CMMLoc: Advancing Text-to-PointCloud Localization with Cauchy-Mixture-Model Based Framework

Supplementary Material

A. Overview

In this supplementary material, we provide additional experiments and visualization results to further demonstrate the effectiveness of our proposed CMMLoc. In Sec. B, we conduct an ablation study on our proposed Cauchy-Mixture-Model-based Transformer with different layer numbers on the KITTI360Pose dataset [18], covering both the coarse and fine stages. Moreover, we present the implementation details of our model in Sec. C and more visualization results in Sec. D.

B. More analysis of Cauchy-Mixture-Model-based Transformer

In this section, we primarily analyze the impact of different layer numbers of the proposed Cauchy-Mixture-Model-based Transformer with spatial consolidation scheme on the performance of our CMMLoc, including the coarse submap retrieval stage and the fine localization stage on the KITTI360Pose dataset [18].

For the retrieval of coarse submap, Tab. 5 shows the performance of CMMLoc with different numbers of CMMT-SC, where '0' means that we use the vanilla attention mechanism in Text2Loc [48] instead of our proposed CMMT-SC. As shown, CMMLoc achieves the best performance with a Cauchy-Mixture-Model-based Transformer with spatial consolidation scheme. When the number is set to 2, the performance drops significantly, falling below that of Text2Loc. The performance in fine localization exhibits a similar trend. As shown in Tab. 6, CMMLoc performs best with a single CMMT-SC, but its performance declines sharply when the number is increased to 2.

The possible explanation for the performance degradation when using 2 layers of CMMT-SC is that the heavy-tailed nature of the Cauchy distribution enhances robustness to outliers during modeling. However, applying it consecutively may excessively blur the differences between the original features, thereby reducing the discriminative capability of the model. As a result, we set the fixed number of CMMT-SC as 1 in our CMMLoc.

Number of CMMT-SC	Submap Retrieval Recall ↑					
	Validation Set			Test Set		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
0	0.32	0.56	0.67	0.28	0.49	0.58
1	**0.35**	**0.61**	**0.73**	**0.32**	**0.53**	**0.63**
2	0.31	0.55	0.66	0.27	0.48	0.58

Table 5. Coarse submap retrieval performance for CMMLoc with different numbers of CMMT-SC on the KITTI360Pose benchmark. CMMT-SC denotes the Cauchy-Mixture-Model-based Transformer with spatial consolidation scheme. '0' means using the vaniila attention architecture in Text2Loc [48].

Number of CMMT-SC	Localization Recall ($\epsilon < 5m$) ↑					
	Validation Set			Test Set		
	$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
0	0.38	0.69	0.80	0.35	0.63	0.73
1	**0.44**	**0.75**	**0.83**	**0.39**	**0.67**	**0.77**
2	0.39	0.70	0.80	0.35	0.64	0.74

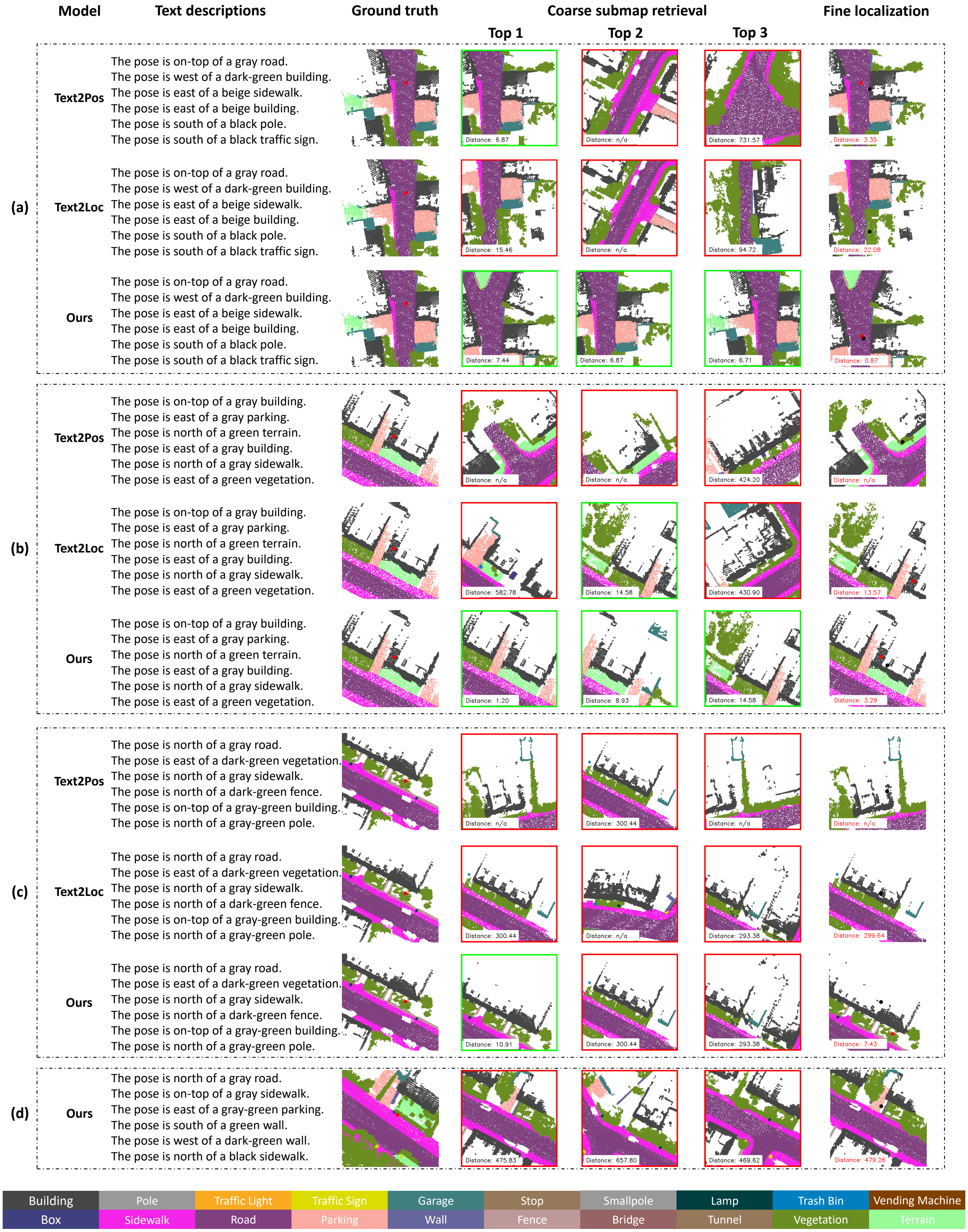
Table 6. Localization performance for Text2Loc with different numbers of CMMT-SC on the KITTI360Pose benchmark. '0' means using the fine localization network from CMMLoc, with submaps retrieved through the coarse submap retrieval framework from Text2Loc [48].

