X-RefSeg3D: Enhancing Referring 3D Instance Segmentation via Structured Cross-Modal Graph Neural Networks

Zhipeng Qian[*], **Yiwei Ma**[*], **Jiayi Ji**, **Xiaoshuai Sun**[†]

Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China,
Xiamen University, 361005, P.R. China
qianzhipeng@stu.xmu.edu.cn, yiweima@stu.xmu.edu.cn, jjyxmu@gmail.com, xssun@xmu.edu.cn

Abstract

Referring 3D instance segmentation is a challenging task aimed at accurately segmenting a target instance within a 3D scene based on a given referring expression. However, previous methods have overlooked the distinct roles played by different words in referring expressions. Additionally, they have failed to incorporate the positional relationship within referring expressions with the spatial correlations in 3D scenes. To alleviate these issues, we present a novel model called X-RefSeg3D, which constructs a cross-modal graph for the input 3D scene and unites textual and spatial relationships for reasoning via graph neural networks. Our approach begins by capturing object-specific text features, which are then fused with the instance features to construct a comprehensive cross-modal scene graph. Subsequently, we integrate the obtained cross-modal features into graph neural networks, leveraging the K-nearest algorithm to derive explicit instructions from expressions and factual relationships in scenes. This enables the effective capture of higher-order relationships among instances, thereby enhancing feature fusion and facilitating reasoning. Finally, the refined feature undergoes a matching module to compute the ultimate matching score. Experimental results on ScanRefer demonstrate the effectiveness of our method, surpassing previous approaches by a substantial margin of ***+3.67%*** in terms of mIOU. The code and models are available at https://github.com/qzp2018/X-RefSeg3D.

Introduction

Understanding natural language and its relationship with visual information forms the fundamental basis for establishing a connection between humans and machines in the realm of artificial intelligence. Given that real-life scenarios inherently exist in three-dimensional (3D) space, the integration of language and spatial information within 3D environments holds immense value across diverse domains, encompassing VR/AR applications, navigation(Wu et al. 2022a), scene understanding (Peng et al. 2022), and intelligent perception (Zheng et al. 2022; Deng et al. 2022). Several tasks are proposed to promote the research in this area, including Multimodal learning (Ma et al. 2022, 2023; Ji et al. 2022b, 2020, 2022a; Yang et al. 2023; Wu et al. 2023), 3D Representation Learning (Wang et al. 2023; Zhao et al. 2023; Huang et al. 2023) and so on. One such task, known as referring 3D instance segmentation, involves the identification and segmentation of specific 3D objects described through natural language. This task is both novel and highly challenging due to the unordered and sparse nature of 3D point clouds, coupled with complex spatial and semantic relationships existing in the scene and expression, making the learning process significantly more demanding.

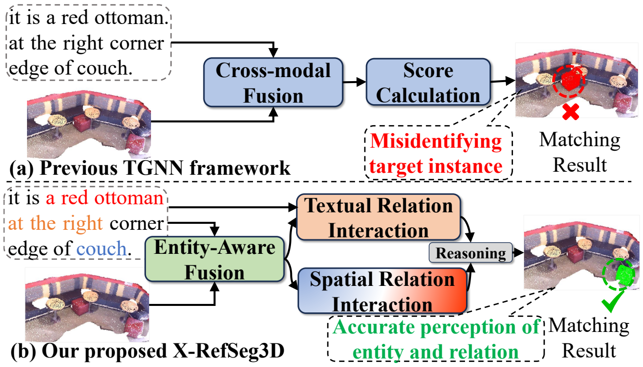
it is a red ottoman. at the right corner edge of couch.
Cross-modal Fusion
Score Calculation
Misidentifying target instance
Matching Result
(a) Previous TGNN framework
it is a red ottoman at the right corner edge of couch.
Entity-Aware Fusion
Textual Relation Interaction
Spatial Relation Interaction
Reasoning
Accurate perception of entity and relation
Matching Result
(b) Our proposed X-RefSeg3D

Figure 1: (a) The previous TGNN (Huang et al. 2022) fuses instance features and textual features and then calculates the matching score of each instance feature to find the target instance. (b) Our proposed X-RefSeg3D first integrates entity-related linguistic information into visual features to construct a cross-modal scene graph and then carries out textual relation interaction and spatial relation interaction. After jointly exploiting beneficial vision-language cues, an accurate target instance matching can thus be obtained.

The TGNN model (Huang et al. 2022) made an initial endeavour to tackle the challenging task of referring 3D instance segmentation. Specifically, as shown in Figure 1(a), TGNN incorporates instance features with textual features and calculates the matching score of each instance feature to identify the target instance referred to in the expression. Despite demonstrating superior performance and employing a straightforward pipeline, TGNN does have certain limitations that hinder its effectiveness. One notable drawback lies in its inability to explicitly prioritize the entity description

[*] These authors contributed equally.

[†] The corresponding author.

Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

words in the expression, which ultimately undermines its results. Additionally, TGNN overlooks the vital spatial relationships hinted at in expressions and their alignment with spatial relationships in 3D scenes. As a result, it falls short of accurately capturing the complex relationships within the 3D scene, leading to suboptimal segmentation outcomes.

