Text2Loc: 3D Point Cloud Localization from Natural Language

Yan Xia*[1,2]† Letian Shi*[1] Zifeng Ding[3] João F. Henriques[4] Daniel Cremers[1,2]
[1]Technical University of Munich [2]Munich Center for Machine Learning (MCML)
[3] LMU Munich [4] Visual Geometry Group, University of Oxford
{yan.xia, letian.shi, cremers}@tum.de, zifeng.ding@campus.lmu.de, joao@robots.ox.ac.uk

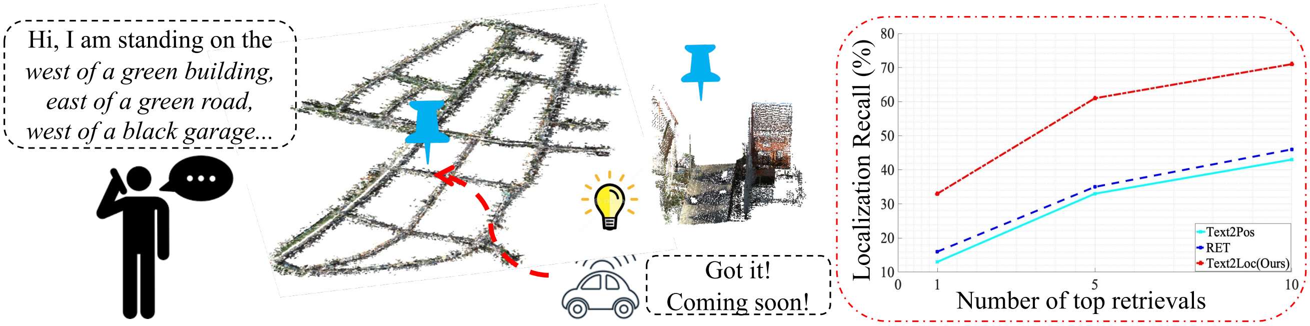


Figure 1. *(Left)* We introduce Text2Loc, a solution designed for city-scale position localization using textual descriptions. When provided with a point cloud representing the surroundings and a textual query describing a position, Text2Loc determines the most probable location of that described position within the map. *(Right)* Localization performance on the KITTI360Pose test set. The proposed Text2Loc achieves consistently better performance across all top retrieval numbers. Notably, it outperforms the best baseline by up to 2 times, localizing text queries below $5\,\mathrm{m}$.

Abstract

We tackle the problem of 3D point cloud localization based on a few natural linguistic descriptions and introduce a novel neural network, Text2Loc, that fully interprets the semantic relationship between points and text. Text2Loc follows a coarse-to-fine localization pipeline: text-submap global place recognition, followed by fine localization. In global place recognition, relational dynamics among each textual hint are captured in a hierarchical transformer with max-pooling (HTM), whereas a balance between positive and negative pairs is maintained using text-submap contrastive learning. Moreover, we propose a novel matching-free fine localization method to further refine the location predictions, which completely removes the need for complicated text-instance matching and is lighter, faster, and more accurate than previous methods. Extensive experiments show that Text2Loc improves the localization accuracy by up to $2\times$ *over the state-of-the-art on the KITTI360Pose dataset. Our project page is publicly available at* `https://yan-xia.github.io/projects/text2loc/`.

1. Introduction

3D localization [18, 28] using natural language descriptions in a city-scale map is crucial for enabling autonomous agents to cooperate with humans to plan their trajectories [11] in applications such as goods delivery or vehicle pickup [36, 38]. When delivering a takeaway, couriers often encounter the "last mile problem". Pinpointing the exact delivery spot in residential neighborhoods or large office buildings is challenging since GPS signals are bound to fail among tall buildings and vegetation [34, 37]. Couriers often rely on voice instructions over the phone from the recipient to determine this spot. More generally, the "last mile problem" occurs whenever a user attempts to navigate to an unfamiliar place. It is therefore essential to develop the capability to perform localization from the natural language, as shown in Fig. 1.

As a possible remedy, we can match linguistic descriptions to a pre-built point cloud map using calibrated depth sensors like LiDAR. Point cloud localization, which focuses on the scene's geometry, offers several advantages over images. It remains consistent despite lighting, weather, and season changes, whereas the same geometric structure in images might appear vastly different.

The main challenge of 3D localization from natural language descriptions lies in accurately interpreting the

†Corresponding author. * Equal contribution.

language and semantically understanding large-scale point clouds. To date, only a few networks have been proposed for language-based localization in a 3D large-scale city map. Text2Pose [12] is a pioneering work that aligns objects described in text with their respective instances in a point cloud, through a coarse-to-fine approach. In the coarse stage, Text2Pose first adopts a text-to-cell cross-model retrieval method to identify the possible regions that contain the target position. In particular, Text2Pose matches the text and the corresponding submaps by the global descriptors from 3D point clouds using PointNet++ [20] and the global text descriptors using a bidirectional LSTM cell [10, 25]. This method describes a submap with its contained instances of objects, which ignores the instance relationship for both points and sentences. Recently, the authors of RET [33] noted this shortcoming and designed Relation-Enhanced Transformer networks. While this results in better global descriptors, both approaches match global descriptors using the pairwise ranking loss without considering the imbalance in positive and negative samples.

Inspired by RET [33], we also notice the importance of effectively leveraging relational dynamics among instances within submaps for geometric representation extraction. Furthermore, there is a natural hierarchy in the descriptions, composed of sentences, each with word tokens. We thus recognize the need to analyze relationships within (intra-text) and between (inter-text) descriptions. To address these challenges, we adopt a frozen pre-trained large language model T5 [23] and design a hierarchical transformer with max-pooling (HTM) that acts as an intra- and inter-text encoder, capturing the contextual details within and across sentences. Additionally, we enhance the instance encoder in Text2Pose [12] by adding a number encoder and adopting contrastive learning to maintain a balance between positive and negative pairs. Another observation is that, when refining the location prediction in the fine localization stage, the widely used text-instance matching module in previous methods should be reduced since the noisy matching or inaccurate offset predictions are a fatal interference in predicting the exact position of the target. To address this issue, we propose a novel matching-free fine localization network. Specifically, we first design a prototype-based map cloning (PMC) module to increase the diversity of retrieved submaps. Then, we introduce a cascaded cross-attention transformer (CCAT) to enrich the text embedding by fusing the semantic information from point clouds. These operations enable one-stage training to directly predict the target position without any text-instance matcher.

To summarize, the main contributions of this work are:

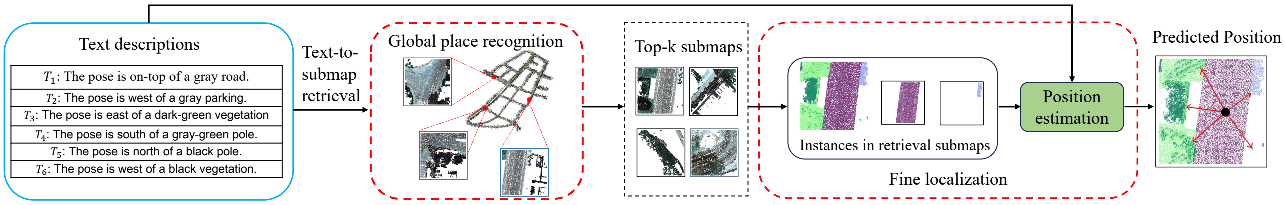
- We focus on the relatively-understudied problem of point cloud localization from textual descriptions, to address the "last mile problem".
- We propose a novel attention-based method that is hierarchical and represents contextual details within and across sentence descriptions of places.
- We study the importance of positive-negative pairs balance in this setting, and show how contrastive learning is an effective tool that significantly improves performance.
- We are the first to completely remove the usage of text-instance matcher in the final localization stage. We propose a lighter and faster localization model while still achieving state-of-the-art performance via our designed prototype-based map cloning (PMC) module in training and cascaded cross-attention transformer (CCAT).
- We conduct extensive experiments on the KITTI360Pose benchmark [12] and show that the proposed Text2Loc greatly improves over the state-of-the-art methods.

2. Related work

To date, only a few networks have been proposed for the natural language based 3D localization in a large-scale outdoor scene. Other tasks that are related to ours include 2D visual localization, 3D point cloud based localization, and 3D understanding with language.

