Text to Point Cloud Localization with Multi-Level Negative Contrastive Learning

Dunqiang Liu[1,2*], **Shujun Huang**[1,2*], **Wen Li**[1,2], **Siqi Shen**[1,2], **Cheng Wang**[1,2†]

[1]Fujian Key Laboratory of Sensing and Computing for Smart Cities, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, School of Informatics, Xiamen University, China
{dqliu, huangshujun, liwen777}@stu.xmu.edu.cn, {cwang, siqishen}@xmu.edu.cn

Abstract

Language-based localization is a crucial task in robotics and computer vision, enabling robots to understand spatial positions through language. Recent methods rely on contrastive learning to establish correspondences between global features of texts and point clouds. However, the inherent ambiguity of textual descriptions makes it difficult to convey geometric information accurately, forcing alignment of them in the feature space may compromise the expressiveness of the point clouds. Unlike previous methods, this paper proposes using language as a filter to distinguish dissimilar locations. To this end, we propose a robust framework of multi-level negative contrastive learning for language-based localization, fully leveraging the descriptive power of language for spatial localization. Our method learns multiple mismatched factors by minimizing the similarity of different locations at different levels, including global-level, instance-level and relation-level, respectively. Extensive experiments conducted on the KITTI360Pose benchmark demonstrate that our method outperforms better that the state-of-the-art methods. Specifically, we achieve a 56.3% improvement in Top-1 retrieval recall and a 45.9% improvement in $5m$ localization recall.

Code — https://github.com/dqliua/MNCL

Introduction

LiDAR localization aims to estimate the position of a LiDAR point cloud in a global map, which is a crucial component of many applications in computer vision and robotics. *e.g.*, autonomous driving and virtual reality.

Retrieval-based methods (Wang et al. 2024; Luo et al. 2023; Xia et al. 2023; Komorowski 2021; Kong et al. 2024; Xia et al. 2021) address this task as a place recognition problem, aiming to identify locations by comparing LiDAR point clouds to stored references in a database. While these methods have achieved satisfactory performance, they require physically reaching the query location. In contrast, language-based LiDAR localization does not face this limitation. By describing the environmental information through natural language, the model can identify its position. This allows robots to understand locations through human language. Developing this technology will benefit human-robot collaboration, fostering applications such as autonomous delivery vehicles and language-guided navigation. However, to date, only a few methods for natural language-based localization have been developed. Text2Pos (Kolmet et al. 2022) is the pioneering work that aligns textual descriptions with point clouds in a coarse-to-fine manner. In the first stage, the goal is to retrieve its Top-k submaps that are likely to contain the described position, while in the later stage, the location is refined through matching. Text2Pos achieves the two objectives within a single step, without considering instances (hints) association. In response, RET (Wang, Fan, and Kankanhalli 2023) first designed a relation enhanced transformer to implicitly learn relations and then proposed a cascaded matching and refinement strategy to reduce noise generated in the matching process, thereby improving performance. Recently, Text2Loc (Xia et al. 2024) introduced the language model T5 (Raffel et al. 2020) to enhance text descriptions, incorporating contrastive learning to achieve optimal performance in a matching-free manner.

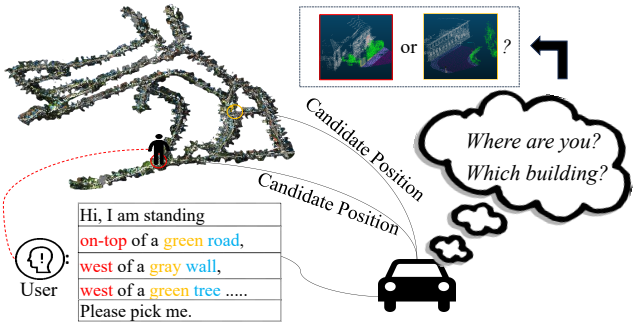

Figure 1: The key observation. City-scale point clouds are typically sparse, resulting in significant repetition of instances, such as the green tree. Moreover, text is inherently ambiguous, while point clouds provide precise geometric information, resulting in informational discrepancies.

We highlight the significant contributions of the aforementioned works, but practical implementation still has a long way to go. As illustrated in Figure 1, on one hand, city-scale point clouds are often sparse and may contain many

*These authors contributed equally.
†Corresponding author.

repeated instances. On the other hand, language descriptions tend to be vague. These issues make it challenging to achieve high localization performance using language.

Existing methods treat language and point clouds equally, aligning them in the feature space. This is nonideal for the task of localization. While point clouds contain precise scene geometry, language is inherently ambiguous. Aligning them through contrastive learning may diminish the representational power of the point clouds. Additionally, these methods typically use pooled features to represent the entire scene, resulting in information loss.

To tackle this issue, we propose a novel language-based localization method. Our method is built upon two key insights: (1) while language may not distinguish highly similar places, it can indicate where the target is not. For example, if the user say `"I am standing by a river"`, locations without a river can be excluded. (2) scenes contain rich localization cues, including landmarks and geometric relationships. Based on these insights, we propose using language as a filtering mechanism instead of direct feature alignment. Specifically, we design modality-specific scene graphs for text and point clouds, where instances are represented as node sets and their relationships as edge sets. This design allows the network to effectively leverage multi-granular information, thereby enhancing the specificity of scene representation. We then introduce Multi-level Negative Contrastive Learning (*MNCL*), which minimizes the similarity between different locations at multiple levels, including global-level, instance-level and relation-level respectively. Notably, our method alleviates the impact of scene conflicts and textual ambiguity, leading to improved localization accuracy. In summary, the main contributions of this work include:

- We propose a novel multi-level negative contrastive learning framework for language-based localization, using language as a filter to enhance its boundary-aware capabilities for place recognition. Experiments show excellent scalability.
- Extensive experiments on the KITTI360Pose benchmark demonstrate that our method outperforms the state-of-the-art methods. Specifically, we achieve a 56.3% improvement in Top-1 retrieval recall and a 45.9% improvement in $5m$ localization recall.

