SFPR: Structural Fingerprinting for LiDAR-to-OpenStreetMap Place Recognition

Jiajun Zheng[1,2,3], **Ziwei Shi**[1,2,3], **Yu Zang**[1,2,3]*, **Chenlu Lin**[1,2,3], **Youbiao Wang**[1,2,3], **Wenjing Xu**[1,2,3], **Cheng Wang**[1,2,3]

[1]Fujian Key Laboratory of Urban Intelligent Sensing and Computing, Xiamen University, Xiamen, China

[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Ministry of Education of China, Xiamen University, Xiamen, China

[3]The Belt and Road Research Institute, Xiamen University, Xiamen, China

zhengjiajun@stu.xmu.edu.cn zangyu7@xmu.edu.cn

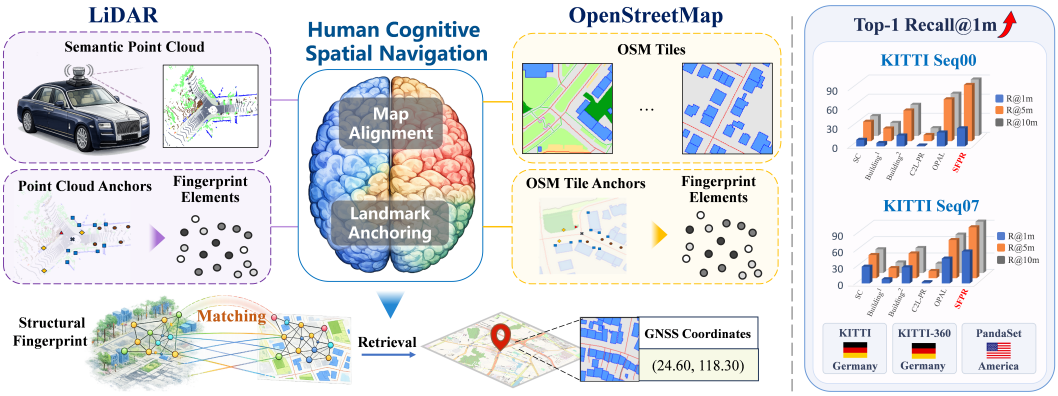

Figure 1: Inspired by human cognitive mechanisms for spatial navigation, we propose SFPR, a structural fingerprinting-based framework for LiDAR-to-OpenStreetMap (L2O) place recognition.

Abstract

Cognitive neuroscience research indicates that human spatial navigation relies not merely on abstract global map representations, but rather on anchoring local observations to spatially stable, decision-relevant landmarks to effectively align the current scene with the cognitive map. Inspired by this mechanism, we propose SFPR, a structural fingerprint-based LiDAR-to-OpenStreetMap (OSM) place recognition framework. This framework goes beyond the traditional coarse retrieval paradigm that relies exclusively on global descriptor similarity by introducing fine-grained discrimination based on structural fingerprints to achieve more accurate place recognition. Specifically, we propose a structural fingerprint-aware attention mechanism that utilizes sparse but spatially stable anchors as saliency prompts, directing the network's attention toward potential fingerprint regions while dynamically generating global descriptors. Subsequently, we formulate the extraction and matching of structural fingerprints as an alternating optimization problem guided by this fingerprint-aware attention. Through iterative optimization, we extract structural fingerprints with high fingerprint-aware attention and geometric consistency from the anchors. These structural fingerprints are further fed back into the training stage as geometric priors, effectively suppressing false matches characterized by similar global features but contradictory local structures. Experiments demonstrate that SFPR effectively overcomes the retrieval bottleneck in

*Corresponding Author

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

macroscopically homogeneous scenes, improving the Top-1 Recall@1m to 28.41% on KITTI sequence 00 compared to current state-of-the-art methods. The code is publicly available at https://github.com/Wiakoel/SFPR.

1 Introduction

Place recognition aims to match current sensor observations with a geo-referenced prior map when GNSS signals are unreliable or unavailable [1]. Given LiDAR's robustness to illumination changes and shadows, LiDAR-based place recognition has attracted significant attention [2]. However, traditional LiDAR place recognition often relies on high-precision yet costly pre-built 3D point cloud maps, making it difficult to meet the practical demands of large-scale, low-cost deployment [3]. To overcome this bottleneck, utilizing low-cost alternative map priors for cross-modal retrieval has become an important research direction. Shi et al. [4] used remote sensing images as reference maps and mitigated cross-view and cross-modal discrepancies through semantic alignment to extract cross-modally aligned global descriptors. Kang et al. [5] constructed a rotation-robust global descriptor based on lightweight OSM, enabling LiDAR place recognition with greater practical value.

Despite the progress achieved by the aforementioned methods, their retrieval paradigm remains highly dependent on global descriptors. These methods indiscriminately compress the entire scene into a single descriptor, inevitably discarding crucial fine-grained local details [6, 7, 8]. However, this retrieval paradigm diverges from findings in cognitive neuroscience on human spatial navigation. Epstein et al. [9] indicate that the effective use of cognitive maps by the human brain also relies on spatially stable landmarks. Neuroimaging studies further reveal that the brain is not only sensitive to physically stable objects but also responds more strongly to cues located at navigational decision points. During navigation, the brain utilizes features with both stability and decision relevance for landmark anchoring, thereby achieving spatial alignment between the local egocentric scene and the broader allocentric map.

