Text to Point Cloud Localization with Relation-Enhanced Transformer

Guangzhi Wang[1], **Hehe Fan**[2], **Mohan Kankanhalli**[2]

[1]Institute of Data Science, National University of Singapore
[2]School of Computing, National University of Singapore
guangzhi.wang@u.nus.edu, hehe.fan@nus.edu.sg, mohan@comp.nus.edu.sg

Abstract

Automatically localizing a position based on a few natural language instructions is essential for future robots to communicate and collaborate with humans. To approach this goal, we focus on the text-to-point-cloud cross-modal localization problem. Given a textual query, it aims to identify the described location from city-scale point clouds. The task involves two challenges. 1) In city-scale point clouds, similar ambient instances may exist in several locations. Searching each location in a huge point cloud with only instances as guidance may lead to less discriminative signals and incorrect results. 2) In textual descriptions, the hints are provided separately. In this case, the relations among those hints are not explicitly described, leading to the difficulties of learning relations. To overcome these two challenges, we propose a unified Relation-Enhanced Transformer (RET) to improve representation discriminability for both point cloud and natural language queries. The core of the proposed RET is a novel Relation-enhanced Self-Attention (RSA) mechanism, which explicitly encodes instance (hint)-wise relations for the two modalities. Moreover, we propose a fine-grained cross-modal matching method to further refine the location predictions in a subsequent instance-hint matching stage. Experimental results on the KITTI360Pose dataset demonstrate that our approach surpasses the previous state-of-the-art method by large margins.

Introduction

Understanding natural language instructions in the 3D real world is a fundamental skill for future artificial intelligence assistants to collaborate with humans. In this paper, we focus on the outdoor environment and study the task of natural language-based localization from city-scale point clouds. As shown in Figure 1, given a linguistic description of a position, which contains several hints, the goal of the task is to find out the target location from a large-scale point cloud. This task can effectively help mobile robots, such as self-driving cars and autonomous drones, cooperate with humans to coordinate actions and plan their trajectories. By understanding the destination from natural language instructions, it reduces the human effort required for manual operation.

However, this task is intrinsically challenging. Precise localization requires both correct language interpretation and

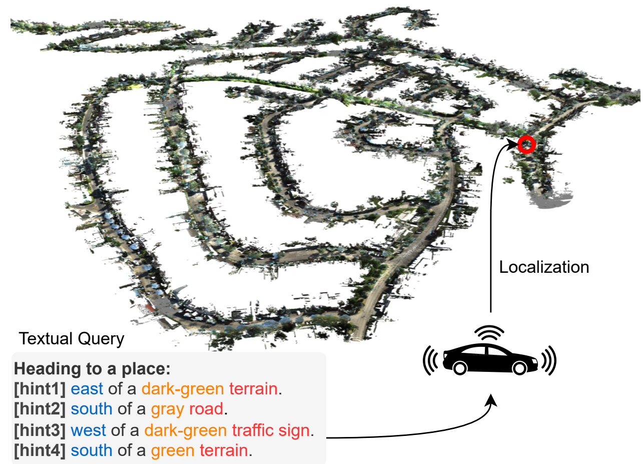

Figure 1: Illustration of the text to point cloud localization task. Given a textual query, which usually contains several independent hints, the goal is to localize the point of interest in a huge city-scale point cloud.

effective large-scale point cloud understanding. Considering the difficulties, an existing method (Kolmet et al. 2022) first divides a city-wide point cloud into several cells, and then solves this task in a *Coarse-to-Fine* manner.

The goal of the 'coarse' stage is to find out the target cell that contains the queried location according to the given natural language descriptions. In this stage, the instances included in point cloud cells and those mentioned in language descriptions are mainly used for text-to-point-cloud retrieval based on their types, without considering their relations. In the 'fine' stage, each object in the textual query is matched with an in-cell point cloud instance, whereby a target location will be predicted from each hint. This pioneering method sets up a significant starting point for tackling the challenging task. However, it fails to consider the intrinsic relations in both stages, resulting in sub-optimal performance.

For the coarse stage, because similar ambient instances may exist in several cells, performing retrieval based on only the cell-contained and query-related instance types without considering their relations may lead to low discriminability for both cell and query representations, which inevitably leads to ambiguity. Based on those low-discriminability representations, it is difficult to find out the correct cell. In the fine stage, we observe that insufficient *cross-modal collaboration* leads to difficulties in location refinement. Given the retrieved cell, precise location prediction requires joint understanding of both point clouds and textual queries. However, in the previous method (Kolmet et al. 2022), the cross-modal collaboration is only performed from textual queries to point clouds in a single step, which results in optimization difficulty for multi-task learning.

In this work, we aim to solve the aforementioned shortcomings in both stages. For the coarse stage, we propose to encode pairwise instance relations to improve representation discriminability for both modalities, which is achieved through a novel Relation-Enhanced Transformer (RET) architecture. In particular, the in-cell point cloud instance relations are modeled as their geometric displacements, while computed as the fusion of hint representations in the linguistic domain. These relations from two modalities are respectively incorporated into their representation in a unified manner, which is achieved through the proposed Relation-enhanced Self-Attention (RSA) mechanism. For the fine stage, we perform Cascaded Matching and Refinement (CMR) to enhance cross-modal collaboration. In particular, different from (Kolmet et al. 2022) which achieves this objective in a single step, we perform description-instance matching and position refinement in two sequential steps. Such formulation allows us to minimize the optimization difficulty of multi-objective learning and noisy intermediate results, thereby improving cross-modal collaboration.