Related Work

LiDAR Localization. The task most relevant to our problem is LiDAR-based localization (Li et al. 2023; Yang et al. 2024). Compared to camera, LiDAR is robust to lighting and weather conditions, leading to more accurate localization. Existing methods can be primarily classified into two categories: 1) regression-based localization and 2) retrieval-based localization. Regression-based methods such as PointLoc (Wang et al. 2021) and several variants (Wang et al. 2023; Yu et al. 2023; Li et al. 2024; Yang et al. 2024; Zhou et al. 2021) directly estimate global poses through a deep regression network in an end-to-end manner. Meanwhile, retrieval-based methods typically use a two-stage pipeline involving place recognition followed by pose estimation. PointNetVlad (Uy and Lee 2018) is the first work that tackles LiDAR place recognition in an end-to-end manner. It uses point-based backbone PointNet (Qi et al. 2017a) and NetVLAD (Arandjelovic et al. 2016) to obtain global features. Subsequently, PCAN (Zhang and Xiao 2019) and SOE-net (Xia et al. 2021) introduce attention mechanism to learn task-relevant features. Minkoc3d (Komorowski 2021) uses sparse 3D convolutions to learn compact global features. Furthermore, various methods (Żywanowski et al. 2021; Vidanapathirana et al. 2023; Komorowski, Wysoczanska, and Trzcinski 2021) use point clouds registration techniques, such as ICP (Segal, Haehnel, and Thrun 2009), to perform pose estimation and thereby enhance localization performance. The previous works primarily focused on retrieval between LiDAR scans. In this work, we attempt to use textual descriptions as queries to retrieve the most matching locations in the LiDAR point clouds database, which requires multimodal understanding and is more challenging to address.

3D Vision and Language. Most existing work focus on 3D vision and language understanding in indoor environments. Chen et al. released the ScanRefer dataset (Chen, Chang, and Nießner 2020), the first 3D object localization benchmark based on natural language descriptions, where point clouds of objects are associated with corresponding text queries in indoor environments. Based on ScanRefer, some existing work (Feng et al. 2021; Yuan et al. 2021) have attempted to localize objects in scenes based on language descriptions, which is known as 3D Vision-Language Grounding. Recently, following Scan2Cap (Chen et al. 2021b), some work (Jin et al. 2024; Wang et al. 2022) have focused on providing textual descriptions of object placements, known as 3D Dense Captioning. Our task is closer to 3D Vision-Language Grounding, but focus on city-scale outdoor point clouds, and the object to be localized is a position rather than an actual object, which introduces greater task difficulty.

Contrastive Learning. As one of the most effective self-supervised learning paradigms, contrastive learning has been widely applied across various fields, such as person re-identification (Chen et al. 2021a; Song et al. 2018) and image-text retrieval (Chen et al. 2020; Wei et al. 2020). The basic idea of contrastive learning is to maximize the similarity between positive pairs while minimizing that of negative pairs in the feature space. CLIP (Radford et al. 2021) leverages contrastive learning as a pre-training objective to align large amounts of image and text data collected from the internet, enabling zero-shot generalization. However, the presence of incorrectly associated pairs can lead to the model collapse when maximizing the similarity of incorrect pairs. ALBEF (Li et al. 2021) proposes using momentum distillation to mitigate this issue. Recently, RCL (Hu et al. 2023) introduces complementary contrastive learning, which aims to minimize the similarity between more reliable negative pairs, thus avoiding issues with incorrect pairs. Inspired by the aforementioned works, our research aims to fully leverage the descriptive power of language for localization from a new perspective.

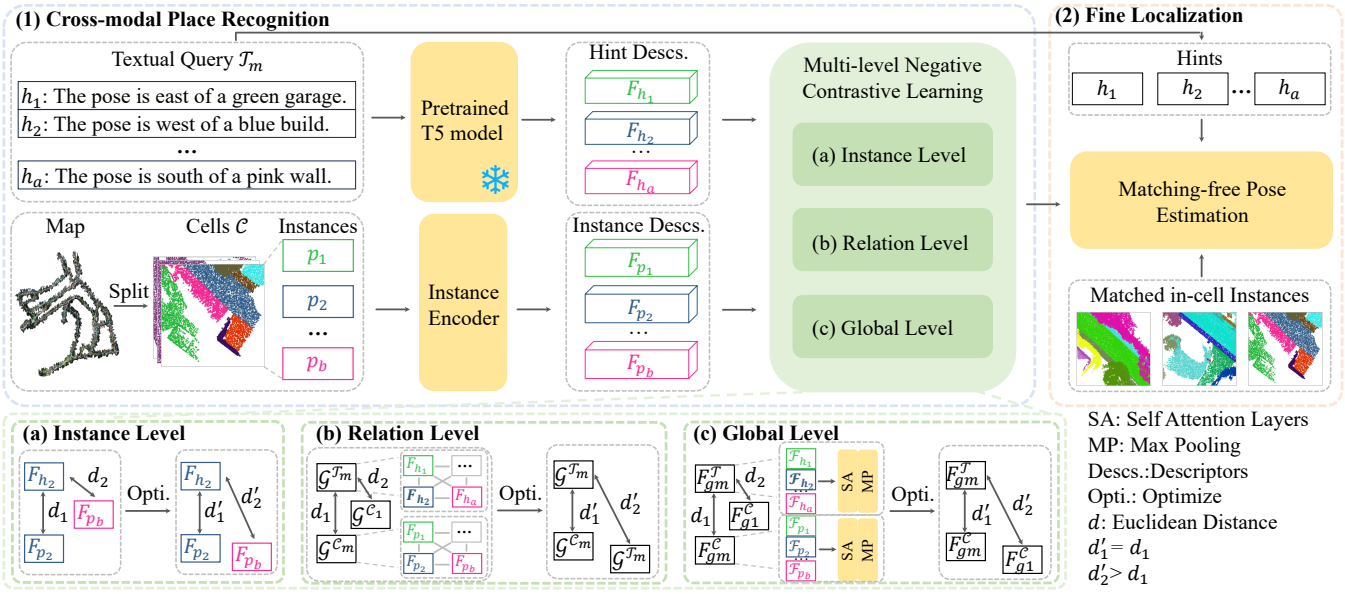

Figure 2: Framework of the proposed method. In the cross-modal place recognition stage, we introduce a multi-level negative contrastive learning framework to minimize the similarity of different locations at global-level, instance-level, and relation-level, respectively. This fully leverages the descriptive power of language for spatial localization. In the fine localization stage, we use the language query and the retrieved cell to regress the corresponding position.