Inspired by these cognitive mechanisms of human spatial navigation, we argue that accurate LiDAR place recognition should not rely solely on global descriptor similarity; rather, stable environmental landmarks facilitate the anchoring of local observations to broader spatial representations. To this end, we propose the concept of the structural fingerprint, which is characterized by a set of spatially stable anchors and their geometric relationships within a scene. Furthermore, as a lightweight and structured map prior, OSM provides a rich semantic foundation for extracting such stable anchors. Building upon this insight, we propose SFPR, a structural fingerprint-based L2O place recognition framework, which primarily consists of: (1) structural fingerprint-aware attention and (2) structural fingerprint extraction and matching.

We define five semantic categories, i.e., poles, traffic signs, tree trunks, building corners, and road forks, as structural anchors. Within the structural fingerprint-aware attention mechanism, the network utilizes these anchors as saliency prompts to transform sparse but spatially stable structural cues into salient responses in the map features. Through spatial context interaction, this mechanism enhances the representational capability of the anchors, thereby directing the network's attention toward potential fingerprint regions. Subsequently, the framework dynamically aggregates features based on their global contextual importance, simultaneously generating the global descriptor and the structural fingerprint-aware attention map.

For structural fingerprint extraction and matching, we formulate the process as an alternating optimization problem guided by the aforementioned attention map. Through iterative optimization, we explicitly extract and match geometrically consistent structural fingerprints composed of spatially stable anchors, thereby achieving fine-grained place recognition. Finally, the geometric consistency of the resulting structural fingerprints is fed back into the training phase as a geometric prior. This imposes rigorous constraints on the embedding space, effectively suppressing false matches characterized by high global similarity but contradictory local structures.

The main contributions of this paper are summarized as follows:

- We propose SFPR, a L2O place recognition framework based on the cognitive mechanisms of human spatial navigation. This framework goes beyond traditional retrieval paradigms relying exclusively on global similarity by introducing structural fingerprints, which are

composed of spatially stable anchors and their geometric structures, to achieve precise recognition.

- We propose a structural fingerprint-aware attention mechanism that utilizes sparse but spatially stable anchors as saliency prompts, and dynamically aggregates these anchor-guided features to guide the network's attention toward potential fingerprint regions. Simultaneously, it generates the global descriptor and extracts the structural fingerprint-aware attention map.
- We formulate structural fingerprint extraction and matching as an alternating optimization problem guided by fingerprint-aware attention. From these anchors, we explicitly extract a set of fingerprints characterized by high fingerprint-aware attention weights and geometric consistency. Finally, this consistency is fed back into the training phase as a geometric prior.

2 Related Work

Uni-modal 3D Place Recognition. An early representative work, PointNetVLAD [10], established the global descriptor paradigm for LiDAR place recognition. DH3D [11] introduced 3D local encoders to learn discriminative local descriptors; SOE-Net [12] incorporated orientation encoding to enhance spatial context modeling; and MinkLoc3D [13] leveraged sparse convolutions and GeM pooling to extract compact global features from sparse voxels. To address viewpoint variations, BEVPlace [14] utilized group convolutions to extract rotation-equivariant features and perform global aggregation, while CVTNet [15] fused range and BEV images to achieve rotation-invariant matching. Furthermore, CASSPR [16] utilized a point-voxel dual-branch architecture and hierarchical cross-attention mechanisms to enhance feature representation for single-frame sparse point clouds.

Cross-Modal 3D Place Recognition. Instead of relying on pre-built 3D maps, recent research has shifted toward utilizing alternative map priors for cross-modal place recognition [4, 17]. Lee et al. [18] reconstructed depth maps from short image sequences to establish a unified representation space with LiDAR range images, enabling recognition in street-view databases. Lip-Loc [19] utilized range image projection to achieve spatial alignment between point clouds and RGB images. C2L-PR [20] combined depth estimation and semantic segmentation to map 2D structures to 3D, mitigating field-of-view discrepancies through orientation voting. VXP [21] introduced a local descriptor similarity constraint to establish a shared voxel-pixel feature space through two-stage training. UniLoc [22] designed a universal retrieval framework supporting queries from any single modality. L2RSI [4] pioneered using remote sensing images as reference maps, mitigating cross-view and cross-modal discrepancies through semantic alignment and performing initial retrieval based on global descriptors.

Among various map priors, OSM demonstrates significant practical value for large-scale, low-cost cross-modal localization due to its global accessibility, lightweight nature, and rich semantic and geometric attributes [23, 24]. Early studies relied on sequential inputs or complex optimization frameworks [25, 26]. These methods aligned sensor trajectories with OSM road networks or building footprints, followed by pose estimation using Monte Carlo Localization [27] or graph-based SLAM [28]. To alleviate reliance on long trajectories and filtering algorithms, subsequent studies shifted toward constructing single-frame descriptors by leveraging handcrafted features, such as encoded street intersections and building gaps [29] or relative sensor-to-building distances [24]. However, these handcrafted features strongly depend on urban layouts. Lee and Ryu [30] proposed a neural network-based descriptor extraction method. Kang et al. [5] proposed OPAL, which constructs a rotation-invariant global descriptor by incorporating cross-modal visibility masks and adaptive radial fusion. Specifically, OPAL employs a visibility-aware polar backbone and learnable radial proposals to adaptively aggregate ring-level features for global descriptor retrieval.

