

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

Mingtao Feng¹ Longlong Mei¹ Zijie Wu^{1*} Jianqiao Luo^{2*} Fenghao Tian¹
Jie Feng¹ Weisheng Dong¹ Yaonan Wang²
¹Xidian University ²Hunan University

Abstract

Text to point cloud cross-modal localization is a crucial vision-language task for future human-robot collaboration. Existing coarse-to-fine frameworks assume that each query text precisely corresponds to the center area of a submap, limiting their applicability in real-world scenarios. This work redefines the task under a more realistic assumption, relaxing the one-to-one retrieval constraint by allowing partially matching query text and submap pairs. To address this challenge, we augment datasets with partially matching submaps and introduce an uncertainty-aware framework. Specifically, we model cross-modal ambiguity in fine-grained location regression by integrating uncertainty scores, represented as 2D Gaussian distributions, to mitigate the impact of challenging samples. Additionally, we propose an uncertainty-aware similarity metric that enhances similarity assessment between query text and submaps by propagating uncertainty into coarse place recognition, enabling the model to learn discriminative features, effectively handle partially matching samples and improve task synergy. Extensive experiments on KITTI360Pose and CityRefer demonstrate that our method achieves state-of-the-art performance across both stages. Our code is available at <https://github.com/Afoolbird/PMSH>

1. Introduction

Understanding natural language instructions in city-scale 3D environments is critical to enable collaboration in applications such as autonomous systems and intelligent logistics [21, 37, 40]. Conventional visual place recognition (VPR) methods, which rely on unimodal sensor data (e.g., cameras or radar) to extract 2D image or point cloud features for database matching, exhibit two key limitations: inefficiency in interactive scenarios and degraded accuracy under seasonal or viewpoint variations [25, 35]. By contrast, text to city-scale point cloud localization circumvents

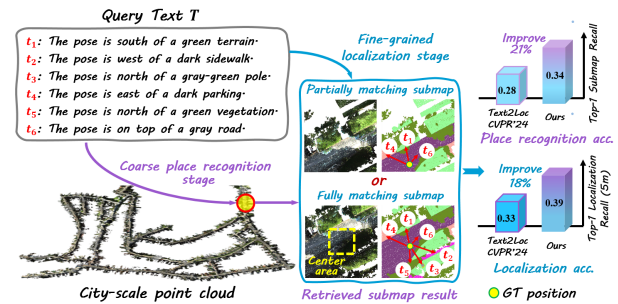


Figure 1. Query text retrieves partially matching submaps, causing cross-modal ambiguity when only a few hints align with scene instances, degrading performances in the coarse-to-fine framework. Our method mitigates this, enhancing across stages’ accuracy.

these issues, achieving precise geolocation without requiring physical user proximity [23, 43].

To date, few methods have addressed the language based localization task in large-scale 3D city maps [40]. Text2Pose [21] pioneers this task by partitioning the proposed KITTI360Pose city-scale point cloud into submaps and aligning textual descriptions with corresponding instances through a coarse-to-fine approach. In coarse stage, it employs cross-modal text-to-submap retrieval to identify potential submaps. The fine stage then refines localization using a text-instance matching module. Building on this framework, RET [37], Text2Loc [40], and IFRP-T2P [38] enhance contextual and relative position modeling, improving localization accuracy. However, existing methods and datasets simply assume each query text has one corresponding submap whose center area fully matches the location of the query text. This assumption is not practical for real-world applications [4], as query texts may appear anywhere within an area of interest, while submaps are captured beforehand [5, 44]. Consequently, multiple submaps may partially overlap with the same query, disrupting the strict one-to-one correspondence. As shown in Fig. 1, the query text is not positioned at the center area of the retrieved partially matching submap, only three hints align with instances, the remaining three fail to establish a match.

Partially matching samples exacerbate semantic gaps and increase cross-modal ambiguity during training, de-

*Corresponding author.

grading both retrieval based coarse place recognition and fine-grained localization regression. This occurs because query texts describe only a subset of instances in partially matching submaps, making their feature embeddings less similar to the query text than fully matching samples. While Text2Loc [40] introduces a prototype-based map cloning module to generate submap variants, it overlooks cross-modal ambiguities caused by minimal instance overlap. However, efficiently handling partially matching submaps presents two challenges: 1) In fine-grained localization, their partial alignment with query text complicates regression training. Treating them as fully matching submaps without adjustment risks overfitting and reduced accuracy. 2) In coarse place recognition, contrastive learning is undermined as their similarity to the query text lies between fully matching positives and mismatched negatives, rather than fitting either category. Addressing these challenges in both stages is crucial for robust localization.

To address this issue, we account for cross-modal ambiguity between the query text and instances of partially matching submaps in fine-grained localization and quantify it through uncertainty modeling. Specifically, we represent the predicted fine-grained location and its uncertainty as a 2D Gaussian distribution, jointly optimizing both via a Gaussian log-likelihood loss. We apply the negative log-likelihood to mitigate the impact of high uncertainty on the regression loss, improving localization performance. In coarse place recognition, we introduce an uncertainty-aware similarity metric to assess query text and submap alignment while accounting for cross-modal ambiguity. By propagating uncertainty scores into retrieval, this metric enhances focus on reliable partially matching submaps, enabling the model to learn discriminative features. By integrating this, the model effectively learns useful information from partially matching samples, promoting adaptive learning and reducing the adverse effects of partially matching samples for practical scenarios. **Our main contributions are:** 1) We augment datasets by incorporating partially matching submaps for text to point cloud task under a more realistic and practical setting and develop an uncertainty modeling and propagation strategy to enhance cross-modal localization performance. 2) We address cross-modal ambiguity in fine-grained location regression through uncertainty modeling, reducing the impact of challenging samples on accuracy. 3) We integrate uncertainty propagation into contrastive learning for partially matching samples, proposing an uncertainty-aware similarity metric to learn discriminative features and improve task synergy. Extensive experiments on KITTI360Pose and CityRefer datasets demonstrate the superiority of our method, improving both tasks.