By observing how humans identify objects in real-world situations using referring expressions, we gain valuable insights to address these challenges. Humans first decode the main objects in scenes based on these descriptions and then understand the relationships the descriptions convey. This comprehension, when combined with the scene's spatial relationships, helps determine the object's identity. For example, when identifying the target instance in Figure 1, people first focus on terms like "red ottoman" and "couch". Next, they merge the scene's spatial indicators with descriptors like "at the right" from the expression. This combination of textual and spatial information enables humans to accurately identify the mentioned instance. Inspired by human cognitive reasoning, we introduce X-RefSeg3D, showcased in Figure 1(b). This model is specifically designed for high-level comprehension and fine-grained feature fusion. X-RefSeg3D consists of two key modules: the Entity-Aware Fusion (EAF) and the Relation-Driven Interaction (RDI) modules. The EAF module selectively extracts textual features that describe entities, subsequently integrating them into the instance feature to craft a cross-modal scene graph. Through this strategy, every graph node embodies both visual and textual descriptors, enhancing clarity and distinctiveness. The RDI module focuses on high-order semantic comprehension. It comprises both textual relation interaction and spatial relation interaction modules. By examining relationships within expressions and 3D scenes, it forms both the textual relation graph and the spatial relation graph, respectively. Merging these graphs enables us to combine explicit instructions from expressions with factual relationships in scenes, yielding a comprehensive representation enriched with both relation and attribute. Ultimately, X-RefSeg3D identifies the target instance by comparing the comprehensive representation of each instance with the expression, ensuring accurate matching. Our main contributions are summarized as follows:

- We propose the Entity-Aware Fusion (EAF) module, which facilitates the modelling of relationships between expressions and objects, enabling the accurate construction of the object-specific cross-modal scene graph, each node of which embodies both visual and textual descriptors, enhancing clarity and distinctiveness.
- We introduce the Relation-Driven Interaction (RDI) module to achieve a high-order semantic comprehension. By examining relationships within expressions and 3D scenes, the model combines explicit instructions from expressions with factual relationships in scenes, yielding a comprehensive representation enriched with both relation and attribute.
- The incorporation of EAF and RDI into the X-RefSeg3D model yields a significant enhancement. In comparison to the previous state-of-the-art (SOTA) method utilizing the same backbone, our approach demonstrates a remarkable improvement of ***+3.67%*** mIOU.

Related Work

3D Instance Segmentation

3D Instance Segmentation has gained increasing attention in the field of 3D vision, and a variety of methods have been proposed. Most methods can be divided into two-stage or single-stage. Two-stage methods (Lee et al. 2018; Hou, Dai, and NieBner 2019; Yang et al. 2019) involve separate object detection and instance segmentation steps, offering advantages such as higher segmentation accuracy and better precision for complex scenes. On the other hand, one-stage methods (Elich et al. 2019; Jean et al. 2019; Schult et al. 2022) directly perform object detection and instance segmentation simultaneously in a single step, which offers advantages such as simplicity, efficiency, real-time processing, and end-to-end learning. Following TGNN, our model also adopts the pre-trained SCN model (Retinskiy 2019) as our visual feature extractor.

3D Visual Grounding

The 3D Visual Grounding task (Wu et al. 2022b; Yuan et al. 2021; Feng et al. 2021; Cai et al. 2022) is instrumental in enabling computers to comprehend natural language instructions. This field has witnessed significant advancements due to the availability of various 3D Visual Grounding datasets, including ScanNetv2 (Dai et al. 2017), ScanRefer (Chen, Chang, and Nießner 2020), Sr3D and Nr3D (Achlioptas et al. 2020). Most existing approaches in this area adopt two-stage methods (Feng et al. 2021; Zhao et al. 2021; Yuan et al. 2021), treating vision grounding as a detection-then-matching task. Besides, TGNN introduces a related task called Referring 3D Instance Segmentation, which aims to segment the target instance in 3D scenes based on a query expression. Referring 3D Instance Segmentation offers several advantages over the 3D Visual Grounding task. For example, it exhibits finer target localization capabilities, provides more precise identification of target instances, and alleviates ambiguity problems. These advantages highlight the great significance and application value of Referring 3D Instance Segmentation task. In this paper, we will continue to explore this task and further exploit its potential.

Graph-Based Reasoning

Graph neural networks (GNNs) play a vital role in content construction by leveraging the graph structure to incorporate contextual information, improving content comprehension. GNNs are widely applied in REC (Yang, Li, and Yu 2019, 2020), RES (Huang et al. 2020), 3D Visual Grounding (Feng et al. 2021), and other tasks. In the task of Referring 3D Instance Segmentation, TGNN utilizes GNN layers to compute an attention map for guiding the aggregation of multimodal node information. However, the usage of GNNs in TGNN is rudimentary, leading to limited performance. In this paper, we propose a more carefully designed GNN, which has more precise context modelling, better cross-modal feature fusion, and interpretability.