Nevertheless, the core paradigm of these existing cross-modal methods indiscriminately compresses the entire scene into a single monolithic global representation, thereby neglecting the crucial information provided by fine-grained local structures.

Structure-Aware Place Recognition. Several studies explicitly model structural, geometric, or topological scene relationships. SegMap [31] learned segment-level descriptors from 3D point clouds and performed localization through geometrically consistent segment correspondences. SGPR [6] formulated the task as a graph matching problem, constructing a semantic graph through instance clustering and computing matching scores. BoxGraph [7] extracted physically meaningful object instances, utilized 3D bounding boxes as shape descriptors, and incorporated optimal assignment

to achieve training-free recognition. These methods demonstrate the value of explicit structural information for place recognition, but primarily operate on uni-modal 3D observations.

Geometric Verification for Place Recognition. Geometric verification complements descriptor-based retrieval by evaluating the spatial consistency of local correspondences to reject false matches and refine candidate rankings. In image retrieval, Stewenius et al. [32] introduced efficient exhaustive geometric verification for large-scale retrieval. Radenovic et al. [33] further exploited structure-from-motion geometry to mine reliable training pairs for descriptor learning. For point cloud retrieval, SpectralGV [34] employed spectral analysis of correspondence compatibility to re-rank candidates based on geometric consistency without explicit registration. Building upon the principle of geometric verification, SFPR performs attention-guided structural fingerprint matching across LiDAR and OSM, while further incorporating fingerprint consistency into descriptor learning.

3 Problem Statement

Let $\mathcal{O}_{\mathrm{ref}} = \{o_i\}_{i=1}^{N_{\mathrm{ref}}}$ denote the database of geo-referenced OSM candidate maps. The LiDAR-to-OSM place recognition task can be formulated as follows: given a query point cloud submap P, the goal is to retrieve the candidate map that is most consistent with the query. Previous works typically rank candidate locations according to global descriptor similarity:

$$o^* = \arg\max_{o_i \in \mathcal{O}_{\mathrm{ref}}} \mathrm{Sim}(F_q(P),\, F_{\mathrm{ref}}(o_i))\,, \tag{1}$$

where $F_q(\cdot)$ and $F_{\mathrm{ref}}(\cdot)$ denote the encoding networks of the point cloud branch and the reference OSM branch, respectively, and $\mathrm{Sim}(\cdot,\cdot)$ denotes global similarity in the shared feature space.

However, relying solely on global descriptors is insufficient to distinguish candidate scenes with similar global appearances but different local structures. To address this limitation, we augment the retrieval formulation with a structural fingerprint consistency term, redefining the objective as:

$$o^* = \arg\max_{o_i \in \mathcal{O}_{\mathrm{ref}}} \left[\mathrm{Sim}(F_q(P),\, F_{\mathrm{ref}}(o_i)) + \alpha\, \Phi(P, o_i)\right], \tag{2}$$

where $\Phi(P, o_i)$ measures the structural fingerprint consistency between the query and the candidate map, and α balances the contributions of global descriptor similarity and the fingerprint term.

4 Methodology

Figure 2 illustrates the overall architecture of the proposed SFPR framework, which primarily consists of four core steps: data preprocessing, structural fingerprint-aware attention, structural fingerprint extraction and matching, and fingerprint-guided embedding space reshaping.

4.1 Data Preprocessing

Semantic Segmentation and OSM Rasterization. For point clouds, we employ Cylinder3D [35], pre-trained on SemanticKITTI [36], to predict 19 semantic classes. For OSM, we rasterize each local tile into a three-channel Cartesian grid that encodes Area, Way, and Node elements [37].

Visibility Mask Generation. We follow OPAL [5] to explicitly model sensor visibility. For point clouds, the visibility mask is determined in polar coordinates by calculating the maximum observable range along each azimuth. For OSM data, buildings and wall-like structures are treated as occluders, and visibility is simulated via ray casting from the center of the local tile.

Spatially Stable Anchor Extraction. For point clouds, trunk, traffic sign, and pole anchors are obtained via Euclidean clustering and centroid estimation; building corners are extracted from the BEV layout using geometry-based corner detection; and road forks are extracted from road-skeleton intersections. For OSM, tree tags are mapped to Trunk; pole-like facilities and traffic signs are mapped to Pole and Traffic Sign, respectively; building corners are mapped to visible polygon vertices; and road forks are mapped to road-graph intersections with a degree of at least three.

