Partially Matching Submap Helps: Uncertainty Modeling and Propagation for Text to Point Cloud Localization

Supplementary Material

Mingtao Feng[1] Longlong Mei[1] Zijie Wu[1*] Jianqiao Luo[2*] Fenghao Tian[1]
Jie Feng[1] Weisheng Dong[1] Yaonan Wang[2]
[1]Xidian University [2]Hunan University

1. Datasets

KITTI360Pose. KITTI360Pose dataset consists of 3D point cloud scenes from 9 urban areas, containing 43,381 paired descriptions and positions and covering a total area of 15.51 km^2. We use five regions for training, one region for validation, and the remaining three regions for testing. For more details, please refer to the supplementary material in [2]. In our training set, we supplement the submaps corresponding to the location descriptions with partially matching submaps. However, the validation and test sets remain unaltered, without additional processing. Details on the selection criteria for partially matching submaps can be found in the main text.

CityRefer. CityRefer dataset [3] is a large-scale dataset built upon SensatUrban [1], featuring manually annotated city-level 3D scene descriptions. This dataset consists of photogrammetric point clouds of two UK cities covering 6 km^2 of the city landscape and provides natural language descriptions for 35,196 3D objects. Following a clipping strategy similar to [2], we partition the map into submaps and designate fully matching and partially matching submaps for each query text. The dataset is split into 23,586 pairs for training, 5,934 pairs for validation, and 5,676 pairs for testing.

2. Implement Details

Our model is trained using the Adam optimizer. For the fine-grained localization network, the model is trained for 35 epochs with a learning rate of 3e-4 and a batch size of 32, decaying the learning rate by a factor of 0.5 every 10 epochs. The constant term in the function $\mathcal{L}_{reg}^{'}$ is excluded. The coarse place recognition phase consists of training the model for 20 epochs with a learning rate of 5e-4 and a batch size of 32. A multi-step training schedule is adopted, where the learning rate decays by a factor of 0.4 every 7 epochs. The temperature coefficient τ is set to 0.1. For the hyperparameters in $\mathcal{L}_{iou}^{'}$, we set α, β and γ to 0.25, 250, and 0.028, respectively. Each submap is assumed to contain a fixed 28 object instances. To make a fair comparison, we set the embedding dimension for both text and submap branches as 256 in coarse place recognition and 128 in fine-grained localization.

Training strategy. A key distinction of our approach from existing SOTA methods is that we incorporate both partially matching and fully matching samples in a coarse-to-fine training framework. In the coarse place recognition stage, we optimize the model using a combination of $\mathcal{L}_{con}$ and $\mathcal{L}_{iou}^{'}$, In the fine-grained localization stage, we optimize the model using $\mathcal{L}_{reg}^{'}$.

Architectural components of the network. Following Text2Loc, we employ the T5 pre-trained model [5] for text feature extraction and PointNet++ [4] for point cloud encoding. In the coarse place recognition stage, the global feature of the query text T is extracted via a two-layer max-pooling operation, while a single max-pooling layer obtains the global feature of the semantic submap. These representations are then optimized through contrastive learning. For fine-grained localization, hint features from the query text T are extracted using a single max-pooling layer and fused with semantic submap instance features via a cascaded Cross Attention Transformer, generating a fused representation for regression-based localization.

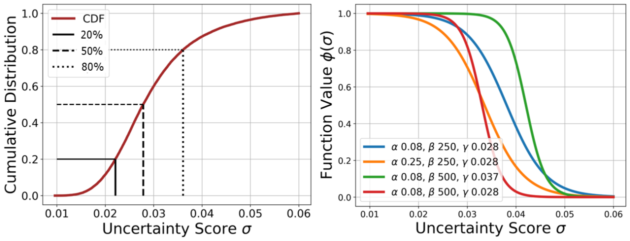

Figure 1. The cumulative distribution of uncertainty score over the partially matching samples on KITTI360Pose dataset(left), and the curves of decaying exponential function $\phi(\sigma)$ using different hyper-parameter combinations

[*] Corresponding author.

