This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

ProxyTransformation: Preshaping Point Cloud Manifold With Proxy Attention For 3D Visual Grounding

Qihang Peng Henry Zheng Gao Huang*

Tsinghua University

Abstract

Embodied intelligence requires agents to interact with 3D environments in real time based on language instructions. A foundational task in this domain is ego-centric 3D visual grounding. However, the point clouds rendered from RGB-D images retain a large amount of redundant background data and inherent noise, both of which can interfere with the manifold structure of the target regions. Existing point cloud enhancement methods often require a tedious process to improve the manifold, which is not suitable for real-time tasks. We propose ***Proxy Transformation*** *suitable for multimodal task to efficiently improve the point cloud manifold. Our method first leverages* ***Deformable Point Clustering*** *to identify the point cloud sub-manifolds in target regions. Then, we propose a* ***Proxy Attention*** *module that utilizes multimodal proxies to guide point cloud transformation. Built upon Proxy Attention, we design a submanifold transformation generation module where textual information globally guides translation vectors for different submanifolds, optimizing relative spatial relationships of target regions. Simultaneously, image information guides linear transformations within each submanifold, refining the local point cloud manifold of target regions. Extensive experiments demonstrate that Proxy Transformation significantly outperforms all existing methods, achieving an impressive improvement of* ***7.49%*** *on easy targets and* ***4.60%*** *on hard targets, while reducing the computational overhead of attention blocks by* ***40.6%****. These results establish a new SOTA in ego-centric 3D visual grounding, showcasing the effectiveness and robustness of our approach.*

1. Introduction

In recent years, the field of embodied AI has gained increasing attention, spurred by 3D visual grounding benchmarks [1, 5, 46] that have led to a surge of research [16, 20, 23, 46, 51, 57, 63]. The 3D visual grounding task, which involves locating target objects in real-world 3D environments based on natural language descriptions, is a core perception capability for embodied agents. This ability is crucial in enabling agents to interpret and interact with their surroundings via language, which supports applications in robotics and human-computer interaction.

Despite these advancements, several significant challenges continue to hinder the performance of 3D visual grounding systems. One key challenge is the limited perception of embodied agents, who rely on ego-centric observations from multiple viewpoints, often lacking a holistic, scene-level understanding. While some methods attempt to improve scene-level perception using preconstructed 3D point clouds [16, 20, 23, 51, 57, 63], following previous 3D perception approaches [24, 50], they are impractical in real-world applications where comprehensive scene-level information is not readily available.

In response, EmbodiedScan [46] was introduced, utilizing multi-view ego-centric RGB-D scans to enable models to process scenes directly from sparse, partial viewpoints. Previous methods[46, 65, 66, 70] decouple the encoding of RGB images and depth-reconstructed point clouds from ego-centric perspectives to extract both semantic and geometric information, which are then projected into 3D space using intrinsic and extrinsic matrices to form a semantically enriched point cloud for bounding box regression. While effective, these methods overlook the noise, such as errors in depth on non-lambertian surfaces, captured by the depth sensors [21, 22, 41] that causes suboptimal performance. Consequently, critical geometric details of corresponding target objects are lost, potentially distorting their original manifold and further compromising the model's grounding performance. Moreover, a substantial portion of the sampled points represents background regions, leading to a diminished density of foreground object points.

In light of such critical challenges, previous works have focused on improving point cloud structure such as point cloud denoising [10, 11, 45] or completion methods [2, 26,

*Corresponding Author.

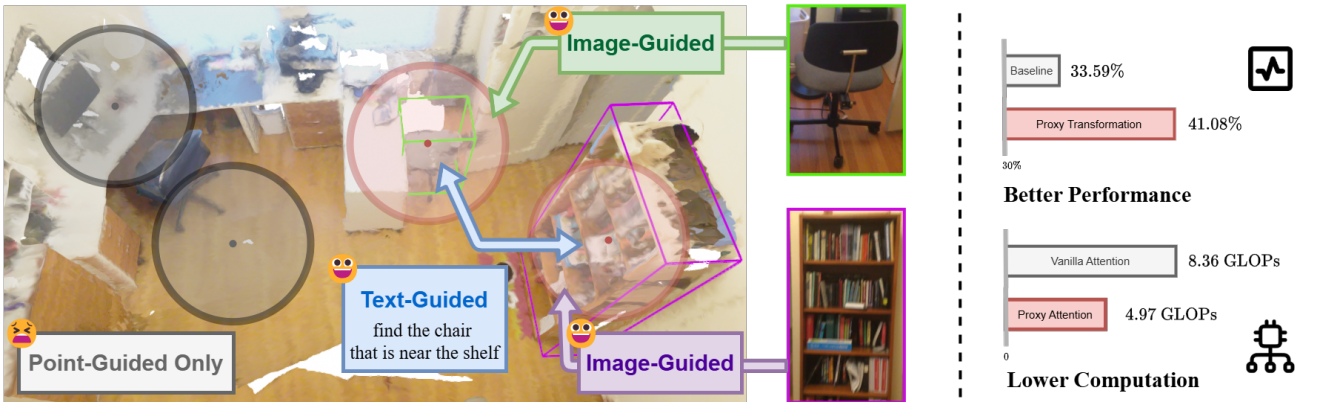

Figure 1. Illustration of our main idea and SOTA results. Ground truth and reference boxes are shown in green and purple respectively. Circular regions represent 3D areas where point cloud enhancement is applied. Black areas indicate regions that do not contribute to grounding performance and would increase unnecessary computation overhead. Traditional single-modality point cloud guidance would enhance these redundant areas. Red regions highlight areas where multimodal-guided point cloud enhancement is efficiently applied. Text modality, containing global relative position information among different critical objects, guides translation vectors for these region, while image modality, with local fine-grained semantic details, guides transformation matrices within each target regions. Our model achieves better results with reduced computation, about which details are in Tabs. 1 and 3.

60, 61]. However, these methods often require extensive preprocessing the point cloud data, making them unsuitable for real-time ego-centric 3D visual grounding. Moreover, these methods rely on traditional statistical [43] methods or learning-based methods [31, 37] to enhance the point cloud structure in single modality, which cannot fully utilize the multimodal information available in our task.

Given the limitations of the above methods in ego-centric 3D visual grounding, we propose **Proxy Transformation**, which enhances point cloud manifolds before feature learning. This approach effectively reduces redundant computation in background point cloud regions and fully leverages the available multimodal information in current context to optimize the submanifolds of target regions. Notably, our method does not require pre-trained offline reconstruction of scene-level point clouds, making it better suited for real-time ego-centric 3D visual grounding task.