2. Related Work

Visual Localization. Visual localization estimates a query image’s pose using environmental data, such as images and

point clouds, and follows two main paradigms: image-based and 3D structure-based localization. Image-based methods [8, 20, 28] treat localization as an image retrieval task, first retrieving relevant images [2, 3, 16, 36] and then establishing pixel correspondences for precise pose estimation [7, 39]. NetVLAD [2] achieves state-of-the-art performance via weakly supervised contrastive learning, while MapLocNet [39] integrates BEV and navigation map features for real-time, high-precision localization without HD maps. In contrast, 3D structure-based methods align 2D keypoints from the query image with 3D keypoints in a pre-built model [27, 31, 32]. HF-Net [31] improves efficiency and accuracy by combining local features with global descriptors for 6-DoF pose estimation.

Language-based 3D Localization. Language-based 3D localization determines a 3D point cloud’s spatial position using natural language descriptions. ScanRefer [6] and Referit3D [1] introduce indoor datasets for 3D visual grounding with free-form object annotations on ScanNet [9]. Later methods [11, 14, 15, 17, 18, 30, 34, 41] localize 3D objects in raw point clouds based on textual queries. While these approaches focus on indoor scene localization, 3D visual grounding in open outdoor environments remains underexplored. Text2Pose [21] introduced large-scale city scene localization by partitioning maps into discrete submaps, establishing the KITTI360Pose dataset. It first identifies coarse locations before refining location estimation but overlooks associations between textual prompts and point cloud instances. RET [37] addresses this by employing a Transformer-based approach to enhance representation discriminability. Building on this, Text2Loc [40] eliminates text-instance matching via contrastive learning with a more efficient strategy, improving localization accuracy. IFRP-T2P [38] proposes a two-stage localization method without real instance input, achieving competitive results. GOTLoc [19] reduces storage and computation by constructing a compact scene graph [13]. However, existing coarse-to-fine frameworks assume that each query text precisely corresponds to the center of a submap, limiting their applicability in real-world scenarios.

3. Partially Matching Sample Augmentation

Partially Matching Samples Definition. The large-scale 3D map $\mathcal{M} = \{S_i : i = 1, \dots, n\}$ is a collection of semantic submaps S_i . Each submap $S_i = \{p_i : i = 1, \dots, m\}$ includes a set of 3D object instances p_i . Let T represent the query text, which comprises a set of spatial hints t_i , each describing the spatial relationship between the position and instances of submap. The KITTI360Pose [21] consists of a series of overlapping square submaps, as illustrated in Fig. 2. Each submap spans $30m$, and the query text in each text-position pair is generated based on instances of submap within a $15m$ radius of the ground truth posi-

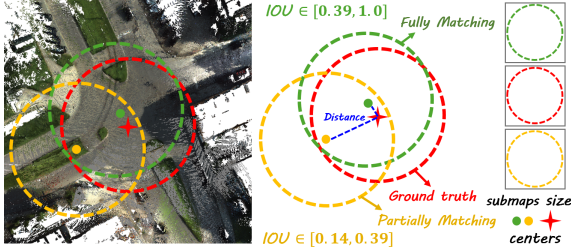


Figure 2. The sampling strategy and IOU definition. Red star: ground truth position; green dot: center of fully matching submap; yellow dot: center of partially matching submap.

tion (the red star), forming a reference area (the red circle). Each position is uniquely associated with a fully matching submap (the submap centered at the green point), denoted as S_f , which contains all the referenced instances of submap in the query text. As shown in Fig. 2, due to the spatial overlap between submaps, there exist partially matching submaps (submaps centered at yellow point), denoted as S_p , which contain only a subset of the referenced instances in the submap.

Collection. In the training set, we filter partially matching submaps based on the submap around each true location. After excluding fully matching submaps, we further discard submaps without any point cloud-instance text match pairs, resulting in all partially matching submaps. Each ground truth position is typically associated with multiple partially matching submaps, with the highest number of submaps having three point cloud instance and text match pairs, followed by two and four. We prioritize selecting the submap with three match pairs as the partially matching submap for a true position. If such a submap is unavailable, we randomly select from submaps with other match pair counts, ensuring that the selected partially matching submap contains at least two point cloud instances relevant to the query text. We exclude fully matching samples without any partially matching submaps, which account for only 1% of the total dataset. Note that information from partially matching samples is utilized only in the training set to improve overall task performance. Evaluation on the validation and test sets follows prior works to ensure fair comparison with only fully matching samples.

Distance and IOU. After obtaining the partially matching submap, its semantic consistency with the query text differs from that of the fully matching submap. To effectively leverage the information contained in partially matching submaps, we require a metric to quantify their discrepancy. A straightforward approach is to measure the L2 distance between the position and the submap center. Let the L2 distance between the centers of S_f/S_p and the position be denoted as D_f and D_p , respectively. Both are less than $15m$, but $D_f < D_p$. The second intuitive method is the Intersection Over Union (IOU) between the circle of position and the submap center circle. As illustrated in Fig. 2, we cal-

culate the IOU between the circle of position with a radius of $15m$ and the center circle of S_f/S_p , denoting the IOU as O_f and O_p , respectively. Empirical analysis shows that O_f consistently exceeds 0.54, while O_p typically falls within the range $[0.10, 0.54]$.

4. Methodology

Preliminaries. Following prior work, we adopt a hierarchical coarse-to-fine strategy for text to point cloud localization. In the coarse place recognition stage, the objective is to identify the most relevant submap corresponding to the query text. Text2Loc encodes the query text T and the instances in the fully matching submap S_f , extracting their respective global feature representations F^T and F^S . These features are then optimized via cross-modal contrastive learning, ensuring their proximity in a shared embedding space. For a given batch of N semantic submap descriptors $\{F_i^S\}_{i=1}^N$ and corresponding query text descriptors $\{F_i^T\}_{i=1}^N$, the contrastive loss \mathcal{L}_{con} for each pair is computed as follows:

$$\mathcal{L}_{con} = -\log \frac{\exp(F_i^T \cdot F_i^S / \tau)}{\sum_{j \in N} \exp(F_i^T \cdot F_j^S / \tau)} - \log \frac{\exp(F_i^S \cdot F_i^T / \tau)}{\sum_{j \in N} \exp(F_i^S \cdot F_j^T / \tau)}, \quad (1)$$

where τ is the temperature coefficient, similar to CLIP [29].