2D visual localization. Visual localization in 2D images has wide-ranging applications from robotics to augmented reality. Given a query image or image sequence, the aim is to predict an accurate pose. One of the early works, Scale-Invariant Feature Transform (SIFT) [15], proposes the use of distinctive invariant features to match objects across different viewpoints, forming a basis for 2D localization. Oriented FAST and Rotated BRIEF (ORB) [24] has been pivotal in achieving robustness against scale, rotation, and illumination changes in 2D localization tasks. Recent learning-based methods [26, 29] commonly adopt a coarse-to-fine pipeline. In the coarse stage, given a query image, place recognition is performed as nearest neighbor search in high-dimensional spaces. Subsequent to this, a pixel-wise correspondence is ascertained between the query and the retrieved image, facilitating precise pose prediction. However, the performance of image-based methods often degrades when facing drastic variations in illumination and appearance caused by weather and seasonal changes. Compared to feature matching in 2D visual localization, in this work, we aim to solve cross-model localization between text and 3D point clouds.

3D point cloud based localization. With breakthroughs in learning-based image localization methods, deep learning of 3D localization has become the focus of intense research. Similar to image-based methods, a two-step pipeline is commonly used in 3D localization: (1) place recognition, followed by (2) pose estimation. PointNetVlad [2] is a pioneering network that tackles 3D place recognition with end-to-end learning. Subsequently, SOE-Net [35] introduces the PointOE module, incorporating orientation encoding into



T_1: The pose is on-top of a gray road.
T_2: The pose is west of a gray parking.
T_3: The pose is east of a dark-green vegetation
T_4: The pose is south of a gray-green pole.
T_5: The pose is north of a black pole.
T_6: The pose is west of a black vegetation.

Figure 2. The proposed Text2Loc architecture. It consists of two tandem modules: Global place recognition and Fine localization. *Global place recognition.* Given a text-based position description, we first identify a set of coarse candidate locations, "submaps," potentially containing the target position. This is achieved by retrieving the top-k nearest submaps from a previously constructed database of submaps using our novel text-to-submap retrieval model. *Fine localization.* We then refine the center coordinates of the retrieved submaps via our designed matching-free position estimation module, which adjusts the target location to increase accuracy.

PointNet to generate point-wise local descriptors. Furthermore, various methods [3, 6, 8, 16, 17, 40, 41] have explored the integration of different transformer networks, specifically stacked self-attention blocks, to learn long-range contextual features. In contrast, Minkloc3D [13] employs a voxel-based strategy to generate a compact global descriptor using a Feature Pyramid Network [14] (FPN) with generalized-mean (GeM) pooling [21]. However, the voxelization methods inevitably suffer from lost points due to the quantization step. CASSPR [37] thus introduces a dual-branch hierarchical cross attention transformer, combining both the advantages of voxel-based approaches with the point-based approaches. After getting the coarse location of the query scan, the pose estimation can be computed with the point cloud registration algorithms, like the iterative closest point (ICP) [30] or autoencoder-based registration [7]. By contrast to point cloud based localization, we use natural language queries to specify any target location.

3D vision and language. Recent work has explored the cross-modal understanding of 3D vision and language. [19] bridges language implicitly to 3D visual feature representations and predicts 3D bounding boxes for target objects. Methods [1, 4, 9, 39] locate the most relevant 3D target objects in a raw point cloud scene given by the query text descriptions. However, these methods focus on real-world indoor scene localization. Text2Pos [12] is the first attempt to tackle the large city-scale outdoor scene localization task, which identifies a set of coarse locations and then refines the pose estimation. Following this, Wang *et al.* [33] propose a Transformer-based method to enhance representation discriminability for both point clouds and textual queries.

3. Problem statement

We begin by defining the large-scale 3D map $M_{\text{ref}} = \{m_i : i = 1, ..., M\}$ to be a collection of cubic submaps m_i. Each submap $m_i = \{P_{i,j} : j = 1, ..., p\}$ includes a set of 3D object instances $P_{i,j}$. Let T be a query text description consisting of a set of hints $\{\vec{h}_k\}_{k=1}^{h}$, each describing the spatial relation between the target location and an object instance. Following [12], we approach this task in a coarse-to-fine manner. The text-submap global place recognition involves the retrieval of submaps based on T. This stage aims to train a function F, which encodes both T and a submap m into a unified embedding space. In this space, matched query-submap pairs are brought closer together, while unmatched pairs are repelled. In fine-grained localization, we employ a matching-free network to directly regress the final position of the target based on T and the retrieved submaps. Thus, the task of training a 3D localization network from natural language is defined as identifying the ground truth position (x, y) (2D planar coordinates w.r.t. the scene coordinate system) from M_{ref}:

$$\min_{\phi, F} \mathop{\mathbb{E}}_{(x,y,T) \sim \mathcal{D}} \left\| (x, y) - \phi\left(T, \operatorname*{argmin}_{m \in M_{\text{ref}}} d\left(F(T), F(m)\right)\right) \right\|^2 \quad (1)$$

where $d(\cdot, \cdot)$ is a distance metric (e.g. the Euclidean distance), $\mathcal{D}$ is the dataset, and ϕ is a neural network that is trained to output fine-grained coordinates from a text embedding T and a submap m.

4. Methodology

Fig. 2 shows our Text2Loc architecture. Given a text-based query position description, we aim to find a set of coarse candidate submaps that potentially contain the target position by using a frozen pre-trained T5 language model [23] and an intra- and inter-text encoder with contrastive learning, described in Section 4.1. Next, we refine the location based on the retrieved submaps via a designed fine localization module, which will be explained in Section 4.2. Section 4.3 describes the training with the loss function.

4.1. Global place recognition

3D point cloud-based place recognition is usually expressed as a 3D retrieval task. Given a query LiDAR scan, the aim is to retrieve the closest match and its corresponding location from the database by matching its global descriptor against the global descriptors extracted from a database of reference scans based on their descriptor distances. Following this general approach, we adopt the *text-submap* cross-modal

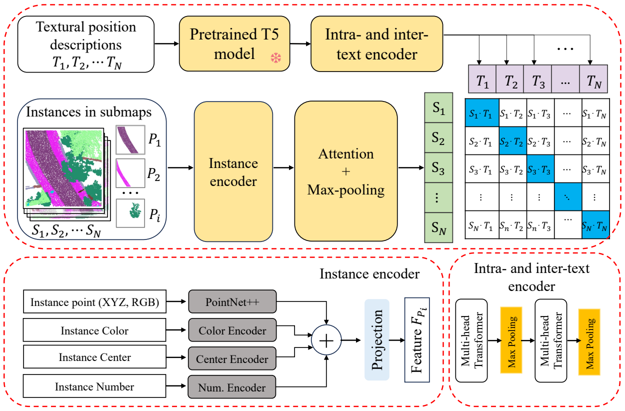


Figure 3. *(top)* The architecture of global place recognition, *(bottom)* instance encoder architecture for point clouds, and the architecture of intra- and inter-text encoder. Note that the pre-trained T5 model is frozen.

global place recognition for coarse localization. With this stage, we aim to retrieve the nearest submap in response to a textual query. The main challenge lies in how to find simultaneously robust and distinctive global descriptors for 3D submaps S and textual queries T. Similar to [12, 33], we employ a dual branch to encode S and T into a shared embedding space, as shown in Fig. 3 (top).

Text branch. We initially use a frozen pre-trained large language model, T5 [23], to extract nuanced features from textual descriptions, enhancing the embedding quality. We then design a hierarchical transformer with max-pooling layers to capture the contextual details within sentences (via self-attention) and across them (via the semantics that are shared by all sentences), as depicted in Fig. 3 (Bottom right). Each transformer is a residual module comprising Multi-Head Self-Attention (MHSA) and FeedForward Network (FFN) sublayers. The feed-forward network comprises two linear layers with the ReLU activation function. More details are in the Supplementary Materials.