Specifically, to generate transformations for point cloud submanifolds, reference points are first initialized as a uniform 3D grid for different scenes. Motivated by the success of deformable offsets in 2D domain [9, 52], a 3D offset network then takes the initial point cloud clusters centered on these reference points as input, producing offsets for each reference cluster center. In this way, these deformable cluster centers shift toward critical regions. Subsequently, they serve as cluster centers in a succeeding stage to select candidate submanifolds on the original point cloud for further transformations. For each submanifold, we employ a novel generalized attention mechanism, **Proxy Attention**, to learn the corresponding transformation matrixes and translation vectors. Specific proxy tokens can be selected based on task requirements (*e.g.* downsampling points after pooling, multi-view image features or textual features).

In the context of online point cloud submanifold enhancement, we utilize text and image features as proxy tokens to guide submanifold transformations. In this approach, text information provides global positional relationships among different submanifolds, while image information offers local semantic details within each submanifold, as illustrated in Fig. 1. Leveraging these designs, our method supplies the subsequent 3D backbone with a higher-quality point cloud manifold, thereby enhancing the model's effectiveness and robustness.

In summary, our contributions are as follows. (1) We propose Proxy Transformation, enabling multimodal point cloud augmentation in the context of 3D visual grounding. (2) To obtain more desirable submanifolds for target regions, we design deformable point clustering, utilizing a 3D offset network to generate flexible and adaptive deformable clusters suited to diverse scenes. (3) We introduce a generalized proxy attention mechanism, allowing the selection of different proxy tokens based on task requirements, achieving linear computational complexity. (4) Our model significantly outperforms all existing methods, achieving an impressive improvement of **7.49%** on easy targets and **4.60%** on hard targets, while reducing the computational overhead of attention blocks by **40.6%**, establishing a new SOTA in ego-centric 3D visual grounding.

2. Related Works

3D Visual Grounding. 3D visual grounding integrates multimodal data to localize target objects in 3D point clouds and it has gained significant attention in recent

years [12, 14, 47]. 3D visual grounding methods are divided into one-stage and two-stage architectures. One-stage approaches [15, 18, 27, 30] fuse textual and visual features to generate predictions in a single step, enabling end-to-end optimization and faster inference. However, they may struggle with complex scene layouts due to the absence of explicit object proposal refinement.In contrast, two-stage approaches [1, 4, 16, 20, 49, 58] follow a sequential process: they first utilize pre-trained object detectors [24, 49] to generate object proposals, which are subsequently matched with the linguistic input to identify the most likely target. This separation of detection and grounding allows for more precise alignment of visual and textual features, enhancing accuracy, particularly in complex scenes, but at the cost of increased computational overhead and inference time.

Recent studies [4, 20, 23, 38, 40] have explored transformer-based structures for 3D visual grounding. BUTD-DETR [23] incorporates outputs from pre-trained detection networks, including predicted bounding boxes and their corresponding class labels, as additional inputs. Multi-View Transformer [20] projects the 3D scene into a multiview space to learn robust feature representations. MiKASA Transformer [4] integrates a scene-aware object encoder and an multi-key-anchor technique, enhancing object recognition accuracy and understanding of spatial relationships. In our paper, we focus on the more challenging one-stage methods with transformer modules.

Point Cloud Enhancement. 3D point clouds provide rich geometric information and dense semantic details, playing a critical role in various domains [46, 54, 56, 60]. However, their inherent sparsity and irregularity limit model performance, and the deficiencies of existing sensors exacerbate issues such as high noise levels, sparsity, and incomplete structures in point clouds [32, 68]. These challenges make 3D point cloud augmentation a critical and challenging problem [35]. Traditional point cloud enhancement methods are often based on interpolation and optimization techniques [43], but their computational and memory overhead limits their applicability to large-scale datasets. Current deep learning-based point cloud augmentation methods can be broadly categorized into three main approaches: point cloud denoising, completion, and upsampling.

Point cloud denoising [6, 10, 11, 45, 64] can eliminate noise in the point cloud while approximating and preserving the underlying surface geometry information and semantic details. Point cloud completion focuses mainly on completion for objects [7, 26, 60, 61] and scenes [48, 53, 55, 62]. And point cloud upsampling [13, 25, 28, 39, 59] aims to improve the resolution of point clouds while maintaining the integrity of the original geometric structure.

However, these methods are still confined to a single point cloud modality and are not well-suited for tasks with high real-time requirements. In the context of ego-centric 3D visual grounding, multimodal information is available. We leverage both textual and visual modalities as guidance to transform point cloud sub-manifolds. This approach enables simultaneous point cloud denoising and densification, providing higher-quality data for subsequent processing. Additionally, our method is designed with efficiency in mind, ensuring it meets the task's real-time requirements.

3. Methodology

In this section, we will introduce our network design. Firstly, we will give a brief introduction of the entire pipeline and what our module does in Sec. 3.1. Then, we introduce our effective yet efficient Proxy Attention in Sec. 3.2 and how deformable offsets are generated for reference points in Sec. 3.3. Finally, details of Proxy Transformation will be described in Sec. 3.4.

3.1. Overall Structure

We improve upon the previous SOTA in ego-centric 3D visual grounding [46]. Here we simply review the overall structure. First, the input RGB-D images from each view is transformed into a partial point cloud, which is then integrated into a holistic 3D point cloud $P \in \mathbb{R}^{N\times 3}$ using global alignment matrices, thereby preserving geometric details. Next, multi-scale image and point features $\{\mathcal{F}^s_{\text{img}}\}^S_{s=1} \in \mathbb{R}^{V\times H_s\times W_s\times C_s}$, $\{\mathcal{F}^s_{\text{point}}\}^S_{s=1} \in \mathbb{R}^{N\times C_s}$, and text features $\mathcal{T} \in \mathbb{R}^{L\times C}$ are learned by respective backbone. Here, N is the total number of points sampled from all perspective views V, S is the number of scale for the encoded feature maps and L is the length of text sequence. The point features are then projected onto the 2D feature planes using intrinsic and extrinsic matrices to get semantic information. Further, they are concatenated to form sparse fusion features $\{\mathcal{F}^s_{\text{fusion}}\}^S_{s=1} \in \mathbb{R}^{N\times 2C_s}$. Finally, they are fed into a DETR-like decoder and then finish 9-degree-of-freedom bounding box regression.