We validated the effectiveness of our method on the KITTI360Pose benchmark (Kolmet et al. 2022). Extensive experiments demonstrate that the proposed method can surpass the previous approach by a large margin, leading to new state-of-the-art results. Our contributions are three-fold:

- We propose a novel Relation-Enhanced Transformer (RET) to improve representation discriminability for both point clouds and textual queries. The core component of RET is the Relation-enhanced Self-Attention (RSA) mechanism, which encodes instance (hint) relations for the two modalities in a unified manner.
- We propose to perform cross-modal instance matching and position refinement in two sequential steps. This formulation allows us to minimize the optimization difficulty of multi-task learning and the influence of noisy intermediate results, thereby improving cross-modal collaboration for fine-grained location prediction.
- We perform extensive experiments on the KITTI360Pose dataset (Kolmet et al. 2022). The results show that our approach can surpass previous method by a large margin, resulting in new state-of-the-art performance. Additional ablation studies further demonstrate the effectiveness of each component in the proposed method.

Related Work

Transformer and Attention Mechanism. Transformer and self-attention mechanism (Vaswani et al. 2017; Fan, Yang, and Kankanhalli 2021) has become increasingly popular in recent years. Although first proposed for natural language processing, with architectural adaptation, Transformer has been widely applied to many vision tasks including visual recognition (Dosovitskiy et al. 2020; Liu et al. 2021), object detection (Carion et al. 2020; Zhu et al. 2020) and semantic segmentation (Cheng, Schwing, and Kirillov 2021). Besides, the transformer-based architectures are also utilized to model cross-modal (*e.g.*, vision and language) relations (Tan and Bansal 2019; Lu et al. 2019; Li et al. 2019; Zhang et al. 2021; Li et al. 2022). In these architectures, the attention mechanism is widely employed to implicitly learn relations among the input tokens. Nevertheless, without explicit relation encoding, the vanilla Transformer can only encode relations implicitly with the help of positional encoding (Dosovitskiy et al. 2020). To facilitate better relation modeling, some works modulate the attention computation process by explicitly incorporating element relations. For example, (Wu et al. 2021) modified the attention mechanism via unified relative position bias to improve visual recognition. For object detection, spatial relations between bounding boxes are introduced to modulate the attention weights (Liu et al. 2022; Gao et al. 2021). For dynamic point cloud analysis, displacement between points (Fan, Yang, and Kankanhalli 2022) is utilized for point-specific attention computation. In this work, we propose to model relations for both point clouds and language queries by explicitly incorporating intra-modality relations in a unified manner.

Visual Localization. The task that is most related to ours is vision-based localization (Arandjelovic et al. 2016; Brachmann et al. 2017; Hausler et al. 2021), which is to estimate a pose based on an image or image sequence. Existing methods mostly solve this task in two stages (Sarlin et al. 2019; Sattler, Leibe, and Kobbelt 2016; Zhou et al. 2020). The first stage finds a subset of all images using image retrieval-based techniques (Arandjelovic et al. 2016; Hausler et al. 2021; Torii et al. 2015), while the second stage establishes pixel-wise correspondence between the query image and the retrieved one to predict the precise pose. In this work, we also study the task of localization in a coarse-to-fine manner, but differ from visual localization in that: 1) we try to infer the location from city-wide point clouds instead of images. 2) we try to estimate the pose from textual query rather than images. Compared to visual localization, our task requires multi-modal understanding and is more challenging to solve.

3D Language Grounding. As we humans live in a 3D world and communicate through natural language, recent work has begun to investigate the tasks on the cross-modal understanding of 3D vision and natural language. Among these tasks, the one that is most related to ours is 3D language grounding, which aims at localizing an object in point clouds from a given natural language query. For example, ScanRefer (Chen, Chang, and Nießner 2020) studies 3D language grounding from real-life in-door scenes. ReferIt3D (Achlioptas et al. 2020) studies a related task under a simpler setting, which assumes the object instances are segmented in advance. InstanceRefer (Yuan et al. 2021) improves previous methods by adopting a 3D panoptic segmentation backbone, utilizing multi-level visual context. Re-