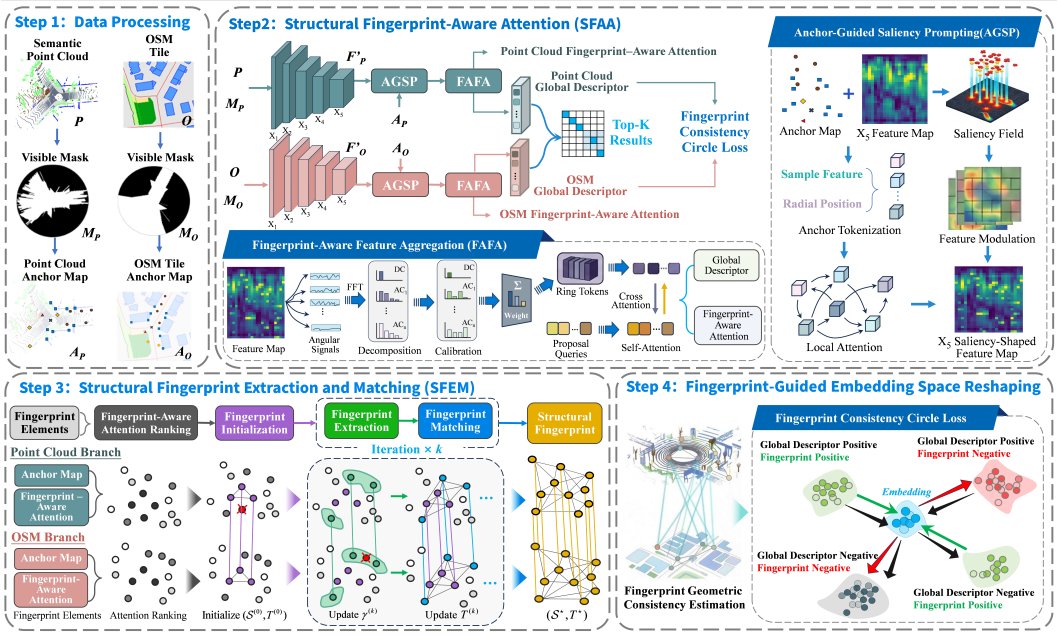

Figure 2: Overview of the proposed SFPR framework.

4.2 Structural Fingerprint-Aware Attention (SFAA)

Feature Extraction. Following the visibility-aware polar representation and backbone design of OPAL [5], we adopt a parallel dual-branch architecture for point cloud and OSM inputs. For the point cloud branch, the semantic point cloud P' is encoded by a lightweight MLP into C_{p}-dimensional features, scattered onto the polar BEV grid, and max-pooled per cell to obtain F_P. Concatenated with the visibility mask M_P, it is fed into the PolarNet [38] encoder to produce $F'_P \in \mathbb{R}^{N_r \times N_\theta \times C}$. For the OSM branch, the rasterized tile O is encoded in Cartesian space as F_O^{cart}, then resampled to the polar BEV grid to obtain F_O. Concatenated with the visibility mask M_O, it is processed by an unshared PolarNet encoder to produce $F'_O \in \mathbb{R}^{N_r \times N_\theta \times C}$, where N_r, N_θ, and C denote the radial, angular, and channel dimensions of the encoded feature maps after backbone downsampling, respectively.

Anchor-Guided Saliency Prompting (AGSP) Module. For each branch $b \in \{P, O\}$, let $F'_b \in \mathbb{R}^{N_r \times N_\theta \times C}$ denote the backbone feature map on the polar BEV grid, and let $\mathcal{A}_b = \{(\mathbf{x}_i^b, s_i^b)\}_{i=1}^{N_b}$ denote the corresponding anchor set, where $\mathbf{x}_i^b \in \mathbb{R}^2$ is the anchor location and $s_i^b \in \{1, \ldots, S\}$ is its semantic type. AGSP converts these sparse structural anchors into a spatial saliency field to modulate features in their neighborhoods:

$$\mathcal{S}_b(\mathbf{u}, c) = \sum_{k=1}^{S} \beta_{k,c} \max_{i:\, s_i^b = k} \mathcal{K}(\mathbf{u}, \mathbf{x}_i^b), \qquad \hat{F}_b(\mathbf{u}, c) = F'_b(\mathbf{u}, c) \left(1 + \kappa_b \tanh\left(\frac{\mathcal{S}_b(\mathbf{u}, c)}{\kappa_b}\right)\right), \tag{3}$$

where $\mathbf{u}$ denotes a spatial location on the encoded polar grid, c denotes the channel index, $\mathcal{K}(\mathbf{u}, \mathbf{x}_i^b)$ is a local Gaussian kernel centered at anchor $\mathbf{x}_i^b$, and $\beta_{k,c}$ is a learnable semantic-channel response coefficient. The positive learnable scale κ_b bounds the modulation response. The maximum is set to zero if no anchor of type k is present.