3D submap branch. Each instance P_i in the submap S_N is represented as a point cloud, containing both spatial and color (RGB) coordinates, resulting in 6D features (Fig. 3 (bottom left)). We utilize PointNet++ [20] (which can be replaced with a more powerful encoder) to extract semantic features from the points. Additionally, we obtain a color embedding by encoding RGB coordinates with our color encoder and a positional embedding by encoding the instance center $\bar{P}_i$ (i.e., the mean coordinates) with our positional encoder. We find that object categories consistently differ in point counts; for example, roads typically (> 1000 points) have a higher point count than poles (< 500 points). We thus design a number encoder, providing potential class-specific prior information by explicitly encoding the point numbers. All the color, positional, and number encoders are 3-layer multi-layer perceptrons (MLPs) with output dimensions matching the semantic point embedding dimension.

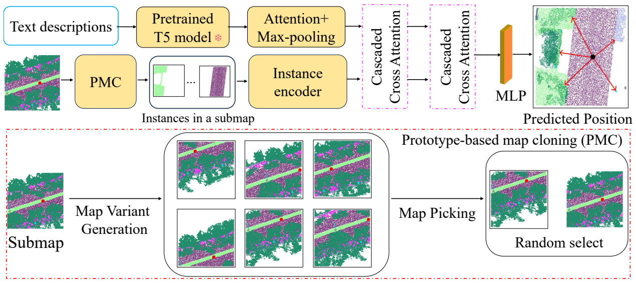


Figure 4. The proposed matching-free fine localization architecture. It consists of two parallel branches: one is extracting features from query text descriptions *(top)* and another is using the instance encoder to extract point cloud features *(bottom)*. Cascaded cross-attention transformers (CCAT) use queries from one branch to look up information in the other branch, aiming to fuse the semantic information from point clouds into the text embedding. The result is then processed with a simple MLP to directly estimate the target position.

We merge the semantic, color, positional, and quantity embeddings through concatenation and process them with a projection layer, another 3-layer MLP. This projection layer produces the final instance embedding F_{p_i}. Finally, we aggregate *in-submap* instance descriptors $\{F_{p_i}\}_{i=1}^{N_p}$ into a global submap descriptor F^S using an attention layer [35] followed by a max pooling operation.

Text-submap Contrastive learning. We introduce a cross-modal contrastive learning objective to address the limitations of the widely used pairwise ranking loss in [12, 33]. This objective aims to jointly drive closer the feature centroids of 3D submaps and the corresponding text prompt. In our overall architecture, illustrated in Figure 3, we incorporate both a text encoder and a point cloud encoder. These encoders serve the purpose of embedding the text-submap pairs into text features denoted as $F^T \in \mathcal{R}^{1\times C}$ and 3D submap features represented as $F^S \in \mathcal{R}^{1\times C}$, respectively. Here, C signifies the embedding dimension. Inspired by CLIP [22], we computer the feature distance between language descriptions and 3D submaps with a contrastive learning loss (See Sec. 4.3 for details).

4.2. Fine localization

Following the text-submap global place recognition, we aim to refine the target location prediction within the retrieved submaps in fine localization. Although the final localization network in previous methods [12, 33] achieved notable success using a text-submap matching strategy, the inherent ambiguity in the text descriptions significantly impeded accurate offset predictions for individual object instances. To address this issue, we propose a novel matching-free fine localization network, as shown in Fig. 4. The text branch (top) captures the fine-grained features by using a frozen pre-trained language model T5 [23] and an attention unit followed by a max pooling layer. The submap branch (bottom) performs a prototype-based map cloning module to increase more map variants and then extracts the point cloud features using an instance encoder, the same as in the global place recognition. We then fuse the text-submap feature with a Cascaded Cross-Attention Transformer and finally regress the target position via a simple MLP.

Cascaded Cross-attention Transformer (CCAT). To efficiently exploit the relationship between the text branch and the 3D submap branch, we propose a CCAT to fuse the features from the two branches. The CCAT consists of two Cross Attention Transformers (CAT), each is the same as in [37]. The CAT1 takes the point cloud features as Query and the text features as Key and Value. It extracts text features with reference to the point features and outputs point feature maps that are informed by the text features. Conversely, CAT2 produces enhanced text features by taking the text features as the Query and the enhanced point cloud features from CAT1 as the Key and Value. Notably, the CAT1 and the CAT2 are a cascading structure, which is the main difference from the HCAT in [37]. In this work, two cascaded CCATs are used. More ablation studies and analyses are in the Supplementary Materials.

Prototype-based Map Cloning (PMC). To produce more effective submap variants for training, we propose a novel prototype-based map cloning module. For each pair $\{T_i, S_i\}$, we hope to generate a collection $\mathcal{G}_i$ of surrounding map variants centered on the current map S_i, which can be formulated as follows:

$$\mathcal{G}_i = \{S_j \mid \left\|\bar{s_j} - \bar{s_i}\right\|_\infty < \alpha,\ \left\|\bar{s_j} - c_i\right\|_\infty < \beta\ \}, \tag{2}$$

where $\bar{s_i}$, $\bar{s_j}$ are the center coordinates of the submaps S_i and S_j respectively. c_i represents the ground-truth target position described by T_i, α and β are the pre-defined thresholds. In this work, we set $\alpha = 15$ and $\beta = 12$.

In practice, we find that certain submaps in $\mathcal{G}_i$ have an insufficient number of object instances corresponding to the textual descriptions T_i. To address this, we introduce a filtering process by setting a minimum threshold $N_m = 1$. This threshold implies that at most one instance mismatch is permissible. After applying this filter, we randomly selected a single submap from the refined $\mathcal{G}_i$ for training.

4.3. Loss function

Global place recognition. Different from the pairwise ranking loss widely used in previous methods [12, 33], we train the proposed method for text-submap retrieval with a cross-model contrastive learning objective. Given an input batch of 3D submap descriptors $\{F_i^S\}_{i=1}^N$ and matching text descriptors $\{F_i^T\}_{i=1}^N$ where N is the batch size, the contrastive loss among each pair is computed as follows,

$$l(i,T,S) = -\log \frac{\exp(F_i^T \cdot F_i^S/\tau)}{\sum\limits_{j\in N} \exp(F_i^T \cdot F_j^S/\tau)} - \log \frac{\exp(F_i^S \cdot F_i^T/\tau)}{\sum\limits_{j\in N} \exp(F_i^S \cdot F_j^T/\tau)}, \tag{3}$$

where τ is the temperature coefficient, similar to CLIP [22]. Within a training mini-batch, the text-submap alignment objective $L(T,S)$ can be described as:

$$L(T,S) = \frac{1}{N}\left[\sum_{i\in N} l(i,T,S)\right]. \tag{4}$$

Fine localization. Unlike previous method [12, 33], our fine localization network does not include a text-instance matching module, making our training more straightforward and faster. Note that this model is trained separately from the global place recognition. Here, our goal is to minimize the distance between the predicted location of the target and the ground truth. In this paper, we use only the mean squared error loss L_r to train the translation regressor.

$$L(C_{gt}, C_{pred}) = \left\|C_{gt} - C_{pred}\right\|_2, \tag{5}$$

where $C_{pred} = (x, y)$ (see Eq. (1)) is the predicted target coordinates, and C_{gt} is the ground-truth coordinates.

5. Experiments

5.1. Benchmark Dataset

We train and evaluate the proposed Text2Loc on the KITTI360Pose benchmark presented in [12]. It includes point clouds of 9 districts, covering 43,381 position-query pairs with a total area of 15.51 km^2. Following [12], we choose five scenes (11.59 km^2) for training, one for validation, and the remaining three (2.14 km^2) for testing. The 3D submap is a cube that is 30m long with a stride of 10m. This creates a database with 11,259/1,434/4,308 submaps for training/validation/testing scenes and a total of 17,001 submaps for the entire dataset. For more details, please refer to the supplementary material in [12].

5.2. Evaluation criteria

Following [12], we use Retrieve Recall at Top k ($k \in \{1, 3, 5\}$) to evaluate text-submap global place recognition. For assessing localization performance, we evaluate with respect to the top k retrieved candidates ($k \in \{1, 5, 10\}$) and report localization recall. Localization recall measures the proportion of successfully localized queries if their error falls below specific error thresholds, specifically $\epsilon < 5/10/15m$ by default.