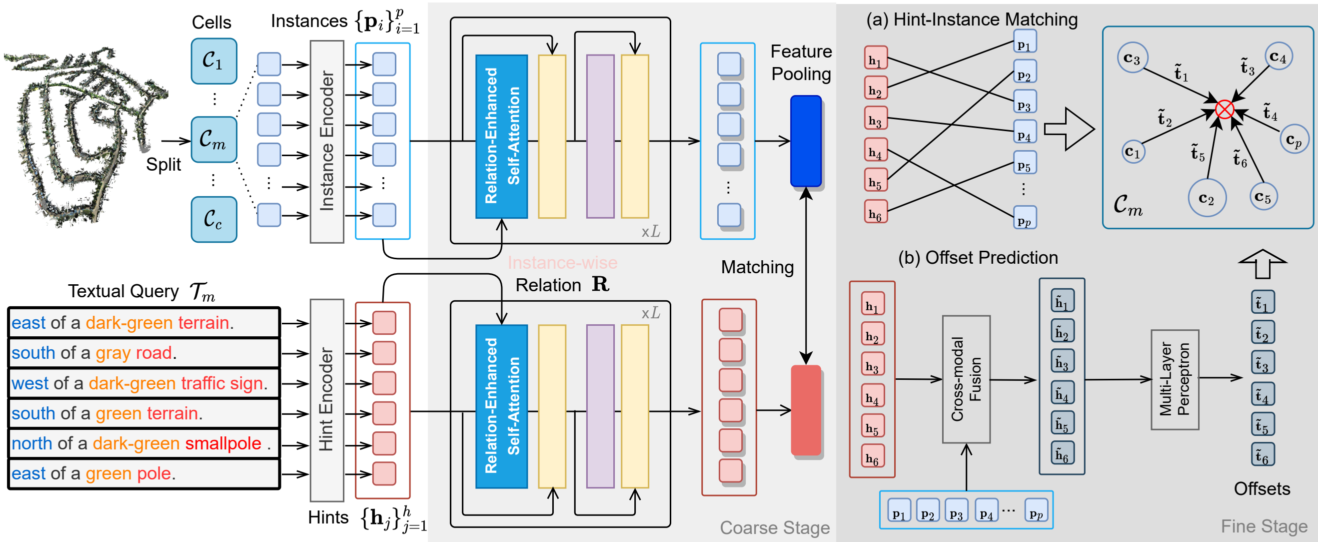

Figure 2: Framework of the proposed method. The city-scale point cloud is first divided into individual cells. Then, in the coarse stage, the cells and the textual query are respectively encoded with the proposed Relation-Enhanced Transformer (RET), which are later used for query-cell matching. In the fine stage, each hint is matched with an in-cell instance. Then, cross-modal fusion dynamically aggregates hints and instance representations for offset prediction. The target location is predicted based on matching results and offset predictions.

cently, graph structure (Feng et al. 2021) is also utilized to improve the representation learning qualities.

Methodology

Preliminaries

Given a textual query, our goal is to identify the position it describes from a city-scale point cloud. To handle the large-scale point cloud, we divide each scene into a set of cubic cells of fixed size by a preset stride. Each cell $\mathcal{C}$ contains a set of p point cloud instances, which are encoded by PointNet++ (Qi et al. 2017) into vector representations $\{\boldsymbol{p}_i\}_{i=1}^{p}$. Following (Kolmet et al. 2022), the textual query $\mathcal{T}$ is represented as a set of hints $\{\boldsymbol{h}_j\}_{j=1}^{h}$, each encoding the direction relation between the target location and an instance.

Inspired by the existing work (Kolmet et al. 2022), given the cell splits, we solve this task in a coarse-to-fine manner with two stages. The coarse stage is formulated as textual query based cell retrieval. The goal of this stage is to train a model that encodes $\mathcal{C}$ and $\mathcal{T}$ into a joint embedding space whereby matched query-cell pairs are close while those unmatched are pulled apart (Kiros, Salakhutdinov, and Zemel 2014). In the fine stage, given a retrieved cell, we aim to refine the position prediction by utilizing fine-grained cross-modal information. In particular, we first match each hint in the query with an in-cell instance by formulating it as an optimal transport problem (Liu et al. 2020). After that, with the matching results, we predict the target location through a cross-modal fusion of point cloud instance and hint representations. Based on the fused representation, we predict the target location for each matched instance. Finally, we obtain the target location prediction based on a weighted combination of the matching and location prediction results. The framework of our method is shown in Figure 2. In the following of this section, we will explain the proposed method for coarse stage and fine stage. After that, our training and inference procedure will be detailed.

Coarse Stage: Relation-Enhanced Transformer

After the cell split, the goal of the coarse stage is to successfully retrieve the cell $\mathcal{C}$ given a textual query $\mathcal{T}$. To approach this objective, we need to encode $\mathcal{C}$ and $\mathcal{T}$ into a joint embedding space. An intuitive solution is to encode both $\mathcal{C}$ and $\mathcal{T}$ based on the instances they contained as is done in (Kolmet et al. 2022). However, with such representations, the low discriminability for cells and textual queries results in poor retrieval performance. We argue that this can be attributed to the following two reasons. On the one hand, the outdoor scenes are often of low diversity, whereby a group of mentioned instances can appear at multiple different locations. Thus, simply describing a cell with its contained instances can result in less discriminative representations. On the other hand, the textual queries often contain limited clues compared to the point clouds, making this cross-modality retrieval especially challenging. To this end, we propose to explicitly encode instance-relations to provide more discriminative representations for both modalities.