In the Fine-grained Location stage, the goal is to localize the ground-truth position within the submap. Text2Loc extracts and fuses features from the hints in the query text and the fully matching submap to predict the position via regression. The model is trained with mean squared error loss \mathcal{L}_{reg} for supervision.

$$\mathcal{L}_{reg} = \|L_{gt} - L_{pred}\|_2, \quad (2)$$

where $L_{pred} = (x, y)$ is the predicted position, and L_{gt} is the ground truth position.

In contrast, our approach integrates the informative content of partially matching samples throughout the entire localization pipeline to enhance both stages' performance. The overall pipeline is shown in Fig. 3.

4.1. Cross-Modal Partial Matching Localization

In the fine-grained localization stage, the misalignment between point cloud instances in partially matching submaps and the query text poses additional challenges to accurate localization. As a result, treating a partially matching submap S_p equivalently to a fully matching submap S_f and applying direct regression-based localization introduces overfitting, ultimately compromising the model's localization performance (see ablation study for details). To mitigate this, we utilize uncertainty to quantify the degree of cross-modal ambiguity between the query text and the semantic submap. This allows the network to adaptively adjust its optimization strategy based on varying levels of uncertainty [12, 22].

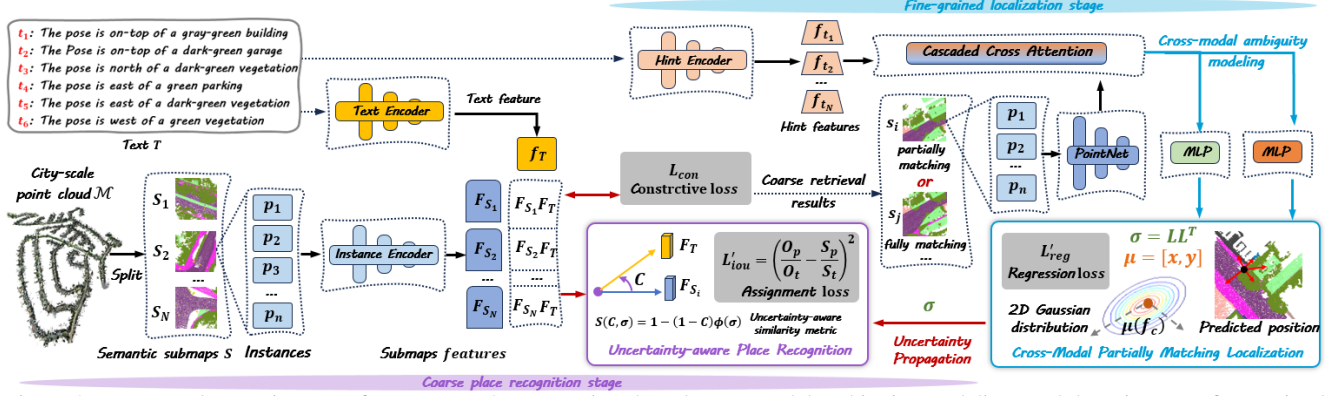


Figure 3. Framework overview. We first propose the uncertainty based cross-modal ambiguity modeling module to improve fine-grained location prediction accuracy. The uncertainty-aware similarity metric is then proposed to enhance retrieval performance by propagation uncertainty and learning discriminative features.

We denote the fused feature obtained in the fine-grained localization stage as f_c , which is then passed through a simple MLP to predict the position. Since the predicted position within the submap is distributed around the ground truth position, we model the predicted position as a Gaussian distribution [42]. Specifically, we represent the predicted coordinates as the mean μ of the Gaussian distribution, while the uncertainty of the predicted position is captured by its covariance matrix σ , which serves as a measure of cross-modal ambiguity. The matrix σ is a symmetric positive-definite matrix of size 2×2 , which has three degrees of freedom. To model this, we pass the fused feature f_c through an additional MLP that decomposes σ into its lower triangular matrix L , such that $LL^T = \sigma$. The MLP consists of two fully connected layers, which derive the three variables that determine L , with the Exponential Linear Unit (ELU) activation function applied to ensure that the diagonal elements of L are positive. By learning σ from the cross-modal fusion features, we effectively capture the discrepancies between the text and point cloud modalities. These discrepancies help us address cross-modal ambiguity with uncertainty.

Given the predicted Gaussian distribution for each sample pair, the likelihood of the predicted position at the ground truth position L_{gt} is formulated as follows:

$$P(L_{gt}|\mu(f_c), \sigma(f_c)) = \frac{\exp(-\frac{1}{2}(L_{gt} - \mu)^T \sigma^{-1} (L_{gt} - \mu))}{2\pi \sqrt{|\sigma|}} \quad (3)$$

where $||$ denotes determinant and $\mu(f_c)$ represents the predicted location L_{pred} . The model represents the predicted position as a 2D Gaussian distribution, parameterized by μ and σ . The objective is to train the network to establish a mapping from fused features f_c to the parameters of the Gaussian distribution. Since maximizing the likelihood function (Eq. 3) is mathematically equivalent to minimizing the negative log-likelihood, we adopt negative log-

likelihood as our new loss function \mathcal{L}'_{reg} .

$$\begin{aligned} \mathcal{L}'_{reg} &= -\ln P(L_{gt}|\mu(f_c), \sigma(f_c)) \\ &= \frac{1}{2}(L_{gt} - \mu)^T \sigma^{-1} (L_{gt} - \mu) + \frac{1}{2} \ln |\sigma| + \ln 2\pi, \end{aligned} \quad (4)$$

The first component penalizes the discrepancy between the predicted and ground truth position, while the second component regularizes high uncertainty.

Guided by Eq. 4, the model utilizes uncertainty estimates to dynamically adjust its optimization based on sample characteristics. When predicting the position for challenging partially matching samples, the network allows a larger covariance σ , signaling high uncertainty to moderate the severe residuals ($L_{gt} - \mu(f_c)$). As a result, the influence of these challenging partially matching samples is reduced during training. This mechanism acts as a strategy for loss attenuation [24], enabling the model to effectively learn from partially matching samples while preserving information from fully matching samples, ultimately enhancing localization accuracy.