Methodology

Problem Definition. Let $\mathcal{M} = \{\mathcal{C}_i\}_{i=1}^{N}$ denotes a city-scale point cloud map. Each cell $\mathcal{C}_i = \{p_b\}_{b=1}^{N_p}$ contains a set of N_p point cloud instances. The goal of this task is to identify the position for a given textual query $\mathcal{T}$ consisting of a set of hints $\{h_a\}_{a=1}^{N_h}$, each describing the spatial relation between the target location (x, y) and an object instance. To this end, we employ a coarse-to-fine manner, *i.e.*, first, retrieving a cell using the textual query (cross-modal place recognition), second, regressing the position based on the textual query and the retrieved cell (fine localization):

$$\min_{\phi,\mathcal{F}} \mathop{\mathbb{E}}_{(x,y,\mathcal{T})\sim\mathcal{D}} \left\| (x, y) - \phi\left(\mathcal{T}, \hat{\mathcal{C}}\right) \right\|_2, \tag{1}$$

where $\hat{\mathcal{C}} = \arg\min d\left(\mathcal{F}(\mathcal{T}), \mathcal{F}(\mathcal{C})\right)$ is the retrieved cell for the text query in the database, $d(\cdot, \cdot)$ is a distance metric, *e.g.*, euclidean distance, the $\mathcal{F}$ and ϕ represent the mapping functions for the stage of cross-modal place recognition and fine localization respectively. Note that the two objective are trained separately.

Overview. As shown in Figure 2, our method consists of two main components: 1) cross-modal place recognition and 2) fine localization. In the cross-model place recognition stage, we first extract features from text and point clouds using modality-specific backbones. Then, we apply multi-level negative contrastive learning to capture the mismatched relationships between locations, effectively separating their features in the feature space. The fine localization phase regresses the final position using the textual query and the retrieved cell.

Multi-Level Negative Contrastive Learning

We propose a multi-level negative contrastive learning framework aimed at minimizing the similarity between different locations across multiple levels. The proposed objective is expressed as follows:

$$\mathcal{L}_{mncl} = \mathcal{L}_{glo} + \mathcal{L}_{ins} + \mathcal{L}_{rel}, \tag{2}$$

where $\mathcal{L}_{glo}$, $\mathcal{L}_{ins}$, $\mathcal{L}_{rel}$ represent the global-level, instance-level and relation-level negative contrastive loss, respectively. In the following section, we will elaborate on the above losses.

Modality-Specific Backbone. Following Text2Loc (Xia et al. 2024), we employ a dual-branch backbone to encode the textual query $\mathcal{T}$ and point cloud $\mathcal{C}$. Specifically, we use the pre-trained language model T5 to extract features from $\mathcal{T}$, and PointNet++ (Qi et al. 2017b) to process each instance in $\mathcal{C}$, capturing point cloud attributes such as color, position, and point counts. These steps yield the feature representations for $\mathcal{T}$ and $\mathcal{C}$:

$$F^{\mathcal{T}} = \{F_{hi}\}_{i=1}^{N_h}, F^{\mathcal{C}} = \{F_{pi}\}_{i=1}^{N_p}, \tag{3}$$

where F_{hi} and F_{pi} represent the i-th hint feature and its corresponding instance feature, respectively.

Global-Level. Text2Loc employs contrastive learning to align textual queries with their corresponding cells. However, due to the presence of numerous similar instances in urban scenes, language descriptions for different locations can often be quite similar. Additionally, since a textual description corresponds to only a local region within the cell, there is an inherent information imbalance. Directly aligning

these features can lead to two main issues: 1) distinguishing between features of similar but different locations becomes difficult, and 2) the expressiveness of the point cloud is diminished.

To address the issue, we propose minimizing the similarity between different locations, enabling the network to distinguish between non-matching pairs. Specifically, given a mini-batch of size K, we first leverage self-attention layers and max pooling to obtain global-level features $\{F_{gi}^{\mathcal{T}}\}_{i=1}^{K}$ and $\{F_{gi}^{\mathcal{C}}\}_{i=1}^{K}$. Then, the global level similarity between positions i and j can be formulated as:

$$S_{ij} = \frac{\exp({F_{gi}^{\mathcal{C}}}^{\top} F_{gj}^{\mathcal{T}}/\tau)}{\sum_k^K \exp({F_{gi}^{\mathcal{C}}}^{\top} F_{gk}^{\mathcal{T}}/\tau)} + \frac{\exp({F_{gi}^{\mathcal{T}}}^{\top} F_{gj}^{\mathcal{C}}/\tau)}{\sum_k^K \exp({F_{gi}^{\mathcal{T}}}^{\top} F_{gk}^{\mathcal{C}}/\tau)}, \tag{4}$$

where τ is the temperature parameter. Based on this, the global-level negative contrastive loss is formulated as:

$$\mathcal{L}_{glo} = -\frac{1}{K}\sum_{i,j}^{K} f_{ij}(1-d)^{\frac{1}{a}} log(1-S_{ij}), \tag{5}$$

where f_{ij} is an indicator, making the loss only apply to negative pairs, *i.e.*, different locations. a is a scale parameter and d is a weighted parameter, this term enables the model to recalibrate the repulsive forces between $\mathcal{T}$ and $\mathcal{C}$. It enhances the network's focus on dissimilar locations, thereby improving its ability to discriminate between them.