However, due to the large number of points in the depth-reconstructed point cloud and computational limitations, only a sparse subset (about 2%) is sampled, resulting in a negative influence on the point cloud manifold. Due to the varying manifold structures in different local regions of scene-level point clouds, applying a global transformation to the entire point cloud is infeasible. To address this issue, we introduce Proxy Transformation to modify point cloud **submanifolds** with Proxy Attention, providing richer geometric information for subsequent feature learning.

3.2. Proxy Attention

For simplicity, we use following formulation to represent Softmax attention:

$$O = \sigma(QK^T)V \triangleq \text{Attn}(Q, K, V) \tag{1}$$

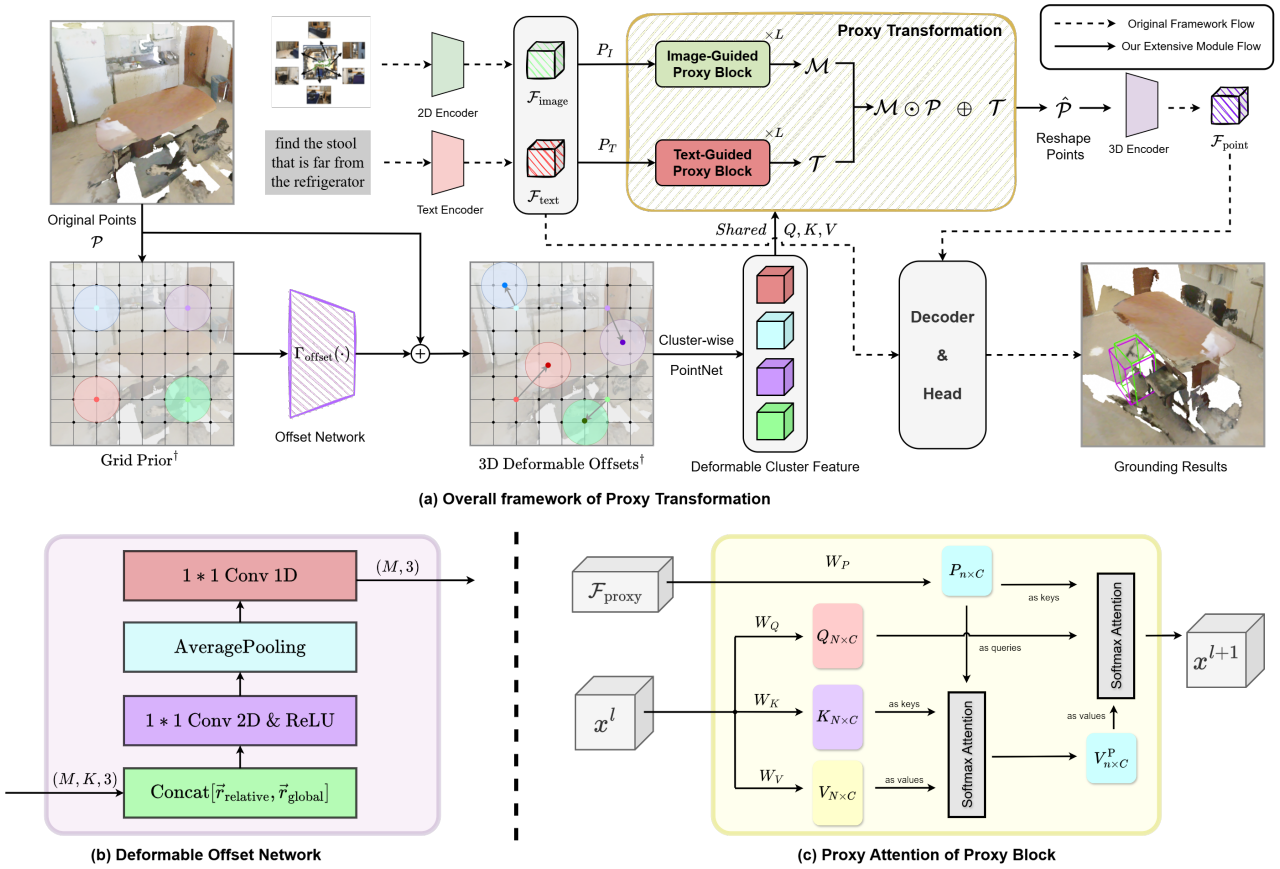

Figure 2. (a) shows the overall framework of Proxy Transforamtion. For simplicity, † indicates that a 2D grid is used to represent the 3D spatial grid, with generated 3D offsets also expressed as 2D vectors for clarity. The Grid Prior spans the entire space, and we illustrate it with only four reference points for clearer visualization. In Proxy Transformation module, $\mathcal{M}$ and $\mathcal{T}$ are sets of transformation matrixs and translation vectors for all clusters. (b) details the structure of our deformable offset network, and indicates the input and output shapes, where M represents the number of clusters and K represents the number of points per cluster. (c) illustrates the information flow in proxy attention. This module combines with FFN and skip connection in a standard Transformer architecture to form the Proxy Block.

where $Q, K, V \in \mathbb{R}^{N\times C}$ denote query, key and value matrices and $\sigma(\cdot)$ represents Softmax function. And our proxy attention consists of three components as follows.

$$\begin{aligned} O^{\text{P}} &= \text{ProxyAttn}(Q, K, V, P) \\ &= \sigma(QP^T)\,\sigma(PK^T)\,V \\ &= \underbrace{\text{Attn}(Q, P, \underbrace{\text{Attn}(P, K, V)}_{\text{Proxy Compression}})}_{\text{Proxy Broadcast}}. \end{aligned} \tag{2}$$

where $P \in \mathbb{R}^{n\times C}$ is our proxy tokens.

Proxy Compression. Previous works [17, 33] have shown significant redundancy in query-key similarity computations within self-attention. To address this, we first compress the attention feature space using proxy tokens, whose number is substantially smaller than N practically. This approach significantly alleviates computational overhead while preserving representational capacity, as shown in Tab. 3. Specifically, we use proxy tokens as queries to compute similarity with initial keys, subsequently aggregating original values to obtain a compressed proxy value representation V^{P} with reduced redundancy.

Proxy Broadcast. Subsequently, we treat proxy tokens P and proxy values V^{P} as keys and values for a second softmax operation with the original queries, allowing proxy features to be broadcast globally. This design preserves global contextual representation while reducing complexity to a linear scale. In practice, proxy tokens are selected based on the model's representational needs (*e.g.* downsampling points after pooing, multi-view image features or textual features). In these contexts, n is much smaller than N. Thus, this approach achieves generalized attention with a linear complexity of $\mathcal{O}(Nnd)$ relative to the number of input features N, balancing computational efficiency with high representational capacity as shown in Tab. 3.

