LESS: Label-Efficient and Single-Stage Referring 3D Segmentation

Xuexun Liu[1]*, **Xiaoxu Xu**[1]*, **Jinlong Li**[2]*, **Qiudan Zhang**[1], **Xu Wang**[1]†, **Nicu Sebe**[2], **Lin Ma**[3]
[1]College of Computer Science and Software Engineering, Shenzhen University,
Shenzhen, 518060, China.
[2]University of Trento, Italy.
[3]Meituan Inc., China.

Abstract

Referring 3D Segmentation is a visual-language task that segments all points of the specified object from a 3D point cloud described by a sentence of query. Previous works perform a two-stage paradigm, first conducting language-agnostic instance segmentation then matching with given text query. However, the semantic concepts from text query and visual cues are separately interacted during the training, and both instance and semantic labels for each object are required, which is time consuming and human-labor intensive. To mitigate these issues, we propose a novel Referring 3D Segmentation pipeline, **L**abel-**E**fficient and **S**ingle-**S**tage, dubbed **LESS**, which is only under the supervision of efficient binary mask. Specifically, we design a Point-Word Cross-Modal Alignment module for aligning the fine-grained features of points and textual embedding. Query Mask Predictor module and Query-Sentence Alignment module are introduced for coarse-grained alignment between masks and query. Furthermore, we propose an area regularization loss, which coarsely reduces irrelevant background predictions on a large scale. Besides, a point-to-point contrastive loss is proposed concentrating on distinguishing points with subtly similar features. Through extensive experiments, we achieve state-of-the-art performance on ScanRefer dataset by surpassing the previous methods about 3.7% mIoU using only binary labels. Code is available at `https://github.com/mellody11/LESS`.

1 Introduction

Referring 3D Segmentation task aims to segment the specific object from a 3D point cloud scene with a free-form natural language expression. It allows users to interact and analyze 3D data through verbal instructions or queries. This approach is particularly beneficial in applications that necessitate direct interaction with 3D environments, such as augmented reality (AR) systems, embodied-AI, robotics, virtual reality (VR) environments, and in fields like architecture and medical imaging, where precise identification and segmentation of objects based on descriptive queries can significantly enhance user experience and operational efficiency.

Previous Referring 3D Segmentation [13; 29] mainly leverage a two-stage workaround, as shown in Fig.1 (a). They typically adopt a 3D instance semantic segmentation network to get the instance proposals at the first stage. The predicted instance proposals will be utilized to match with the queries and finally get the final prediction mask according to the matching score. Although those method have achieved remarkable performance, there still exist some problems. First, owing to the large-scale and

*Equal contributions.
†Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0356

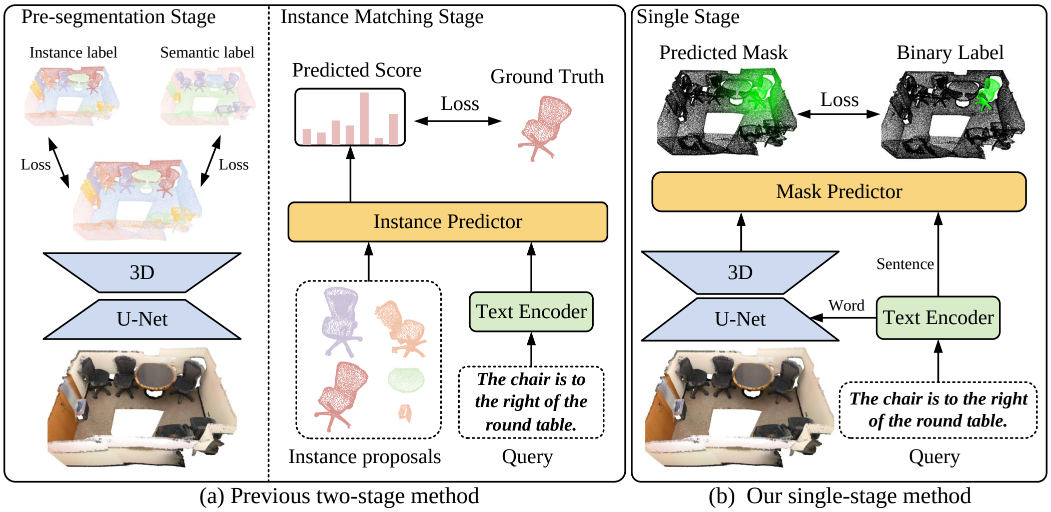

Figure 1: Comparison between the two-stage method and our single-stage method. (a) The two-stage method initially performs instance segmentation with instance labels then semantic labels to get the instance proposals and bases on the provided query to match the most relevant instance proposal. (b) Our single-stage method only utilizes the binary mask of the described object for training and integrates language and vision features during feature extraction.

irregular 3D point clouds, some instance proposals may leave out the target in the pre-segmentation stage. Besides, lacking of linguistic guidance in the segmentation stage fails to focus on the objects that are more essential to the referring task. Moreover, existing Referring 3D Segmentation utilizes both instance labels and semantic labels to segment target proposal rather than binary mask used in referring image segmentation, which is more time consuming and labor intensive.

To address the aforementioned problems, we propose a **L**abel-**E**fficient and **S**ingle-**S**tage Referring 3D Segmentation method, namely **LESS**, which is under the supervision of binary mask, as shown in Fig.1 (b). We first process the query with text encoder to get the word-level feature and sentence-level feature. Then we extract the multi-modal feature with the guidance of text feature through a 3D sparse U-Net. Finally, the mask predictor aligns multi-modal features with language features, and directly predicts the mask of the described object. Here only object mask serves as the label to supervise the whole training procedure.