For each anchor, we construct a token $\mathbf{z}_i^b$ by concatenating the sampled feature $\hat{F}_b(\mathbf{x}_i^b)$ with its normalized radial coordinate. To capture local structural relationships, we define the geometric relation between neighboring anchors (i, j) as:

$$\mathbf{r}_{ij}^b = \left[d_{ij}^b,\ \Delta\rho_{ij}^b,\ \cos(\Delta\theta_{ij}^b),\ \sin(\Delta\theta_{ij}^b),\ \mathbf{1}[s_i^b = s_j^b]\right], \tag{4}$$

where d_{ij}^b, $\Delta\rho_{ij}^b$, and $\Delta\theta_{ij}^b$ denote the Euclidean distance, radial offset, and angular offset between anchors i and j, respectively. Let $\mathcal{N}_b(i) = \{j : \|\mathbf{x}_i^b - \mathbf{x}_j^b\|_2 < r_n\}$ denote the local anchor neighborhood within radius r_n in Cartesian coordinates, with a cap on its size for computational

efficiency. AGSP then performs relation-biased local attention over this neighborhood:

$$\tilde{\mathbf{z}}_i^b = \mathrm{Attn}\big(\mathbf{z}_i^b,\, \{\mathbf{z}_j^b\}_{j\in\mathcal{N}_b(i)},\, \{\psi(\mathbf{r}_{ij}^b)\}_{j\in\mathcal{N}_b(i)}\big)\,, \tag{5}$$

where $\psi(\mathbf{r}_{ij}^b)$ denotes the relation bias induced by the geometric relation feature.

The refined anchor representations produce difference-based residuals $\boldsymbol{\delta}_i^b$, which are scattered back to the polar feature map using local Gaussian kernels:

$$\boldsymbol{\delta}_i^b = \tanh\Big(W_{\mathrm{res}}^b(\tilde{\mathbf{z}}_i^b - \mathbf{z}_i^b)\Big)\,, \qquad \Delta F_b(\mathbf{u}) = \lambda_b \tanh\left(\sum_{i=1}^{N_b} \mathcal{G}(\mathbf{u}, \mathbf{x}_i^b)\boldsymbol{\delta}_i^b\right),$$

where W_{res}^b is a learnable projection, $\mathcal{G}$ is the local Gaussian scatter kernel, and λ_b is a bounded learnable gate controlling the residual magnitude. The final AGSP output is:

$$F_b^{\mathrm{AGSP}} = \hat{F}_b + \Delta F_b, \tag{6}$$

Fingerprint-Aware Feature Aggregation (FAFA) Module. Given the AGSP-enhanced feature map F_b^{AGSP}, we apply a Fast Fourier Transform (FFT) [39] along the angular dimension of each radial ring. Let $\mathbf{F}_{b,r}^{\mathrm{AGSP}} \in \mathbb{R}^{N_\theta \times C}$ denote the feature sequence on the r-th radial ring, where $r \in \{1, \ldots, N_r\}$. The calibrated spectral responses are compressed into a ring token:

$$\mathbf{h}_r^b = \sum_{\omega=0}^{K} \alpha_\omega^b \Big(\gamma_\omega\, g_\omega\big(\mathcal{F}_\varphi(\mathbf{F}_{b,r}^{\mathrm{AGSP}})\big) + \beta_\omega\Big), \tag{7}$$

where $\mathcal{F}_\varphi(\cdot)$ denotes the FFT along the angular dimension, $g_\omega(\cdot)$ extracts the response of the ω-th frequency component, K is the maximum retained frequency index, γ_ω and β_ω are learnable frequency calibration parameters, and α_ω^b denotes the sample-adaptive spectral weight. Collecting all radial ring tokens yields $R_b = [\mathbf{h}_1^b; \ldots; \mathbf{h}_{N_r}^b] \in \mathbb{R}^{N_r \times C}$.

For each branch, we introduce M learnable proposal queries [5, 40, 41], denoted by $Q \in \mathbb{R}^{M\times C}$ (with the branch index omitted for brevity). These queries are randomly initialized and optimized during training to aggregate ring-level structural information. Let $\bar{R}_b = R_b + E_{\mathrm{pos}}$, where $E_{\mathrm{pos}} \in \mathbb{R}^{N_r \times C}$ is the sinusoidal positional encoding [42] for radial order. The proposal queries are first refined by self-attention and then cross-attend to the position-enhanced ring tokens:

$$Q' = Q + \mathrm{softmax}\left(\frac{QQ^\top}{\sqrt{C}}\right) Q, \qquad \widetilde{Q}_b = Q' + \mathrm{softmax}\left(\frac{Q'\bar{R}_b^\top}{\sqrt{C}}\right) \bar{R}_b. \tag{8}$$

The global descriptor is obtained by projecting and vectorizing the aggregated query representation:

$$\mathbf{d}_b = \mathrm{vec}\Big(\Pi(\widetilde{Q}_b)\Big)\,, \tag{9}$$

where $\Pi(\cdot)$ denotes a learnable projection head and $\mathrm{vec}(\cdot)$ denotes vectorization.

The query-to-ring attention is also used to derive fingerprint-aware attention for structural anchor ranking. For the i-th anchor in branch b, let $r(i)$ denote its host ring index. Its fingerprint-aware attention weight c_i^b is defined as:

$$A_b = \mathrm{softmax}\left(\frac{Q'\bar{R}_b^\top}{\sqrt{C}}\right), \qquad c_i^b = \frac{1}{M}\sum_{m=1}^{M} (A_b)_{m,r(i)}, \tag{10}$$

where $A_b \in \mathbb{R}^{M\times N_r}$ is the query-to-ring attention matrix, with softmax over the ring dimension. Each anchor inherits the attention weight of its corresponding ring. The resulting scores guide subsequent fingerprint extraction and matching to prioritize promising correspondences, reducing the computational cost of exhaustive matching while maintaining comparable retrieval accuracy.