5.3. Results

5.3.1 Global place recognition

We compare our Text2Loc with the state-of-the-art methods: Text2Pos [12] and RET [33]. We evaluate global place

Methods	Localization Recall ($\epsilon < 5/10/15m$) ↑					
	Validation Set			Test Set		
	$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
Text2Pos [12]	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 [33]	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 (Ours)	**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**

Table 1. Performance comparison on the KITTI360Pose benchmark [12].

Methods	Submap Retrieval Recall ↑					
	Validation Set			Test Set		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
Text2Pos [12]	0.14	0.28	0.37	0.12	0.25	0.33
RET [33]	0.18	0.34	0.44	-	-	-
Text2Loc (Ours)	**0.32**	**0.56**	**0.67**	**0.28**	**0.49**	**0.58**

Table 2. Performance comparison for gloabl place recognition on the KITTI360Pose benchmark [12]. Note that only values that are available in RET [33] are reported.

recognition performance on the KITTI360Pose validation and test set for a fair comparison. Table 2 shows the top-1/3/5 recall of each method. The best performance on the validation set reaches the recall of 0.32 at top-1. Notably, this outperforms the recall achieved by the current state-of-the-art method RET by a wide margin of **78%**. Furthermore, Text2Loc achieves recall rates of 0.56 and 0.67 at top-3 and top-5, respectively, representing substantial improvements of 65% and 52% relative to the performance of RET. These improvements are also observed in the test set, indicating the superiority of the method over baseline approaches. Note that we report only the values available in the original publication of RET. These improvements demonstrate the efficacy of our proposed Text2Loc in capturing cross-model local information and generating more discriminative global descriptors. More qualitative results are given in Section 6.2.

5.3.2 Fine localization

To improve the localization accuracy of the network, [12, 33] further introduce fine localization. To make the comparisons fair, we follow the same setting in [12, 33] to train our fine localization network. As illustrated in Table 1, we report the top-k ($k = 1/5/10$) recall rate of different error thresholds $\epsilon < 5/10/15m$ for comparison. Text2Loc achieves the top-1 recall rate of 0.37 on the validation set and 0.33 on the test set under error bound $\epsilon < 5m$, which are **95**% and **2** × higher than the previous state-of-the-art RET, respectively. Furthermore, our Text2Loc performs consistently better when relaxing the localization error constraints or increasing k. This demonstrates that Text2Loc can accurately interpret the text descriptions and semantically understand point clouds better than the previous state-of-the-art methods. We also show some qualitative results in Section 6.2.

6. Performance analysis

6.1. Ablation study

The following ablation studies evaluate the effectiveness of different components of Text2Loc, including both the text-submap global place recognition and fine localization.

Global place recognition. To assess the relative contribution of each module, we remove the frozen pre-trained large language model T5, hierarchical transformer with max-pooling (HTM) module in the text branch, and number encoder in the 3D submap branch from our network one by one. We also analyze the performance of the proposed text-submap contrastive learning. All networks are trained on the KITTI360Pose dataset, with results shown in Table. 3. Utilizing the frozen pre-trained LLM T5, we observed an approximate 8% increase in retrieval accuracy at top 1 on the test set. While the HTM notably enhances performance on the validation set, it shows marginal improvements on the test set. Additionally, integrating the number encoder has led to a significant 6% improvement in the recall metric at top 1 on the validation set. Notably, the performance on the validation/test set reaches 0.32/0.28 recall at top 1, exceeding the same model trained with the pairwise ranking loss by 52% and 40%, respectively, highlighting the superiority of the proposed contrastive learning approach.

Fine localization. To analyze the effectiveness of each proposed module in our matching-free fine-grained localization, we separately evaluate the Cascaded Cross-Attention Transformer (CCAT) and Prototype-based Map Cloning (PMC) module, denoted as Text2Loc_CCAT and Text2Loc_PMC. For a fair comparison, all methods utilize the same submaps retrieved from our global place recognition. The results are shown in Table. 4. Text2Pos* significantly outperforms the origin results of Text2Pos [12], indicating the superiority of our proposed global place recognition. Notably, replacing the matcher in Text2Pos [12] with our CCAT results in about 7% improvements at top 1 on the test set. We also observe the inferior performance of Text2Loc_PMC to the proposed method when interpreting only the proposed PMC module into the Text2Pos [12] fine

Methods	Submap Retrieval Recall ↑					
	Validation Set			Test Set		
	$k=1$	$k=3$	$k=5$	$k=1$	$k=3$	$k=5$
w/o T5	0.29	0.53	0.65	0.26	0.45	0.54
w/o HTM	0.30	0.54	0.65	**0.28**	0.48	0.57
w/o CL	0.21	0.42	0.53	0.20	0.36	0.45
w/o NE	0.30	0.52	0.63	0.27	0.47	0.56
Full (Ours)	**0.32**	**0.56**	**0.67**	**0.28**	**0.49**	**0.58**

Table 3. Ablation study of the global place recognition on KITTI360Pose benchmark. ”w/o T5” indicates replacing the frozen pre-trained T5 model with the LSTM in [12]. ”w/o HTM” indicates removing the proposed hierarchical transformer with max-pooling (HTM). ”w/o CL” indicates replacing the proposed contrastive learning with the widely used pairwise ranking loss. ”w/o NE” indicates reducing the number encoder in the instance encoder of 3D submap branch.

localization network. The results are consistent with our expectations since PMC can lead to the loss of object instances in certain submaps (See Supp.). Combining both modules achieves the best performance, improving the performance by 10% at top 1 on the test set. This demonstrates adding more training submaps by PMC is beneficial for our matching-free strategy without any text-instance matches.

6.2. Qualitative analysis

In addition to quantitative results, we show some qualitative results of two correctly point cloud localization from text descriptions and one failure case in Fig. 5. Given a query text description, we visualize the ground truth, top-3 retrieved submaps, and our fine localization results. In text-submap global place recognition, a retrieved submap is defined as positive if it contains the target location. Text2Loc excels in retrieving the ground truth submap or those near in most cases. However, there are instances where negative submaps are retrieved, as observed in (b) with the top 3. Text2Loc showcases its ability to predict more accurate locations based on positively retrieved submaps in fine localization. We also present one failure case in (c), where all retrieved submaps are negative. In these scenarios, our fine localization struggles to predict accurate locations, highlighting its reliance on the coarse localization stage. An additional observation is that despite their distance from the target location, all these negative submaps contain instances similar to the ground truth. These observations show the challenge posed by the low diversity of outdoor scenes, emphasizing the need for highly discriminative representations to effectively disambiguate between submaps.

6.3. Computational cost analysis

In this section, we analyze the required computational resources of our coarse and matching-free fine localization network regarding the number of parameters and time efficiency. For a fair comparison, all methods are tested on the KITTI360Pose test set with a single NVIDIA TITAN X (12G) GPU. Text2Loc takes 22.75 ms and 12.37 ms to obtain a global descriptor for a textual query and a submap respectively, while Text2Pos [12] achieves it in 2.31 ms and 11.87 ms. Text2Loc has more running time for the text query due to the extra frozen T5 (21.18 ms) and HTM module (1.57 ms). Our text and 3D networks have 13.65 M (without T5) and 1.84 M parameters respectively. For fine localization, we replace the proposed matching-free CCAT module with the text-instance matcher in [12, 33], denoted as Text2Loc_Matcher. From Table. 5, we observe that Text2Loc is nearly two times more parameter-efficient than the baselines [12, 33] and only uses their 5% inference time. The main reason is that the previous methods adopt Superglue [27] as a matcher, which resulted in a heavy and time-consuming process. Besides, our matching-free architecture prevents us from running the Sinkhorn algorithm [5]. These improvements significantly enhance the network's efficiency without compromising its performance.

Methods	Localization Recall ($\epsilon < 5m$) ↑					
	Validation Set			Test Set		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
Text2Pos [12]	0.14	0.36	0.48	0.13	0.33	0.43
Text2Pos*	0.33	0.65	0.75	0.30	0.58	0.67
Text2Loc_CCAT	0.32	0.64	0.74	0.32	0.60	0.70
Text2Loc_PMC	0.32	0.64	0.74	0.29	0.56	0.66
Text2Loc (Ours)	**0.37**	**0.68**	**0.77**	**0.33**	**0.61**	**0.71**

Table 4. Ablation study of the fine localization on the KITTI360Pose benchmark. * indicates the fine localization network from Text2Pose [12], and the submaps retrieved through our global place recognition. Text2Loc_CCAT indicates the removal of only the PMC while retaining the CCAT in our network. Conversely, Text2Loc_PMC keeps the PMC but replaces the CCAT with the text-instance matcher in Text2Pos.