Proxy Bias. Additionally, to compensate for the lack of po-

sitional information and the diversity of features, we introduce a novel proxy bias. To minimize parameter overhead, we merge two low-dimensional subspaces into a higher-dimensional space through a union operation of linear space. Specifically, we define two sets of low-dimensional learnable parameters, representing two subspaces V_c and V_r within feature space V that satisfy $V = V_c \cup V_r$. After broadcasting these two parameter sets, they match the original feature dimensionality, enriching the feature space with positional information, and guiding Proxy Attention to focus on diverse regions.

3.3. Deformable Point Clustering

To transform submanifolds, it is essential first to obtain suitable clusters. Motivated by [9, 19, 52, 71], we design a deformable clustering approach to identify optimal cluster center locations, capturing the desired target submanifolds.

Grid Prior. As mentioned in Sec. 3.1, the randomly sampled sparse point cloud has already lost a significant amount of geometric information. To compensate for this loss, we first generate a 3D uniform grid and use its grid centers as reference points for the initial clustering:

$$\begin{aligned} \mathcal{Q} &\triangleq \{q_{(i,j,k)}\}_{i,j,k=1}^{x_s,y_s,z_s} \sim U[\mathcal{C}_{min}, \mathcal{C}_{max}], \\ \mathcal{N}_{(i,j,k)} &\triangleq \{\, p_l^{i,j,k}\}_{l=1}^{m} = \gamma\,(\, \mathcal{Q}_{(i,j,k)},\ \{p_l\}_{l=1}^{N}\,), \end{aligned} \tag{3}$$

where reference points $\mathcal{Q}$ are also the initial cluster centers. U is a uniform distribution. $\mathcal{C}$ is the 3D coordinate of points. $p_l^{i,j,k}$ is a specific point in corresponding cluster $\mathcal{N}_{(i,j,k)}$ and γ is clustering function, *e.g.* ball query or kNN. N and m are respective numbers of point cloud and local cluster.

To simplify, we regard $\mathcal{N}_{(i,j,k)}$ as $\mathcal{N}_t$, where $t \in [0, n-1]$, where $n = x_s \times y_s \times z_s$ and x_s, y_s, z_s are predefined as hyperparameter. Experimentally, we find that this approach does not affect performance, but provides greater stability in the training process compared to stochastic sampling.

Deformable Offsets. To increase the diversity of submanifold selection, an offset network outputs the offset values for reference points respectively, allowing them to shift toward more relevant surrounding regions based on the current submanifold structure. Considering that each reference point covers a local s ball region (s is the largest value for offset), the offset network is able to percept local features to learn reasonable offsets. In addition, we also provide global information to further enhance offset generation.

$$\begin{aligned} \hat{q}_t &= q_t + \Gamma_{\text{offset}}(\mathcal{N}_t), \\ \hat{\mathcal{N}}_t &= \gamma\,(\,\hat{q}_t,\ \{\,p_l\}_{l=1}^{N}\,), \end{aligned} \tag{4}$$

where $\Gamma_{\text{offset}}(\cdot)$ is our offset network. $\hat{q}_t \in \hat{\mathcal{Q}}$ is a new cluster center and $\hat{\mathcal{N}}_t$ is a new neighborhood.

Specifically, for $\mathcal{N}_t$, we first concatenate the relative coordinates of sub-points with respect to the center and their global coordinates. It's then passed through 1×1 Conv2D and ReLU activation to extract point-wise features. We apply average pooling to obtain high-dimensional local-global features for each cluster, which are then fed into another 1×1 convolution to generate the final offsets, as shown in Fig. 2. Notably, we set the bias of the 1×1 convolution to zero to avoid introducing additional perturbations. Finally, we discard some clusters according to a predefined drop ratio β with a certain sampling method(*e.g.* random sampling or FPS), allowing us to obtain a more diverse neighborhood based on the new cluster centers.

3.4. Proxy Transformation

For each submanifold, we leverage the Proxy Block to learn a unique manifold transformation. Any arbitrary transformation in 3D space can be decomposed into a linear transformation and a translation. Then, these transformations are applied to each deformable cluster. Details about the transformation generation module are described below.

Proxy Block. Following the standard block design in Transformer architectures [42], we developed a Proxy Block based on proxy attention, aimed at cross-modal learning for subsequent manifold transformations. We denote P_0 and F_0 as a set of proxy tokens and cluster features. And F_0 is extracted by a simplified PointNet [34]. Then a transformer block comprises of proxy attention module and feedforward network:

$$F = F_0 + B^{\mathrm{P}}, \tag{5}$$

$$Q = FW_Q, K = FW_K, V = FW_V, P = P_0 W_P, \tag{6}$$

$$O^{\mathrm{P}} = F + \mathrm{ProxyAttn}(Q, K, V, P), \tag{7}$$

$$O = O^{\mathrm{P}} + \mathrm{FFN}(O^{\mathrm{P}}), \tag{8}$$

where B^{P} is our proxy bias introduced in Sec. 3.2. W_Q, W_K, W_V are are projections for query, key and value. $\mathrm{ProxyAttn}(\cdot)$ is a proxy attention module in Eq. (2). $\mathrm{FFN}(\cdot)$ is a position-wise feedforward network.

In the following sections, for simplicity, we use

$$O = \mathrm{ProxyBlock}(F, P), \tag{9}$$

to represent the basic proxy block (Eqs. (5) to (8)).

Text Guided Translation. In ego-centric 3D visual grounding, the input text contains global relative positional information between the target object and other reference objects, while the remaining parts naturally correspond to background point clouds. Therefore, text features can effectively guide the model in learning translation vectors that establish spatial relationships between various submanifolds, thereby enhancing the global relational information embedded within the point cloud structure. Specifically, we leverage text features as proxies within the Proxy Block, allowing it to generate text-guided point cloud features that ultimately yield accurate and context-aware final translation

vectors for more precise grounding results.

$$F^{l+1} = \text{ProxyBlock}(F^l, P_T), l = 0, .., L-1, \quad (10)$$

$$T = F^L U_{\text{text}}, \quad (11)$$

where P_T are text features $\mathcal{F}_{\text{text}} \in \mathbb{R}^{S \times C}$. L is number of layers and U_{text} is a fully connected layer to generate final translation vectors $T \in \mathbb{R}^{n \times 3}$.