4.3 Structural Fingerprint Extraction and Matching (SFEM)

4.3.1 Optimization Problem Formulation

We formulate structural fingerprint extraction and matching as a joint optimization problem guided by fingerprint-aware attention. Specifically, we regard each anchor, together with its fingerprint-aware

attention weight, as a structural fingerprint element. Given the candidate correspondence set $\mathcal{U}$ between point-cloud and OSM fingerprint elements, we aim to select a subset with high fingerprint-aware attention weights and low geometric residuals. Thus, $\Phi(P, o_i)$ in Eq. (2) is formulated as:

$$(\mathcal{S}^*, T^*) = \arg \max_{\mathcal{S} \subseteq \mathcal{U},\, T \in SE(2)} \sum_{(i,j) \in \mathcal{S}} w_{ij} \left(1 - \lambda \left\| T\mathbf{x}_i^p - \mathbf{x}_j^o \right\|_2^2\right), \tag{11}$$

where $\mathcal{S}$ is the selected subset of fingerprint element correspondences, T is the relative transformation, w_{ij} is the fingerprint-aware matching weight for the candidate pair (i, j), λ controls the tolerance to fingerprint element geometric residuals, and $\mathbf{x}_i^p$ and $\mathbf{x}_j^o$ denote the 2D locations of the corresponding point-cloud and OSM fingerprint elements, respectively.

Truncated Quadratic Robust Reformulation. In cross-modal structural fingerprint matching, the candidate fingerprint element correspondence set $\mathcal{U}$ typically contains false matches with large geometric residuals. To suppress their influence on the fingerprint matching objective, we introduce a truncated quadratic kernel:

$$E(T) \triangleq \frac{1}{2} \sum_{(i,j) \in \mathcal{U}} w_{ij}\, \phi_\tau\left(\|T\mathbf{x}_i^p - \mathbf{x}_j^o\|_2^2\right), \qquad \phi_\tau(u) = \min\{u, \tau^2\} = \begin{cases} u, & u \le \tau^2, \\ \tau^2, & u > \tau^2. \end{cases} \tag{12}$$

Here, τ is the truncation threshold induced by the residual tolerance in Eq. (11).

Discrete Binary-Gated Reformulation. The hard truncation introduces a non-smooth transition at $u = \tau^2$, and the rigid-motion constraint $T \in SE(2)$ makes the fingerprint pose estimation non-convex. We therefore introduce binary gating variables to make structural fingerprint element selection explicit:

$$\widetilde{E}(T, \gamma) = \frac{1}{2} \sum_{(i,j) \in \mathcal{U}} w_{ij} \left[\gamma_{ij} \|T\mathbf{x}_i^p - \mathbf{x}_j^o\|_2^2 + (1 - \gamma_{ij})\tau^2\right], \tag{13}$$

where $\gamma_{ij} \in \{0, 1\}$ indicates whether the candidate fingerprint element correspondence (i, j) is selected as a geometrically consistent fingerprint match. This reformulation leads to two coupled subproblems, fingerprint extraction and matching, solved by alternating optimization.

4.3.2 Alternating Optimization

We initialize the optimization by selecting a small set of structural fingerprint elements with the highest fingerprint-aware attention to construct the initial pose estimate $T^{(0)}$. Starting from this initialization, the optimization alternates between Fingerprint Extraction and Fingerprint Matching.

Fingerprint Extraction. With the current fingerprint pose estimate $T^{(k)}$ fixed, we first perform residual-based fingerprint element selection:

$$\bar{\gamma}^{(k+1)} \in \arg \min_{\gamma \in \{0,1\}^{|\mathcal{U}|}} \widetilde{E}(T^{(k)}, \gamma), \tag{14}$$

where $\bar{\gamma}^{(k+1)}$ identifies candidate correspondences that satisfy the truncated geometric residual criterion. We then apply a greedy one-to-one assignment in ascending order of geometric residuals, ensuring that each anchor participates in at most one selected correspondence. The resulting binary selection is denoted by $\gamma^{(k+1)}$.

Fingerprint Matching. With the binary gating variables $\gamma^{(k+1)}$ fixed, the selected fingerprint element correspondence set is defined as $\mathcal{C}^{(k+1)} = \{(i, j) \in \mathcal{U} \mid \gamma_{ij}^{(k+1)} = 1\}$. The pose estimate is then updated by estimating the rigid transformation that best matches the selected structural fingerprint:

$$T^{(k+1)} = \arg \min_{T \in SE(2)} \sum_{(i,j) \in \mathcal{C}^{(k+1)}} w_{ij} \left\| T\mathbf{x}_i^p - \mathbf{x}_j^o \right\|_2^2. \tag{15}$$

These two steps are alternated until convergence or until the maximum number of iterations is reached. After optimization, the final selected fingerprint element correspondence set

$$\mathcal{S}^* = \{(i, j) \in \mathcal{U} \mid \gamma_{ij}^* = 1\}, \tag{16}$$

is defined as the structural fingerprint for the current query–map pair, while the final pose T^* is used to compute the cross-modal fingerprint similarity.