However, 3D point cloud inherently provide a higher level of complexity and a large scale. There exists numerous different objects in a single 3D scene compared to the referring image segmentation task. Besides, binary mask has less semantic meanings compared to instance labels and semantic labels. These challenges make it difficult to supervise our model to localize and segment target objects with only binary mask. Therefore, we propose to alleviate these problems by some ways, as shown in Fig.2. Firstly, to facilitate fine-grained alignment between points and words, we propose Point-Word Cross-Modal Alignment module. The PWCA module utilizes cross-modal attention in the multi-modal context to align textual features extracted by the text encoder with point cloud features, followed by further extraction of useful multi-modal information using robust 3D sparse convolutional layers. Thus we can extract a more semantic meaning fused feature. Meanwhile, we employ Query Mask Predictor, which utilizes the extracted multi-modal features to decode learnable query embeddings, generating candidate masks. By introducing the Query-Sentence Alignment modules, we can compute similarity between the decoded query embeddings and sentence features, using the similarity as weights to perform weighted summation of the candidate masks to produce the final mask. To address the significant interference caused by multiple objects and backgrounds, we propose an area regularization loss and a point-to-point contrastive loss. Area regularization loss reduces irrelevant background predictions by constraining the probabilities of the predicted mask, while point-to-point contrastive loss constrains the distances between positive and negative points in the latent space to achieve better segmentation.

To summarize, our contributions are as follows:

https://doi.org/10.52202/079017-0356

- We propose a new Referring 3D Segmentation method LESS, which directly performs Referring 3D Segmentation at a single stage to bridge the gap between detection and matching under the supervision of binary mask. To the best of our knowledge, LESS is the first work investigating label-efficient and single-stage in Referring 3D Segmentation task.
- Our LESS utilize a Point-Word Cross-Modal Alignment module to align fined-grained point and word features. Besides, we employed Query Mask Predictor module and Query-Sentence Alignment modules for coarse-grained alignment between masks and sentences. Moreover, the area regularization loss and the point-to-point contrastive loss are introduced to better support to eliminate interference caused by multiple objects and backgrounds.
- Extensive experiments confirm the effectiveness of our method. Our method outperforms the existing state-of-the-art method on ScanRefer dataset with only the supervision of binary mask. Our LESS and its results provide valuable insights to improve further research of label-efficient and single-stage Referring 3D Segmentation.

2 Related Works

2.1 Referring 3D Segmentation

Referring 3D Segmentation has previously received limited exploration, however, with the advancements in multi-modal learning and embodied AI, it is set to attract increasing interest in the future. TGNN [13] is the first to introduce Referring 3D Segmentation task. They initially trained an instance segmentation network, followed by a Graph Neural Network (GNN) to learn features of instances and their relationships guided by linguistic information. Building on TGNN, X-RefSeg [29] developed a cross-modal graph. They employed an GNN to model the texture and spatial relationships of instances, and refining the results through inference and matching processes.

2.2 3D Visual Grounding

3D visual grounding aims to locate the object within point clouds mentioned by free-form natural language descriptions. Most methods follow a two-stage detection-then-matching pipeline. Initially, they utilize a pre-trained 3D detector [28; 23] or segmenter [15; 33] to extract object representations. Subsequently, these methods align text features with object features to identify the best-matched object. Researchers primarily concentrate on the second stage which involves modeling object relationships and exploring feature fusion between language and objects. Methods employed include multi-modal feature concatenation [3; 1], attention-based multi-modal feature alignment [44; 11], graph neural network-based reasoning [13; 41], and the alignment of visual and language features aided by 2D images [40; 42; 38; 37].Other researchers have investigated single-stage approaches for 3D visual grounding. 3D-SPS [25] views the task as key point selection, progressively identifying keypoints with the guidance of language and directly locates the target. BUTD-DETR [14] employs a transformer decoder [2] to identify described objects using language cues and proposal boxes. Building on this, EDA [36] enhances dense alignment between objects and point clouds by explicitly decoupling textual attributes from sentences.

The distinction between 3D visual grounding and Referring 3D Segmentation lies in the latter's enhanced localization precision, offering significant value in applications like robotic grasping. While traditional single-stage 3D visual grounding methods regress a 3D bounding box, our single-stage approach decodes a binary mask for the entire point cloud scene, presenting a more complex challenge.

2.3 Referring Image Segmentation

Referring image segmentation is a visual task that involves pixel-level segmentation of an image's target object based on a referring expression. Early approaches [12; 20; 18; 4; 17] utilize CNNs and RNNs to extract features and then perform simple concatenation for multi-modal feature fusion. Subsequent research [16; 19; 9] adopt transformer models for more effective feature extraction and fusion. Recent studies [39; 10; 43] focus on identifying optimal positions for language-vision alignment. Additionally, some methods [34; 5] have enhanced alignment between language and pixel features by designing specialized loss functions. Other approaches [21; 45] treat referring image