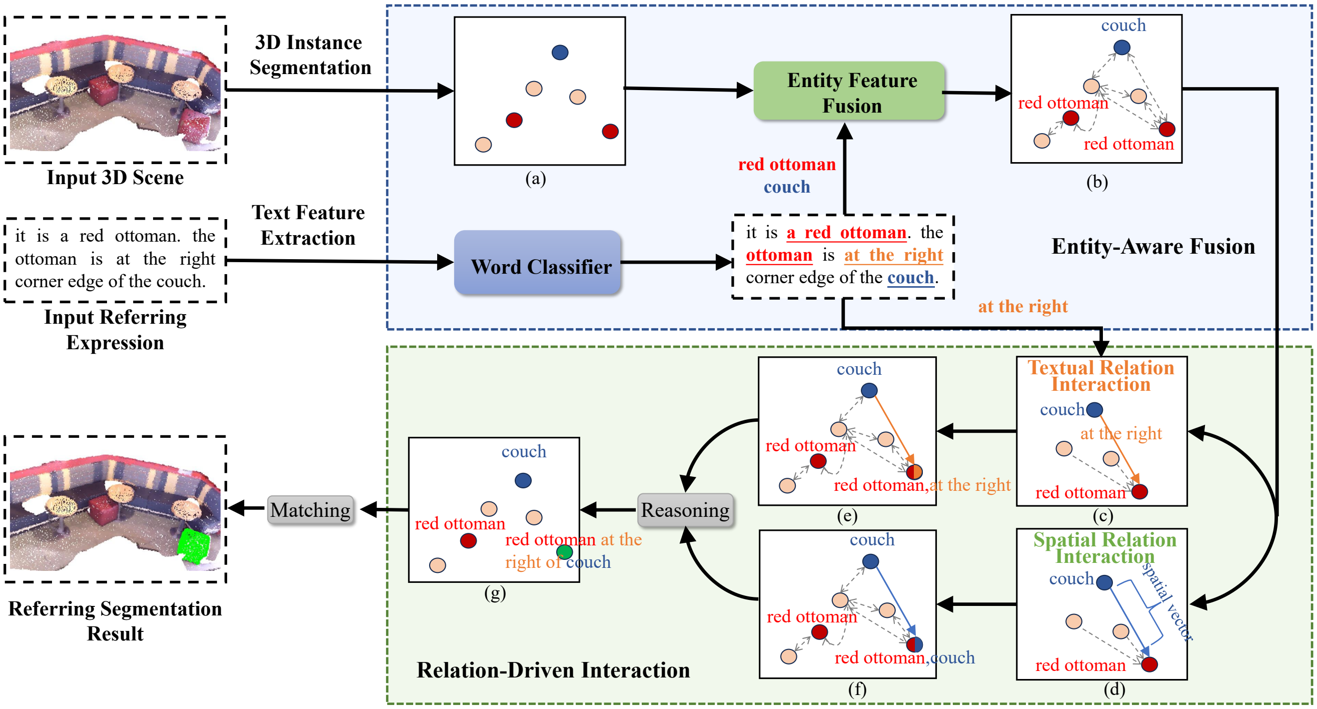
3D Instance Segmentation
Input 3D Scene
Text Feature Extraction
it is a red ottoman. the ottoman is at the right corner edge of the couch.
Input Referring Expression
Entity Feature Fusion
Word Classifier
red ottoman
couch
at the right
couch
red ottoman
red ottoman
(a)
(b)
Entity-Aware Fusion
Textual Relation Interaction
Spatial Relation Interaction
spatial vector
red ottoman, at the right
red ottoman, couch
(c)
(d)
(e)
(f)
(g)
Reasoning
Matching
red ottoman at the right of couch
Referring Segmentation Result
Relation-Driven Interaction

Figure 2: An overview of our proposed network. Different colored circles in the picture represent different objects, i.e. “red” means ottoman, and “blue” means “couch”. Our method consists of the Entity-Aware Fusion (EAF) module and the Relation-Driven Interaction (RDI) module.

Method

Referring 3D instance segmentation is the task that aims to segment the target instance out of the 3D scene according to the expression. Following TGNN (Huang et al. 2022), we transform the task into a matching problem. First, we use a pre-trained 3D instance segmentation model to extract point features and predict instance masks and corresponding centre coordinates. Second, We introduce a novel Entity-Aware Fusion (EAF) module that generates a directed cross-modal scene graph $\hat{G}$ by fusing entity description language features with corresponding instance features and utilizing relative spatial relations. Third, a Relation-Driven Interaction (RDI) module is proposed, it enhances the cross-modal features by integrating the surrounding instances’ features, helping the model better understand the relative position relationship contained within space and expression. Lastly, the model is trained on two complementary loss functions and generates the final prediction score based on refined features.

Visual and Linguistic Feature Extraction

Our model adopts the 3D instance segmentation model and mask prediction algorithm consistent with TGNN to extract instance masks, $I = \{I_1, I_2, ..., I_C\}$, and the corresponding visual features, $F = \{f_1, f_2, ..., f_C\}$, where C is the number of instances segmented within the scene and $f_i \in \mathbb{R}^{C_f}(i \in \{1, 2, ..., C\})$. For linguistic feature representation, a pre-trained GloVE model (Pennington, Socher, and Manning 2014) is used to embed each word into a 300-d vector and then encoded by a language encoder, such as GRU (Cho et al. 2014) and BERT (Devlin et al. 2019), into linguistic features $L = \{l_1, l_2, ..., l_T\}$, where T is the length of expression and $l_t \in \mathbb{R}^{C_l\times 1}(t \in \{1, 2, ..., T\})$ denotes feature of the t-th word.

Entity-Aware Fusion

In this section, we introduce the Entity-Aware Fusion Module, which effectively integrates entity information from the expression with corresponding objects. Additionally, this module utilizes the fused cross-modal features and spatial positions of instance to construct a comprehensive cross-modal scene graph (Fei et al. 2023), which is used for further inference.

Cross-Modal Entity Perception and Fusion. Words of different parts of speech play various roles in the expression, and entity-related words are the key to identifying the target instance. Following (Yu et al. 2018; Yang, Li, and Yu 2019; Huang et al. 2020), we divide the words into four different types: entity, attribute, relation, and unnecessary part words, and predict the weight of each type for each word as follows:

$$w_t = \mathrm{softmax}(W_2\sigma(W_1 l_t + b_1) + b_2), \qquad (1)$$

where $W_1 \in \mathbb{R}^{C_w\times C_l}$, $W_2 \in \mathbb{R}^{4\times C_w}$, $b_1 \in \mathbb{R}^{C_w\times 1}$ and $b_2 \in \mathbb{R}^{4\times 1}$ are learnable parameters, $\sigma(\cdot)$ is sigmoid function, $w_t = [w_t^{ent}, w_t^{attr}, w_t^{rel}, w_t^{un}]$, denotes the weight of entity, attribute, relation and unnecessary part for t-th word.