The equivalence of the discrete binary-gated reformulation, detailed derivations for Fingerprint Extraction and Fingerprint Matching, and the monotonicity proof of the optimization problem are deferred to Appendices A.10, A.11, and A.12, respectively.

4.4 Fingerprint-Guided Embedding Space Reshaping

During training, the standard Circle Loss [43] supervises the learning of global descriptors. Given the positive similarity s_{ii} between query i and its paired ground-truth map sample, and the negative similarity s_{ij} between query i and a non-matching candidate j, the loss is formulated as:

$$\begin{aligned} \mathcal{L}_{\mathrm{CL}} &= \frac{1}{B}\sum_{i=1}^{B}\log\left[1+\sum_{j\neq i}\exp\Big(\gamma\alpha_{ij}^{-}(s_{ij}-\Delta_n)\Big)\exp\Big(-\gamma\alpha_{ii}^{+}(s_{ii}-\Delta_p)\Big)\right], \\ \alpha_{ii}^{+} &= \max(0, O_p - s_{ii}), \qquad \alpha_{ij}^{-} = \max(0, s_{ij} - O_n), \end{aligned} \tag{17}$$

where $O_p = 1+m$, $O_n = -m$, $\Delta_p = 1-m$, and $\Delta_n = m$ denote optimal points and margins.

We propose structural fingerprint consistency as a geometric prior to reshape pairwise similarity relations in the embedding space. Given a fingerprint consistency score $G_{ij} \in [0,1]$ between query i and map sample j, where G_{ii} denotes the consistency of the ground-truth pair, we modulate both the adaptive weights and the decision margins of Circle Loss:

$$\begin{aligned} \widehat{\alpha}_{ii}^{+} &= \alpha_{ii}^{+}(1+\lambda_p G_{ii}), & \widehat{\Delta}_{p,i} &= \Delta_p + \eta_p G_{ii}, \\ \widehat{\alpha}_{ij}^{-} &= \alpha_{ij}^{-}(1+\lambda_n(1-G_{ij}))(1-\rho G_{ij}), & \widehat{\Delta}_{n,ij} &= \Delta_n + \eta_n(1-G_{ij}), \end{aligned} \tag{18}$$

Here, λ_p and λ_n control adaptive-weight modulation, ρ reduces the contribution of structurally consistent negative pairs, and η_p and η_n adjust the corresponding margins. Together, these factors regulate the influence of structural consistency on pairwise supervision. The resulting Fingerprint Consistency Circle Loss(FCCL) is:

$$\mathcal{L}_{\mathrm{FCCL}} = \frac{1}{B}\sum_{i=1}^{B}\log\left[1+\sum_{j\neq i}\exp\Big(\gamma\widehat{\alpha}_{ij}^{-}(s_{ij}-\widehat{\Delta}_{n,ij})\Big)\exp\Big(-\gamma\widehat{\alpha}_{ii}^{+}(s_{ii}-\widehat{\Delta}_{p,i})\Big)\right]. \tag{19}$$

To further mitigate high-similarity false positives, we introduce a Top-K structural penalty. In Eqs. (17)–(19), we assume $g(i) = i$. Here, we write $g(i)$ explicitly for generality:

$$\mathcal{L}_{\mathrm{FP}} = \frac{1}{B}\sum_{i=1}^{B}\frac{1}{K}\sum_{j\in\mathcal{N}_K(i)}(1-G_{ij})\max\big(0,\, s_{ij} - s_{i,g(i)} + m_{\mathrm{fp}}\big), \tag{20}$$

where $\mathcal{N}_K(i)$ denotes the Top-K non-matching candidates ranked by descriptor similarity, and m_{fp} is the false-positive margin. The factor $(1-G_{ij})$ reduces the penalty for structurally consistent negative pairs.

The final loss function is:

$$\mathcal{L} = \mathcal{L}_{\mathrm{FCCL}} + \lambda_{\mathrm{fp}}\mathcal{L}_{\mathrm{FP}}, \tag{21}$$

where λ_{fp} scales the false-positive suppression term.

5 Experiments

Table 1: Statistics of the test sets in KITTI, KITTI-360, and PandaSet.

Dataset	KITTI		KITTI-360				PandaSet
Sequence	00	07	00	05	06	09	San Francisco
Point Cloud Frames	4541	1101	10514	6291	9186	13247	3920
OSM Tiles	8782	3332	12491	15080	12730	13060	8457
Trajectory Length (m)	8478	3226	11612	14541	12201	12570	8095

Datasets. We evaluate SFPR on three public datasets: KITTI [45], KITTI-360 [46], and PandaSet [47]. The corresponding reference maps are obtained from the official OSM platform. The KITTI dataset was collected in Karlsruhe, Germany. KITTI-360 can be regarded as a large-scale extension of KITTI, covering mixed urban and suburban environments. PandaSet was mostly collected within the Silicon Valley region. The test sets for all datasets are summarized in Table 1, including the number of query point cloud frames, the number of OSM map tiles, and the total trajectory length.