Methods	Parameters (M)	Runtime (ms)	Localization Recall
Text2Loc_Matcher	2.08	43.11	0.30
Text2Loc (Ours)	**1.06**	**2.27**	**0.33**

Table 5. Computational cost requirement analysis of our fine localization network on the KITTI360Pose test dataset.

6.4. Robustness analysis

In this section, we analyze the effect of text changes on localization accuracy. For a clear demonstration, we only change one sentence in the query text descriptions, denoted as Text2Loc_modified. All networks are evaluated on the KITTI360Pose test set, with results shown in Table. 6. Text2Loc_modified only achieves the recall of 0.15 at top-1 retrieval, indicating our text-submap place recognition network is very sensitive to the text embedding. We also observe the inferior performance of Text2Loc_modified in the

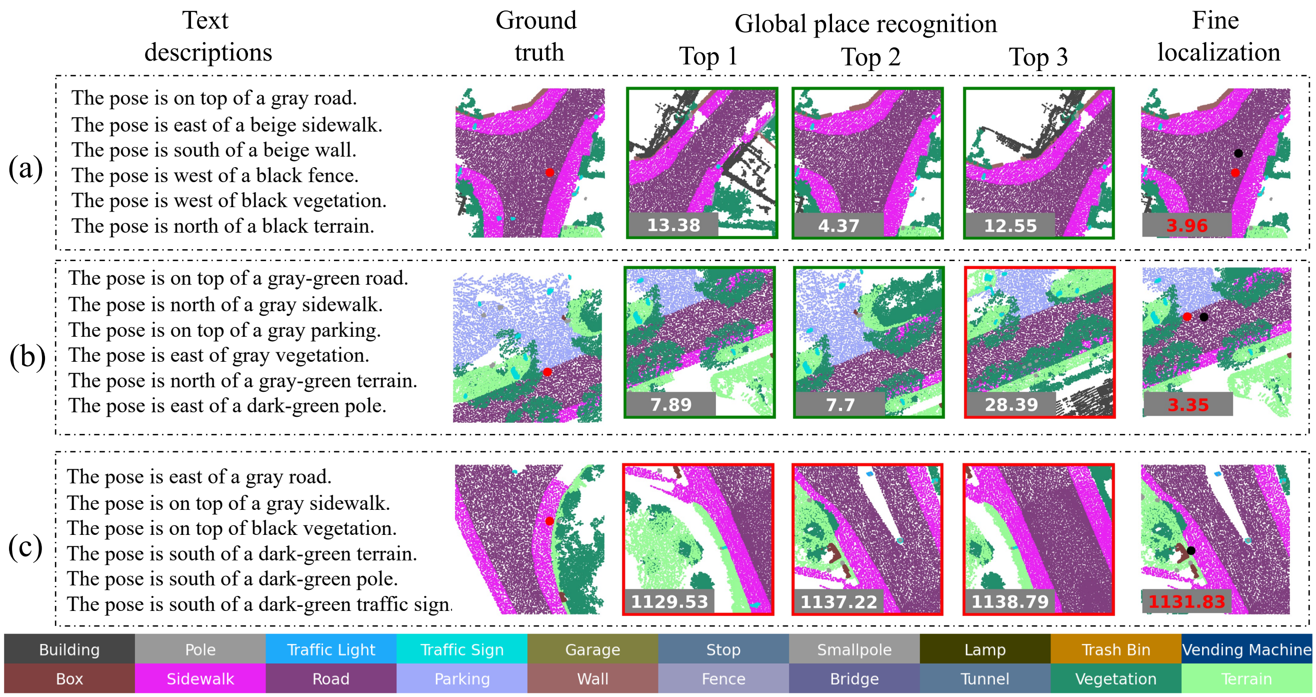

Text descriptions
Ground truth
Global place recognition
Top 1
Top 2
Top 3
Fine localization
(a)
The pose is on top of a gray road.
The pose is east of a beige sidewalk.
The pose is south of a beige wall.
The pose is west of a black fence.
The pose is west of black vegetation.
The pose is north of a black terrain.
13.38
4.37
12.55
3.96
(b)
The pose is on top of a gray-green road.
The pose is north of a gray sidewalk.
The pose is on top of a gray parking.
The pose is east of gray vegetation.
The pose is north of a gray-green terrain.
The pose is east of a dark-green pole.
7.89
7.7
28.39
3.35
(c)
The pose is east of a gray road.
The pose is on top of a gray sidewalk.
The pose is on top of black vegetation.
The pose is south of a dark-green terrain.
The pose is south of a dark-green pole.
The pose is south of a dark-green traffic sign.
1129.53
1137.22
1138.79
1131.83
Building
Pole
Traffic Light
Traffic Sign
Garage
Stop
Smallpole
Lamp
Trash Bin
Vending Machine
Box
Sidewalk
Road
Parking
Wall
Fence
Bridge
Tunnel
Vegetation
Terrain

Figure 5. Qualitative localization results on the KITTI360Pose dataset: In global place recognition, the numbers in top3 retrieval submaps represent center distances between retrieved submaps and the ground truth. Green boxes indicate positive submaps containing the target location, while red boxes signify negative submaps. For fine localization, red and black dots represent the ground truth and predicted target locations, with the red number indicating the distance between them.

Methods	Test set					
	Submap Retrieval Recall			Localization Recall		
	$k = 1$	$k = 3$	$k = 5$	$k = 5$ ($\epsilon < 5/10/15m$)		
Text2Loc_modified	0.15	0.30	0.38	0.39	0.54	0.58
Text2Loc (Ours)	**0.28**	**0.49**	**0.58**	**0.53**	**0.68**	**0.71**

Table 6. Performance comparisons of changing one sentence in the queries on the KITTI360Pose test set.

fine localization. More qualitative results are in the Supplementary Materials.

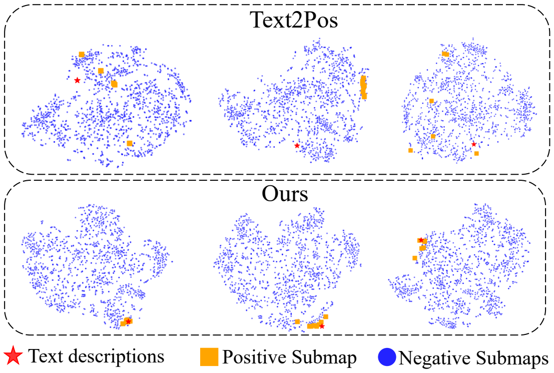


Figure 6. T-SNE visualization for the global place recognition.

6.5. Embedding space analysis

We employ T-SNE [31] to visually represent the learned embedding space, as illustrated in Figure 6. The baseline method Text2Pos [12] yields a less discriminative space, with positive submaps often distant from the query text descriptions and even scattered across the embedding space. In contrast, our method brings positive submaps and query text representations significantly closer together within the embedding distance. It shows that the proposed network indeed results in a more discriminative cross-model space for recognizing places.

7. Conclusion

We proposed Text2Loc for 3D point cloud localization based on a few natural language descriptions. In global place recognition, we capture the contextual details within and across text sentences with a novel attention-based method and introduce contrastive learning for the text-submap retrieval task. In addition, we are the first to propose a matching-free fine localization network for this task, which is lighter, faster, and more accurate. Extensive experiments demonstrate that Text2Loc improves the localization performance over the state-of-the-art by a large margin. Future work will explore trajectory planning in real robots.

Acknowledgements. This work was supported by the ERC Advanced Grant SIMULACRON, by the Munich Center for Machine Learning, and by the Royal Academy of Engineering (RF\201819\18\163).