Next, an attention map is calculated between textual features $L = \{l_1, l_2, ..., l_T\}$ and instance features $F = \{f_1, f_2, ..., f_C\}$ as follows:

$$
\begin{aligned}
f_i^j &= \text{MLP}_1^{\text{j}}(f_i), & (2)\\
m_{t,i}^j &= \frac{exp((W_3 l_t + b_3)^T (W_4 f_i^j + b_4))}{\sum_{n=1}^{C} exp((W_3 l_t + b_3)^T (W_4 f_n^j + b_4))}, & (3)
\end{aligned}
$$

where $\text{MLP}_1^j(\cdot)$ are multilayer perceptrons consisting of several linear and a LeakyReLU layer (Maas et al. 2013), $f_i^j \in \mathbb{R}^{C_f \times 1}$, W_3 and $W_4 \in \mathbb{R}^{C_m \times C_f}$, b_3 and $b_4 \in \mathbb{R}^{C_m \times 1}$ are learnable parameters, $j \in \{1, 2, \cdots, r\}$, r is a hyperparameter which sets 3. Each attention weight $m_{t,i}^j$ represents the relevance score between t-th word and i-th instance in the scene. We then use the attention map combined with words' entity weights $\{w_t^{ent}\}_{t=1}^T$ and attribute weights $\{w_t^{attr}\}_{t=1}^T$ to compute the probability $\alpha_{t,i}^j$, indicating the word l_t is the entity-related word refers to the instance I_i:

$$
\alpha_{t,i}^j = (w_t^{ent} + w_t^{attr}) \frac{exp(m_{t,i}^j)}{\sum_{n=1}^{C} exp(m_{t,n}^j)}. \quad (4)
$$

After obtaining the entity-related matching degree between words and instance features, the entity description information of the entire sentence for each instance β_i can be aggregated. Specifically, we adopt a simplified bilinear fusion strategy (Ben-younes et al. 2017), combining semantic information with the corresponding instance feature f_i to obtain the cross-modal features μ_i:

$$
\begin{aligned}
\beta_i^j &= \sum_{t=1}^{T} \alpha_{t,i}^j l_t, & (5)\\
\mu_i &= \sum_{j=1}^{r} (f_i^j \odot \beta_i^j), & (6)
\end{aligned}
$$

where $\odot$ denotes element-wise product and $\mu_i \in \mathbb{R}^{C_l \times 1}$.

Directed Scene Graph Relationship Construction. Exploring relative spatial relations among instances within the scene is necessary for referring segmentation. Not only because spatial information commonly appears in referring expressions, but also because spatial relationships constitute a crucial aspect of visual relationships within scenes (Feng et al. 2021). To fully explore the spatial relationship of the instances in the scene, we firstly construct the cross-modal graph $G = (V, E, U)$, where $V = \{v_i\}_{i=1}^C$ is the set of vertices corresponding to the central coordinates of instance masks, $I = \{I_1, I_2, \cdots, I_C\}$; $E = \{e_{ij}\}_{i,j=1}^C$ is the set of edges, which represent the type of spatial relationship between different instances; $U = \{u_i\}_{i=1}^C$ represents the fused cross-modal features. Specifically, to get the type of edge, we divide the three-dimensional space into 8 quadrants ranging from I to VIII, and the type of edge is determined by computing the relative position vector $(x_i - x_j, y_i - y_j, z_i - z_j)$ based on the instance masks' central coordinates, where $o_i = (x_i, y_i, z_i)$ and $o_j = (x_j, y_j, z_j)$ denote the i-th and j-th instance mask's central coordinate. Then the edge type can correspond to spatial relations like "left", "right", "in front of" and "behind" according to different directions. Considering the relative distance also matters in spatial relationships, we first compute the longest distance between instance centers in the scene, and then use it to normalize the distance between all instances. If the normalized distance ratio is larger than 0.25, we take the edge type as "No Relation", which is the ninth spatial relation.

While the model has identified the type of each edge, it is important to note that spatial relations can vary across different referring expressions. Therefore, it is logical to devise specific gates for different types of edges, enabling the capture of spatial relationships in referring expressions. The possibility for each word's edge type is computed as follows:

$$
q_t = w^{rel}(W_7 \sigma(W_6 l_t + b_6) + b_7), \quad (7)
$$

where $W_6 \in \mathbb{R}^{C_m \times C_l}, b_6 \in \mathbb{R}^{C_m \times 1}, W_7 \in \mathbb{R}^{N \times C_m}$ and $b_7 \in \mathbb{R}^{N \times 1}$ are learnable parameters, $\sigma(\cdot)$ is LeakyReLU function, and N is the number of edge types (*i.e.,* 9), which represents the number of edge types. $q_{t,s}$ is the s-th element of q_t, which represents the probability of the word l_t relates to the edge type s. To determine the likelihood that the expression corresponds to a specific type of edge, our model sums up the probability of words associated with each edge, which is formulated as follows:

$$
p_s = \sum_{t=1}^{T} q_{t,s}. \quad (8)
$$

So the cross-modal scene graph is expanded as $\hat{G} = (V, E, U, P)$, where $P = \{p_s\}_{s=1}^N$ is the weighted edge gate.

Relation-Driven Interaction

Combining expressions related to relative position with factual relationships in scenes is pivotal for identifying the target instance. For example, in the case of referring *"it is a red ottoman. the ottoman is at the right corner edge of the couch"*, there are several ottomans in the scene. By extracting the relative position relationship (*"at the right"*) with other objects (*"couch"*) from the referring expression and corresponding it to the position in the 3D scene, the target ottoman can stand out from other candidates. It is understandable that the objects utilized as auxiliary references are neighbors of the referred instance. Distant objects, on the other hand, offer little assistance in positioning and may even cause interference. So we further introduce the Relation-Driven Interaction (RDI) module, which employs a K-nearest algorithm to carry out local perception. This module facilitates the aggregation of positional information within expressions and scenes, contributing to better semantic modeling and finer cross-modal fusion.