Image Guided Transformation. The input multi-view image contains rich fine-grained semantic information, compensating for the contextual loss caused by point cloud downsampling. Multi-view images provide pose information of the target object from various perspectives, guiding the model to learn internal transformations of each subpoints relative to the cluster center. Thus, we use pooled image features as proxies in the Proxy Block to generate reshaped image-guided point cloud features, which in turn yield the linear transformations for each submanifold:

$$F^{l+1} = \text{ProxyBlock}(F^l, P_I), l = 0, .., L-1, \quad (12)$$

$$M = F^L U_{\text{image}}, \quad (13)$$

where $P_I \in \mathbb{R}^{V \times C}$ are pooled image features from an additional attention pooling layer. And L is the number of layers and U_{image} is a fully connected layer to generate final transformation matrices $M \in \mathbb{R}^{n \times 3 \times 3}$.

Submanifold Reshape. Once the transformation matrixs, containing fine-grained semantic information from the image, and the translation vectors, encompassing precise global relative information from the text, are obtained, we can apply a Proxy Transformation to each submanifold and finally get a diverse enhanced point cloud manifold. To represent this process, we use concise set operations:

$$\hat{\mathcal{P}} = \mathcal{M} \odot \mathcal{P} \ \oplus \ \mathcal{T}, \quad (14)$$

where $\mathcal{M}$ and $\mathcal{T}$ are sets of transformation matrixs and translation vectors for all clusters. $\mathcal{P}$ is a union set of submanifolds of different neighborhoods $\{\mathcal{N}_i\}_{i=1}^n$, namely a subset of original point cloud. $\odot$ and $\oplus$ represent element-wise multiplication and addition for sets.

Equivalently, for a specific submanifold:

$$\hat{P}_i = P_i M_i^T + T_i, \quad (15)$$

where $P_i \in \mathcal{P} \cap \mathbb{R}^{m \times 3}$ represent sub-points in $\mathcal{N}_i$. $M_i \in \mathcal{M} \cap \mathbb{R}^{3 \times 3}$ and $T_i \in \mathcal{T} \cap \mathbb{R}^3$ are the specific transformation matrix and translation vector for this submanifold.

Finally, we replace the corresponding parts of the original point cloud with transformed submanifolds to obtain final enhanced point cloud, consisting of rich diversity in manifold structure, which is then fed into 3D backbone during training, enabling efficient multimodal-guided point cloud manifold enhancements.

4. Experiments

Dataset and Benchmark. The EmbodiedScan dataset [46] used in our experiments is a large-scale, multi-modal, egocentric resource tailored for comprehensive 3D scene understanding. It comprises **5,185** scene scans from widely-used datasets like ScanNet [8], 3RScan [44], and Matterport3D [3]. This diverse dataset offers a rich foundation for 3D visual grounding, covering a broader range of scenes than previous datasets. The training set includes **801,711** language prompts, while the validation set contains **168,322** prompts, making EmbodiedScan notably larger and more challenging, thus providing a rigorous benchmark for egocentric 3D visual grounding tasks.

Experimental Settings. Due to limited resources, we train on the official mini dataset. Despite this, our model outperforms the official baseline trained on the full training set, as shown in Tab. 1. For analysis experiments in Sec. 4.2, we also use the mini training and validation sets available through the official EmbodiedScan release [46].

Our reported accuracy uses the official IoU metric, focusing on cases where IoU exceeds 25% and 50%. We also assess model performance on both "Easy" and "Hard" scenes, where a "Hard" scene contains three or more target objects of the same category. The "View-Dependent" and "View-Independent" metrics further test the model's spatial reasoning by evaluating performance with and without perspective-specific descriptions.

Implementation Details. This work builds upon the strong baseline from EmbodiedScan for ego-centric 3D visual grounding. We applied several techniques to enhance the original baseline. First, we replace the RoBERTa language encoder [29] with the CLIP encoder [36] to achieve superior alignment between language and vision. Additionally, we incorporate Class-Balanced Grouping and Sampling (CBGS) [67] during pretraining to address data imbalance and enhance detection accuracy across rare and common object categories. We follow EmbodiedScan's approach by using pretrained visual encoders. Finally, we also adopt text augmented by LLM to provide richer semantic context when training, proposed by [65]. Moreover, our Proxy Transformation is trained using the AdamW optimizer with a learning rate of 5e-4, weight decay of 5e-4, and a batch size of 48. Training spans 12 epochs, with the learning rate reduced by 0.1 at epochs 8 and 11.

4.1. Main Results

We evaluate the 3D visual grounding performance of our proposed method, Proxy Transformation, with results presented in Tab. 1, comparing it against established SOTA methods from the dataset benchmark. In addition, the upper methods with † are trained on the full dataset, while the lower methods are trained on a mini dataset, approximately 20% of the full dataset.

Method	Easy AP_{25}	Hard AP_{25}	Indep AP_{25}	Dep AP_{25}	Overall AP_{25}	Easy AP_{50}	Hard AP_{50}	Indep AP_{50}	Dep AP_{50}	Overall AP_{50}
ScanRefer† [5]	13.78	9.12	13.44	10.77	12.85	-	-	-	-	-
BUTD-DETR† [23]	23.12	18.23	22.47	20.98	22.14	-	-	-	-	-
L3Det† [69]	24.01	18.34	23.59	21.22	23.07	-	-	-	-	-
EmbodiedScan†	39.82	31.02	40.30	38.48	39.10	18.79	14.93	18.03	18.71	18.48
EmbodiedScan [46]	33.87	30.49	33.55	33.61	33.59	14.58	12.41	13.92	14.65	14.40
DenseG [65]	40.17	34.38	38.79	40.18	39.70	18.52	**15.88**	17.47	18.75	18.31
ProxyTransformation	**41.66**	**34.38**	**41.57**	**40.81**	**41.08**	**19.43**	14.09	**18.65**	**19.18**	**19.00**

Table 1. Main Results on the official validation set. The table displays accuracy performance considering instances where IoU exceeds 25% and 50% under different circumstance as mentioned in Sec. 4. 'Indep' and 'Dep' mean the targets are view-independent and view-dependent. And † denotes that models are trained on full train dataset. Although our model is trained on mini train dataset (approximately 20% of the full data), but it still surpasses previous methods trained on the full train dataset.