Instance-Level. While language may not distinguish highly similar places, it can indicate where the target is not. For example, if the user says "`I am standing by a river`", locations without a river can be excluded. Based on this insight, we employ negative contrastive learning at the instance level to fully exploit the descriptive power of language for localization. To this end, we first use a 3-layer fully connected layers to map $F^{\mathcal{T}}$ and $F^{\mathcal{C}}$ to a subspace. Then, given a mini-batch K, we can obtain the instance-level features $\{F_{oi}^{\mathcal{T}}\}_{i=1}^{K}$ and $\{F_{oi}^{\mathcal{C}}\}_{i=1}^{K}$. The instance level similarity between positions i and j can be defined as:

$$S_{ij} = \frac{\exp(s_{ij}^{c2t}/\tau)}{\sum_{k=1}^{K}\exp(s_{ik}^{c2t}/\tau)} + \frac{\exp(s_{ij}^{t2c}/\tau)}{\sum_{k=1}^{K}\exp(s_{ik}^{t2c}/\tau)}, \tag{6}$$

where s_{ij}^{c2t} represents the average of the maximum similarity scores between instances in $F_{oi}^{\mathcal{C}}$ and $F_{oi}^{\mathcal{T}}$. Accordingly, the instance-level negative contrastive loss can be formulated as:

$$\mathcal{L}_{ins} = -\frac{1}{K}\sum_{i,j}^{K} f_{ij}(1-d)^{\frac{1}{a}} log(1-S_{ij}). \tag{7}$$

Relation-Level. To address the challenge of potentially repetitive scene instances, we construct a scene graph to capture relation-level features. The insight here is that while instances may be similar, their relationships can significantly enhance the specificity of scene representation. To fully leverage the geometric information of point clouds, we take the geometric displacement between pairs of instances as relations, which can be formulated as:

$$R_{ij}^{\mathcal{C}} = \texttt{MLP}(c_i - c_j), \tag{8}$$

where $c_i \in \mathbb{R}^3$ represents the center coordinate of the i-th instance and $\texttt{MLP}(\cdot)$ denotes a fully connected layer. We then concatenate the hints descriptors as their relations, which is formulated as:

$$R_{ij}^{\mathcal{T}} = \texttt{MLP}([F_{hi}; F_{hj}]). \tag{9}$$

We then adopt an m-th layer GCN (Wu et al. 2020) for encoding and updating. Based on this, we can generate cell scene graph $\mathcal{G}^{\mathcal{C}}$ and language scene graph $\mathcal{G}^{\mathcal{T}}$:

$$\mathcal{G}^{\mathcal{C}} = \{\mathcal{V}^{\mathcal{C}}, \mathcal{E}^{\mathcal{C}}\}, \mathcal{G}^{\mathcal{T}} = \{\mathcal{V}^{\mathcal{T}}, \mathcal{E}^{\mathcal{T}}\}, \tag{10}$$

where $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ denote a modality-specific scene graph, and $\mathcal{V} = \{V_v\}_{v=1}^{N_v}$ is the nodes-set, $\mathcal{E} = \{E_e\}_{e=1,}^{N_e}$ is the edges-set, N_v and N_e are the numbers of the corresponding set. Inspired by the external space attention (ESA) (Zhu et al. 2023), we utilize a novel attention mechanisms to capture and integrate both local and global realtional features from language and point clouds. To be specific, we first using learnable parameters $W \in \mathbb{R}^{N_e \times N_e}$ and E_e to construct the query $\mathbf{Q}$, and take E_e as value $\mathbf{V}$. Unlike traditional self-attention mechanisms that operate at the sample level, our method uses an external learnable maping space $\mathbf{M} \in \mathbb{R}^{D\times D}$ as key $\mathbf{K}$ to capture general pattens at the dataset level. The process is formulated as follows:

$$\begin{aligned} \mathbf{M}_E &= \sigma(\mathbf{Q}\cdot\mathbf{M}), \\ E_e &= \sum_{i=1}^{N_v} \mathbf{M}_E \odot \mathbf{V}, \end{aligned} \tag{11}$$

where $\mathbf{M}_E \in \mathbb{R}^{D\times D}$ represents the attention matrix, σ is the softmax normalization along the second dimension. $\odot$ represent Hadamard product. After that, given a mini-batch K, Similar to instance-level features, we can obtain the relation-level features $\{F_{ri}^{\mathcal{T}}\}_{i=1}^{K}$ and $\{F_{ri}^{\mathcal{C}}\}_{i=1}^{K}$. The relation level similarity between positions i and j can be defined as:

$$S_{ij} = \frac{\exp(s_{ij}^{c2t}/\tau)}{\sum_{k=1}^{K}\exp(s_{ik}^{c2t}/\tau)} + \frac{\exp(s_{ij}^{t2c}/\tau)}{\sum_{k=1}^{K}\exp(s_{ik}^{t2c}/\tau)}, \tag{12}$$

where s_{ij}^{c2t} represents the average of the maximum similarity scores between $F_{ri}^{\mathcal{C}}$ and $F_{ri}^{\mathcal{T}}$. The relation-level negative contrastive loss can be formulated as:

$$\mathcal{L}_{rel} = -\frac{1}{K}\sum_{i,j}^{K} f_{ij}(1-d)^{\frac{1}{a}} log(1-S_{ij}). \tag{13}$$