Textual Relation Interaction. To fully leverage positional-related information in the referring expression, we fuse the positional text features with previously obtained instances' features, similar to the entity description feature fusion:

$$
m_{t,i}^e = \frac{exp((W_3 l_t + b_3)^T (W_4 \mu_i + b_4))}{\sum_{n=1}^{C} exp((W_3 l_t + b_3)^T (W_4 \mu_n + b_4))}, \quad (9)
$$

$$\gamma_{t,i} = w_t^{rel}\frac{exp(m_{t,i}^e)}{\sum_{n=1}^{C} exp(m_{t,n}^e)}, \quad (10)$$

$$\beta_i^e = \sum_{t=1}^{T}\gamma_{t,i}l_t, \quad (11)$$

$$\nu_i = \mathrm{MLP}_2(f_i^e \odot \beta_i^e), \quad (12)$$

where $m_{t,i}^e$ represents the relevance score between t-th word and i-th instance in the scene, $\gamma_{t,i}$ is the probability that the word l_t is positional related to the instance I_i. β_i^e represents the sentence's positional information for instance I_i, and $\nu_i \in \mathbb{R}^{C_n\times 1}$ is the location-dependent cross-modal feature for instance I_i. Then weighted edge gate p is applied to help aggregate the nearby positional features, enabling the model to exploit the expression's positional relationships in different directions:

$$\hat{\nu}_i = \sum_{k=1}^{K} p_i^k \nu_i^k, \quad (13)$$

where $\hat{\nu}_i \in \mathbb{R}^{C_n\times 1}$ represents i-th instance's refined positional feature, p_i^k represents the weighted edge gate between i-th instance and its k-th neighbour, ν_i^k represents its k-th neighbour's positional related feature. A textual relation graph G_t is then obtained through the above operations, whose vertices contain rich location information extracted from the expression.

Spatial Relation Interaction. The relative position of objects in the scene plays a crucial role in accurately identifying the target instance. So it is not just the referred instance itself that matters, auxiliary references provided by other instances within its nearby space also hold significant importance. Consequently, the key to achieving precise localization lies in effectively extracting each instance's spatial relations. Inspired by TGNN (Huang et al. 2022), we find relative position vectors are rich in positional information, which can help instances perceive their relative positional relations in the scene. So we encode the relative position vector between i-th instance and its k-th neighbor as follows:

$$r_i^k = \mathrm{MLP}_3\left([o_i; o_i^k; (o_i - o_i^k); \|o_i - o_i^k\|]\right), \quad (14)$$

where $r_i^k \in \mathbb{R}^{C_n\times 1}$, o_i^k represents the k-th neighbor's center coordinate of the i-th instance, the notation $[;]$ denotes the concatenation operation, and $\|\cdot\|$ is the Euclidean distance. Subsequently, the spatial position information of entities is fused through aggregation with K neighbours. The process is accomplished by element-wise products with their corresponding relative position vectors as follows:

$$\widetilde{\mu}_i = \mathrm{MLP}_4(\mu_i), \quad (15)$$

$$\hat{\mu}_i = \sum_{k=1}^{K}(r_i^k \odot \widetilde{\mu}_i^k), \quad (16)$$

where $\hat{\mu}_i$ represents the refined i-th instance's cross-modal feature, $\widetilde{\mu}_i^k \in \mathbb{R}^{C_n\times 1}$ represents the k-th neighbour's feature of the i-th instance, and r_i^k represents the corresponding relative position vector. And then the spatial relation graph G_s is obtained.

Finally, we get the refined cross-modal feature x_i by combining the features of two relation graphs' corresponding vertices, which is rich in context information with high-level semantics:

$$x_i = \mathrm{LayerNorm}(\hat{\nu}_i + \hat{\mu}_i + \nu_i + \widetilde{\mu}_i). \quad (17)$$

Matching Module

Given the refined cross-modal features, our model takes two complementary objectives to predict the matching score between instances and corresponding expressions, and the final score is obtained by combining the two predicted scores. Specifically, although the model obtains fine-grained cross-modal features in earlier stages, it overlooks the significance of global textual features. These global textual features contain rich contextual information and are vital for correct matching. Therefore we obtain the first matching score by calculating the cosine similarity between the instances' refined features and the global context feature. We use max-pooling strategy to get the global textual feature $g \in \mathbb{R}^{C_l\times 1}$:

$$g = \mathrm{MaxPool}(l_t), \quad t = \{1, 2, \cdots, T\}. \quad (18)$$

Then the cosine similarity score is computed as follows:

$$s_i^{cos} = \mathrm{L2Norm}(W_8 x_i) \odot \mathrm{L2Norm}(W_9 g), \quad (19)$$

where $W_8 \in \mathbb{R}^{C_l\times C_n}$ and $W_9 \in \mathbb{R}^{C_l\times C_l}$ are transformation matrices, $\mathrm{L2Norm}(\cdot)$ is the L2 normalization. Additionally, the refined cross-modal feature itself contains rich information, which refers to the target instance. So another score is obtained by directly passing the refined features into a fully-connected layer:

$$s_i^{pdt} = W_{10}x_i + b_{10}, \quad (20)$$

where $W_{10} \in \mathbb{R}^{1\times C_n}, b_{10} \in \mathbb{R}^{1\times 1}$ are learnable parameters.

The final score S_i is obtained by simply adding together the cosine similarity score and the instance prediction score, which is then used to predict the target instance:

$$S_i = s_i^{cos} + s_i^{pdt}. \quad (21)$$

Loss Function. The training loss is a linear combination of the cosine loss $\mathcal{L}_{cos}$ and the prediction loss $\mathcal{L}_{pdt}$:

$$\mathcal{L} = \mathcal{L}_{cos} + \mathcal{L}_{pdt}. \quad (22)$$

Specifically, the loss functions $\mathcal{L}_{cos}$ and $\mathcal{L}_{pdt}$ are computed as cross-entropy losses based on the cosine similarity score and the instance prediction score, respectively.

Experiments

Dataset

Following TGNN, we conduct extensive experiments on the ScanRefer dataset, which is built on top of the ScanNet dataset. ScanNet is an RGB-D video dataset containing 2.5 million views in more than 1,500 scans, annotated using 3D camera poses, surface reconstruction, and instance-level semantic segmentation.