4.2. Uncertainty-Aware Place Recognition

During the coarse place recognition stage, the valid information from partially matching submaps also contributes to enhanced retrieval performance. However, directly incorporating these submaps into cross-modal contrastive learning is inappropriate, not only due to the differences between S_f and S_p , but also because of their overlapping regions. Consequently, their similarity to the query text should be closer than that of mismatched pairs in contrastive learning, but not as close as matched pairs. Thus, a new approach is required to effectively leverage the information from partially matching samples. Since the point cloud instances in partially matching submaps only partially align with the query text, their similarity in the feature embedding space should be lower than that of fully matching pairs. As explained above, we assign embedding similarity based on the IOU

or L2 distance between the ground truth position and the submap. The corresponding loss functions for these assignments are defined as follows:

$$\mathcal{L}_{\text{dis}} = \left(\frac{D_f}{D_p} - \frac{C_p}{C_f} \right)^2, \mathcal{L}_{\text{iou}} = \left(\frac{O_p}{O_f} - \frac{C_p}{C_f} \right)^2, \quad (5)$$

where C_p and C_f denote the cosine similarities between the query text and the partially matching and fully matching submaps, respectively. This loss forces the ratio of the similarities in the embedding space to be close to the ratio of L2 distance/IOU. As \mathcal{L}_{iou} exhibited superior performance during training, we adopt IOU as our method for addressing partially matching samples. Further details are provided in the ablation study. We observe that \mathcal{L}_{iou} exhibits instability during training. This is primarily because, in some partially matching submaps, the IOU value fails to accurately reflect the correspondence between point cloud instances and the query text. Specifically, instances occur where the IOU value is high despite the actual number of matched point cloud instances being low. A comparable issue arises when employing \mathcal{L}_{dis} . Furthermore, certain partially matching samples exhibit challenges in identification. It is necessary to further optimize assignment loss.

Accurately predicting the positions of these challenging partially matching samples is also difficult due to inherent ambiguities in their alignment with the query text, a challenge analogous to those encountered in fine-grained localization. Therefore, we propagate the uncertainty estimated during fine-grained localization into \mathcal{L}_{iou} . To prevent the model from excessively optimizing on challenging samples, we modulate the cosine similarity based on the degree of cross-modal ambiguity. Based on the cosine similarity C (include C_p and C_f), we define an uncertainty-aware similarity metric $S(C, \sigma)$, which is adaptively adjusted using the uncertainty σ . Leveraging this metric, we propose a novel assignment loss function.

$$\mathcal{L}'_{\text{iou}} = \left(\frac{O_p}{O_f} - \frac{S(C_p, \sigma_p)}{S(C_f, \sigma_f)} \right)^2,$$

$$S(C, \sigma) = 1 - (1 - C)\phi(\sigma), \quad \phi(\sigma) = \frac{1}{1 + \alpha e^{\beta(|\sigma| - \gamma)}}, \quad (6)$$

where $\phi(\sigma)$ is a decaying exponential function [22] with three positive parameters: α , β , and γ . When the uncertainty σ is low, the similarity metric $S(C, \sigma)$ closely approximates the original cosine similarity C . As uncertainty σ approaches zero, the function $\phi(\sigma)$ converges to 1, resulting in $S(C, \sigma) = C$, then the original assignment loss is applied for contrastive learning. Conversely, when the uncertainty σ is high, $\phi(\sigma)$ approaches 0, and $S(C, \sigma)$ takes a value close to 1, effectively excluding the challenging sample from optimization. This mechanism filters out challenging partially matching samples, reducing their impact on

training. By focusing on samples with lower cross-modal ambiguity, the network can extract more reliable information, thereby enhancing retrieval performance.

As fully matching submaps typically exhibit strong certainty and high similarity with text queries, the network focuses on enhancing the embedding similarity for partially matching submaps. Through gradient analysis, we demonstrate how the uncertainty measure $S(C, \sigma)$ contributes to discriminative learning in the embedding space. By replacing the similarity measure in the assignment loss function \mathcal{L}_{iou} with $S(C, \sigma)$, we derive the relationship $\frac{\partial \mathcal{L}_{\text{iou}}}{\partial C} = \frac{\partial \mathcal{L}'_{\text{iou}}}{\partial S}$, which is influenced by the specific form of the loss function. Let the network parameters and feature embeddings be denoted as θ and f , respectively; the gradient of the loss with respect to parameter $\frac{\partial \mathcal{L}'_{\text{iou}}}{\partial \theta}$ can be expressed as:

$$\begin{aligned} \frac{\partial \mathcal{L}'_{\text{iou}}}{\partial \theta} &= \frac{\partial \mathcal{L}'_{\text{iou}}}{\partial S} \frac{\partial S}{\partial C} \frac{\partial C}{\partial f} \frac{\partial f}{\partial \theta} = \frac{\partial \mathcal{L}_{\text{iou}}}{\partial C} \frac{\partial S}{\partial C} \frac{\partial C}{\partial f} \frac{\partial f}{\partial \theta} \\ &= \frac{\partial \mathcal{L}_{\text{iou}}}{\partial \theta} \frac{\partial S}{\partial C} = \frac{\partial \mathcal{L}_{\text{iou}}}{\partial \theta} \phi(\sigma), \end{aligned} \quad (7)$$

We observe that $\frac{\partial \mathcal{L}_{\text{iou}}}{\partial \theta}$ is modulated by the decay function $\phi(\sigma)$, meaning that, compared to the original gradient $\frac{\partial \mathcal{L}_{\text{iou}}}{\partial \theta}$, higher uncertainty results in a smaller gradient. Additionally, the design of $S(C, \sigma)$ enables the model to focus on specific partially matching submaps, providing more informative feature embeddings for the coarse place recognition task. This approach not only mitigates the issue of cross-modal ambiguity but also ensures proper gradient alignment during training [33], which aids in achieving more effective optimization for the coarse place recognition task.

5. Experiments

Datasets. Our experiments are primarily conducted on the **KITTI360Pose** [21] dataset. Through our data augmentation strategy, we train the model with both partially matching and fully matching submaps corresponding to the query texts. We also use the text to point cloud dataset **CityRefer** [26] to evaluate the generalization of the method. Following [21], we also segment the large-scale 3D city maps into comparable semantic submaps. Further details can be found in the Appendix.