The Transformer (Vaswani et al. 2017) has been widely utilized for relation-based representation learning in various tasks (Hu et al. 2018; Liu et al. 2021; Fan, Yang, and Kankanhalli 2022). The key component of the Transformer is the Self-Attention (SA) operation:

$$\texttt{Attn}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \texttt{Softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d})\boldsymbol{V}, \quad (1)$$

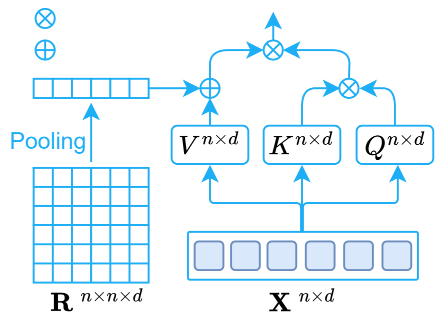

Figure 3: Illustration of the proposed Relation-enhanced Self-Attention (RSA) mechanism. Pairwise relations are explicitly encoded into the value computation process.

where d is the representation dimension and $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{N\times d}$ are the query, key and value matrices by transforming in-cell instances (or hints for textual queries) with corresponding linear transformations:

$$\boldsymbol{Q} = \boldsymbol{W}^Q\boldsymbol{X}, \boldsymbol{K} = \boldsymbol{W}^K\boldsymbol{X}, \boldsymbol{V} = \boldsymbol{W}^V\boldsymbol{X}, \quad (2)$$

with $\boldsymbol{W}^* \in \mathbb{R}^{d\times d}$ are learnable matrices and $\boldsymbol{X} = \boldsymbol{P} \in \mathbb{R}^{p\times d}$ or $\boldsymbol{H} \in \mathbb{R}^{h\times d}$ represents stacked instances[1].

Despite its generality, the vanilla SA lacks explicit relations in both modalities, thus is less informative to represent the cell and query. To this end, we propose a novel Relation-Enhanced Transformer (RET) to model explicit instance relations in both point clouds and textual descriptions. Our RET is a stack of multiple Transformer encoder layers, except that, in place of SA, we propose a Relation-enhanced Self-Attention (RSA) to explicitly incorporate relation information into value computation. The computation process is shown as follows and illustrated in Figure 3.

$$\texttt{RSA}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}, \boldsymbol{R}) = \texttt{Softmax}(\boldsymbol{Q}\boldsymbol{K}^T/\sqrt{d})(\boldsymbol{V}+\texttt{Pool}(\boldsymbol{R}, 1)), \quad (3)$$

where $\boldsymbol{R} \in \mathbb{R}^{N\times N\times d}$ captures pairwise relations with $\boldsymbol{R}_{ij} \in \mathbb{R}^d$ representing the relation between the i-th and j-th instance (hint). $\texttt{Pool}(\boldsymbol{R}, 1)$ indicates pooling tensor $\boldsymbol{R}$ along dimension 1. In this way, our model can explicitly encode instance relations through this computation process, leading to more informative representations.

The definition of relation varies flexibly with task objective and input modality. For point cloud data, we take the geometric displacement of two instances as their relations, as direction is often mentioned in textual queries and thus informative for retrieval:[2]

$$\boldsymbol{R}^V_{ij} = \boldsymbol{W}^V(\boldsymbol{c}_i - \boldsymbol{c}_j), \quad (4)$$

where $\boldsymbol{c}_i \in \mathbb{R}^3$ represents the center coordinate of the i-th instance and $\boldsymbol{W}^v \in \mathbb{R}^{d\times 3}$ transforms the displacement into embedding space. For the linguistic description, we compute the hint relation as the concatenation of their embeddings:

$$\boldsymbol{R}^L_{ij} = \boldsymbol{W}^L[\boldsymbol{h}_i; \boldsymbol{h}_j], \quad (5)$$

where $\boldsymbol{W}^L \in \mathbb{R}^{d\times 2d}$ transforms the linguistic feature into representation space. With the computation of RSA, the instance-wise relations for different modalities can be uniformly incorporated into query or cell representations

Finally, the cell (description) representations $\mathcal{C}_m$ ($\mathcal{T}_m$) are obtained via a pooling operation over all instances (hints) output from the RET for cross-modal retrieval.

Fine Stage: Cascaded Matching and Refinement

Following the coarse stage, we aim to refine the location prediction within the retrieved cell in the fine stage. Inspired by (Kolmet et al. 2022), we perform instance matching and location refinement to utilize the fine-grained visual and linguistic information, which involves the following two objectives: (1) For each hint, we find the in-cell instance it refers to via a matching process. (2) For each matched pair (i, j), a regressor predicts an offset $\hat{\boldsymbol{t}}_i \in \mathbb{R}^2$ for each matched hint $\boldsymbol{h}_j$, which represents the offset from the instance center $\boldsymbol{c}_i$ to the target location.[3]

Previous method (Kolmet et al. 2022) achieves the two objectives within a single step. However, given the objective of both hint-instance matching and offset prediction, the multi-task learning process introduces optimization difficulty. Furthermore, in the early training steps, the matcher is only partially trained, which produces noisy matching results. The regressor learns and makes predictions based on this noisy results, leading to unstable learning process and sub-optimal performance.

To this end, we propose a Cascaded Matching and Refinement (CMR) strategy for the fine stage, where hint-instance matching and offset regression are sequentially performed. Specifically, following (Kolmet et al. 2022), we first train the SuperGlue (Sarlin et al. 2020) matcher for hint-instance matching, which is formulated as an optimal-transport problem. Given the trained matcher, we obtain a set of hint-instance matching results $\{\boldsymbol{p}_i, \boldsymbol{h}_j, w_i\}_{j=1}^h$, where w_i represents the confidence of the match. Then, to reduce the noise for regression, we predict the target location according to matched instances only.