Method	mIOU(%)	Acc@0.25	Acc@0.5	Memory Usage(GB)
TGNN(GRU)	26.10	35.00	29.00	22.69
Ours(GRU)	**29.77**	**39.85**	**33.52**	**10.43**
TGNN(BERT)	27.80	37.50	31.40	37.09
Ours(BERT)	**29.94**	**40.33**	**33.77**	**22.29**

Table 1: Comparison with state-of-the-art on ScanRefer.

	EAF	TRI	SRI	mIOU(%)	Acc@0.25	Acc@0.5
(a)				24.66	33.02	27.64
(b)	✓			26.78	35.84	30.29
(c)	✓	✓		28.65	38.28	32.50
(d)	✓		✓	29.19	39.25	32.65
(e)	✓	✓	✓	**29.77**	**39.85**	**33.52**

Table 2: The results obtained after ablating different network modules on the ScanRefer validation set. EAF represents the Entity-Aware Fusion module, TRI represents the Textual Relation Interaction module and SRI represents the Spatial Relation Interaction module.

	mIOU(%)	Acc@0.25	Acc@0.5
w/o edge gate	29.39	39.39	32.97
w edge gate	**29.77**	**39.85**	**33.52**

Table 3: Ablation study on the weighted edge gate.

	$+\mathcal{L}_{cos}$	$+\mathcal{L}_{ref}$	mIOU(%)	Acc@0.25	Acc@0.5
(a)	✓		28.52	38.24	32.11
(b)		✓	28.24	38.18	31.64
(c)	✓	✓	**29.77**	**39.85**	**33.52**

Table 4: Loss ablation on the ScanRefer dataset.

ScanRefer. ScanRefer comprises 51,583 descriptions of 11,046 3D objects from 800 real-world scenes captured using the ScanNet dataset. This dataset is groundbreaking as it is the first large-scale collection to facilitate 3D object grounding in point clouds through complex and diverse natural language descriptions. Each scene contains 13.81 objects and 64.48 descriptions on average, while each object is associated with 4.67 descriptions, providing rich and varied data for object-reference association tasks.

Implementation Details

For 3D instance segmentation, we adopt the pre-trained 3D UNet feature extractor proposed in TGNN, whose parameters are fixed during the training and testing. In our experiments utilizing GRU as the language extractor, we employ the ADAM optimizer with an initial learning rate of 1e-3 and adopt CosineAnnealingLR as the learning rate decay strategy. The training process consisted of 48 rounds with a batch size of 8, the total training time is around 8 hours. For experiments utilizing BERT as the textual extractor, we adopt the same optimizer and learning rate decay strategy with the GRU mode. However, we update the parameters of the BERT model and our model separately. The initial learning rate is set to 1e-5 for BERT and 1e-3 for our model. Both BERT and our proposed model are trained for 64 epochs with a batch size of 16, the total training and validation time is around 11 hours. All experiments are implemented on PyTorch and a single 24-GB NVIDIA RTX-3090 GPU.

Quantitative Comparisons

To the best of our knowledge, this is the second work addressing the task of referring 3D instance segmentation. We conduct a comparison with the first work, TGNN, utilizing the evaluation metrics proposed earlier: mean IOU and Acc@kIOU. As shown in Table 1, our method exhibits a significant performance boost, achieving an impressive increase of **3.67% mIOU** with GRU and **2.14% mIOU** with BERT over TGNN. Notably, our approach demonstrates substantial improvements, with **4.85%** improvement in Acc@0.25 and **4.52%** improvement in Acc@0.5 for the GRU mode, and **2.83%** improvement in Acc@0.25 and **2.37%** improvement in Acc@0.5 for the BERT mode. Moreover, our method not only enhances performance but also achieves a substantial reduction of over ten gigabytes in memory usage during the training process, owing to our more efficient feature fusion method.

Ablation Study

Dense Components Ablation. To investigate the impacts of various modules, we conducted extensive ablation experiments on the ScanRefer dataset, all experiments were conducted using GRU as the language encoder, and the results are displayed in Table 2. The analyses of experiments are as follows: **i)** Following TGNN, we established a baseline model in (a) by simply concatenating the instance features with the processed language features, and then directly using the fusion feature vector to predict the matching score. **ii)** Dense-aligned sub-methods (b)-(e) outperform the baseline (a), demonstrating the validity of the three modules. **iii)** Specifically, (b) demonstrates the necessity of fusing entity-related textual features with instances' visual features for locating the referred instance. **iv)** (c)-(d) conduct the high-order semantic modelling through the interaction with nearby instances, which effectively integrate the features to align with the referring expression. Notably, the better performance of (d) relative to (c) suggests that entity description words within the expression combined with spatial relationship contain more contextual information than those positional words. **v)** (e) integrates all components and achieves the highest performance gains, indicating the efficacy of entity-related feature extraction and the synergy between textual and spatial positional relationships. The above experimental results demonstrate the effectiveness of our method.

We also conducted an ablation experiment on the weighted edge gate to explore its effectiveness. As depicted in Table 3, the disappearance of the weighted edge gate leads to a mIOU reduction of around **0.4%**. This illustrates the weighted edge gate's contribution in better incorporating details of relative positional relationships within expressions.

Loss Ablation. We also conduct ablation experiments on cosine similarity loss and direct referring loss, results are shown in Table 4. Both separate losses yield comparable performance, with the cosine loss slightly outperforming the

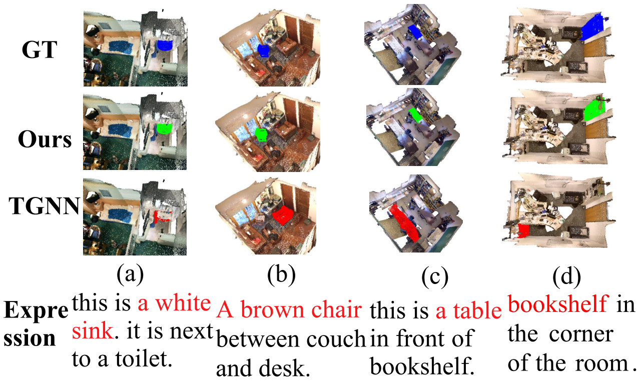
GT
Ours
TGNN
(a)
(b)
(c)
(d)
Expression
this is a white sink. it is next to a toilet.
A brown chair between couch and desk.
this is a table in front of bookshelf.
bookshelf in the corner of the room.