https://doi.org/10.52202/079017-0356

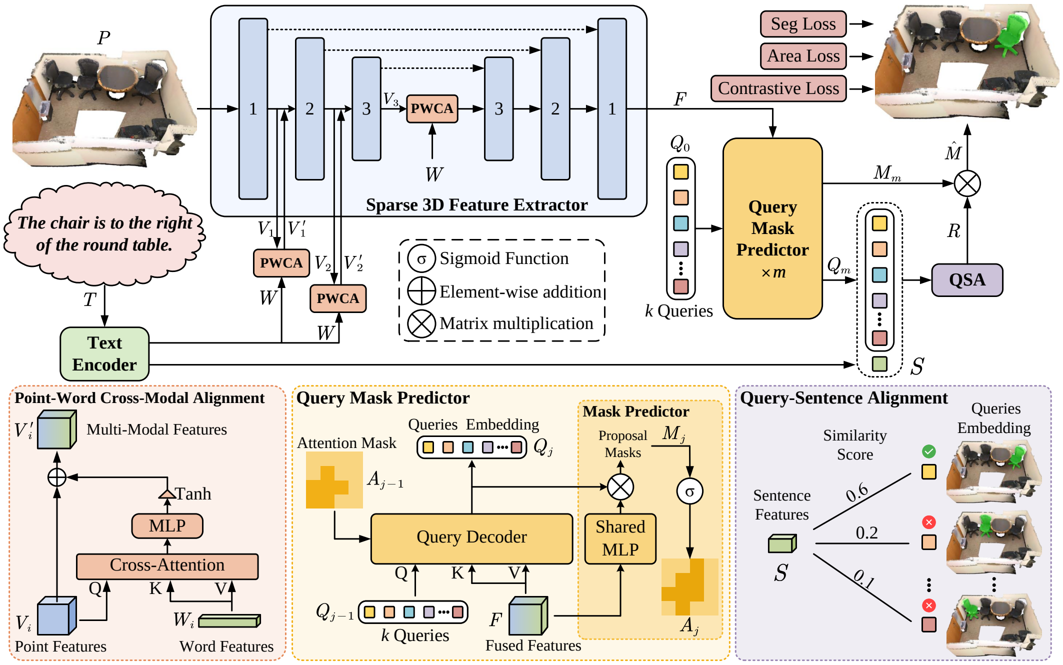

Figure 2: Overview of our **LESS** framework. Given a point cloud scene P, we use a sparse 3D feature extractor to extract multi-scale feature V_i. The query T is sent to a text encoder and we obtain the word features W and sentence features S. Meanwhile, we introduce a PWCA module aligns the word features W with the multi-scale point cloud features V_i. After that, an m-layer QMP module is adopted to decode K learnable queries Q_0 base on the fused feature F, and output query embeddings Q_m and proposal masks M_m. Finally, QSA module aligns the query embeddings Q_m with sentence features S, i.e., computes the similarity scores R that filter the proposal masks M_m to the final mask prediction $\hat{M}$.

segmentation as an auto-regressive vector generation problem, creating masks from generated closed vectors.

Methods for referring image segmentation have been extensively explored, yet they cannot be directly applied to Referring 3D Segmentation due to the inherent challenges of point cloud scenes. Additionally, unlike referring image segmentation methods that typically employ pre-trained visual encoders, Referring 3D Segmentation lacks such resources, compelling reliance on limited supervisory signals for model training.

3 Method

The overall framework of our proposed LESS is shown in Fig.2, which leverages Point-Word Cross-Modal Alignment and Query-Sentence Alignment to facilitate multi-modal interaction. In this section, we start by introducing our visual and text feature extractor in Sec.3.1. Then Query Mask Predictor and Query-Sentence Alignment is detailed in Sec.3.2 and Sec.3.3. Finally in Sec.3.4, we introduce our area regularization loss and point-to-point contrastive loss.

3.1 Visual and Text Feature Extractor

Sparse 3D Feature Extractor. The point cloud scene $P \in \mathbb{R}^{N\times 6}$ contains N points in the scene, and each point is represented with six dimensions of RGBXYZ attributes. We first voxelize points into regular voxels and adopt a sparse 3D U-Net [31; 15] to extract point-wise fused feature $F \in \mathbb{R}^{N\times C}$. Here the encoder part of the U-Net has 5 stages and the feature from the i-th encoder stage is denoted as V_i.

Text Encoder. Given the query sentence T with L words, a text encoder is used to embed the query into C-dimensional feature vectors. In this paper we choose GRU [6], BETR [8] and RoBERTa [22] as our text encoder respectively and fine-tune the BERT or RoBERTa during training. Finally, we can get both word features $W \in \mathbb{R}^{L\times C}$ and sentence features $S \in \mathbb{R}^{C}$ after the text encoder.

Point-Word Cross-Modal Alignment. Due to the lack of such rich annotations as [13; 29], it is crucial for our model to learn the relationship between fine-grained word-level features and point-level features in such a point-level segmentation task. Meanwhile, we notice that leveraging the rich convolutional layers in the encoder to excavate multi-modal context is effective way to extract language-aware visual feature. Therefore, we design Point-Word Cross-Modal Alignment (PWCA) module, as shown in the lower left part in Fig.2, which contains a standard cross-attention module [32] and an MLP-Tanh gate. Cross-attention module aligns point-wise and word-wise feature to get the language-aware visual feature. A nonlinear tanh gate is adopted to prevent the fused signal from overwhelming the original signal. Given the multi-scale point cloud features $\{V_i \in \mathbb{R}^{N_i\times C_i}\}_{i=1}^{5}$, we simply project the word feature $W \in \mathbb{R}^{L\times C}$ to $W_i \in \mathbb{R}^{L\times C_i}$. PWCA can be formulated as follow:

$$V_i' = \text{Tanh}(\text{MLP}(\text{CrossAttn}(V_i, W_i)) + V_i, \quad i \in \{1, ..., 5\}, \tag{1}$$

where i indicates the i-th stage of the encoder part of our sparse 3D feature extractor. Here we use V_i as the query and W_i as the key and value for cross attention.

3.2 Query Mask Predictor

Inspired by [31; 30], Query Mask Predictor (QMP) module, as shown in Fig.2, takes fused feature $F \in \mathbb{R}^{N\times C}$ and learnable queries $Q_0 \in \mathbb{R}^{K\times C}$ as input and progressively distinguishes the referring target by multi-layer cross-modal transformers. Finally, we extract the proposal masks $M_m \in \mathbb{R}^{K\times N}$ based on queries embeddings $Q_m \in \mathbb{R}^{K\times C}$ and fused feature F.

Query Decoder. As illustrated in the lower middle section of Fig.2. Here query decoder is comprised m-layer masked cross-attention modules [32], where the fused features F are served as keys and values and $A_{j-1} \in \{0,1\}^{K\times N}$ is utilized as the attention mask. Therefore, in the j-th layer, the learnable queries Q_{j-1} capture multi-modal contextual information from fused feature F via a query decoder to obtain the query embeddings $Q_j \in \mathbb{R}^{K\times C}$.

Mask Predictor. First the fused feature $F \in \mathbb{R}^{N\times C}$ is processed by a Shared MLP, which indicates each Query Mask Predictor layer shares the same MLP. We perform the matrix multiplication on the new F and $Q_j \in \mathbb{R}^{K\times C}$ to generate proposal mask predictions $M_j \in \mathbb{R}^{K\times N}$. After applying a sigmoid function whose threshold of 0.5, we can get the new binary attention mask $A_j \in \{0,1\}^{K\times N}$.

Finally, the Query Mask Predictor is formally described as follows:

$$Q_j = \text{Query Decoder}_j(Q_{j-1}, A_{j-1}, F), \tag{2}$$

$$M_j = Q_j \otimes \text{Shared-MLP}(F)^\top, \tag{3}$$

$$A_j = \begin{cases} 1 & \sigma(M_j) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}, \tag{4}$$

where $\top$ denotes the transpose operation, $\otimes$ represents matrix multiplication, and σ refers to the sigmoid function. Here we need to note that the Q_0 is randomly initialized. Therefore we can leverage the function (3-4) to initialize the attention mask A_0.