Precise location prediction requires proper understanding on both point cloud (what and where the referred instance is, *e.g.*, `dark-green terrain`) and language description (what is the relation between the matched instance and the target location, *e.g.*, `east of`). For this, we propose to facilitate cross-modal collaboration via the Cross-Attention (CA) mechanism, which is commonly used for cross-modality information fusion.

$$\texttt{CA}(\boldsymbol{H}, \boldsymbol{P}) = \texttt{Attn}(\boldsymbol{W}^Q\boldsymbol{H}, \boldsymbol{W}^K\boldsymbol{P}, \boldsymbol{W}^V\boldsymbol{P}), \quad (6)$$

where $\boldsymbol{H}, \boldsymbol{P}$ represent hints and instances, respectively, and $\boldsymbol{W}^*$ are learnable transformation matrices. Shortcut connection and layer normalization (Ba, Kiros, and Hinton 2016)

[1] Note that the attention operation is often performed in different subspaces with multiple heads, which is omitted for simplicity.

[2] We have also tried other features such as number of points and bounding boxes of instances but didn't observe performance improvement.

[3] For position prediction, we ignore the height information and considers 2D coordinates only.

Method	Localization Recall ($\epsilon < 5/10/15m$) ↑					
	Validation Set			Test Set		
	$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
Text2Pos (Kolmet et al. 2022)	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 (Ours)	**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**

Table 1: Performance comparison on the KITTI360Pose.

follows the cross-attention operation. With these operations, the hint representation $\boldsymbol{h}_i$ is accordingly updated to $\tilde{\boldsymbol{h}}_i$ by dynamically fusing visual information. As such, the information in the two modalities are joint utilized with the help of cross-modal collaboration.

Then, we predict the offset (the direction vector from instance center to target location) from the updated hint:

$$\hat{\boldsymbol{t}}_i = \texttt{MLP}(\tilde{\boldsymbol{h}}_j). \tag{7}$$

To utilize the matching results, the final prediction is obtained via a weighted combination of each hint's prediction:

$$\hat{\boldsymbol{g}} = \sum_i \frac{w_i}{\sum_m w_m}(\boldsymbol{c}_i + \hat{\boldsymbol{t}}_i), \tag{8}$$

where $w_i \in [0, 1]$ is the confidence score of the match $(\boldsymbol{p}_i, \boldsymbol{h}_j, w_i)$ and is set to 0 for non-matched instances. To filter out noisy matches, we consider only matches with confidence score greater than 0.2.

Training and Inference

Training. For the coarse stage, we train the proposed RET for cross-modal retrieval with pairwise ranking loss (Kiros, Salakhutdinov, and Zemel 2014):

$$\begin{aligned} \mathcal{L}_{coarse} = & \sum_{m=1}^{N_b} \sum_{n \neq m}^{N_b} [\alpha - \langle \mathcal{C}_m, \mathcal{T}_m \rangle + \langle \mathcal{C}_m, \mathcal{T}_n \rangle]_+ \\ & + \sum_{m=1}^{N_b} \sum_{n \neq m}^{N_b} [\alpha - \langle \mathcal{T}_m, \mathcal{C}_m \rangle + \langle \mathcal{T}_m, \mathcal{C}_n \rangle]_+, \end{aligned} \tag{9}$$

where N_b is the batch size, α is a hyper-parameter to control the separation strength and $\langle \cdot, \cdot \rangle$ represents inner product between vectors. This loss function encourages the representation of matched description-cell pair to be by a margin α closer than those unmatched. For the fine stage, we employ the loss in (Sarlin et al. 2020) to train the matcher, while L_2 loss is applied to train the offset regressor.

Inference. We first encode all cells and queries into a joint embedding space with the proposed Relation-Enhanced Transformer. Then, for each query representation, we retrieve top-k cells with highest similarity. For each retrieved cell, we use the SuperGlue matcher trained in the fine stage to match each hint with an in-cell instance, which is followed by offset prediction based on the fused representations. Finally, the position prediction is given by Eq. 8.

Experiments

Dataset and Implementation Details

Dataset Details. We evaluate our method on the recently proposed *KITTI360Pose* dataset (Kolmet et al. 2022), which is built upon the KITTI360 dataset (Liao, Xie, and Geiger 2021) with sampled locations and generated hints. It contains point clouds of a total of 9 scenes, covering 14,934 positions with a total area of $15.51km^2$. We follow (Kolmet et al. 2022) to use five scenes for training, one for validation, and the remaining three for testing. We sample the cells of size 30m with a stride of 10m. For more details on the dataset preprocessing, please refer to our supplementary material.

Implementation Details For the coarse stage, we trained the model with AdamW optimizer (Loshchilov and Hutter 2018) with a learning rate of 2e-4. The models are trained for a total of 18 epochs while the learning rate is decayed by 10 at the 9-th epoch. The α is set to 0.35. For the fine stage, we first train the matcher with a learning rate of 5e-4 for a total of 16 epochs. Afterwards, we fix the matcher and train the regressor based on the matching results for 10 epochs with a learning rate of 1e-4. The regressor is formulated as a 3 layer Multi-Layer Perceptron. Both of the two steps adopt an Adam (Kingma and Ba 2014) optimizer. The RET has 2 encoder layers for both point cloud part and linguistic part, each utilizing the Relation-enhanced Attention (RSA) mechanism with 4 heads and hidden dimension 2048. For the two stages, we encode each instance in the cell with PointNet++ (Qi et al. 2017) provided by Text2Pos (Kolmet et al. 2022) for a fair comparison. The hint representations are obtained by concatenating learned word embeddings. More details are provided in our appendix.[4]