Evaluation Metrics. Following [40], we adopt top-k retrieval recall ($k \in \{1, 3, 5\}$) for text-to-submap retrieval evaluation in coarse place recognition. To evaluate fine-grained localization, we assess performance based on the top-k retrieved candidates and report localization recall, where k is set to 1, 5, and 10. Localization recall is defined as the fraction of queries that are successfully localized when their localization error falls below certain thresholds, specifically $\epsilon < 5/10/15m$ by default. Additionally, we further evaluate the normalized Euclidean distance error between the ground truth and predicted positions.

Methods	Reference	Localization Recall ($\epsilon < 5/10/15m$) \uparrow					
		Validation Set			Test Set		
		$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
Text2Pos [21]	CVPR'22	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 [37]	AAAI'23	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 [40]	CVPR'24	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
IFRP-T2P [38]	ACM MM'24	0.23/0.45/0.53	0.53/0.70/0.81	0.64/0.86/0.89	0.22/0.40/0.46	0.47/0.68/0.73	0.58/0.78/0.82
Ours	-	0.42/0.62/0.68	0.75/0.89/0.90	0.83/0.94/0.95	0.39/0.55/0.59	0.68/0.81/0.83	0.78/0.89/0.90

Table 1. Performance comparison on the KITTI360Pose. Both Validation and test sets are used for fair comparison with SOTA.

Methods	Validation Error \downarrow	Test Error \downarrow
Text2Pos [21]	0.120	0.127
Text2Loc [40]	0.091	0.090
IFRP-T2P [38]	0.118	0.119
Ours	0.085	0.081

Table 2. Fine-grained localization comparison on KITTI360Pose.

Methods	Submap Retrieval Recall \uparrow					
	Validation Set			Test Set		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 5$
Text2Pos [21]	0.14	0.28	0.37	0.12	0.25	0.33
RET [37]	0.18	0.34	0.44	-	-	-
Text2Loc [40]	0.32	0.56	0.67	0.28	0.49	0.58
IFRP-T2P [38]	0.24	0.46	0.57	0.23	0.39	0.48
GOTLoc [19]	0.36	0.58	0.67	0.30	0.49	0.57
Ours	0.37	0.63	0.73	0.34	0.56	0.65

Table 3. Coarse place recognition comparison on KITTI360Pose.

5.1. Comparison with State-of-the-Art

We compare our method with state-of-the-art methods [19, 21, 37, 38, 40] across both stages.

Fine-grained localization. Tab. 1 presents the top- k recall rates ($k = 1/5/10$) under different localization error thresholds ($\epsilon < 5/10/15m$). IFRP-T2P [38] eliminates the reliance on ground truth semantic submaps by leveraging its own semantic segmentation approach [10], but lacks open-source code; thus, we conduct direct comparisons under different settings. To ensure consistency, we use the semantic submap input provided by the original dataset, similar to other methods. However, unlike previous approaches that trained solely on fully matching samples, our method leverages both partially and fully matching samples (all tabulated data are sourced from the original papers). Our approach surpasses all existing methods across all recall levels, outperforming the previous state-of-the-art, Text2Loc [40], by 14% and 18% in top-1 recall on the validation and test sets, respectively. Moreover, our method consistently demonstrates superior performance as the threshold or k increases. These results validate the effectiveness of our uncertainty modeling and propagation in leveraging partially matching samples throughout the coarse-to-fine framework, enhancing overall localization accuracy. Additionally, we analyze the impact of integrating partially matching samples into the open-source method Text2Loc [40] in Tab. 4 as part of our

ablation study.

To independently assess our fine-grained location stage, we evaluate the trained model using only fully matching samples for a consistent and fair comparison with others. Normalized Euclidean distance serves as the evaluation metric. As shown in Tab. 2, our method reduces error by 7% and 10% on the validation and test sets, respectively, compared to Text2Loc [40]. These results validate the effectiveness of our uncertainty modeling in leveraging both partially and fully matching samples during training, enhancing fine-grained localization accuracy.

Coarse place recognition. In the coarse place recognition stage, we evaluated top-1/3/5 recall rates, with our method surpassing the SOTA Text2Loc by 16%/13%/9% (Tab.3), showing similar gains on the test set. It also outperforms the retrieval-only method GOTLoc[19] (arXiv 2025). These results highlight the importance of partially matching samples in enhancing retrieval and validate the effectiveness of our proposed IOU and uncertainty propagation approach.

Qualitative Analysis. We provide a qualitative comparison of coarse-to-fine localization between Text2Loc and our method, as shown in Fig. 4. Given a query text, we visualize the ground truth, top-3 retrieved submaps, and top-1 fine-grained localization result. Each submap spans $30m$; thus, in coarse place recognition, a retrieved submap is classified as a partially matching submap if its center distance is within $15m$ of the ground truth and it is not a fully matching submap. In coarse place recognition, both Text2Loc and our method retrieve fully or partially matching submaps near the position. However, our method is more likely to retrieve a fully matching submap under stricter metrics, such as top-1. As shown in the first column, while Text2Loc retrieves only a partially matching submap at top-2, our method achieves a fully matching submap at top-1. This demonstrates that our uncertainty-aware assignment loss \mathcal{L}'_{iou} in the coarse stage helps distinguish fully and partially matching samples, improving retrieval performance. The second column highlights a key challenge for Text2Loc: despite failing to retrieve the correct fully matching submap at top-3, the retrieved submaps contain similar object instances, indicating difficulty in resolving inter-submap ambiguities. In contrast, our method retrieves the fully matching submap at top-1, followed by a partially matching submap. This