Fine Localizaiton

In the fine localization stage, the goal is to accurately locate the coordinates corresponding to the text description within the top-k candidate cells. Text2Loc innovatively addresses this issue by using a translation regressor. It obtains fused features by applying cross-attention between the candidate cells and the textual description, and finally regresses the target position via a simple MLP. The regression loss can be formulated as:

$$\mathcal{L}_{fine} = \left\| C_{gt} - C_{pred} \right\|_2, \tag{14}$$

Method	Localization Recall ($\epsilon < 5/10/15m$) ↑					
	Validation Set			Test Set		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
Text2Pos	0.14/0.25/0.31	0.36/0.55/0.61	0.48/0.68/0.74	0.13/0.21/0.25	0.33/0.48/0.52	0.43/0.61/0.65
RET	0.19/0.30/0.37	0.44/0.62/0.67	0.52/0.72/0.78	0.16/0.25/0.29	0.35/0.51/0.56	0.46/0.65/0.71
Text2Loc	0.37/0.57/0.63	0.68/0.85/0.87	0.77/0.91/0.93	0.33/0.48/0.52	0.61/0.75/0.78	0.71/0.84/0.86
Ours	**0.54/0.75/0.79**	**0.83/0.94/0.95**	**0.89/0.97/0.98**	**0.50/0.66/0.68**	**0.76/0.87/0.89**	**0.83/0.93/0.94**

Table 1: Performance of complete pipeline comparison on the KITTI360Pose. The best results are indicated in black bold.

Method	Submap Retrieval Recall ↑					
	Validation Set			Test Set		
	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
Text2Pos	0.14	0.28	0.37	0.12	0.25	0.33
RET	0.18	0.34	0.44	-	-	-
Text2Loc	0.32	0.56	0.67	0.28	0.49	0.58
Ours	**0.50**	**0.75**	**0.84**	**0.44**	**0.67**	**0.75**

Table 2: Performance of retrieval stage comparison on the KITTI360Pose benchmark. Note that only values that are available in RET are reported.

Method	Submap Retrieval Recall ↑					
	Validation Set			Test Set		
	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
$\mathcal{L}_{glo}$	0.40	0.67	0.75	0.38	0.58	0.66
$\mathcal{L}_{ins}$+$\mathcal{L}_{glo}$	0.44	0.70	0.79	0.42	0.62	0.70
$\mathcal{L}_{rel}$+$\mathcal{L}_{glo}$	0.46	0.72	0.81	0.42	0.64	0.73
Ours	**0.50**	**0.75**	**0.84**	**0.44**	**0.67**	**0.75**

Table 3: Ablation study of the multi-level negative contrastive loss on the KITTI360Pose benchmark.

where C_{gt} is the corresponding ground-truth coordinates and C_{pred} is the predicted target coordinates.

Notably, although previous studies (Yu et al. 2022) have shown that adding regression head parameters can improve performance, we maintain the same architecture as Text2Loc in this paper. This choice clearly validates our insight that language can serve as a filter to narrow down the localization range, leading to more accurate localization results.

Experiments

Benchmark Dataset

We train and evaluate the proposed method on the KITTI360Pose benchmark(Kolmet et al. 2022). It includes point clouds of 9 districts, covering 43,381 position-query pairs with a total area of 15.51 km^2. Following (Kolmet et al. 2022), we sample the cells of size 30m with a stride of 10m. Five scenes are used for training ($11.59km^2$), one for validation, and the remaining three for testing ($2.14km^2$).

Evaluation Metric

As the evaluation metric, we use Retrieve Recall at Top-k (k $\in \{1, 3, 5\}$) in the cross-modal place recognition stage. Additionally, in the fine localization stage, we assess the Top-k (k $\in \{1, 5, 10\}$) retrieved candidates and report localization recall. Here, localization recall measures the proportion of queries successfully localized queries when their errors are below specific thresholds ($\epsilon \in \{5m, 10m, 15m\}$). A higher value indicates better performance.

Implementation Details

For hyper-parameters, we set the number of hidden layers in GCNs to 3, and set the parameter α in the loss function to 2. We train our model 32 epochs with AdamW optimizer (Loshchilov and Hutter 2017). The learning rate is set to $1e^{-3}$ and decays by half every 5 epochs. All the experiments are conducted on an NVIDIA RTX 4090 GPU.

Results

Quantitative Analysis. To validate the performance of the proposed method, we conduct a comparative analysis with several state-of-the-art language-based methods, such as Text2Pos (Kolmet et al. 2022), RET (Wang, Fan, and Kankanhalli 2023), and Text2Loc (Xia et al. 2024). For the cross-modal place recognition, Table 2 shows the performance for submap retrieval recall of each method. Our method outperforms Text2Loc by $56\%/34\%/25\%$ and $57\%/37\%/29\%$ in Top-1/3/5 retrieval performance on the validation and test sets, respectively. These results demonstrate that our method can perform localization well by language query in large-scale outdoor scenes.

For the fine localization, we report the Top-1/5/10 localization recall rates with different thresholds ($\epsilon \in \{5m, 10m, 15m\}$). As shown in Table 1, Text2Loc achieves a recall rate of 0.37 on the validation set for Top-1 recall within $5m$. In contrast, our method gets a 45.9% significant improvement. This improvement remains evident even as the number of candidates k increases or when evaluated on the test set. Moreover, as shown in the last column of Table 1, we achieve a Top-10 localization recall of 93% with a $10m$ threshold for the first time, which bridges the gaps in practical applications.