α β γ	Top-1	Top-3	Top-5
0.25, 285, 0.028	0.33	0.55	0.64
0.50, 285, 0.028	0.31	0.51	0.60
0.50, 250, 0.028	0.32	0.54	0.63
0.25, 250, 0.022	0.32	0.53	0.62
0.25, 250, 0.037	0.31	0.52	0.60
0.25, 250, 0.028	**0.34**	**0.56**	**0.65**

Table 1. Comparisons of retrieval performance using different hyper-parameter combinations for the decaying exponential function $\phi(\sigma)$

3. Hyper-parameter Setting.

Based on the definition in Equ.6, we design an uncertainty-aware similarity metric to enable effective similarity assessment within the feature embedding space, utilizing a decaying exponential function $\phi(\sigma)$ to regulate similarity weighting. In this section, we describe the methodology for selecting the three hyperparameters α, β, and γ in $\phi(\sigma)$. Fig. 1 (right) visualizes function curves under different parameter settings. From these curves, we observe that the function $\phi(\sigma)$ exhibits a monotonic decrease with increasing uncertainty score σ, indicating that high-uncertainty samples have less influence on model training. The parameter γ controls the uncertainty threshold beyond which samples are suppressed. As indicated by the green curve with $\gamma = 0.037$, samples with uncertainty higher than 0.037 will exert little influence for training, meaning that 20% of the data is suppressed. Reducing γ makes more samples to be suppressed and the proportion of the suppressed samples can be visualized through the cumulative distribution of the uncertainty, as illustrated on the left of Fig. 1. The parameters α and β control the smoothness of the function curve, as reflected in the differences among the red and orange curves. Specifically, increasing α and decreasing β lead to a smoother curve.

In Table 1, we report the quantitative retrieval performance corresponding to different function curves. The results show that compared to $\gamma = 0.022$ or $\gamma = 0.028$, setting $\gamma = 0.028$, which retains 50% of the samples, achieves optimal performance. This finding highlights the importance of mitigating cross-modal ambiguity introduced by partially matching samples. Furthermore, we observe that increasing α and decreasing β consistently improve performance, suggesting that a smoother curve contributes to better model generalization. Based on these observations, we set α, β, and γ to 0.25, 250 and 0.028, respectively. It is important to note that the value of γ varies across the KITTI360Pose and CityRefer datasets, as the quantile threshold is dataset-dependent, determined by the cumulative distribution specific to each dataset.

4. Embedding Space Analysis.

Fig. 2 illustrates the learned embedding space using T-SNE, comparing our method with Text2Loc in the coarse place recognition stage. As observed, in Text2Loc, the features of query texts remain relatively distant from their corresponding fully matching submap features. Additionally, in some cases, the feature distance between partially matching samples is even smaller than that between fully matching samples, which is consistent with the findings from the qualitative analysis presented in the main text. In contrast, our method effectively reduces intra-space distances for both fully matching and partially matching samples, while also preserving the distinction between them. This results in a more structured and discriminative embedding space. Moreover, these findings highlight the complementary role of partially matching sample in refining spatial retrieval and representation learning.

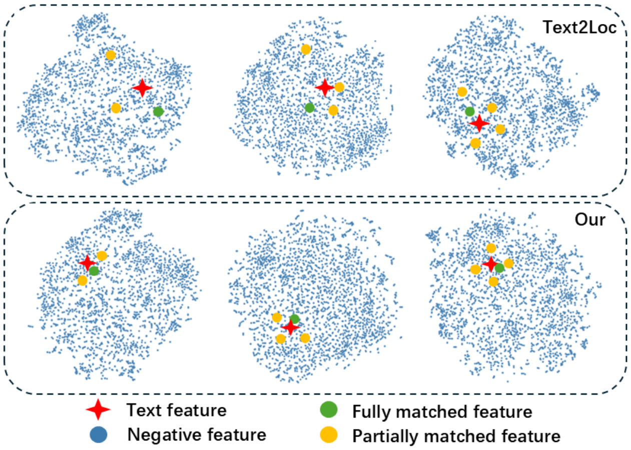

Figure 2. T-SNE visualization for the text query features and submap features in the coarse place recognition stage.

5. More Visualization Results

In this section, we provide additional visualizations of point cloud localization from query text in Fig. 3. Unlike the visualizations in the main text, Fig. 3 highlights the advantage of our method more clearly by using green boxes to indicate both partially matching and fully matching submaps.