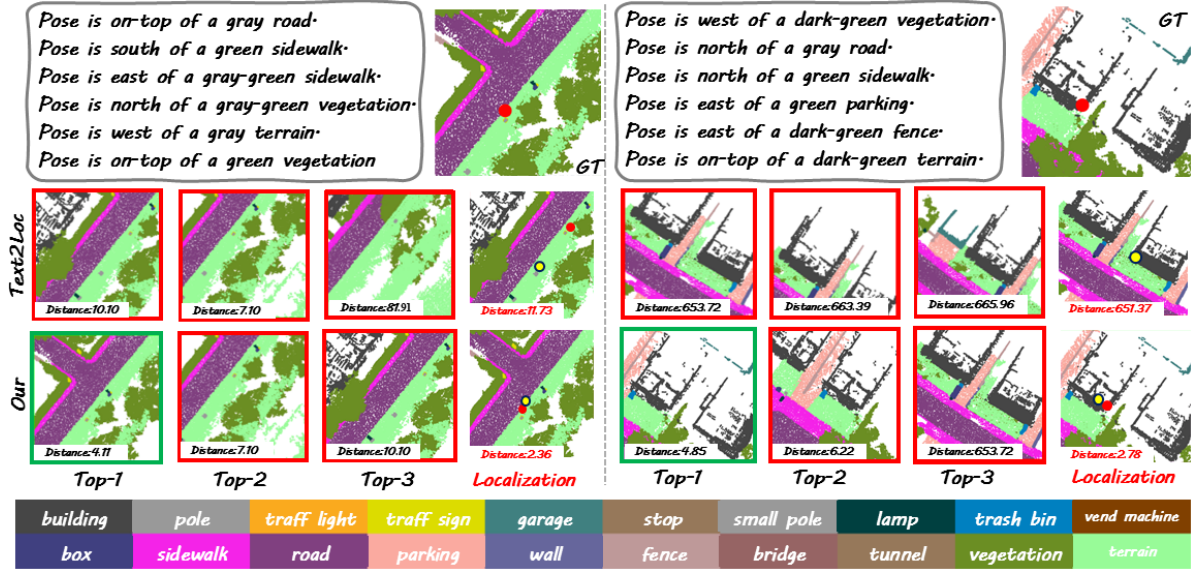


Figure 4. Qualitative comparisons on KITTI360Pose: In coarse place recognition, the numbers in top-3 retrieved submaps show their center to ground truth distances. Green boxes denote fully matching submaps containing the position, while red indicates retrieval failures. In fine-grained localization, red and yellow dots mark the ground truth and predicted locations, with red numbers showing their distances.

Methods	Localization Recall ($\epsilon < 5m$) \uparrow					
	Validation Set			Test Set		
	$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
Text2Loc	0.37	0.68	0.77	0.33	0.61	0.71
Text2Loc*	0.40	0.71	0.80	0.36	0.65	0.74
w/o \mathcal{L}'_{reg}	0.25	0.43	0.51	0.21	0.38	0.47
Ours	0.42	0.75	0.83	0.39	0.68	0.78

Table 4. Ablation study on fine localization (KITTI360Pose). * denotes the Text2Loc fine localization network with submaps retrieved by our coarse recognition. ‘w/o \mathcal{L}'_{reg} ’ indicates training with \mathcal{L}_{reg} using partially matching submaps.

demonstrates that integrating \mathcal{L}_{con} with \mathcal{L}'_{iou} helps the model focus on distinctive features of fully and partially matching submaps, enhancing its ability to differentiate from other semantically similar submaps. The top-1 fine-grained localization results in Fig. 4 confirm our method’s ability to refine predictions using the retrieved fully matching submap. Additional visualizations are provided in the Appendix.

5.2. Ablation Study

Fine-grained localization. To further assess our method’s effectiveness in fine-grained localization, we present a quantitative analysis in Tab. 4. For fair comparison, all methods use the same submaps retrieved by our coarse place recognition. Text2Loc* significantly outperforms the original Text2Loc, demonstrating the benefits of our coarse place recognition. However, incorporating partially matching submaps into Text2Loc’s training degrades performance, consistent with our earlier analysis (see Sec. 5.1). By introducing uncertainty modeling as a loss function to handle partially matching samples, our method achieves a

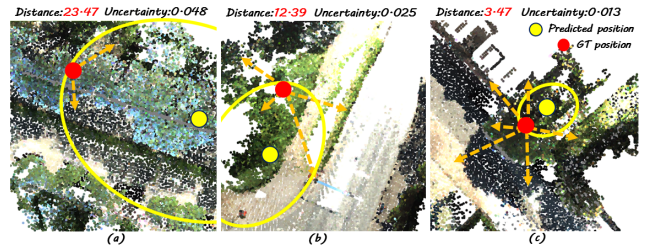


Figure 5. Visualization of cross-modal ambiguity modeling. Estimated uncertainty is shown as yellow circles, while the orange dashed line links the text query to its corresponding point cloud instance in the submap.

top-1 accuracy of 0.39 on the test set at $\epsilon < 5m$, an 8% improvement over Text2Loc. These results confirm that our framework effectively models cross-modal ambiguity, enhancing the model’s ability to distinguish fully from partially matching samples and improving text to submap localization accuracy.

To further validate this finding, we visualize predicted location, uncertainty scores, and prediction errors. As shown in Fig. 5(a), higher uncertainty scores correspond to fewer matched text–point cloud instances and larger prediction errors, indicating that uncertainty reflects cross-modal ambiguity arising from imperfect text–point cloud alignment. Moreover, Fig. 5(b) and (c) reveal that as the number of matched instances increases, both uncertainty and prediction errors decrease, demonstrating the model’s ability to learn discriminative feature representations for improved localization accuracy.

We further analyze the results by binning test samples based on uncertainty values using histogram binning and computing mean and median localization errors within each

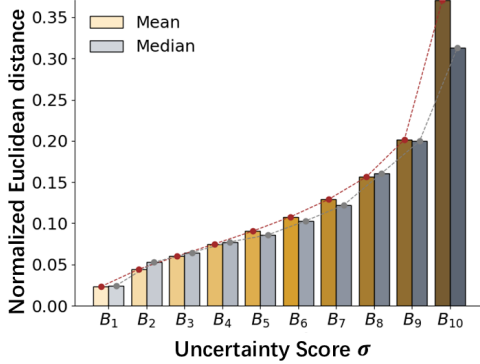


Figure 6. The mean and median errors of fine-grained predictions are evaluated across uncertainty intervals, with the range [0.01, 0.06] divided into 10 segments, corresponding to bins B_i .