3.3 Query-Sentence Alignment

The Referring 3D Segmentation task involves the segmentation of a solitary target object according to the query. Previous modules focus on extracting language-aware visual feature from aligning point-wise feature and word-wise feature, which lacks a comprehensive perception of the entire query sentence. Therefore, we adopt Query-Sentence Alignment (QSA) to better align the query feature with sentence-level feature. We perform the matrix multiplication on $Q_m \in \mathbb{R}^{K\times C}$ and $S \in \mathbb{R}^{C}$ to get the their similarity score $R \in [0,1]^K$. Formally, Query-Sentence Alignment can be represented as:

$$R = \text{Softmax}(Q_m \otimes S). \tag{5}$$

https://doi.org/10.52202/079017-0356

The final mask prediction $\hat{M} \in \mathbb{R}^N$ is produced by weighted sum of similarity score R and proposed mask prediction $M_m \in \mathbb{R}^{K \times N}$. Finally, we use a sigmoid function and a threshold of 0.5 to produce the final predicted binary mask $\hat{Y} \in \{0, 1\}^N$:

$$\hat{M} = R \otimes M_m, \qquad \hat{Y} = \begin{cases} 1 & \sigma(\hat{M}) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

3.4 Loss Function

Segmentation Loss. Different from previous work [13; 29], we take the Referring 3D Segmentation task as segmentation task with only binary mask $Y \in \{0, 1\}^N$. Here we utilize the Binary Cross-Entropy (BCE) loss function to compute the segmentation loss, which can be formulated as:

$$\mathcal{L}_{\text{seg}} = \text{BCE}(\sigma(\hat{M}), Y). \tag{7}$$

Area Regularization Loss. For Referring 3D Segmentation task, each query always corresponds to one target object in the point cloud scene. The target objects occupy a smaller area in the large scale of 3D point cloud. As a result, the predicted mask often includes backgrounds or other objects. To address this, we propose a region regularization loss, which promotes the network to predict a smallest mask by minimize the output probability of each point, formulated as:

$$\mathcal{L}_{\text{area}} = \frac{1}{N} \sum_{i=1}^{N} \sigma(\hat{M}_i). \tag{8}$$

By combining with the segmentation loss, we intend to segment only the most probable regions while reducing segmentation of large-scale irrelevant background areas.

Point-to-Point Contrastive Loss. Area regularization loss uniformly penalizes the predicted probabilities of all points, which can reduce the majority of the background points. However, the network struggles to differentiate between objects that possess characteristics similar to those described target object in the latent space. Therefore, we propose a point-to-point contrastive loss [26] that pull the points from the described object together and push away the rest points:

$$\mathcal{L}_{\text{p2p}} = -\frac{1}{|\mathcal{P}|} \sum_{i=1}^{|\mathcal{P}|} \frac{\exp(\mathbf{P}_i \cdot \mathbf{P}_{\text{avg}} / \tau)}{\exp(\mathbf{P}_i \cdot \mathbf{P}_{\text{avg}} / \tau) + \sum_{j=1}^{|\mathcal{N}|} \exp(\mathbf{P}_i \cdot \mathbf{N}_j / \tau)}, \tag{9}$$

where $\mathcal{P}$ is the positive point set from the described object, $\mathcal{N}$ is the negative point set from the background, $\mathbf{P}_i$ denotes the L2-normalized feature vector of i -th positive points from F, while $\mathbf{N}_j$ denotes the L2-normalized feature vector of j-th negative points from F, $\mathbf{P}_{\text{avg}}$ is the average feature vector of positive point $\mathbf{P}_{\text{avg}} = \frac{1}{\mathcal{P}} \sum_{i=1}^{|\mathcal{P}|} \mathbf{P}_i$, and τ is the hyper-parameter. The contrastive loss promotes the network to distinguish the described object from the adjacent background points in a fined-grained manner.

The overall loss function is the weighted sum of the aforementioned three loss functions:

$$\mathcal{L} = \lambda_{\text{seg}} \mathcal{L}_{\text{seg}} + \lambda_{\text{area}} \mathcal{L}_{\text{area}} + \lambda_{\text{p2p}} \mathcal{L}_{\text{p2p}}, \tag{10}$$

where $\{\lambda_{\text{seg}}, \lambda_{\text{area}}, \lambda_{\text{p2p}}\}$ is set to $\{1, 1, 0.05\}$ in practice to balance the contrastive loss because of the large amount of points.

4 Experiments

4.1 Dataset and Experiment Settings

ScanRefer. ScanRefer [3] is a dataset for 3D referring expression comprehension tasks such as 3D visual grounding and 3D referring instance segmentation. It contains $51,583$ queries of $11,046$ objects from 800 ScanNet [7] scenes. Each scene contains 13.81 objects and 64.48 queries on average.

Table 1: Quantitative results of different methods on ScanRefer [3] validation set. "Supervision" indicates the type of supervision. **Ins.** denotes instance labels and **Sem.** indicates semantic labels. **Mask** represents binary labels. **Bold** indicates the best.

	Method	Backbone	Label Effort‡	Supervision	mIoU	Acc@0.25	Acc@0.5
Two Stage	TGNN	GRU	> 20 min	Ins.+ Sem.	26.10	35.00	29.00
	TGNN	BERT		Ins.+ Sem.	27.80	37.50	31.40
	X-RefSeg	GRU		Ins.+ Sem.	29.77	39.85	33.52
	X-RefSeg	BERT		Ins.+ Sem.	29.94	40.33	33.77
Single Stage	LESS (ours)	GRU	< 2 min	Mask	32.19	51.00	26.41
	LESS (ours)	BERT		Mask	32.44	51.41	29.02
	LESS (ours)	RoBERTa		Mask	**33.74**	**53.23**	**29.88**

‡ The evaluate of label effort is base on a single sample.

Table 2: Module ablation on ScanRefer dataset.

	PWCA	QSA	mIoU	A@25	A@50
(a)			32.66	51.71	27.20
(b)	✓		33.44	52.73	28.92
(c)	✓	✓	**33.74**	**53.23**	**29.88**

Table 3: Loss ablation on ScanRefer dataset.