In the coarse place recognition stage, our method successfully retrieves a submap containing the ground truth within the top-3 results in most cases. As observed in the first three rows, when both partially matching and fully matching submaps are retrieved, our method ranks the fully matching submap higher than the partially matching one. This suggests that our uncertainty-aware assignment loss enables the model to learn more information about submaps containing the ground truth, and differentiate feature information between partially matching samples and fully matching samples, improving retrieval accuracy. In the last three rows, the model fails to retrieve a fully matching submap within the top-3 results. This issue is particularly evident in the last row, where no submap containing the ground truth is found. The primary reason for these failures is that the point cloud instances mapped from the

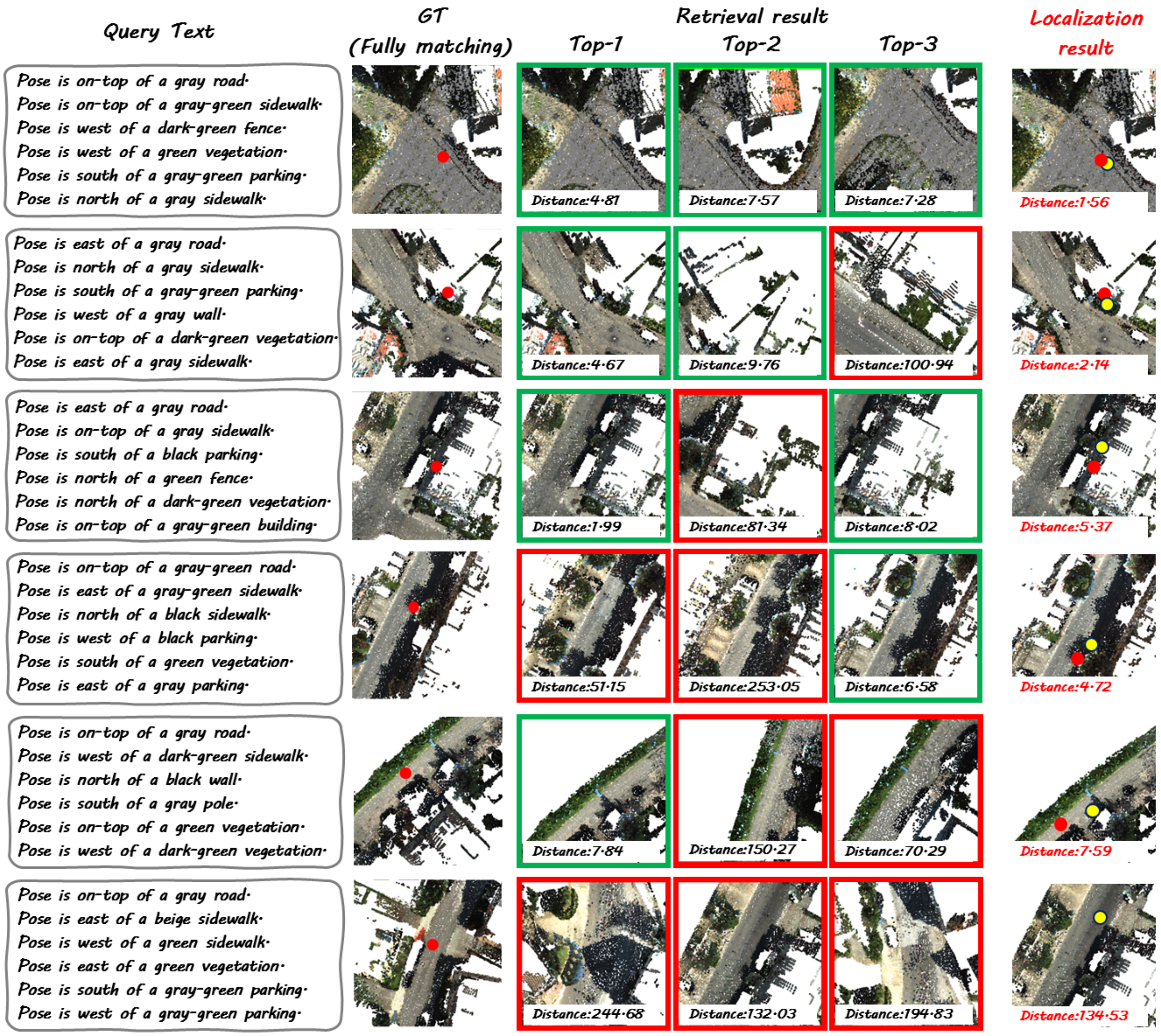
Query Text
GT (Fully matching)
Retrieval result
Top-1
Top-2
Top-3
Localization result
Pose is on-top of a gray road.
Pose is on-top of a gray-green sidewalk.
Pose is west of a dark-green fence.
Pose is west of a green vegetation.
Pose is south of a gray-green parking.
Pose is north of a gray sidewalk.
Distance:4.81
Distance:7.57
Distance:7.28
Distance:1.56
Pose is east of a gray road.
Pose is north of a gray sidewalk.
Pose is south of a gray-green parking.
Pose is west of a gray wall.
Pose is on-top of a dark-green vegetation.
Pose is east of a gray sidewalk.
Distance:4.67
Distance:9.76
Distance:100.94
Distance:2.14
Pose is east of a gray road.
Pose is on-top of a gray sidewalk.
Pose is south of a black parking.
Pose is north of a green fence.
Pose is north of a dark-green vegetation.
Pose is on-top of a gray-green building.
Distance:1.99
Distance:81.34
Distance:8.02
Distance:5.37
Pose is on-top of a gray-green road.
Pose is east of a gray-green sidewalk.
Pose is north of a black sidewalk.
Pose is west of a black parking.
Pose is south of a green vegetation.
Pose is east of a gray parking.
Distance:51.15
Distance:253.05
Distance:6.58
Distance:4.72
Pose is on-top of a gray road.
Pose is west of a dark-green sidewalk.
Pose is north of a black wall.
Pose is south of a gray pole.
Pose is on-top of a green vegetation.
Pose is west of a dark-green vegetation.
Distance:7.84
Distance:150.27
Distance:70.29
Distance:7.59
Pose is on-top of a gray road.
Pose is east of a beige sidewalk.
Pose is west of a green sidewalk.
Pose is east of a green vegetation.
Pose is south of a gray-green parking.
Pose is west of a gray-green parking.
Distance:244.68
Distance:132.03
Distance:194.83
Distance:134.53