Comparison with the State-of-the-art

We compared our method with Text2Pos (Kolmet et al. 2022) on the KITTI360Pose dataset. Following (Kolmet et al. 2022), we report top-k ($k = 1/5/10$) recall rate of different error ranges $\epsilon < 5/10/15m$ for comprehensive comparison. The results are shown in Table 1. Text2Pos gives a recall of 0.14 when $k = 1$ and $\epsilon < 5m$. In contrast, our method can significantly improve the recall rate to 0.19, which amounts to 35.7% relative improvement upon the baseline. Furthermore, when we relax the localization error constraints or increase k, consistent improvements upon the baseline can also be observed. For example, with $\epsilon < 5m$, our method achieves top-5 recall rate of 0.44, which is 8% higher than previous state-of-the-art. Similar improvements can also be seen on the test set, showing our method is superior to the baseline method.

[4]Code available at: https://github.com/daoyuan98/text2pos-ret

Method	$k = 1 \uparrow$	$k = 3 \uparrow$	$k = 5 \uparrow$
w/o both relations	0.11	0.24	0.32
w/o linguistic relation	0.14	0.28	0.37
w/o visual relation	0.16	0.30	0.40
Full (Ours)	**0.18**	**0.34**	**0.44**

Table 2: Ablation study of the Relation-Enhanced Transformer (RET) on KITTI360Pose validation set. "wo X relation" indicates replacing the proposed RSA with the vanilla Self-Attention in the corresponding modality.

Ablation Studies

In this section, we perform ablation studies for both stages to investigate the effectiveness of each proposed component in our method. The ablation studies for coarse stage and fine stage are provided separately for clear investigation.

Coarse Stage. We study the importance of explicit relation incorporation in the coarse stage. Since the coarse stage is formulated as a retrieval task, we use top-1/3/5 recall rate as evaluation metric, whereby the cell that contains the ground truth location is defined as positive.

Relation Incorporation. We first study the necessity of explicit relation modeling for both point cloud and textual queries. The results are shown in Table 2. It can be observed that relation modeling contributes significantly to successful retrieval. In particular, without any relation incorporation, the top-5 recall rate is 0.32. With the explicit fusion of linguistic relation, we observe an increase of 0.05 recall rate under same condition. Besides, with the incorporation of visual (point cloud instance) relations only, the top-5 recall rate can be improved by 0.08, indicating explicit relations in the point clouds play a more important role. Finally, with both relations, we achieve an improvement of 0.12 at top-5 recall rate upon that without any relation, showing that both visual and linguistic relations are necessary and complementary to improve the cell retrieval performance.

RET Hyper-parameters. We also studied the importance of the hyper-parameters involved in RET, namely the number of layers of RET and the number of heads of RSA. The results are shown in Table 3. It can be observed that, thanks to the strong relation modeling capacity of the proposed RET, we can obtain the best performance with 2 layers and 4 heads in the RSA. Decreasing and increasing the number of layers both lead to worse performance, which may be attributed to underfitting and overfitting, respectively.

Fine Stage. The objective of the fine stage is to correctly *match* linguistic hints and point cloud instances and *regress* the target location. Thus, we study the performance of the *matcher* and *regressor*, respectively.

Matcher. Following (Sarlin et al. 2020), we take precision and recall as the the evaluation metric of the matcher. With an identical matcher architecture, we investigate the impact of training strategy on the matcher performance. The results are shown in Table 4. It can be seen that compared with joint training (Kolmet et al. 2022), our cascaded training achieves not only high precision and recall in the training set, but also stronger generalization on the validation set. The results demonstrate that the cascade training strategy is able to mitigate the multi-task optimization difficulty.

#Layers	#Heads	$k = 1 \uparrow$	$k = 3 \uparrow$	$k = 5 \uparrow$
1	4	0.16	0.31	0.40
1	8	0.16	0.30	0.40
2	2	0.17	0.32	0.42
2	4	**0.18**	**0.34**	**0.44**
2	8	0.16	0.31	0.40
3	4	0.16	0.32	0.39
3	8	0.15	0.29	0.37

Table 3: The effects of #layers of RET and #heads of RSA.

Strategy	Train		Validation	
	Precision ↑	Recall ↑	Precision ↑	Recall ↑
joint	98.12	98.16	86.67	87.59
cascade(ours)	**98.89**	**99.04**	**92.18**	**93.01**

Table 4: Comparison of training strategy and matcher performance on the KITTI360Pose dataset.

Regressor. The regressor predicts the target location based on the the matching results. We study the effects of cascaded training, cross-attention based cross-modal fusion and confidence weighting for final location prediction. We use regression error as evaluation metric and compare different versions on both KITTI360Pose training and validation set. The results are shown in Table. 5. Without cascaded training strategy, the regressor achieves an error of 10.24 and 10.01 on the training and validation set, respectively, which is 1.72 and 0.86 higher than that with cascaded training. This result suggests that our cascaded training strategy also alleviates the optimization difficulty of the regressor, which was caused by the noisy intermediate results. Furthermore, without cross-attention mechanism, the regression error also increases by a considerable margin, showing that cross-modal collaboration is important for precise location prediction. Finally, with confidence-based weighting, we can further reduce the regression error on both the training and validation set, suggesting this information from the trained matcher can be further utilized to improve performance.