As shown in Tab. 1, Proxy Transformation achieves a notable 7.49% and 4.60% improvement over the previous strongest baseline, EmbodiedScan, on the Overall@0.25 and Overall@0.50. Although only trained on the mini dataset, our method even surpasses the baseline trained on the full dataset. Through deformable point clustering, our model focuses on the most crucial target regions in the scene, reducing the extra computation overhead caused by redundant point clouds and improving efficiency. Additionally, a grid prior preserves essential original spatial information, mitigating early training instability. Recognizing that text information provides global positional relationships among different submanifolds and image information offers local semantic details within each submanifold, we designed generalized proxy attention to guide local transformations of the selected point cloud submanifolds. These transformations optimize local point cloud structures for each submanifold, ultimately providing higher-quality data for subsequent feature extraction and fusion.

In summary, **Proxy Transformation** consistently outperforms the previous state-of-the-art across multiple evaluation metrics, including easy and hard samples as well as view-independent and view-dependent tasks. These consistent improvements across diverse metrics highlight not only the robustness and generalizability of our approach for 3D visual grounding tasks but also its adaptability to complex scenes and dynamic environments, making it an effective solution for advancing multimodal grounding capabilities in real-world embodied AI applications.

4.2. Analysis Experiments

Ablation on Proposed Modules. We conducted ablation studies on each module to verify their effectiveness as shown in Tab. 2. To ensure fair comparisons, the baseline is enhanced with several modules described in the implementation details. First, we introduce Proxy Transformation without deformable point clustering, resulting in a substantial improvement of 3.23%, which demonstrates the effectiveness and necessity of our multimodal point cloud enhancement. Then, we add deformable offsets, yielding an additional improvement on easy samples, indicating that offsets effectively guide cluster centers toward target regions. Finally, we incorporate grid prior, which improves

Grid Prior	Offsets	PT	Easy	Hard	Overall
			37.05	30.60	36.53
		✓	40.39	32.60	39.76
	✓	✓	40.59	32.18	39.91
✓	✓	✓	**41.66**	**34.38**	**41.08**

Table 2. Ablation of Proposed Modules. 'PT' represents 'Proxy Transformation'. The reported values are mAP for predictions greater than 25% IoU.

Attn	FLOPs	#Params	Easy	Hard	Overall
Vanilla	8.36G	2.52M	40.71	33.65	40.14
Proxy	**4.97G**	2.71M	**41.66**	**34.38**	**41.08**

Table 3. Ablation on Proxy Attention. Here, we compare the vanilla self-attention block with our Proxy Block. The FLOPs and parameters are computed over 3 transformer blocks with cluster features as input. We use a grid size of 12 and a drop radio of 0.6. The reported values are mAP for predictions greater than 25% IoU.

Grid Size	Easy	Hard	Overall
w/o	40.59	32.18	39.91
10	40.94	32.81	40.29
12	**41.66**	**34.38**	**41.08**
14	41.11	34.28	40.56

Table 4. Ablation on the size of Grid Prior. The reported values are mAP for predictions greater than 25% IoU.

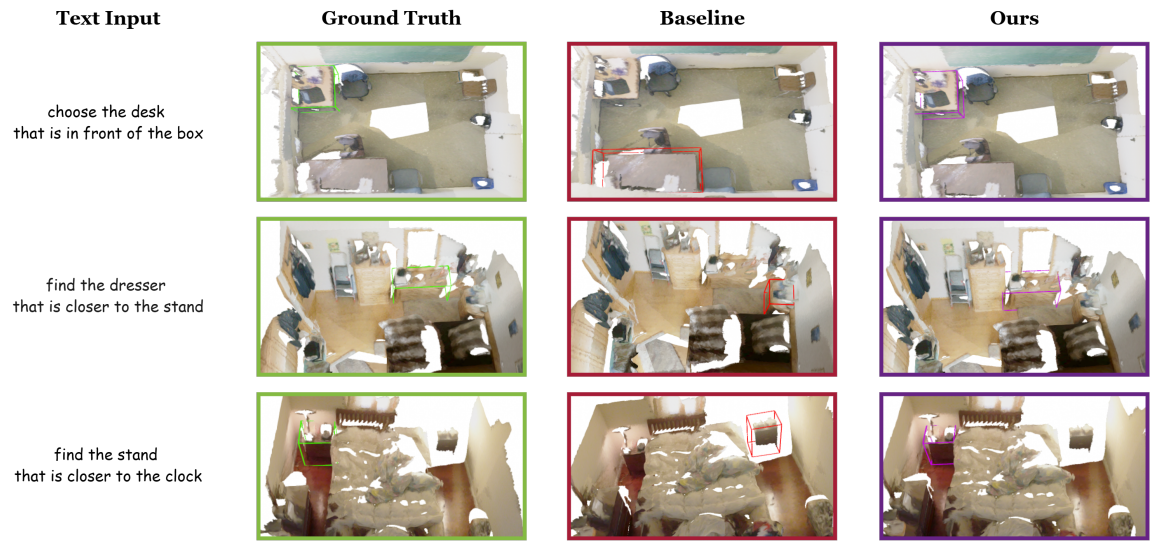

Figure 3. Visualization of ground truth and predictions. Ground truth boxes are shown in green, baseline in red, and ProxyTransformation's predictions in violet.

Layers	Easy	Hard	Overall
1	41.60	34.07	40.99
2	41.57	31.97	40.79
3	41.66	**34.38**	**41.08**
4	**41.72**	31.55	40.90

Table 5. Ablation on the number of Proxy Layers. The reported values are mAP for predictions greater than 25% IoU.

training stability and achieves an increase of 2.20% on hard samples. This result shows that the grid prior mitigates the randomness of offsets in the early stages of training, preserving essential scene information for the model.

Ablation on Proxy Attention. We validate the effectiveness and efficiency of proxy attention by comparing the vanilla attention block with the proxy block. The input cluster features are aligned across both blocks. As discussed in Sec. 3.2, the number of proxy tokens is significantly smaller than the input sequence length. This reduction in proxy tokens compresses redundancy within the attention module, enhancing computational efficiency, which aligns with our experiment results. As shown in Tab. 3, our approach achieves higher performance than vanilla attention, while reducing the computation overhead by **40.6%**.

Ablation on Main Hyperparameters. Then, we ablate several hyperparameters in deformable point clustering and proxy transformation. As shown in Tab. 4, the size of the grid prior also influences performance by reducing instability during the early training stages, which can result from the randomness of the initial transformations. As shown in Tab. 5, we conduct several experiments to determine the optimal layers of our proxy attention block for efficient learning. Based on these findings, we selected the optimal hyperparameters for our experiments. Moreover, further ablation studies are given in the supplementary material.