Figure 3: Qualitative results from TGNN and our model.

N	mIOU(%)	Acc@0.25	Acc@0.5
1	**29.77**	**39.85**	**33.52**
2	28.85	38.65	32.74
3	28.00	37.40	31.73
4	27.43	36.66	30.68

Table 5: Ablation study on the number of inference layers in the RDI module.

K	mIOU(%)	Acc@0.25	Acc@0.5
1	28.22	37.67	31.78
5	29.20	39.19	32.94
10	**29.77**	**39.85**	**33.52**
15	29.05	39.08	32.79

Table 6: Ablation study on the number of nearby objects during local perception.

other. However, there is still a large gap compared with the best performance. Upon combining both losses for supervision, the model demonstrates a significant improvement in mIOU and accuracy, suggesting the two loss functions are mutually reinforcing.

Inference Layer Ablation. We also explore the influence of the number of inference layers used in our X-RefSeg3D model. As Table 5 shows, the performance decreases gradually with the increase of Relation-Driven Interaction (RDI) module's inference layers. One plausible explanation is that the majority of expressions primarily involve first-order relationships, whereas the occurrence of second-order, third-order, or more complex associations gradually decreases.

K Neighbours Ablation. The results in Table 6 demonstrate an initial increase and subsequent decrease in model performance as K-neighbors increase. This trend indicates that excessive or inadequate fusion of local features may hinder model inference. A limited perception range might cause the model to overlook crucial reference information from the surroundings. Conversely, an overly broad perception range could introduce interfering noise and adversely affect performance. Consequently, we selected K=10 as the optimal configuration for our model.

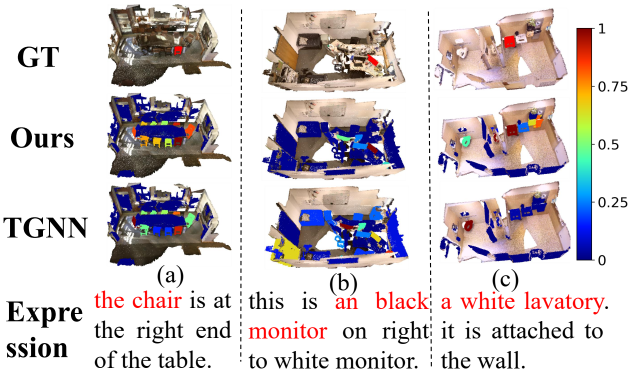
GT
Ours
TGNN
1
0.75
0.5
0.25
0
(a)
(b)
(c)
Expression
the chair is at the right end of the table.
this is an black monitor on right to white monitor.
a white lavatory. it is attached to the wall.

Figure 4: Visualization of affinity maps between instances and expressions in our model and TGNN.

Visualization. We visualize the referring segment results and the affinity maps between instances and expressions in Figure 3 and Figure 4. Figure 3 shows four qualitative referring segment results produced by the TGNN method and our method on the ScanRefer dataset. As illustrated in Figure 3 (a-b), compared with TGNN, our method has a superior perception of the referred entity, thereby reducing the occurrence of misidentification among other object categories. Figure 3 (c-d) indicates that our method can more accurately extract relative positional relationships within 3D scenes and expressions, enabling the exact identification of the referred instance among several candidates of the same class.

In Figure 4 (a-b), compared with TGNN, our method enables a finer attention distribution among instances of the same type located in different locations. Figure 4 (c) shows that our method can accurately distinguish instances whose categories are incorrectly identified by TGNN. These improvements are made possible by our method's accurate perception of the entity and relative positional relationship within the scene and expression.

Conclusions

In this paper, we introduce X-RefSeg3D, a novel end-to-end model for referring 3D instance segmentation. Our method achieves accurate identification and segmentation via reasonable high-level semantic modelling and fine-grained cross-modal feature fusion. Specifically, X-RefSeg3D utilizes a scene graph and structured graph neural networks to facilitate entity-related information extraction and relative position relationship correspondence. This refinement notably contributes to a substantial enhancement in segmentation performance for the referred expressions. Extensive experiments and ablation studies validate the effectiveness of our approach and demonstrate the superiority of each module in the X-RefSeg3D model. In the future, we will continue to explore and improve our method, including investigating alternative visual feature extraction techniques. With the rapid development of 3D vision, we believe our approach will contribute significantly to advancing this field.

Acknowledgements

This work was supported by National Key R&D Program of China (No.2023YFB4502804), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), the National Natural Science Fund for Young Scholars of China (No. 62302411), China Postdoctoral Science Foundation (No. 2023M732948), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

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*, 422–440.

Ben-younes, H.; Cadene, R.; Cord, M.; and Thome, N. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *2017 IEEE International Conference on Computer Vision (ICCV)*.

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*, 202–221.

Cho, K.; van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Niessner, M. 2017. ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Deng, J.; Shi, S.; Li, P.; Zhou, W.; Zhang, Y.; and Li, H. 2022. Voxel R-CNN: Towards High Performance Voxel-based 3D Object Detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 1201–1209.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North*.

Elich, C.; Engelmann, F.; Kontogianni, T.; and Leibe, B. 2019. *3D-BEVIS: Birds-Eye-View Instance Segmentation.*, 48–61.