	$\mathcal{L}_{area}$	$\mathcal{L}_{p2p}$	mIoU	A@25	A@50
(a)			25.86	40.85	16.81
(b)	✓		31.04	49.61	24.72
(c)	✓	✓	**33.74**	**53.23**	**29.88**

Evaluation Metric. Following previous work [13; 29], we use the mean intersection-over-union (**mIoU**), and **Acc@kIOU** as the evaluation metrics. The mIoU is the average of the IoU over all test samples and the Acc@kIOU measures the accuracy of test samples with an IoU higher than the threshold k, where $k \in \{0.25, 0.5\}$. We use A@25 and A@50 for brevity in some of the following tables.

Implementation Details. We adopt the 3D spares U-Net [31] as our 3D feature extractor. we explore multiple text encoders, i.e., GRU [6], BERT [8] and RoBERTa [22] in our experiments for comparative analysis. For BERT and RoBERTa, we use the official pre-trained weights and fine-tune them during training. We set the number of queries K to 20 and use a single layer for QMP module. We set an initial learning rate of 2e-5 for the text encoder and 1e-4 for the others. We reduce the learning rate by a multiplicative factor of 0.95 each epoch and adopt Adam [24] as our optimizer. The weights of our loss function $\{\lambda_{\text{seg}}, \lambda_{\text{area}}, \lambda_{\text{p2p}}\}$ is set to $\{1, 1, 0.05\}$. We train for 64 epochs with a batchsize of 14, and all experiments are implemented on PyTorch [27].

4.2 Comparison with State-of-the-Art Methods

As shown in Tab.1, we evaluate our LESS against the previous two-stage methods. LESS outperforms the previous SOTA method using GRU [6] and BERT [8] by an impressive progress of 2.42%, and 2.50% on mIoU, and 11.15% and 11.08% on Acc@0.25 respectively. Moreover, we conduct an extra experiment using RoBERTa [22] , outperforming the best method of 3.8% and 12.9% on mIoU and Acc@0.25 respectively. Such results demonstrate the potential of our method. We also reported the comparison between our method and the two-stage approach in terms of label effort. We find that our label-efficient method saves more than 90% label effort compared to existing methods.

However we find that the performance of our LESS has a gap on Acc@0.5 compared to previous methods. Previous methods employed a segmentation-matching strategy. Once matching successfully, the IoU between predicted mask and ground truth is mostly higher than 0.5, which is beneficial for the Acc@0.5. In contrast, our single-stage method without instance labels and semantic labels can not extract the more accurate instance candidates as prior knowledge to assist referring 3D segmentation task. Therefore, it is acceptable for our method to perform lower than previous methods on Acc@0.5 with fewer supervisory signals.

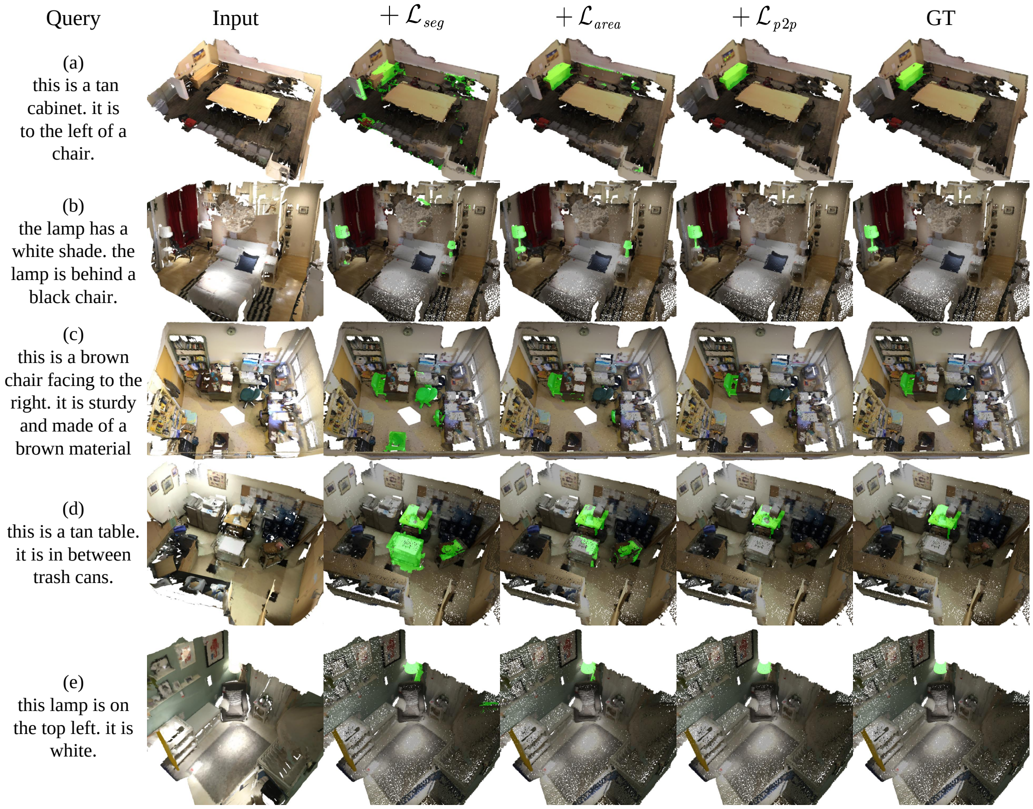
Query
Input
$+\mathcal{L}_{seg}$
$+\mathcal{L}_{area}$
$+\mathcal{L}_{p2p}$
GT
(a)
this is a tan cabinet. it is to the left of a chair.
(b)
the lamp has a white shade. the lamp is behind a black chair.
(c)
this is a brown chair facing to the right. it is sturdy and made of a brown material
(d)
this is a tan table. it is in between trash cans.
(e)
this lamp is on the top left. it is white.

Figure 3: Final predictions using different combinations of loss functions. The queries and input scenes are shown in column 1 and 2. Columns 3 to 5 indicate the gradual addition of loss functions.

4.3 Ablation Studies.

We conducted several ablation studies to evaluate the effectiveness of the key components in our proposed network. In these studies, all other components and hyper-parameter settings are kept consistent with the aforementioned experiments, except for the component being ablated.