4.3. Qualitative Results and Discussion

With our deformable point clustering and Proxy Transformation, the point cloud structure in target regions is optimized, providing higher-quality data for subsequent feature learning and fusion. As shown in Fig. 3, reference objects in these regions are small and difficult to distinguish, but with the enhanced manifold structure, our model effectively captures the relationships between target and reference objects, achieving improved grounding performance.

5. Conclusion

In this work, we propose Proxy Transformation, a real-time multimodal point cloud enhancement model, to address issues of geometric loss and unexpected noise in 3D points within ego-centric 3D visual grounding. This approach yields a high-quality point cloud structure for feature learning. Extensive experiments show that our model not only outperforms all existing methods but also improves the computational efficiency of attention modules, setting a new SOTA on the EmbodiedScan benchmark. These results demonstrate the effectiveness and robustness of using multimodal information for real-time point cloud enhancement in ego-centric 3D visual grounding, offering fresh insights for advancing embodied AI and enabling human-computer interaction in real-world environments.

Acknowledgement

The work is supported in part by the National Natural Science Foundation of China under Grant U24B2017, and Beijing Natural Science Foundation under Grant QY24257.

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. 1, 3

[2] Alexandre Boulch and Renaud Marlet. Poco: Point convolution for surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6302–6314, 2022. 1

[3] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 6

[4] Chun-Peng Chang, Shaoxiang Wang, Alain Pagani, and Didier Stricker. Mikasa: Multi-key-anchor & scene-aware transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14131–14140, 2024. 3

[5] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*, pages 202–221. Springer, 2020. 1, 7

[6] Honghua Chen, Zeyong Wei, Xianzhi Li, Yabin Xu, Mingqiang Wei, and Jun Wang. Repcd-net: Feature-aware recurrent point cloud denoising network. *International Journal of Computer Vision*, 130(3):615–629, 2022. 3

[7] Zhikai Chen, Fuchen Long, Zhaofan Qiu, Ting Yao, Wengang Zhou, Jiebo Luo, and Tao Mei. Learning 3d shape latent for point cloud completion. *IEEE Transactions on Multimedia*, 2024. 3

[8] 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. 6

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2, 5

[10] Dasith de Silva Edirimuni, Xuequan Lu, Zhiwen Shao, Gang Li, Antonio Robles-Kelly, and Ying He. Iterativepfn: True iterative point cloud filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13530–13539, 2023. 1, 3

[11] Dasith de Silva Edirimuni, Xuequan Lu, Gang Li, Lei Wei, Antonio Robles-Kelly, and Hongdong Li. Straightpcf: Straight point cloud filtering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20721–20730, 2024. 1, 3

[12] John David N Dionisio, William G Burns Iii, and Richard Gilbert. 3d virtual worlds and the metaverse: Current status and future possibilities. *ACM computing surveys (CSUR)*, 45 (3):1–38, 2013. 3

[13] Hang Du, Xuejun Yan, Jingjing Wang, Di Xie, and Shiliang Pu. Point cloud upsampling via cascaded refinement network. In *Proceedings of the Asian Conference on Computer Vision*, pages 586–601, 2022. 3

[14] Qi Feng, Vitaly Ablavsky, and Stan Sclaroff. Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. *arXiv preprint arXiv:2101.04741*, 2021. 3

[15] Liang Geng and Jianqin Yin. Viewinfer3d: 3d visual grounding based on embodied viewpoint inference. *IEEE Robotics and Automation Letters*, 2024. 3

[16] Zoey Guo, Yiwen Tang, Ray Zhang, Dong Wang, Zhigang Wang, Bin Zhao, and Xuelong Li. Viewrefer: Grasp the multi-view knowledge for 3d visual grounding. In *ICCV*, pages 15372–15383, 2023. 1, 3

[17] Dongchen Han, Tianzhu Ye, Yizeng Han, Zhuofan Xia, Siyuan Pan, Pengfei Wan, Shiji Song, and Gao Huang. Agent attention: On the integration of softmax and linear attention. *arXiv preprint arXiv:2312.08874*, 2023. 4

[18] Shuting He and Henghui Ding. Refmask3d: Language-guided transformer for 3d referring segmentation. *arXiv preprint arXiv:2407.18244*, 2024. 3

[19] Gao Huang. Dynamic neural networks: advantages and challenges. *National Science Review*, 11(8):nwae088, 2024. 5

[20] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *CVPR*, pages 15524–15533, 2022. 1, 3

[21] Kei Ikemura, Yiming Huang, Felix Heide, Zhaoxiang Zhang, Qifeng Chen, and Chenyang Lei. Robust depth enhancement via polarization prompt fusion tuning. In *CVPR*, 2024. 1

[22] A.B.M. Tariqul Islam, Christian Scheel, Renato Pajarola, and Oliver Staadt. Robust enhancement of depth images from depth sensors. *Comput. Graph.*, 68(C), 2017. 1

[23] 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. 1, 3, 7

[24] Li Jiang, Hengshuang Zhao, Shaoshuai Shi, Shu Liu, Chi-Wing Fu, and Jiaya Jia. Pointgroup: Dual-set point grouping for 3d instance segmentation. In *CVPR*, 2020. 1, 3

[25] Akash Kumbar, Tejas Anvekar, Tulasi Amitha Vikrama, Ramesh Ashok Tabib, and Uma Mudenagudi. Tp-node: Topology-aware progressive noising and denoising of point clouds towards upsampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2272–2282, 2023. 3

[26] Siyi Leng, Zhenxin Zhang, and Liqiang Zhang. A point contextual transformer network for point cloud completion. *Expert Systems with Applications*, 249:123672, 2024. 1, 3

[27] Yue Liao, Si Liu, Guanbin Li, Fei Wang, Yanjie Chen, Chen Qian, and Bo Li. A real-time cross-modality correlation filtering method for referring expression comprehension. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10880–10889, 2020. 3

[28] Hao Liu, Hui Yuan, Junhui Hou, Raouf Hamzaoui, and Wei Gao. Pufa-gan: A frequency-aware generative adversarial network for 3d point cloud upsampling. *IEEE Transactions on Image Processing*, 31:7389–7402, 2022. 3

[29] 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, 2019. 6

[30] 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. 3

[31] Shitong Luo and Wei Hu. Score-based point cloud denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4583–4592, 2021. 2

[32] AAM Muzahid, Wanggen Wan, Ferdous Sohel, Lianyao Wu, and Li Hou. Curvenet: Curvature-based multitask learning deep networks for 3d object recognition. *IEEE/CAA Journal of Automatica Sinica*, 8(6):1177–1187, 2020. 3