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. 3

[2] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4470–4479, 2018. 2

[3] Tiago Barros, Luís Garrote, Ricardo Pereira, Cristiano Premebida, and Urbano J Nunes. Attdlnet: Attention-based deep network for 3d lidar place recognition. In *Iberian Robotics conference*, pages 309–320. Springer, 2022. 3

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

[5] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2013. 7

[6] Haowen Deng, Tolga Birdal, and Slobodan Ilic. Ppfnet: Global context aware local features for robust 3d point matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2018. 3

[7] Gil Elbaz, Tamar Avraham, and Anath Fischer. 3d point cloud registration for localization using a deep neural network auto-encoder. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2017. 3

[8] Zhaoxin Fan, Zhenbo Song, Hongyan Liu, Zhiwu Lu, Jun He, and Xiaoyong Du. Svt-net: Super light-weight sparse voxel transformer for large scale place recognition. AAAI, 2022. 3

[9] Mingtao Feng, Zhen Li, Qi Li, Liang Zhang, XiangDong Zhang, Guangming Zhu, Hui Zhang, Yaonan Wang, and Ajmal Mian. 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*, pages 3722–3731, 2021. 3

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[11] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, Lewei Lu, Xiaosong Jia, Qiang Liu, Jifeng Dai, Yu Qiao, and Hongyang Li. Planning-oriented autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[12] Manuel Kolmet, Qunjie Zhou, Aljoša Ošep, and Laura Leal-Taixé. Text2pos: Text-to-point-cloud cross-modal localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6687–6696, 2022. 2, 3, 4, 5, 6, 7, 8, 1

[13] Jacek Komorowski. Minkloc3d: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1790–1799, 2021. 3

[14] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 3

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 2

[16] Junyi Ma, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. Overlaptransformer: An efficient and yaw-angle-invariant transformer network for lidar-based place recognition. *IEEE Robotics and Automation Letters*, 7(3): 6958–6965, 2022. 3

[17] Junyi Ma, Guangming Xiong, Jingyi Xu, and Xieyuanli Chen. Cvtnet: A cross-view transformer network for place recognition using lidar data. *arXiv preprint arXiv:2302.01665*, 2023. 3

[18] Zhixiang Min, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Enrique Dunn, and Manmohan Chandraker. Neurocs: Neural nocs supervision for monocular 3d object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21404–21414, 2023. 1

[19] Mihir Prabhudesai, Hsiao-Yu Fish Tung, Syed Ashar Javed, Maximilian Sieb, Adam W Harley, and Katerina Fragkiadaki. Embodied language grounding with implicit 3d visual feature representations. 2019. 3

[20] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 2, 4

[21] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1655–1668, 2018. 3

[22] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4, 5

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 2, 3, 4

[24] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[25] Hasim Sak, Andrew W. Senior, and Françoise Beaufays. Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition. *CoRR*, abs/1402.1128, 2014. 2

[26] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical

localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12716–12725, 2019. 2

[27] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 7

[28] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21632–21642, 2023. 1

[29] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Efficient & effective prioritized matching for large-scale image-based localization. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1744–1756, 2016. 2

[30] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, page 435. Seattle, WA, 2009. 3

[31] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *JMLR*, 2008. 8

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[33] Guangzhi Wang, Hehe Fan, and Mohan Kankanhalli. Text to point cloud localization with relation-enhanced transformer. *arXiv preprint arXiv:2301.05372*, 2023. 2, 3, 4, 5, 6, 7

[34] Yan Xia. *Perception of vehicles and place recognition in urban environment based on MLS point clouds*. PhD thesis, Technische Universität München, 2023. 1

[35] Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, and Uwe Stilla. 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*, pages 11348–11357, 2021. 2, 4

[36] Yan Xia, Yusheng Xu, Cheng Wang, and Uwe Stilla. Vpc-net: Completion of 3d vehicles from mls point clouds. *ISPRS Journal of Photogrammetry and Remote Sensing*, 174:166–181, 2021. 1

[37] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, João F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8461–8472, 2023. 1, 3, 5

[38] Yan Xia, Qiangqiang Wu, Wei Li, Antoni B Chan, and Uwe Stilla. A lightweight and detector-free 3d single object tracker on point clouds. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 1

[39] Zhihao Yuan, Xu Yan, Yinghong Liao, Ruimao Zhang, Sheng Wang, Zhen Li, and Shuguang Cui. Instancerefer: Cooperative holistic understanding for visual grounding on point clouds through instance multi-level contextual referring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1791–1800, 2021. 3

[40] Wenxiao Zhang, Huajian Zhou, Zhen Dong, Qingan Yan, and Chunxia Xiao. Rank-pointretrieval: Reranking point cloud retrieval via a visually consistent registration evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 3

[41] Zhicheng Zhou, Cheng Zhao, Daniel Adolfsson, Songzhi Su, Yang Gao, Tom Duckett, and Li Sun. Ndt-transformer: Large-scale 3d point cloud localisation using the normal distribution transform representation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5654–5660. IEEE, 2021. 3

Text2Loc: 3D Point Cloud Localization from Natural Language

Supplementary Material

A. Overview

In this supplementary material, we provide more experiments on the KITTI360Pose dataset [12] to demonstrate the effectiveness of our Text2Loc and show more insights we gathered during the development. We first present thorough ablation experiments to study the impact of the proposed CCAT on the fine localization performance in Sec. B. In Sec. C, we provide qualitative results of top-3 candidate submaps retrieved and localization performance when changing one sequence in the query textural descriptions. Next, we describe implementation details about our network architecture in Sec. D and analysis of the proposed PMC module in Sec. E. Finally, Sec. F shows more visualizations of point cloud localization from text descriptions.

B. More analysis of Cascaded Cross-Attention Transformers

In this section, we first explore the performance of different numbers of Cascaded Cross-Attention Transformers (CCAT) in our fine localization network. We further provide a comparison to study the difference between our CCAT and Hierarchical Cross-Attention Transformer (HCAT) in [37].

Number of CCAT. We insert CCAT one by one before the MLP layer in Text2Loc. '0' means using a single Cross Attention Transformer (CAT) to fuse text and 3D point cloud features. Table 7 shows the localization performance of our Tex2Loc with different numbers of CCAT units. As seen from the table, Text2Loc achieves the best performance with 2 CCAT units. When the number expands to 3, the performance degrades. This implies that the text-submap feature fusion is sufficient with fewer CCAT units. On the other hand, when the number is set to 1, the performance decreases. Therefore, we set the fixed number of CCAT as 2 in our network.

Number of CCAT	Localization Recall ($\epsilon < 5m$) ↑					
	Validation Set			Test Set		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
0	0.28	0.57	0.66	0.26	0.51	0.60
1	0.36	0.67	0.77	0.32	0.59	0.69
2	**0.37**	**0.68**	**0.77**	**0.33**	**0.61**	**0.71**
3	0.35	0.67	0.77	0.32	0.59	0.69

Table 7. Localization performance for Text2Loc with different numbers of CCAT on the KITTI360Pose benchmark. '0' means using a single Cross Attention Transformer (CAT) to fuse text and 3D point cloud features.

Methods	Localization Recall ($\epsilon < 5m$) ↑					
	Validation Set			Test Set		
	$k=1$	$k=5$	$k=10$	$k=1$	$k=5$	$k=10$
HCAT [37]	0.35	0.66	0.75	0.32	0.59	0.68
CCAT (Ours)	**0.37**	**0.68**	**0.77**	**0.33**	**0.61**	**0.71**