Fei, H.; Liu, Q.; Zhang, M.; Zhang, M.; and Chua, T.-S. 2023. Scene Graph as Pivoting: Inference-time Image-free Unsupervised Multimodal Machine Translation with Visual Scene Hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5980–5994.

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. *Cornell University - arXiv,Cornell University - arXiv*.

Hou, J.; Dai, A.; and NieBner, M. 2019. 3D-SIS: 3D Semantic Instance Segmentation of RGB-D Scans. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Huang, L.; Wang, H.; Zeng, J.; Zhang, S.; Cao, L.; Ji, R.; Yan, J.; and Li, H. 2023. Geometric-aware Pretraining for Vision-centric 3D Object Detection.

Huang, P.-H.; Lee, H.-H.; Chen, H.-T.; and Liu, T.-L. 2022. Text-Guided Graph Neural Networks for Referring 3D Instance Segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 1610–1618.

Huang, S.; Hui, T.; Liu, S.; Li, G.; Wei, Y.; Han, J.; Liu, L.; and Li, B. 2020. Referring Image Segmentation via Cross-Modal Progressive Comprehension. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Jean, L.; Bernard, G.; Martin, R.; and Marc, P. 2019. 3D Instance Segmentation via Multi-Task Metric Learning. *Cornell University - arXiv*.

Ji, J.; Huang, X.; Sun, X.; Zhou, Y.; Luo, G.; Cao, L.; Liu, J.; Shao, L.; and Ji, R. 2022a. Multi-branch distance-sensitive self-attention network for image captioning. *IEEE Transactions on Multimedia*.

Ji, J.; Ma, Y.; Sun, X.; Zhou, Y.; Wu, Y.; and Ji, R. 2022b. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31: 4321–4335.

Ji, J.; Sun, X.; Zhou, Y.; Ji, R.; Chen, F.; Liu, J.; and Tian, Q. 2020. Attacking image captioning towards accuracy-preserving target words removal. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4226–4234.

Lee, K.; Zhao, W.; Simchi-Levi, D.; Sung, M.; and Guibas, L. 2018. GSPN: Generative Shape Proposal Network for 3D Instance Segmentation in Point Cloud. *Cornell University - arXiv*.

Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; and Ji, R. 2022. X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*, 638–647.

Ma, Y.; Zhang, X.; Sun, X.; Ji, J.; Wang, H.; Jiang, G.; Zhuang, W.; and Ji, R. 2023. X-Mesh: Towards Fast and Accurate Text-driven 3D Stylization via Dynamic Textual Guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2749–2760.

Maas, A. L.; Hannun, A. Y.; Ng, A. Y.; et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, 3. Atlanta, GA.

Peng, S.; Genova, K.; Jiang, C.; Tagliasacchi, A.; Pollefeys, M.; and Funkhouser, T. 2022. OpenScene: 3D Scene Understanding with Open Vocabularies.

Pennington, J.; Socher, R.; and Manning, C. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Retinskiy, D. 2019. Submanifold Sparse Convolutional Networks. In *Submissions to the 2019 Kidney Tumor Segmentation Challenge: KiTS19*.

Schult, J.; Engelmann, F.; Hermans, A.; Litany, O.; Tang, S.; and Leibe, B. 2022. Mask3D for 3D Semantic Instance Segmentation.

Wang, H.; Tang, J.; Ji, J.; Sun, X.; Zhang, R.; Ma, Y.; Zhao, M.; Li, L.; Zhao, Z.; Lv, T.; et al. 2023. Beyond first impressions: Integrating joint multi-modal cues for comprehensive 3d representation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 3403–3414.

Wu, H.; Wen, C.; Li, W.; Li, X.; Yang, R.; and Wang, C. 2022a. Transformation-Equivariant 3D Object Detection for Autonomous Driving.

Wu, S.; Fei, H.; Cao, Y.; Bing, L.; and Chua, T.-S. 2023. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14734–14751.

Wu, Y.; Cheng, X.; Zhang, R.; Cheng, Z.; and Zhang, J. 2022b. EDA: Explicit Text-Decoupling and Dense Alignment for 3D Visual Grounding.

Yang, B.; Wang, J.-a.; Clark, R.; Hu, Q.; Wang, S.; Markham, A.; and Trigoni, N. 2019. Learning Object Bounding Boxes for 3D Instance Segmentation on Point Clouds. *arXiv: Computer Vision and Pattern Recognition*.

Yang, D.; Ji, J.; Sun, X.; Wang, H.; Li, Y.; Ma, Y.; and Ji, R. 2023. Semi-Supervised Panoptic Narrative Grounding. In *Proceedings of the 31st ACM International Conference on Multimedia*.

Yang, S.; Li, G.; and Yu, Y. 2019. Cross-Modal Relationship Inference for Grounding Referring Expressions. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, S.; Li, G.; and Yu, Y. 2020. Graph-Structured Referring Expression Reasoning in The Wild. *arXiv: Computer Vision and Pattern Recognition,arXiv: Computer Vision and Pattern Recognition*.

Yu, L.; Lin, Z.; Shen, X.; Yang, J.; Lu, X.; Bansal, M.; and Berg, T. L. 2018. MAttNet: Modular Attention Network for Referring Expression Comprehension. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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 *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Zhao, L.; Cai, D.; Sheng, L.; and Xu, D. 2021. 3DVG-Transformer: Relation Modeling for Visual Grounding on Point Clouds. *International Conference on Computer Vision,International Conference on Computer Vision*.

Zhao, Y.; Fei, H.; Ji, W.; Wei, J.; Zhang, M.; Zhang, M.; and Chua, T.-S. 2023. Generating Visual Spatial Description via Holistic 3D Scene Understanding. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7960–7977.

Zheng, W.; Tang, W.; Chen, S.; Jiang, L.; and Fu, C.-W. 2022. CIA-SSD: Confident IoU-Aware Single-Stage Object Detector From Point Cloud. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3555–3562.