer: Grasp the multi-view knowledge for 3d visual grounding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 372–15 383.

[19] Y. Zhang, Z. Gong, and A. X. Chang, "Multi3drefer: Grounding text description to multiple 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 15 225–15 236.

[20] E. Bakr, Y. Alsaedy, and M. Elhoseiny, "Look around and refer: 2d synthetic semantics knowledge distillation for 3d visual grounding," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, 2022, pp. 37 146–37 158.

[21] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems (NeurlPS)*, vol. 30, 2017.

[22] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3d instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4867–4876.

[23] J. Luo, J. Fu, X. Kong, and S. Liu, "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 (CVPR)*, 2022, pp. 16 454–16 463.

[24] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5828–5839.

[25] L. van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 86, pp. 2579–2605, 2008.

[26] T. Zhang, S. He, D. Tao, B. Chen, Z. Wang, and S.-T. Xia, "Vision-language pre-training with object contrastive learning for 3d scene understanding," *arXiv preprint arXiv:2305.10714*, 2023.