Qualitative Analysis. Figure 3 shows a comparision of retrieval results between our method and Text2Loc. For a given query, we display the Top-2 retrieved cells. Here, A cell is considered a positive sample if it contains the target location. In the first row, when the scene contains salient instances (*e.g.*, poles and buildings), both models accurately retrieve the correct cell. The second row illustrates a scene

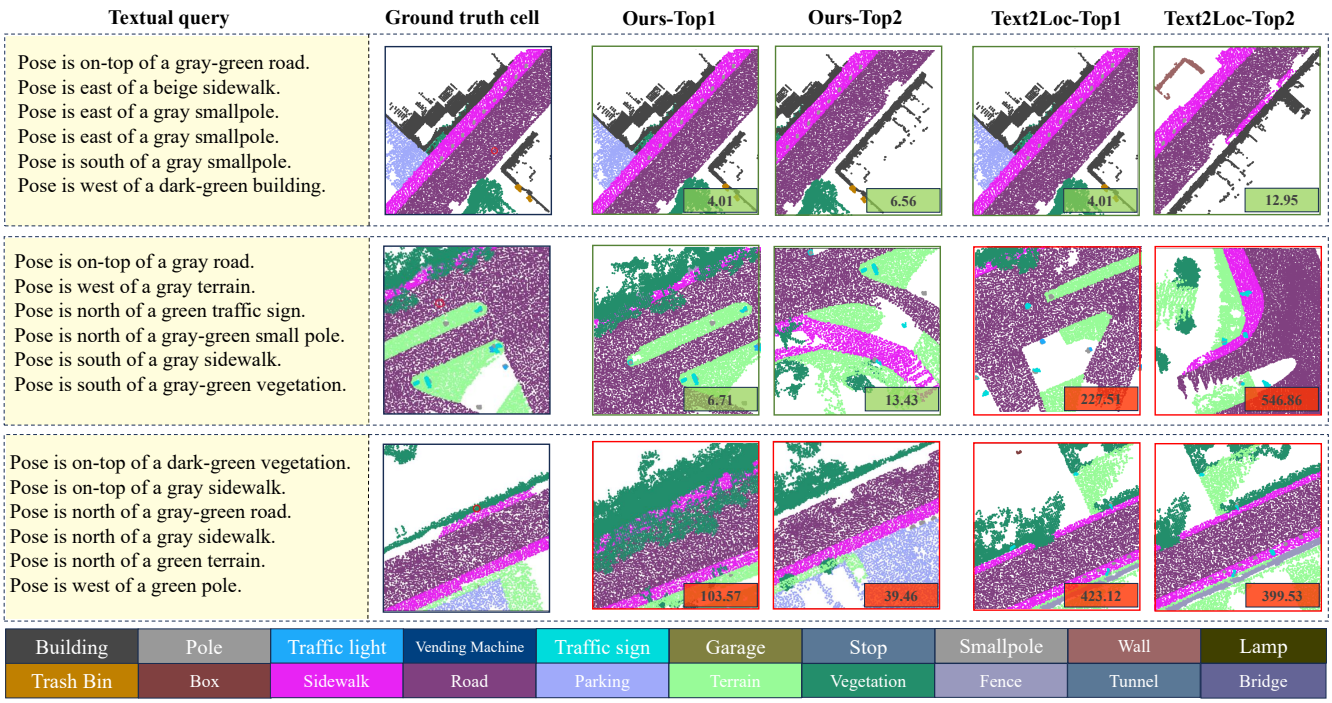

Figure 3: Qualitative retrieval results on the KITTI360Pose dataset. The red circle in the ground truth cell indicates the target location. Each retrieved cell displays the center distance from the ground truth in the bottom-right corner. Correct matches are shown with a green background, while incorrect matches have a red background. Green boxes highlight positive cells containing the target location, whereas red boxes indicate negative cells.

with repetitive instances, such as roads and vegetation. The results show that our method successfully retrieves the relevant cell, while Text2Loc fails. This may be attributed to our model's ability to capture and focus on prominent instances and their relationships, which enhances its ability to correctly identify locations. The last row shows a challenging case with low-distinctiveness instances. In this scenario, although neither model successfully retrieves the correct cell, our model achieves a smaller localization error than Text2Loc, further demonstrating its effectiveness.

Ablation Study

In this section, we perform ablation studies to investigate the effectiveness of our method. The analysis is divided into three parts: multi-level feature learning mechanism, negative contrastive loss, and hyper-parameters.

Multi-level Feature Learning Mechanism. We analyze the benefits of multi-level feature learning by keeping other components unchanged. The results are shown in Table 3. Here, $\mathcal{L}_{glo}$ represents using only global-level features, $\mathcal{L}_{ins} + \mathcal{L}_{glo}$ refers to using both global-level and instance-level features, and $\mathcal{L}_{rel} + \mathcal{L}_{glo}$ refers to using both global-level and relation-level features. Compared to Row 1 and Row 2, instance-level features raise the Top-1 retrieval recall rate from 0.40 to 0.44 on the validation set. This result indicates that incorporating more local features, which may represent landmark information, enhances feature discriminability. Furthermore, adding relation-level features to global-level alignment boosts this metric to 0.46, demonstrating that the topological relationships between instances provide valuable scene-specific information in urban contexts. Finally, our method, which integrates all three level features, effectively leverages multiple relational cues within the scene, achieving optimal performance.

Method	Submap Retrieval Recall ↑					
	Validation Set			Test Set		
	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
w/o $\mathcal{L}_{mncl}$	0.45	0.70	0.78	0.41	0.63	0.71
Ours	**0.50**	**0.75**	**0.84**	**0.44**	**0.67**	**0.75**

Table 4: Ablation study of the negative contrastive learning on the KITTI360Pose benchmark. "w/o $\mathcal{L}_{mncl}$" indicates the vanilla contrastive loss.

Negative Contrastive Loss. We conduct experiments by replacing the proposed negative contrastive loss with the vanilla loss. The results, shown in Table 4, indicate that after replacing the loss function with the original one, the Top-1/3/5 recall rates on the validation set drop from 0.50/0.75/0.84 to 0.45/0.70/0.78, The trend is also observed on the test set. These results indicate that by combining with negative contrastive loss, replacing the task of pulling text-point positive pairs closer with pushing negative pairs fur-

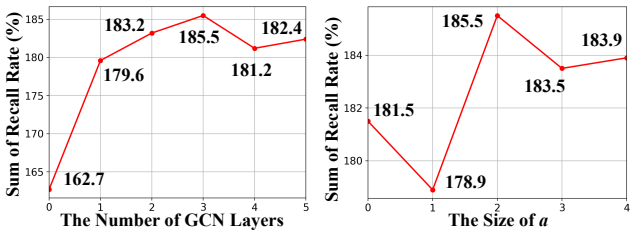

Figure 4: The sum of retrieval recall rates for Top-1/3/5 (rsum) of the (a) number of GCN layers and (b) the size of a in the multi-level negative contrastive loss.