Figure 3. Qualitative localization results on KITTI360Pose: In coarse place recognition stage, the numerical values within the top-3 retrieved submaps indicate the distance between the center of the retrieved submap and the ground truth. Green boxes highlight submaps that contain the ground truth, whereas red boxes denote incorrectly retrieved submaps. In fine-grained localization stage, red and yellow dots represent the ground truth and the predicted position, respectively. The red numerical values indicate the distance between the predicted position and the ground truth.

query text appear in numerous submaps, leading to incorrect retrieval results.

For fine-grained localization, we focus on visualizing the predicted positions within the submap nearest to the ground truth. The first three rows demonstrate that our method achieves high precision in predicting the ground truth within fully matching submaps. In the fourth and fifth rows, we observe that even within partially matching submaps, our method maintains high localization accuracy. However, in the last row, localization prediction fails inevitably due to the retrieval failure in the coarse stage.

These results indicate that the overall localization performance in the coarse-to-fine text to point cloud localization task is highly dependent on retrieval performance in the coarse stage. Moving forward, we aim to improve retrieval accuracy by: enhancing feature discriminability, and incorporating richer textual descriptions, such as street names and nearby landmarks.

References

[1] Qingyong Hu, Bo Yang, Sheikh Khalid, Wen Xiao, Niki Trigoni, and Andrew Markham. Towards semantic segmentation of urban-scale 3d point clouds: A dataset, benchmarks and challenges. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4977–4987, 2021. 1

[2] 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 (CVPR)*, pages 6687–6696, 2022. 1

[3] Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: geography-aware 3d visual grounding dataset on city-scale point cloud data. *arXiv preprint arXiv:2310.18773*, 2023. 1

[4] 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. 1

[5] 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. 1