Methods	Submap Retrieval Recall \uparrow					
	Validation Set			Test Set		
	$k = 1$	$k = 3$	$k = 5$	$k = 1$	$k = 3$	$k = 15$
Text2Loc	0.32	0.56	0.67	0.28	0.49	0.58
Text2Loc*	0.19	0.28	0.41	0.16	0.25	0.37
w/ \mathcal{L}_{dis}	0.33	0.59	0.69	0.31	0.50	0.61
w/ \mathcal{L}_{iou}	0.35	0.60	0.71	0.32	0.53	0.63
Ours	0.37	0.63	0.73	0.34	0.56	0.65

Table 5. Ablation study of coarse stage on KITTI360Pose. * denotes training with partially matching submaps using \mathcal{L}_{con} . ‘w/ \mathcal{L}_{dis} ’ denotes training both partially and fully matching samples using \mathcal{L}_{con} and \mathcal{L}_{dis} , ‘w/ \mathcal{L}_{iou} ’ follows a similar strategy.

bin (Fig. 6). Lower uncertainty corresponds to smaller errors, while higher uncertainty correlates with increased errors, confirming its positive relationship with localization accuracy. These findings validate the effectiveness of our uncertainty modeling in capturing cross-modal ambiguity and serving as a reliable indicator of localization accuracy.

Coarse place recognition. To evaluate the effectiveness of our method in coarse place recognition, we compare strategies for handling partially matching samples (Tab. 5). Consistent with fine-grained localization, naively incorporating partially matching samples into contrastive learning reduces retrieval performance. Based on this, we explore alternative approaches using \mathcal{L}_{iou} , \mathcal{L}_{dis} , and our final proposed \mathcal{L}'_{iou} . The results show that both \mathcal{L}_{iou} and \mathcal{L}_{dis} outperform the baseline, with IOU-based similarity surpassing L2 distance. The IOU ratio more effectively distinguishes between partially and fully matching samples. Our final approach achieves the best performance, demonstrating that the uncertainty-aware similarity loss filters out hard samples, allowing the model to better utilize informative partially matching samples and enhance retrieval accuracy.

Further Training Analysis. We analyze the training process, as shown in Fig. 7. The left panel displays loss curves for the coarse place recognition stage with \mathcal{L}_{iou} and \mathcal{L}'_{iou} . Propagating uncertainty guidance improves stability and accelerates convergence, suggesting that \mathcal{L}'_{iou} allows

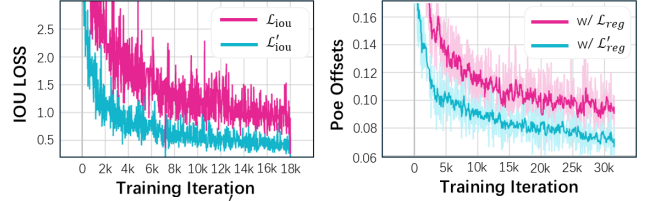


Figure 7. \mathcal{L}_{iou} and \mathcal{L}'_{iou} loss curves (left), the Normalized Euclidean distances when utilize \mathcal{L}_{reg} or \mathcal{L}'_{reg} loss (right).

Methods	Localization Recall ($\epsilon < 10m$) \uparrow					
	Validation Set			Test Set		
	$k = 1$	$k = 5$	$k = 10$	$k = 1$	$k = 5$	$k = 10$
Text2Pos	0.13	0.32	0.55	0.10	0.28	0.47
Text2Loc	0.25	0.46	0.69	0.19	0.38	0.57
Ours	0.29	0.54	0.78	0.24	0.47	0.66

Table 6. Performance comparison on the CityRefer.

dynamic adjustment, mitigating the impact of hard samples within partially matched pairs and enhancing training efficiency. The right panel shows the normalized Euclidean distance error curve in the fine-grained localization stage. Using \mathcal{L}'_{reg} improves convergence, indicating that uncertainty based cross-modal ambiguity modeling effectively utilizes discriminative features from partially matching samples, boosting accuracy for fully matching samples.

Details in Appendix: The hyperparameter ablation in Eq. 6 and its impact on model performance. The effectiveness of our method in structuring the embedding space and refining spatial retrieval.

5.3. Generalization to CityRefer dataset

We evaluate our method against new CityRefer dataset and report the fine-grained localization performance in Tab. 6. We implement the open-source methods on CityRefer with the same settings as KITTI360Pose. Due to the increased linguistic variability of CityRefer compared to the fixed templates in KITTI360Pose, existing methods show degraded performance. In contrast, our method maintains a significant advantage and outperforms Text2Loc by 16% and 26% in top-1 recall on the validation and test sets, respectively, highlighting its robustness in handling diverse query text and improving localization accuracy.

6. Conclusion

In this paper, we pioneered a more realistic setting for text to large-scale point cloud task by relaxing the one-to-one retrieval constraint by allowing partially matching submap and query text pairs. To address the associated challenges, we first modeled cross-modal ambiguity in fine-grained location regression by integrating uncertainty scores and then proposed an uncertainty-aware similarity metric to propagate uncertainty into coarse place recognition. Experiments show our method’s superiority. In the future, we will design more efficient similarity metrics with the modeled uncertainty for overall performance improvement.

Acknowledgments

This work was supported in part by Science and Technology Project of Jiangxi province (20232ACC01007), in part by Ji'an Science and Technology Project (20233TGV06020), in part by the National Natural Science Foundation of China under Grant (62373293, 62463020, 62403189), in part by Jiangxi Provincial Natural Science Foundation (20242BAB20050), in part by Ji'an Science and Technology Plan Natural Science Foundation (20244018591).