Module Ablation. In this ablation study, we evaluate the effectiveness of the QSA and PWCA modules. As indicated in Table2, the ablation model (a) only retain the sparse 3D feature extractor and query mask predictor. Here we perform element-wise addition of the word features W to the multi-scale features $\{V_i\}_{i=1}^{5}$ instead of PWCA and project the proposed mask prediction $M_m \in \mathbb{R}^{K\times N}$ to the final mask prediction $\hat{M} \in \mathbb{R}^{N}$ via an MLP instead of QSA. We set model (a) as the baseline of our experiment. Compared to model (a), model (b) adopts PWCA module and we find that the model performance increase greatly from 32.66% to 33.44%. This observation proves that PWCA facilitate fine-grained cross-modal alignment between points and words, which is more effectively to leverage the rich convolutional layers in the encoder to excavate multi-modal context. When we introduce QSA to model (b), as shown in model (c), we can find that the performance of mIoU is improved from 33.44% to 33.74%. Such results indicates that the fine-grained point-word alignment of PWCA and the coarse-grained query-sentence alignment of QSA effectively coupled to enhance the capability in capturing multi-modal context.

Loss Ablation. As demonstrated in 3.4, we refine predicted masks from coarse to fine by introducing two loss functions, i.e., area regularization loss $\mathcal{L}_{area}$ and point-to-point contrastive loss $\mathcal{L}_{p2p}$. We successively add the loss functions and the results are shown in Tab.3. The model (a) is only supervised by segmentation loss $\mathcal{L}_{seg}$. When we introduce the $\mathcal{L}_{area}$ into model (a), as shown in model (b), the performance greatly increase from 25.86% to 31.04% on mIoU, which indicates our

https://doi.org/10.52202/079017-0356

Table 4: The impact of linguistic features at different granularities. **Word** represent word-level features, and **Sentence** represent sentence-level features.

Word	Sentence	mIoU	Acc@0.25	Acc@0.5
✓		31.56	49.50	25.84
	✓	33.02	52.33	27.75
✓	✓	**33.74**	**53.23**	**29.88**

Table 5: The impact of the number of queries.

Num of queries	mIoU	A@25	A@50
20	**33.74**	**53.23**	**29.88**
60	32.62	51.96	27.70
100	33.16	52.98	28.82

Table 6: The impact of the number of layers.

Num of layers	mIoU	A@25	A@50
1	**33.74**	**53.23**	**29.88**
3	32.45	51.09	27.70
6	32.66	52.18	28.33

area loss can exclude a significant number of irrelevant background points. Moreover, we successively add the contrastive loss $\mathcal{L}_{p2p}$, as shown in the model (c). The performance is improved from 31.04% to 33.74%, which proves that the contrastive loss can make the model more focused on the target area rather than others.

Qualitative results are shown in Fig.3, it indicates that: **i)** When only the segmentation loss $\mathcal{L}_{seg}$ is applied *(column 3)*, the predicted masks include many points from other regions. **ii)** After adding the $\mathcal{L}_{area}$, most of the irrelevant points disappear *(row a, row d)*. **iii)** After incorporating the $\mathcal{L}_{p2p}$, objects that were previously difficult to distinguish due to their similarity are successfully separated, and the predictions are close to the ground-truth.

Both quantitative and qualitative experiment demonstrate that our proposed loss function effectively reduces large-scale background misclassifications and distinguish objects or points with the similar characteristics.

4.4 Extension Experiments

Linguistic Features at Different Levels of Granularity. In this section, we will investigate the impact of linguistic features at different levels of granularity, as shown in Table.4. The first row represent that we leverage word-level features in both PWCA module and QSA module as text features. The second row indicates we utilize sentence-level features in both modules. We can find that sentence-only method even outperforms the word-only one. Referring 3D Segmentation task involves the segmentation of a solitary target object. This mandates a more profound and thorough comprehension of the semantic information conveyed by the sentence, extending beyond a mere focus on its individual words. The last row indicates the word-level features are utilized in PWCA module and sentence-level features are leveraged in QSA module. We can find that the performance of it outperforms two introduced above, which proves that fine-grained word-level feature is also helpful in extracting the 3D language-aware visual feature.

Layer and Query Number of QMP. We also investigate the performance on different query and layer numbers in the QMP module. As shown in Tab.5 and Tab.6, we find that too many queries and layers do not bring performance gains for Referring 3D Segmentation task. As a result, we set 20 queries and 1 QMP layer as the default configuration after balancing performance and efficiency.

4.5 Limitations

The limitations of LESS due to the inherent complexity of 3D point clouds and the ambiguous queries, although we have made significant improvements on previous methods. The scarcity of detailed semantic annotations are still challenging our model from distinguishing multiple similar objects. These limitations could guide our future work.

https://doi.org/10.52202/079017-0356

5 Conclusions

In this paper, we propose LESS, a label-efficient and single-stage approach for Referring 3D Segmentation. Specifically, our LESS enhances feature extraction by integrating multi-modal features and employs progressive constraints on predicted masks, achieving fine-grained alignment between points and words and distinguishing between points or objects with similar characteristics. Comprehensive experiments demonstrate that our single-stage method outperforms existing two-stage approaches on ScanRefer dataset, using only the binary labels as supervision. Though our framework still has some limitations, we believe that solving the Referring 3D Segmentation task only using the binary labels is a new and promising research, and we hope that LESS can serve as a simple but powerful baseline to inspire future research on Referring 3D Segmentation.

6 Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grants 62371310, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515011236, in part by the Stable Support Project of Shenzhen (Project No.20231122122722001).

We also thank the support by the MUR PNRR project FAIR (PE00000013) funded by the NextGenerationEU, the PRIN project CREATIVE (Prot. 2020ZSL9F9), and the EU Horizon project ELIAS (No. 101120237). We also acknowledge the CINECA award under the ISCRA initiative, for the availability of partial HPC resources support.

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[3] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[4] Ding-Jie Chen, Songhao Jia, Yi-Chen Lo, Hwann-Tzong Chen, and Tyng-Luh Liu. See-through-text grouping for referring image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7454–7463, 2019.

[5] Yong Xien Chng, Henry Zheng, Yizeng Han, Xuchong Qiu, and Gao Huang. Mask grounding for referring image segmentation. *arXiv preprint arXiv:2312.12198*, 2023.

[6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.