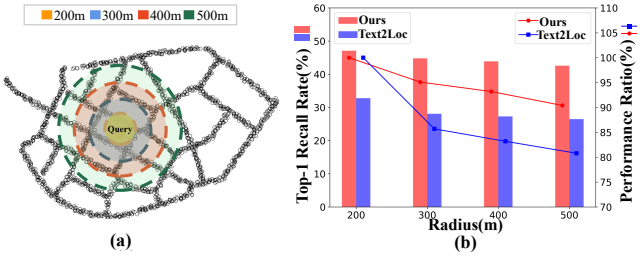

Figure 5: (a): Visualization for database and query on trajectory Scene-0009. (b): Performance of our model with Text2Loc on the same set of queries across four different scales of databases, reporting the Top-1 retrieval recall rates. The bar chart shows the absolute values of the retrieval recall rates, while the line chart represents the performance ratio relative to the standard setting of the database and query with a 200m radius.

ther apart effectively solves text ambiguity, achieving significant improvements in retrieval recall rates.

Hyper-Parameters. We further investigate the effects of different parameter settings on our method. Specifically, this section verifies the impact of the number of GCN layers (Figure 4 (a)) and the size of a in multi-level negative contrastive loss (Figure 4 (b)). We report the sum of retrieval recall rates for Top-1/3/5 (rsum) on the test set.

In Figure 4 (a), we set the layer number from zero to five, and report its metric. Here, the zero indicates removing the instance and relation-level alignment, and only using global-level alignment. It shows that when a GCN layer is added, the performance is significantly improved from 162.7% to 179.6%. This demonstrates the importance of alignment at both instance and relation levels. It can also be seen that when the number is set to two and further increased, the performance improvement is not obvious, and the optimal value is achieved when the number is three. Notably, when the number of hidden layers exceeds three, the sum of retrieval recall decreases. This may be due to the model starting to capture noise and irrelevant details in the instances as the network depth increases.

In Figure 4 (b), we set the value of a from zero to four, and report its metric. Here, the zero indicates the removal of the proposed multi-level negative contrastive loss, replaced by a vanilla contrastive learning loss function (Radford et al. 2021). Results shows that our model demonstrates good robustness to different parameters. On the other hand, we observe that either excessively ignoring or focusing on the reliability of negative pairs by increasing or decreasing the value of a leads to a decline in retrieval recall. The best result is achieved when a is set to two.

Scalability Analysis

In this section, we analyze the scalability of our method. We evaluate the performance difference between our model and the state-of-the-art method Text2Loc by restricting the query to a specific area and progressively increasing the size of the database. Specifically, we choose the geometric center of trajectory Scene-0009 as the circle's center, define a circle with a radius of $200m$, and sample 2373 target positions within it, using their textual descriptions as a set of queries. For the database, we use the same center and select radiuses of $200m$/$300m$/$400m$/$500m$, resulting in 819/1799/2278/2549 cells as the database. As shown in Figure 5, using the $200m$ radius database as a standard setting, when increasing the database radius to $300m$, $400m$, and $500m$, our model's performance decreased by 4.9%, 6.8% and 9.6%, respectively. The results of our method are significantly better than the performance drop of 14.3%, 16.8%, and 19.2% observed with Text2Loc. Experiments demonstrate that as the data scale expands and more interfering scenes are introduced into the database, performance is impacted. However, unlike existing methods, our proposed model does not aim to learn the similarity between language and point clouds but rather to minimize incorrect locations in the database. This allows it to maintain better performance in distinguishing and matching correct locations, showing minimal performance degradation as the data scale increases. This demonstrates excellent robustness and scalability.

Conclusion

In this paper, we propose a robust language-based localization method that avoids directly aligning text and point cloud features. Instead, we leverage the proposed multi-level negative contrastive learning to minimize the similarity between different locations across multi-granularity features, including global, instance, and relational levels. Our method effectively utilizes landmarks and relational information within scenes, enhancing the capability of language for spatial localization. Quantitative and qualitative results demonstrate that our method outperforms state-of-the-art methods and exhibits excellent scalability. Notably, we are the first to achieve a $10m$ localization recall rate of over 90% in city-scale environments using language. We believe this method provides a novel perspective for language-based localization by filtering out irrelevant locations and reducing the search space. However, due to the lack of additional available datasets, our experiments were conducted on a single benchmark. Future work could focus on developing more diverse datasets to enable comprehensive and accurate analysis, further advancing this field.

References

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5297–5307.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, 202–221. Springer.

Chen, H.; Ding, G.; Liu, X.; Lin, Z.; Liu, J.; and Han, J. 2020. IMRAM: Iterative Matching with Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, H.; Wang, Y.; Lagadec, B.; Dantcheva, A.; and Bremond, F. 2021a. Joint generative and contrastive learning for unsupervised person re-identification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2004–2013.

Chen, Z.; Gholami, A.; Nießner, M.; and Chang, A. X. 2021b. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 3193–3203.

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

Hu, P.; Huang, Z.; Peng, D.; Wang, X.; and Peng, X. 2023. Cross-modal retrieval with partially mismatched pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9595–9610.

Jin, B.; Zheng, Y.; Li, P.; Li, W.; Zheng, Y.; Hu, S.; Liu, X.; Zhu, J.; Yan, Z.; Sun, H.; et al. 2024. TOD3Cap: Towards 3D Dense Captioning in Outdoor Scenes. *arXiv preprint arXiv:2403.19589*.