Visualizations

Embedding Space Visualization. We visualize the learned embedding space via T-SNE (Van der Maaten and Hin-

Method	Train Error ↓	Validation Error ↓
w/o cascade training	10.24 (+1.72)	10.01 (+0.86)
w/o cross-attention	9.57 (+1.05)	9.56 (+0.41)
w/o confidence weighting	9.02 (+0.50)	9.23 (+0.08)
Ours	**8.52**	**9.15**

Table 5: Ablation study on the regression error of the fine-stage on the KITTI360Pose dataset.

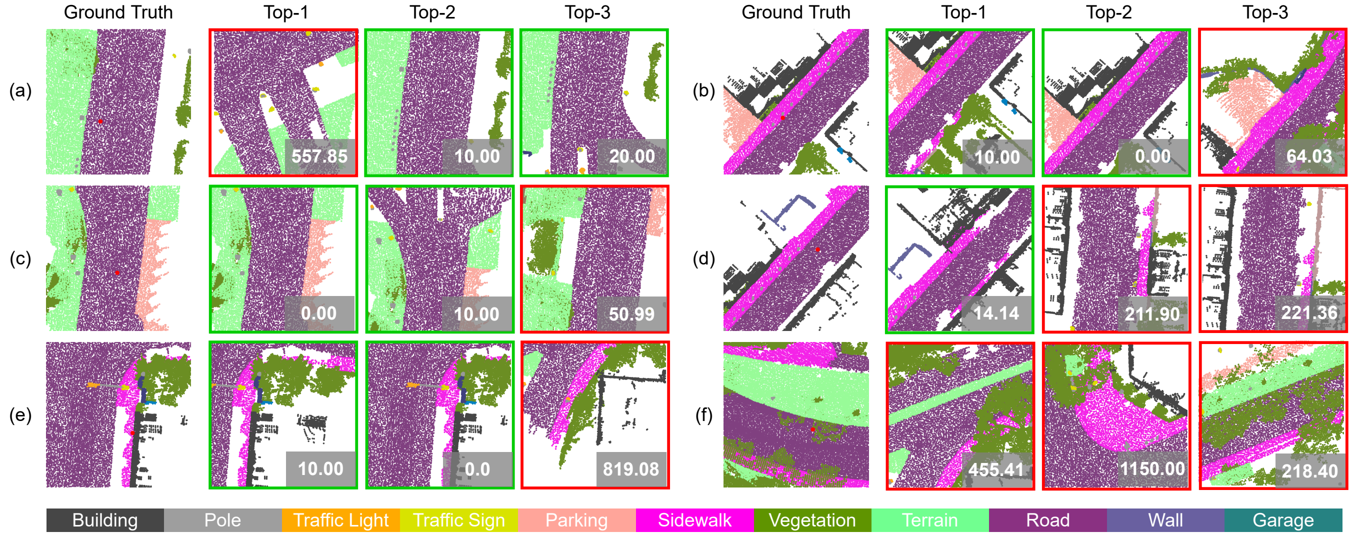

Figure 4: Qualitative retrieval results on KITTI360Pose validation set. The red dot in the ground truth cell indicates the target location. In each retrieved cell, the number in the lower right indicates the center distance between this cell and the ground truth. Green box indicates positive cell which contains the target location, while red box indicates negative cells.

ton 2008) in Figure 5. It can be observed that the baseline method Text2Pos (Kolmet et al. 2022) results in a less discriminative space, where positive cells are relatively far away from the query and sometimes separated across the embedding space. In contrast, our method draw positive cell and query representations closer in the embedding space, resulting in a more informative embedding space for retrieval.

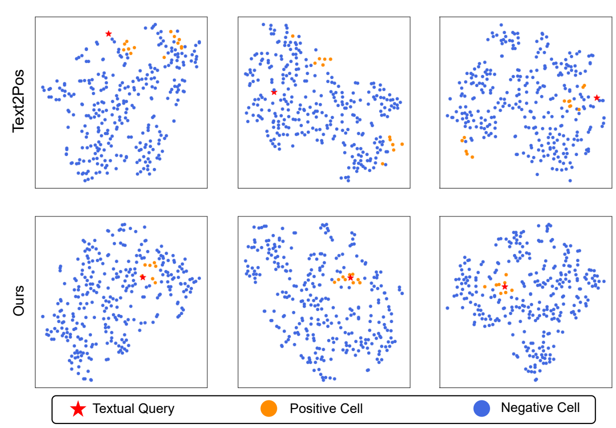

Figure 5: T-SNE visualization of embedding space for the coarse stage. A cell is considered as positive if it contains the location described by the query. Compared with baseline method (Kolmet et al. 2022), our method can produce better representation where positive cells are closer to the target.

Qualitative Cell Retrieval Results. We show some example text to point cloud retrieval results in Figure. 4. For a given query, we visualize the top-3 retrieved cells. A retrieved cell is defined as positive if it contains the target location. It can be observed that, our method can retrieve the ground truth cell or those close in most cases. Sometimes, negative cells can also be retrieved, *e.g.*, top-1 in (a) and top-3 in (e). It can be seen that these retrieved negative cells exhibit high semantic similarity with the ground truth cell, even though far away from it. We also show a failure case (f), where the retrieved cells are all negative. It can be seen that even though far away from the target location, all these negative cells have instances similar to the ground truth. These observations suggest that outdoor scenes are indeed of low diversity, indicating that successful retrieval requires highly discriminative representations to disambiguate the cells.