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. 2
- [2] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Padjla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2
- [3] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 2
- [4] Jinghao Bian, Mingtao Feng, Weisheng Dong, Zijie Wu, Jianqiao Luo, Fangfang Wu, Yaonan Wang, and Guangming Shi. Locally aware visual state space for small defect segmentation in complex component images. *IEEE Transactions on Industrial Informatics*, 2025. 1
- [5] Xudong Cai, Yongcai Wang, Zhe Huang, Yu Shao, and Deying Li. Voloc: Visual place recognition by querying compressed lidar map. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10192–10199. IEEE, 2024. 1
- [6] 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. 2
- [7] Mengzhao Chen, Mingbao Lin, Ke Li, Yunhang Shen, Yongjian Wu, Fei Chao, and Rongrong Ji. Cf-vit: A general coarse-to-fine method for vision transformer. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7042–7052, 2023. 2
- [8] Zihao Chen, Kunhong Li, Haoran Li, Zhiheng Fu, Hanmo Zhang, and Yulan Guo. Metric localization for lunar rovers via cross-view image matching. *Visual Intelligence*, 2(1):12, 2024. 2
- [9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 2
- [10] Mingtao Feng, Liang Zhang, Xuefei Lin, Syed Zulqarnain Gilani, and Ajmal Mian. Point attention network for semantic segmentation of 3d point clouds. *Pattern Recognition*, 107:107446, 2020. 6
- [11] 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. 2
- [12] Mingtao Feng, Fenghao Tian, Jianqiao Luo, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Saeed Mian. Semantic ambiguity modeling and propagation for fine-grained visual cross view geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2978–2986, 2025. 3
- [13] Mingtao Feng, Chenbo Yan, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. History-enhanced 3d scene graph reasoning from rgb-d sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025. 2
- [14] Mingtao Feng, Chenbo Yan, Zijie Wu, Weisheng Dong, Yaonan Wang, and Ajmal Mian. Hyperrectangle embedding for debiased 3d scene graph prediction from rgb sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2
- [15] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xi-anzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xi-anzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv preprint arXiv:2309.00615*, 2023. 2
- [16] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14141–14152, 2021. 2
- [17] Dailan He, Yusheng Zhao, Junyu Luo, Tianrui Hui, Shaofei Huang, Aixi Zhang, and Si Liu. Transrefer3d: Entity-and-relation aware transformer for fine-grained 3d visual grounding. In *Proceedings of the 29th ACM international conference on multimedia*, pages 2344–2352, 2021. 2
- [18] Shijia Huang, Yilun Chen, Jiaya Jia, and Liwei Wang. Multi-view transformer for 3d visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15524–15533, 2022. 2
- [19] Donghwi Jung, Keonwoo Kim, and Seong-Woo Kim. Gotloc: General outdoor text-based localization using scene graph retrieval with openstreetmap. *arXiv preprint arXiv:2501.08575*, 2025. 2, 6
- [20] Hyo Jin Kim, Enrique Dunn, and Jan-Michael Frahm. Predicting good features for image geo-localization using per-bundle vlad. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1170–1178, 2015. 2
- [21] 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, 2, 5, 6
- [22] Abhinav Kumar, Tim K Marks, Wenxuan Mou, Chen Feng, and Xiaoming Liu. Uglli face alignment: Estimating uncer-

- tainty with gaussian log-likelihood loss. In *ICCV Workshops on Statistical Deep Learning in Computer Vision*, 2019. 3, 5
- [23] Jinpeng Li, Haiping Wang, Yuan Liu, Zhiyang Dou, Yuexin Ma, Sibe Yang, Yuan Li, Wenping Wang, Zhen Dong, Bisheng Yang, et al. Cityanchor: City-scale 3d visual grounding with multi-modality llms. In *The Thirteenth International Conference on Learning Representations*. 1
- [24] Xiaobin Li, Kai Wu, Xiaoyu Zhang, and Handing Wang. B2opt: Learning to optimize black-box optimization with little budget. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(17):18502–18510, 2025. 4
- [25] Yuhao Li, Jianping Li, Zhen Dong, Yuan Wang, and Bisheng Yang. Saliency2ploc: saliency-guided image-point cloud localization using contrastive learning. *Information Fusion*, page 103015, 2025. 1
- [26] 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. 5
- [27] Uzair Nadeem, Mohammad AAK Jalwana, Mohammed Bennamoun, Roberto Togneri, and Ferdous Sohel. Direct image to point cloud descriptors matching for 6-dof camera localization in dense 3d point clouds. In *International Conference on Neural Information Processing*, pages 222–234. Springer, 2019. 2
- [28] Filip Radenović, Giorgos Toliás, and Ondřej Chum. Cnn image retrieval learns from bow: Unsupervised fine-tuning with hard examples. In *Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 3–20. Springer, 2016. 2
- [29] 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. 3
- [30] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. Languagerefer: Spatial-language model for 3d visual grounding. In *Conference on Robot Learning*, pages 1046–1056. PMLR, 2022. 2
- [31] 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
- [32] 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
- [33] Dmitry Senushkin, Nikolay Patakin, Arseny Kuznetsov, and Anton Konushin. Independent component alignment for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20083–20093, 2023. 5
- [34] Xiangxi Shi, Zhonghua Wu, and Stefan Lee. Aware visual grounding in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14056–14065, 2024. 2
- [35] Yujiao Shi, Hongdong Li, Akhil Perincherry, and Ankit Vora. Weakly-supervised camera localization by ground-to-satellite image registration. In *European Conference on Computer Vision*, pages 39–57. Springer, 2024. 1
- [36] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1808–1817, 2015. 2
- [37] Guangzhi Wang, Hehe Fan, and Mohan Kankanhalli. Text to point cloud localization with relation-enhanced transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2501–2509, 2023. 1, 2, 6
- [38] Lichao Wang, Zhihao Yuan, Jinke Ren, Shuguang Cui, and Zhen Li. Instance-free text to point cloud localization with relative position awareness. *ACM International Conference on Multimedia*, 2024. 1, 2, 6
- [39] Hang Wu, Zhenghao Zhang, Siyuan Lin, Xiangru Mu, Qiang Zhao, Ming Yang, and Tong Qin. Maplocnet: Coarse-to-fine feature registration for visual re-localization in navigation maps. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13198–13205. IEEE, 2024. 2
- [40] Yan Xia, Letian Shi, Zifeng Ding, Joao F. Henriques, and Daniel Cremers. Text2loc: 3d point cloud localization from natural language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14958–14967, 2024. 1, 2, 5, 6
- [41] Li Yang, Ziqi Zhang, Zhongang Qi, Yan Xu, Wei Liu, Ying Shan, Bing Li, Weiping Yang, Peng Li, Yan Wang, et al. Exploiting contextual objects and relations for 3d visual grounding. *Advances in Neural Information Processing Systems*, 36:49542–49554, 2023. 2
- [42] Qitong Yang, Mingtao Feng, Zijie Wu, Weisheng Dong, Fangfang Wu, Yaonan Wang, and Ajmal Mian. Hierarchical gaussian mixture model splatting for efficient and part controllable 3d generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11104–11114, 2025. 4
- [43] Huan Yin, Xuecheng Xu, Sha Lu, Xieyanli Chen, Rong Xiong, Shaojie Shen, Cyrill Stachniss, and Yue Wang. A survey on global lidar localization: Challenges, advances and open problems. *International Journal of Computer Vision*, 132(8):3139–3171, 2024. 1
- [44] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 1