Kolmet, M.; Zhou, Q.; Ošep, A.; and Leal-Taixé, L. 2022. Text2pos: Text-to-point-cloud cross-modal localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6687–6696.

Komorowski, J. 2021. Minkloc3d: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1790–1799.

Komorowski, J.; Wysoczanska, M.; and Trzcinski, T. 2021. Egonn: Egocentric neural network for point cloud based 6dof relocalization at the city scale. *IEEE Robotics and Automation Letters*, 7(2): 722–729.

Kong, D.; Li, X.; Xu, Q.; Hu, Y.; and Ni, P. 2024. SC_LPR: Semantically Consistent LiDAR Place Recognition Based on Chained Cascade Network in Long-Term Dynamic Environments. *IEEE Transactions on Image Processing*.

Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34: 9694–9705.

Li, W.; Yang, Y.; Yu, S.; Hu, G.; Wen, C.; Cheng, M.; and Wang, C. 2024. DiffLoc: Diffusion Model for Outdoor LiDAR Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15045–15054.

Li, W.; Yu, S.; Wang, C.; Hu, G.; Shen, S.; and Wen, C. 2023. SGLoc: Scene geometry encoding for outdoor LiDAR localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9286–9295.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Luo, L.; Zheng, S.; Li, Y.; Fan, Y.; Yu, B.; Cao, S.-Y.; Li, J.; and Shen, H.-L. 2023. BEVPlace: Learning LiDAR-based place recognition using bird's eye view images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8700–8709.

Qi, C. R.; Su, H.; Mo, K.; and Guibas, L. J. 2017a. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 652–660.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140): 1–67.

Segal, A.; Haehnel, D.; and Thrun, S. 2009. Generalized-icp. In *Robotics: science and systems*, 4, 435. Seattle, WA.

Song, C.; Huang, Y.; Ouyang, W.; and Wang, L. 2018. Mask-guided contrastive attention model for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1179–1188.

Uy, M. A.; and Lee, G. H. 2018. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4470–4479.

Vidanapathirana, K.; Moghadam, P.; Sridharan, S.; and Fookes, C. 2023. Spectral geometric verification: Re-ranking point cloud retrieval for metric localization. *IEEE Robotics and Automation Letters*, 8(5): 2494–2501.

Wang, G.; Fan, H.; and Kankanhalli, M. 2023. Text to point cloud localization with relation-enhanced transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2501–2509.

Wang, H.; Zhang, C.; Yu, J.; and Cai, W. 2022. Spatiality-guided transformer for 3d dense captioning on point clouds. *arXiv preprint arXiv:2204.10688*.

Wang, S.; Kang, Q.; She, R.; Wang, W.; Zhao, K.; Song, Y.; and Tay, W. P. 2023. HypLiLoc: Towards effective LiDAR pose regression with hyperbolic fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5176–5185.

Wang, S.; She, R.; Kang, Q.; Jian, X.; Zhao, K.; Song, Y.; and Tay, W. P. 2024. DistilVPR: Cross-Modal Knowledge Distillation for Visual Place Recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 10377–10385.

Wang, W.; Wang, B.; Zhao, P.; Chen, C.; Clark, R.; Yang, B.; Markham, A.; and Trigoni, N. 2021. Pointloc: Deep pose regressor for lidar point cloud localization. *IEEE Sensors Journal*, 22(1): 959–968.

Wei, X.; Zhang, T.; Li, Y.; Zhang, Y.; and Wu, F. 2020. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10941–10950.

Wu, Z.; Pan, S.; Chen, F.; Long, G.; Zhang, C.; and Philip, S. Y. 2020. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 32(1): 4–24.

Xia, Y.; Gladkova, M.; Wang, R.; Li, Q.; Stilla, U.; Henriques, J. F.; and Cremers, D. 2023. Casspr: Cross attention single scan place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8461–8472.

Xia, Y.; Shi, L.; Ding, Z.; Henriques, J. F.; and Cremers, D. 2024. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14958–14967.

Xia, Y.; Xu, Y.; Li, S.; Wang, R.; Du, J.; Cremers, D.; and Stilla, U. 2021. SOE-Net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 11348–11357.

Yang, B.; Li, Z.; Li, W.; Cai, Z.; Wen, C.; Zang, Y.; Muller, M.; and Wang, C. 2024. LiSA: LiDAR Localization with Semantic Awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15271–15280.

Yu, S.; Sun, X.; Li, W.; Wen, C.; Yang, Y.; Si, B.; Hu, G.; and Wang, C. 2023. NIDALoc: Neurobiologically Inspired Deep LiDAR Localization. *IEEE Transactions on Intelligent Transportation Systems*.

Yu, S.; Wang, C.; Wen, C.; Cheng, M.; Liu, M.; Zhang, Z.; and Li, X. 2022. LiDAR-based localization using universal encoding and memory-aware regression. *Pattern Recognition*, 128: 108685.

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

Zhang, W.; and Xiao, C. 2019. PCAN: 3D attention map learning using contextual information for point cloud based retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12436–12445.

Zhou, K.; Chen, C.; Wang, B.; Saputra, M. R. U.; Trigoni, N.; and Markham, A. 2021. Vmloc: Variational fusion for learning-based multimodal camera localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 6165–6173.

Zhu, H.; Zhang, C.; Wei, Y.; Huang, S.; and Zhao, Y. 2023. Esa: External space attention aggregation for image-text retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(10): 6131–6143.

Żywanowski, K.; Banaszczyk, A.; Nowicki, M. R.; and Komorowski, J. 2021. Minkloc3d-si: 3d lidar place recognition with sparse convolutions, spherical coordinates, and intensity. *IEEE Robotics and Automation Letters*, 7(2): 1079–1086.