Conclusion

In this work, we proposed a novel method for precise text-based localization from large-scale point clouds. Our method employs a coarse-to-fine principle and pipelines this process into two stages. For the coarse stage which is formulated as a textual query based cell retrieval task, we aim to improve representation discriminability for both point cloud and query representations. This is achieved through explicit modeling of instance relations and implemented via a newly proposed Relation-Enhanced Transformer (RET). The core of RET is a novel Relation-enhanced Self-Attention (RSA) mechanism, whereby the instance relations for the two modalities are explicitly incorporated into the value computation process in a unified manner. For the fine stage, our method performs description-instance matching and position refinement in a cascaded way, whereby cross-modal information collaboration is enhanced through the cross-attention mechanism. Extensive experiments on the KITTI360Pose dataset validated the effectiveness of the proposed method, which achieves new state-of-the-art performance. Additional ablation studies further corroborate the effectiveness of each component in the proposed method.

Acknowledgements

This research is supported by the National Research Foundation, Singapore under its Strategic Capability Research Centres Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

References

Achlioptas, P.; Abdelreheem, A.; Xia, F.; Elhoseiny, M.; and Guibas, L. 2020. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *ECCV*. Springer.

Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*.

Ba, J. L.; Kiros, J. R.; and Hinton, G. E. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Brachmann, E.; Krull, A.; Nowozin, S.; Shotton, J.; Michel, F.; Gumhold, S.; and Rother, C. 2017. Dsac-differentiable ransac for camera localization. In *CVPR*.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-end object detection with transformers. In *ECCV*. Springer.

Chen, D. Z.; Chang, A. X.; and Nießner, M. 2020. Scanrefer: 3d object localization in rgb-d scans using natural language. In *ECCV*.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per-pixel classification is not all you need for semantic segmentation. *NeurIPS*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fan, H.; Yang, Y.; and Kankanhalli, M. 2022. Point spatio-temporal transformer networks for point cloud video modeling. *TPAMI*.

Fan, H.; Yang, Y.; and Kankanhalli, M. S. 2021. Point 4D Transformer Networks for Spatio-Temporal Modeling in Point Cloud Videos. In *CVPR*.

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 *ICCV*.

Gao, P.; Zheng, M.; Wang, X.; Dai, J.; and Li, H. 2021. Fast Convergence of DETR With Spatially Modulated Co-Attention. In *ICCV*.

Hausler, S.; Garg, S.; Xu, M.; Milford, M.; and Fischer, T. 2021. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *CVPR*.

Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2018. Relation Networks for Object Detection. In *CVPR*.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kiros, R.; Salakhutdinov, R.; and Zemel, R. S. 2014. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*.

Kolmet, M.; Zhou, Q.; Osep, A.; and Leal-Taixe, L. 2022. Text2Pos: Text-to-Point-Cloud Cross-Modal Localization. In *CVPR*.

Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019. Entangled Transformer for Image Captioning. In *ICCV*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*.

Liao, Y.; Xie, J.; and Geiger, A. 2021. KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D. *arXiv preprint arXiv:2109.13410*.

Liu, S.; Li, F.; Zhang, H.; Yang, X.; Qi, X.; Su, H.; Zhu, J.; and Zhang, L. 2022. DAB-DETR: Dynamic Anchor Boxes are Better Queries for DETR. In *ICLR*.

Liu, Y.; Zhu, L.; Yamada, M.; and Yang, Y. 2020. Semantic Correspondence as an Optimal Transport Problem. In *CVPR*.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *ICLR*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *NeurIPS*.

Qi, C. R.; Yi, L.; Su, H.; and Guibas, L. J. 2017. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*.

Sarlin, P.-E.; Cadena, C.; Siegwart, R.; and Dymczyk, M. 2019. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*.

Sarlin, P.-E.; DeTone, D.; Malisiewicz, T.; and Rabinovich, A. 2020. Superglue: Learning feature matching with graph neural networks. In *CVPR*.

Sattler, T.; Leibe, B.; and Kobbelt, L. 2016. Efficient & effective prioritized matching for large-scale image-based localization. *TPAMI*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP-IJCNLP*.

Torii, A.; Arandjelovic, R.; Sivic, J.; Okutomi, M.; and Pajdla, T. 2015. 24/7 place recognition by view synthesis. In *CVPR*.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *JMLR*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *NeurIPS*.

Wu, K.; Peng, H.; Chen, M.; Fu, J.; and Chao, H. 2021. Rethinking and improving relative position encoding for vision transformer. In *ICCV*.

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 *ICCV*.

Zhang, H.; Sun, A.; Jing, W.; Nan, G.; Zhen, L.; Zhou, J. T.; and Goh, R. S. M. 2021. Video Corpus Moment Retrieval with Contrastive Learning. In *SIGIR*.

Zhou, Q.; Sattler, T.; Pollefeys, M.; and Leal-Taixe, L. 2020. To learn or not to learn: Visual localization from essential matrices. In *ICRA*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2020. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In *ICLR*.