[7] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-language transformer and query generation for referring segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16321–16330, 2021.

https://doi.org/10.52202/079017-0356

[10] Guang Feng, Zhiwei Hu, Lihe Zhang, and Huchuan Lu. Encoder fusion network with co-attention embedding for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15506–15515, 2021.

[11] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2344–2352, 2021.

[12] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 108–124. Springer, 2016.

[13] Pin-Hao Huang, Han-Hung Lee, Hwann-Tzong Chen, and Tyng-Luh Liu. Text-guided graph neural networks for referring 3d instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 1610–1618, 2021.

[14] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds. In *European Conference on Computer Vision*, pages 417–433. Springer, 2022.

[15] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and Pattern recognition*, pages 4867–4876, 2020.

[16] Namyup Kim, Dongwon Kim, Cuiling Lan, Wenjun Zeng, and Suha Kwak. Restr: Convolution-free referring image segmentation using transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18145–18154, 2022.

[17] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. In *Advances in Neural Information Processing Systems*, volume 35, pages 16037–16051, 2022.

[18] Jinlong Li, Zequn Jie, Xu Wang, Yu Zhou, Xiaolin Wei, and Lin Ma. Weakly supervised semantic segmentation via progressive patch learning. *IEEE Transactions on multimedia*, 2022.

[19] Muchen Li and Leonid Sigal. Referring transformer: A one-step approach to multi-task visual grounding. *Advances in neural information processing systems*, 34:19652–19664, 2021.

[20] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018.

[21] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18663, 2023.

[22] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[23] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2949–2958, 2021.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[25] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si 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*, pages 16454–16463, 2022.

https://doi.org/10.52202/079017-0356

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[27] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

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

[29] Zhipeng Qian, Yiwei Ma, Jiayi Ji, and Xiaoshuai Sun. X-refseg3d: Enhancing referring 3d instance segmentation via structured cross-modal graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 4551–4559, 2024.

[30] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023.

[31] Jiahao Sun, Chunmei Qing, Junpeng Tan, and Xiangmin Xu. Superpoint transformer for 3d scene instance segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2393–2401, 2023.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[33] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2708–2717, 2022.

[34] Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. Cris: Clip-driven referring image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11686–11695, 2022.

[35] Changli Wu, Yiwei Ma, Qi Chen, Haowei Wang, Gen Luo, Jiayi Ji, and Xiaoshuai Sun. 3d-stmn: Dependency-driven superpoint-text matching network for end-to-end 3d referring expression segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5940–5948, 2024.

[36] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19231–19242, 2023.

[37] Xiaoxu Xu, Yitian Yuan, Jinlong Li, Qiudan Zhang, Zequn Jie, Lin Ma, Hao Tang, Nicu Sebe, and Xu Wang. 3d weakly supervised semantic segmentation with 2d vision-language guidance. *arXiv preprint arXiv:2407.09826*, 2024.

[38] Xiaoxu Xu, Yitian Yuan, Qiudan Zhang, Wenhui Wu, Zequn Jie, Lin Ma, and Xu Wang. Weakly-supervised 3d visual grounding based on visual linguistic alignment. *arXiv preprint arXiv:2312.09625*, 2023.

[39] Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip HS Torr. Lavt: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18155–18165, 2022.

[40] Zhengyuan Yang, Songyang Zhang, Liwei Wang, and Jiebo Luo. Sat: 2d semantics assisted training for 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1856–1866, 2021.

[41] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. 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*, pages 1791–1800, 2021.

https://doi.org/10.52202/079017-0356

[42] Yiming Zhang, ZeMing Gong, and Angel X Chang. Multi3drefer: Grounding text description to multiple 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15225–15236, 2023.

[43] Zicheng Zhang, Yi Zhu, Jianzhuang Liu, Xiaodan Liang, and Wei Ke. Coupalign: Coupling word-pixel with sentence-mask alignments for referring image segmentation. *Advances in Neural Information Processing Systems*, 35:14729–14742, 2022.

[44] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021.

[45] Chaoyang Zhu, Yiyi Zhou, Yunhang Shen, Gen Luo, Xingjia Pan, Mingbao Lin, Chao Chen, Liujuan Cao, Xiaoshuai Sun, and Rongrong Ji. Seqtr: A simple yet universal network for visual grounding. In *European Conference on Computer Vision*, pages 598–615. Springer, 2022.

https://doi.org/10.52202/079017-0356

Appendix

A Semantic and Instance Labels vs. Binary Labels

The differences between semantic labels, instance labels, and binary labels are shown in Fig.4.

Semantic labels assign a category ID to each point in a scene, with points belonging to the same category (e.g., chairs) sharing the same ID. Instance labels assign a object ID to each point within the same object (e.g., a specific chair), distinguishing different objects by assigning different instance IDs. Binary labels assign a Boolean value to each point in the scene; points that are part of the described object are assigned a value of 1, and points that are not are assigned a value of 0.

Therefore, compared to binary labels, the annotations of semantic labels and instance labels are more time consuming and labor intensive.

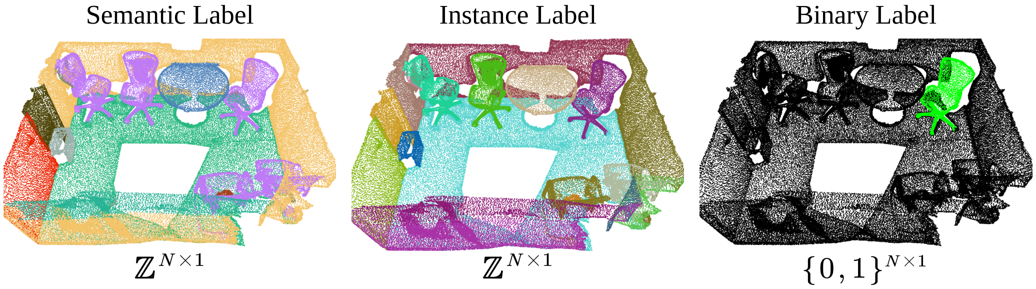

Figure 4: There different types of labels.

B More Qualitative Results

We present our success cases and failure cases in Fig.5. Our method accurately segments objects with clear queries. However, ambiguous descriptions can still confuse our model and leads to segment both the referred object and other similar objects.