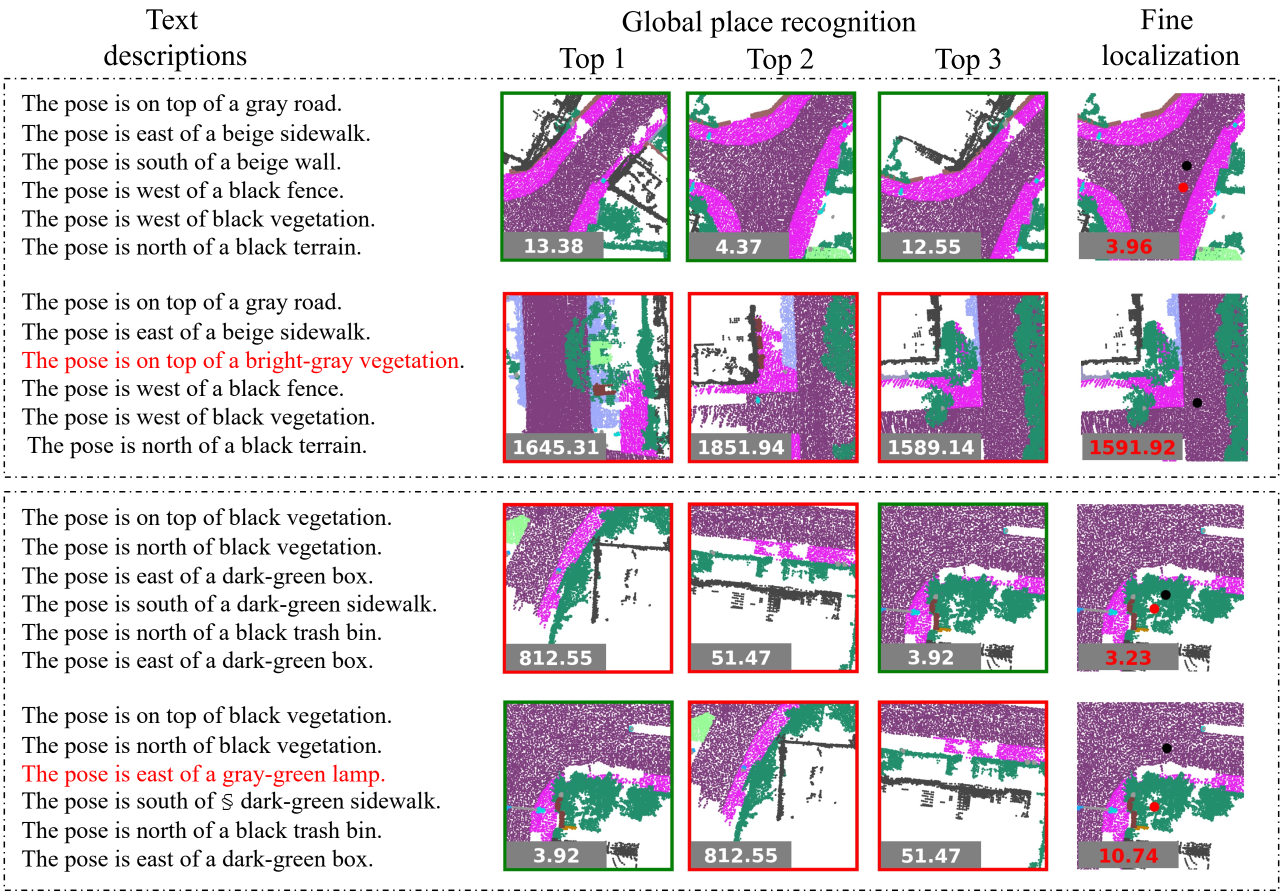
Table 8. Performance comparison of different modules within our Text2Loc architecture on the KITTI360Pose benchmark.

Difference with HCAT. Recent work CASSPR [37] has explored the integration of 3D point-wise features with voxelized representations through a designed Hierarchical Cross-Attention Transformer (HCAT). In HCAT, two parallel Cross Attention Transformers (CAT1 and CAT2) process inputs from different branches (point and voxel), each serving as query and key respectively. In contrast, our Cascaded Cross-Attention Transformer (CCAT) employs a sequential, cascaded structure to merge text and point cloud cross-modal information. Notably, in our CCAT, the second CAT utilizes the output of the first CAT as its key and value, distinguishing it from the parallel architecture of HCAT. Table. 8 presents a performance comparison of different modules within our Text2Loc architecture. Utilizing the proposed CCAT, we observed an approximate 4% increase in retrieval accuracy at top 10 on the test set. This table demonstrates a consistently superior performance of our CCAT compared to the HCAT used in [37].

Motivation of CCAT. The motivation for the CCAT module in fine localization arose from the challenge of target position regression based on the text descriptions. Encoding accurate textual features is crucial for regression since the model directly predicts target positions, without any text-instance matcher. We thus design a cascade structure to enhance text features with the information from retrieved point clouds. The HCAT [37] module, in contrast, aims to compensate for the quantization losses for the LiDAR-based place recognition task. HCAT should ensure that each branch is useful in isolation, thus preventing one branch from dominating over the other.

C. Visualization of robustness analysis

Fig. 7 visualizes some qualitative results for Sec. 6.4. For each instance, we display the original query text descriptions along with the top 3 retrieved submaps and their final predicted locations at the top, followed by modified queries (highlighted in red) and their results at the bottom. In the first example, we cannot find the positive submaps in the top-3 matches, leading to a complete localization failure. In the second example, even though we identify the positive



The pose is on top of a gray road.
The pose is east of a beige sidewalk.
The pose is south of a beige wall.
The pose is west of a black fence.
The pose is west of black vegetation.
The pose is north of a black terrain.

The pose is on top of a gray road.
The pose is east of a beige sidewalk.
The pose is on top of a bright-gray vegetation.
The pose is west of a black fence.
The pose is west of black vegetation.
The pose is north of a black terrain.

The pose is on top of black vegetation.
The pose is north of black vegetation.
The pose is east of a dark-green box.
The pose is south of a dark-green sidewalk.
The pose is north of a black trash bin.
The pose is east of a dark-green box.

The pose is on top of black vegetation.
The pose is north of black vegetation.
The pose is east of a gray-green lamp.
The pose is south of § dark-green sidewalk.
The pose is north of a black trash bin.
The pose is east of a dark-green box.

Figure 7. Robust analysis of our Text2Loc on the KITTI360Pose Benchmark. We present the top-3 retrieved submaps in global place recognition and the final predicted location for both the original query text descriptions and the modified queries (in red).

submaps in the global place recognition, the exact localization is still off. The results are consistent with our expectation that accurate text embedding is essential for predicting the target location in fine localization.

D. Implementation Details

We train the model with Adam optimizer for the text-submap global place recognition with a learning rate (LR) of 5e-4. The model is trained for a total 20 epochs with batch size 64, and we follow a multi-step training schedule wherein we decay LR by a factor of 0.4 at each 7 epoches. The temperature coefficient τ is set to 0.1. We consider each submap to contain a constant 28 object instances. The intra- and inter-text encoder in the text branch has 1 encoder layer respectively. We utilize PointNet++ [20] from [12] to encode every individual instance within the submap. In all quantitative results relating to global place recognition, we adopt the definition of the ground truth (GT) submap as [12], where it refers to the submap in the database that contains textual descriptions of targets, with its center point closest to the target. For the fine localization network, we train the model with an LR of 3e-4 for 35 epochs with batch size 32. To make a fair comparison, we set the embedding dimension for both text and submap branch as 256 in global place recognition and 128 in fine localization. The code is available for reproducibility.

Transformer in global place recognition. Formally, each transformer with max-pooling in the proposed intra- and inter-text encoder can be formulated as follows:

$$\begin{aligned} \mathbf{F}_T &= \text{Max-pooling} \circ \text{Transformer}(\mathbf{Q},\mathbf{K},\mathbf{V}) \\ &= \text{Max-pooling}\circ \left[\widetilde{\mathbf{F}}_T + \text{FFN}\left(\widetilde{\mathbf{F}}_T\right)\right], \\ \widetilde{\mathbf{F}}_T &= \mathbf{Q} + \text{MHSA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right), \end{aligned} \tag{6}$$

where $\mathbf{Q} = \mathbf{K} = \mathbf{V} = F_t \in \mathbb{R}^{N_t \times d}$ represent the query, key, and value matrices.

Within the MHSA layer, self-attention is conducted by projecting $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ using h heads, with our choice being $h = 4$. More precisely, we initially calculate the weight matrix using scaled dot-product attention [32], as in Eq. 7:

$$\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \text{Softmax}\left(\frac{\mathbf{Q}\cdot\mathbf{K}^{T}}{\sqrt{d_k}}\right)\mathbf{V}, \tag{7}$$

Subsequently, we compute the values for the h heads and concatenate them together as follows:

$$\text{Multi-Head}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \left[\,\text{head}1,\ldots,\text{head}h\right]\mathbf{W}^{O}, \tag{8}$$

$$\text{head}i = \text{Attention}\left(\mathbf{Q}\mathbf{W}i^{Q},\mathbf{K}\mathbf{W}i^{K},\mathbf{V}\mathbf{W}i^{V}\right), \tag{9}$$

where $\mathbf{W}_i^{Q,K,V,O}$ denote the learnable parameters.