C. Implementation Details

We conduct our experiments on an NVIDIA A800 GPU. For the coarse submap retrieval, we train the model with Adam optimizer with a learning rate of 5e-4 for 20 epochs. We set the batch size to 64 and utilize a multi-step training schedule wherein the learning rate is decayed by 0.4 every 7 epochs. The temperature coefficient τ in the contrastive loss function is set to 0.1. Each submap contains a maximum of 28 objects. We employ PointNet++ [31] in [18] to extract the semantic feature of every object in the submap, followed by a single Cauchy-Mixture-Model-based Transformer with spatial consolidation scheme for local modeling to achieve better submap representation. For the fine localization, we first pre-train the text encoder and object encoder with a learning rate of 3e-4 with batch size 32 to get a well-initialized state for localization refinement. Then we train the fine localization network with the same learning rate for 45 epochs. For a fair comparison, we set the embedding dimension to 256 for both the text and submap branches in coarse submap retrieval and 128 in fine localization, following the same configuration as in previous work.

D. More Visualization results

In this section, we present additional visualizations to compare our CMMLoc with previous methods, as shown in Fig. 10, including an analysis of a failure case. For (a) and (b), CMMLoc successfully retrieves all positive submaps within the top-3 results during coarse submap retrieval,



(a) Text2Pos / Text2Loc / Ours:

The pose is on-top of a gray road.
The pose is west of a dark-green building.
The pose is east of a beige sidewalk.
The pose is east of a beige building.
The pose is south of a black pole.
The pose is south of a black traffic sign.

(b) Text2Pos / Text2Loc / Ours:

The pose is on-top of a gray building.
The pose is east of a gray parking.
The pose is north of a green terrain.
The pose is east of a gray building.
The pose is north of a gray sidewalk.
The pose is east of a green vegetation.

(c) Text2Pos / Text2Loc / Ours:

The pose is north of a gray road.
The pose is east of a dark-green vegetation.
The pose is north of a gray sidewalk.
The pose is north of a dark-green fence.
The pose is on-top of a gray-green building.
The pose is north of a gray-green pole.

(d) Ours:

The pose is north of a gray road.
The pose is on-top of a gray sidewalk.
The pose is east of a gray-green parking.
The pose is south of a green wall.
The pose is west of a dark-green wall.
The pose is north of a black sidewalk.

Figure 10. Qualitative localization results on the KITTI360Pose dataset: In coarse submap retrieval, green boxes indicate positive submaps containing the target location, while red boxes represent negative submaps. For fine localization, red and black dots correspond to the ground truth and predicted target locations. We label the distances between the prediction and ground truth in the submap, with "n/a" indicating submaps from different scenes.

whereas most retrievals from Text2Loc and Text2Pos are incorrect. When the positive submap is retrieved by other methods, our CMMLoc achieves more accurate localization performance. In case (c), although some of the top-3 submaps retrieved by our coarse submap retrieval are negative, CMMLoc effectively localizes the text queries within a 10m range after applying the fine localization network. Moreover, we present a failure case in (d), where all retrieved submaps are negative. In this case, the retrieved submap contains objects with semantic labels and categories that closely resemble those in the ground truth. Consequently, the submap features modeled using the Cauchy Mixture Model are insufficiently discriminative, highlighting the importance of developing more robust representations to better distinguish between submaps.