[33] Yifan Pu, Zhuofan Xia, Jiayi Guo, Dongchen Han, Qixiu Li, Duo Li, Yuhui Yuan, Ji Li, Yizeng Han, Shiji Song, et al. Efficient diffusion transformer with step-wise dynamic attention mediators. *arXiv preprint arXiv:2408.05710*, 2024. 4

[34] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017. 5

[35] Siwen Quan, Junhao Yu, Ziming Nie, Muze Wang, Sijia Feng, Pei An, and Jiaqi Yang. Deep learning for 3d point cloud enhancement: A survey. *arXiv preprint arXiv:2411.00857*, 2024. 3

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 6

[37] Marie-Julie Rakotosaona, Vittorio La Barbera, Paul Guerrero, Niloy J Mitra, and Maks Ovsjanikov. Pointcleannet: Learning to denoise and remove outliers from dense point clouds. In *Computer graphics forum*, pages 185–203. Wiley Online Library, 2020. 2

[38] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 3

[39] Yi Rong, Haoran Zhou, Kang Xia, Cheng Mei, Jiahao Wang, and Tong Lu. Repkpu: Point cloud upsampling with kernel point representation and deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21050–21060, 2024. 3

[40] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Viewpoint-aware visual grounding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14056–14065, 2024. 3

[41] Tommi Tykkälä, Cédric Audras, and Andrew I Comport. Direct iterative closest point for real-time visual odometry. IEEE, 2011. 1

[42] Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 5

[43] Ignacio Vizzo, Benedikt Mersch, Rodrigo Marcuzzi, Louis Wiesmann, Jens Behley, and Cyrill Stachniss. Make it dense: Self-supervised geometric scan completion of sparse 3d lidar scans in large outdoor environments. *IEEE Robotics and Automation Letters*, 7(3):8534–8541, 2022. 2, 3

[44] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance relocalization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 6

[45] Junbo Wang and Ying Li. Transformer-based point cloud denoising network. In *2023 4th International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*, pages 384–388. IEEE, 2023. 1, 3

[46] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, Xihui Liu, Cewu Lu, Dahua Lin, and Jiangmiao Pang. Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai. In *CVPR*, 2024. 1, 3, 6, 7

[47] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6629–6638, 2019. 3

[48] Zifan Wang, Zhuorui Ye, Haoran Wu, Junyu Chen, and Li Yi. Semantic complete scene forecasting from a 4d dynamic point cloud sequence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5867–5875, 2024. 3

[49] Tung-Yu Wu, Sheng-Yu Huang, and Yu-Chiang Frank Wang. Dora: 3d visual grounding with order-aware referring. *arXiv preprint arXiv:2403.16539*, 2024. 3

[50] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *CVPR*, 2024. 1

[51] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *CVPR*, pages 19231–19242, 2023. 1

[52] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4794–4803, 2022. 2, 5

[53] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17642–17651, 2023. 3

[54] Chao Xiang, Chen Feng, Xiaopo Xie, Botian Shi, Hao Lu, Yisheng Lv, Mingchuan Yang, and Zhendong Niu. Multi-sensor fusion and cooperative perception for autonomous driving: A review. *IEEE Intelligent Transportation Systems Magazine*, 2023. 3

[55] Jinfeng Xu, Xianzhi Li, Yuan Tang, Qiao Yu, Yixue Hao, Long Hu, and Min Chen. Casfusionnet: A cascaded network for point cloud semantic scene completion by dense feature fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3018–3026, 2023. 3

[56] Jiaqi Yang, Zhiqiang Huang, Siwen Quan, Zhiguo Cao, and Yanning Zhang. Ransacs for 3d rigid registration: A comparative evaluation. *IEEE/CAA Journal of Automatica Sinica*, 9 (10):1861–1878, 2022. 3

[57] Li Yang, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, et al. Exploiting contextual objects and relations for 3d visual grounding. *NeurIPS*, 36, 2024. 1

[58] Sibei Yang, Guanbin Li, and Yizhou Yu. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4644–4653, 2019. 3

[59] Lequan Yu, Xianzhi Li, Chi-Wing Fu, Daniel Cohen-Or, and Pheng-Ann Heng. Pu-net: Point cloud upsampling network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2790–2799, 2018. 3

[60] Xumin Yu, Yongming Rao, Ziyi Wang, Zuyan Liu, Jiwen Lu, and Jie Zhou. Pointr: Diverse point cloud completion with geometry-aware transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12498–12507, 2021. 2, 3

[61] Mohang Zhang, Yushi Li, Rong Chen, Yushan Pan, Jia Wang, Yunzhe Wang, and Rong Xiang. Walkformer: Point cloud completion via guided walks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3293–3302, 2024. 2, 3

[62] Zhaoxuan Zhang, Xiaoguang Han, Bo Dong, Tong Li, Baocai Yin, and Xin Yang. Point cloud scene completion with joint color and semantic estimation from single rgb-d image. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11079–11095, 2023. 3

[63] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021. 1

[64] Tianming Zhao, Peng Gao, Tian Tian, Jiayi Ma, and Jinwen Tian. From noise addition to denoising: A self-variation capture network for point cloud optimization. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 3

[65] Henry Zheng, Hao Shi, Yong Xien Chng, Rui Huang, Zanlin Ni, Tianyi Tan, Qihang Peng, Yepeng Weng, Zhongchao Shi, and Gao Huang. Denseg: Alleviating vision-language feature sparsity in multi-view 3d visual grounding. 2024. 1, 6, 7

[66] Henry Zheng, Hao Shi, Qihang Peng, Yong Xien Chng, Rui Huang, Yepeng Weng, Gao Huang, et al. Densegrounding: Improving dense language-vision semantics for ego-centric 3d visual grounding. In *The Thirteenth International Conference on Learning Representations*, 2025. 1

[67] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 6

[68] Chen Zhu, Jianyu Yang, Zhanpeng Shao, and Chunping Liu. Vision based hand gesture recognition using 3d shape context. *IEEE/CAA Journal of Automatica Sinica*, 8(9):1600–1613, 2019. 3

[69] Chenming Zhu, Wenwei Zhang, Tai Wang, Xihui Liu, and Kai Chen. Object2scene: Putting objects in context for open-vocabulary 3d detection, 2023. 7

[70] Chenming Zhu, Tai Wang, Wenwei Zhang, Kai Chen, and Xihui Liu. Scanreason: Empowering 3d visual grounding with reasoning capabilities, 2024. 1

[71] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 5