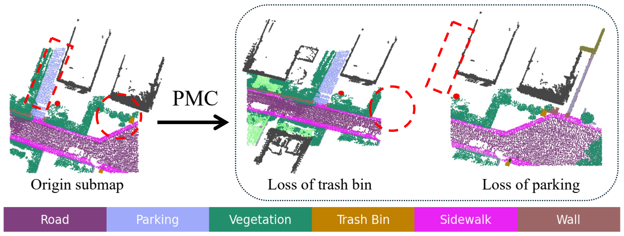


Figure 8. Visualization of lost instances due to our PMC.

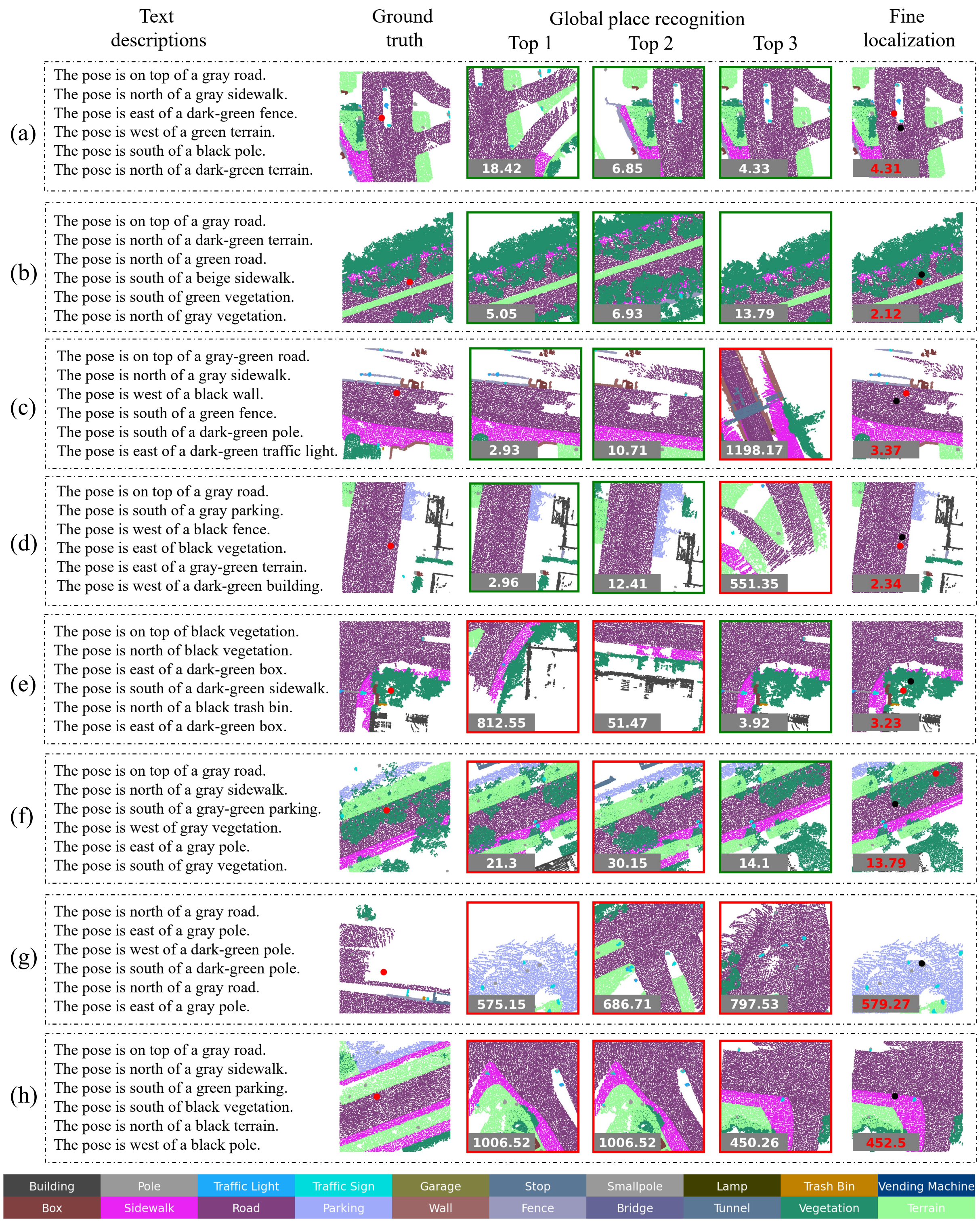
E. Analysis of PMC module

PMC can be seen as a data augmentation. However, this augmentation is not suitable for the previous text-instance matcher in Text2Pos [12] and RET [33] since PMC can lead to the loss of object instances in certain submaps (see Fig. 8 above); thereby, solely integrating the PMC into Text2Pos results in performance degradation. Conversely, adding more training submaps by PMC benefits our Text2Loc since we adopt a matching-free strategy without any text-instance matches.

F. More visualization results

In this section, we visualize more examples of correct point cloud localization from text descriptions and failure cases in Fig. 9. For (a) and (b), Text2Loc successfully retrieves all positive submaps within the top-3 results during global place recognition. We observe that these top-3 retrieved submaps display a high degree of semantic similarity to both the ground truth and each other. In cases of (c) - (e), despite some of the top-3 submaps being negatives retrieved by our text-submap place recognition, Text2Loc effectively localizes the text queries within a $5\,\mathrm{m}$ range after applying the fine localization network. It demonstrates our fine localization network can improve the localization recall, which turns such wrong cases in place recognition into a successful localization.

We also present some failure cases where all retrieved submaps are negative. For example, in case (g), the query text description contains an excessive number of objects of the same category ’Pole’. This description ambiguity poses a significant challenge to our place recognition network, leading to the retrieval of incorrect submaps. In the future, We hope to investigate more precise and accurate text descriptions, like integrating specific landmark information, including street names, zip codes, and named buildings, into text-based localization networks.



(a)
The pose is on top of a gray road.
The pose is north of a gray sidewalk.
The pose is east of a dark-green fence.
The pose is west of a green terrain.
The pose is south of a black pole.
The pose is north of a dark-green terrain.

(b)
The pose is on top of a gray road.
The pose is north of a dark-green terrain.
The pose is north of a green road.
The pose is south of a beige sidewalk.
The pose is south of green vegetation.
The pose is north of gray vegetation.

(c)
The pose is on top of a gray-green road.
The pose is north of a gray sidewalk.
The pose is west of a black wall.
The pose is south of a green fence.
The pose is south of a dark-green pole.
The pose is east of a dark-green traffic light.

(d)
The pose is on top of a gray road.
The pose is south of a gray parking.
The pose is west of a black fence.
The pose is east of black vegetation.
The pose is east of a gray-green terrain.
The pose is west of a dark-green building.

(e)
The pose is on top of black vegetation.
The pose is north of black vegetation.
The pose is east of a dark-green box.
The pose is south of a dark-green sidewalk.
The pose is north of a black trash bin.
The pose is east of a dark-green box.

(f)
The pose is on top of a gray road.
The pose is north of a gray sidewalk.
The pose is south of a gray-green parking.
The pose is west of gray vegetation.
The pose is east of a gray pole.
The pose is south of gray vegetation.

(g)
The pose is north of a gray road.
The pose is east of a gray pole.
The pose is west of a dark-green pole.
The pose is south of a dark-green pole.
The pose is north of a gray road.
The pose is east of a gray pole.

(h)
The pose is on top of a gray road.
The pose is north of a gray sidewalk.
The pose is south of a green parking.
The pose is south of black vegetation.
The pose is north of a black terrain.
The pose is west of a black pole.

Figure 9. Qualitative localization results on the KITTI360Pose dataset: In global place recognition, the numbers in top3 retrieval submaps represent center distances between retrieved submaps and the ground truth. Green boxes indicate positive submaps containing the target location, while red boxes signify negative submaps. For fine localization, red and black dots represent the ground truth and predicted target locations, with the red number indicating the distance between them.