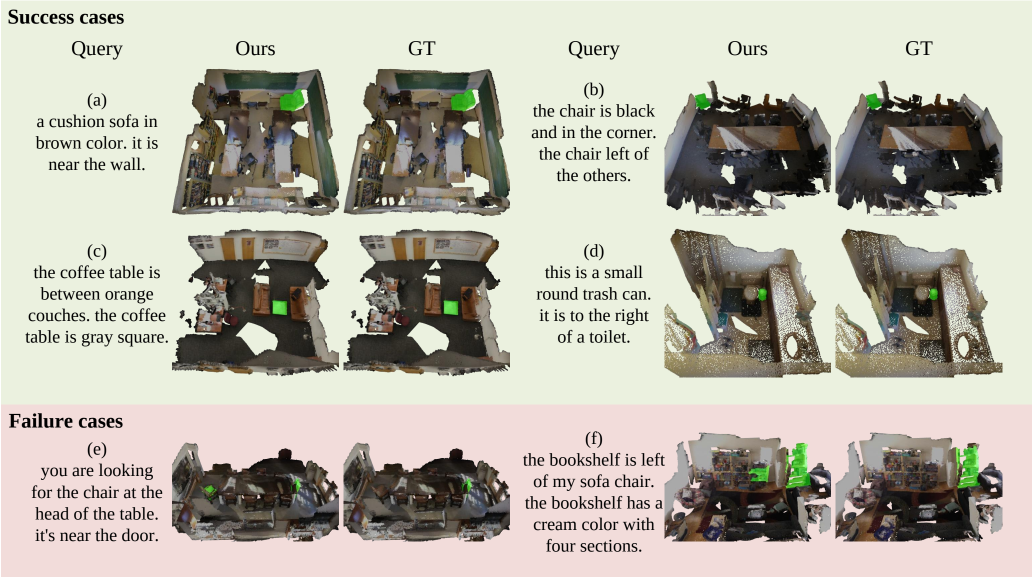
Success cases
Query
Ours
GT
Query
Ours
GT
(a)
a cushion sofa in brown color. it is near the wall.
(b)
the chair is black and in the corner. the chair left of the others.
(c)
the coffee table is between orange couches. the coffee table is gray square.
(d)
this is a small round trash can. it is to the right of a toilet.
Failure cases
(e)
you are looking for the chair at the head of the table. it's near the door.
(f)
the bookshelf is left of my sofa chair. the bookshelf has a cream color with four sections.

Figure 5: Qualitative results of both the success cases and failure cases of our LESS.

C Time Consumption Comparison

As shown in Tab.7, we evaluate the training and inference time both of TGNN[13] and X-RefSeg[29]. All experiments are conducted on an NVIDIA 4090 GPU and the number of batch sizes and epoch of three methods are kept the same. For the two-stage training and inference of TGNN and X-RefSeg, we followed the settings in their open source codes. We can find that our LESS consumes less time than both of TGNN and X-RefSeg in both training and inference.

Table 7: The comparison of inference time and training time with previous work.

Method	Inference (Whole Dataset) (min)	Inference (Per Scan) (ms)	Training (Stage 1) (h)	Training (Stage 2) (h)	Training (All) (h)
TGNN	27.98	176.57	156.02	8.53	164.55
X-RefSeg	20.00	126.23	156.02	7.59	163.61
Ours	**7.09**	**44.76**	-	-	**40.89**

D More Quantitative Results

A concurrent work 3D-STMN [35] utilizes a pre-trained 3D feature extractor [31] to perform referring 3D segmentation. Given that their backbone is pre-trained on an instance segmentation task with semantic and instance label, it is reasonable to conclude that their approach cannot be considered a label-efficient and single-stage method. For fair comparison, we follow the settings in their open source code and train their network from scratch, except for BERT [8] module. As shown in Tab.8, our method overtakes 3D-STMN by 11.34%, 18.27% and 12.65% on mIoU, Acc@0.25 and Acc@0.5 respectively.

Table 8: Quantitative results of 3D-STMN [35] and ours.

Method	mIoU	Acc@0.25	Acc@0.5
3D-STMN (from scratch)	22.40	34.96	17.23
Ours	**33.74**	**53.23**	**29.88**

E More Ablation Results

E.1 Query Number

We further conduct more ablation studies in terms of the number of queries on ScanRefer. As shown in Tab.9, we ablate the number of queries as 5, 15, 20, it can be observed that keeping 20 queries brings higher accuracy while lower accuracy when using fewer queries. We suppose that fewer queries can not help to learn comprehensive feature patterns while an appropriate number of queries is enough to cover the needed feature patterns.

Table 9: The impact of the number of queries.

Num of queries	mIoU	Acc@0.25	Acc@0.5
5	32.75	51.81	28.68
15	33.27	52.56	28.97
20	**33.74**	**53.23**	**29.88**

E.2 Mask Selection Strategy

We also conduct the ablation study which selects the mask with the highest score during QSA process as shown in Tab.10. The results demonstrate that the aggregation of multiple masks yields superior

https://doi.org/10.52202/079017-0356

performance compared to the selection of a single, highest-ranked mask. This approach facilitates the capture of subtle nuances and intricate details that may be overlooked when relying on a single mask alone. The incorporation of multiple masks offers a more comprehensive and precise representation, ultimately enhancing the accuracy of the final model prediction.

Table 10: The impact of mask selection strategy.

Num of queries	mIoU	Acc@0.25	Acc@0.5
Top-1	33.18	52.93	28.65
Ours	**33.74**	**53.23**	**29.88**

F Potential Negative Societal Impacts

Our method has no ethical risk on datasets usage and privacy violation because all the datasets and tools are publicly available and transparent.

https://doi.org/10.52202/079017-0356

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please see the Abstract and Introduction sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see the Limitations subsection precede the Conclusion section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-0356

Justification: Our paper does not include theoretical results

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Please see the Method section and Experiments section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-0356

Answer: [No]

Justification: Once the paper is accepted , we will release our code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please see the Experiments section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The experiments are conducted under a fixed seed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/079017-0356

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see the Appendix section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research is with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see the Appendix section.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

https://doi.org/10.52202/079017-0356

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please see the Experiment and Reference sections.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

https://doi.org/10.52202/079017-0356

Answer: [NA]

Justification: Our paper does not release new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0356