Table 2: Top-1 Recall@Km on KITTI and PandaSet. SFPR-Rotate and SFPR-Offset denote SFPR evaluated with random point cloud rotations and OSM spatial offsets, respectively.

Method	KITTI						PandaSet		
	00			07			San Francisco		
	1m	5m	10m	1m	5m	10m	1m	5m	10m
SC [44]	10.31	30.54	31.16	30.61	42.42	43.42	1.61	14.01	15.03
Building^1 [24]	5.31	19.89	20.52	7.63	18.53	19.35	1.53	11.96	14.43
Building^2 [24]	17.09	48.23	48.93	29.43	45.32	45.78	4.03	22.91	25.53
C2L-PR [20]	1.39	9.69	12.20	2.27	13.17	17.26	0.24	4.39	5.93
OPAL [5]	21.82	65.78	66.40	45.41	69.85	70.30	2.50	13.34	15.33
SFPR(w/o SFEM)	27.66	82.36	83.22	46.50	74.57	75.57	3.70	21.89	23.32
SFPR	**28.41**	**88.37**	**88.72**	**58.40**	**93.19**	**94.01**	**4.57**	**26.68**	**28.29**
SFPR-Rotate	27.73	88.61	88.95	58.31	93.46	94.10	4.62	26.91	28.49
SFPR-Offset	28.28	88.70	88.97	57.13	93.10	93.82	4.77	26.84	28.75

Table 3: Top-1 Recall@Km on KITTI-360. SFPR-Rotate and SFPR-Offset denote SFPR evaluated with random point cloud rotations and OSM spatial offsets, respectively.

Method	KITTI-360											
	00			05			06			09		
	1m	5m	10m	1m	5m	10m	1m	5m	10m	1m	5m	10m
SC [44]	15.14	39.61	40.66	3.69	16.69	17.18	4.14	14.22	14.59	13.92	27.45	28.07
Building^1 [24]	5.22	15.76	17.07	0.87	3.91	4.61	0.60	3.18	3.79	4.21	12.16	13.27
Building^2 [24]	17.12	49.61	51.29	4.23	15.94	16.29	3.00	12.82	13.39	18.28	41.92	42.64
C2L-PR [20]	1.70	8.23	10.93	0.81	4.72	6.36	0.45	3.35	4.50	1.69	9.59	12.45
OPAL [5]	14.92	42.82	44.18	7.74	30.49	31.55	7.71	36.38	37.54	27.89	60.96	61.92
SFPR(w/o SFEM)	21.44	52.25	53.30	10.81	44.32	45.54	9.04	42.54	43.27	37.61	75.77	76.80
SFPR	**25.60**	**64.34**	**65.19**	**11.87**	**49.12**	**50.17**	**10.85**	**49.83**	**50.44**	**39.10**	**81.43**	**82.23**
SFPR-Rotate	25.67	64.66	65.50	11.75	49.58	50.66	10.95	48.07	49.76	39.13	81.77	82.54
SFPR-Offset	25.38	64.89	65.86	12.05	49.61	50.85	10.77	48.35	49.07	38.58	81.96	82.68

Baseline Methods. We compare SFPR with the following representative methods: (1) **Building**[24], a pioneering LiDAR-to-OSM localization framework, including a simplified single-stage variant (Building^1) and the original two-stage version (Building^2); (2) **SC**[44], a classical point cloud place recognition method based on handcrafted descriptors; (3) **C2L-PR**[20], a hybrid cross-modal place recognition framework for image-to-point-cloud retrieval; and (4) **OPAL**[5], the current state-of-the-art method for LiDAR-to-OSM place recognition, which extracts robust global descriptors with cross-modal visibility masks and adaptive radial fusion.

Evaluation Criteria.
We adopt standard Recall@Km as the primary evaluation metric, defined as the proportion of queries whose Top-1 retrieved result falls within a radius of K meters from the ground-truth location, thereby quantifying retrieval accuracy.

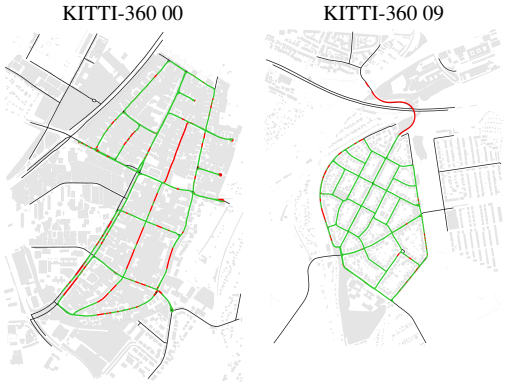

Figure 3: Top-1 retrieved results at the 5 m threshold. Black points (•) denote OSM tile locations, while red points (•) and green points (•) indicate incorrect and corrected results, respectively.

5.1 Results

Evaluation on Test Sets. Tables 2 and 3 show that SFPR consistently outperforms all baselines on Sequences 00 and 07 of KITTI, achieving a Top-1 retrieval accuracy of 94% on Sequence 07. Moreover, SFPR attains the best performance in the zero-shot generalization experiments on KITTI-360 and PandaSet, demonstrating robustness under challenging cross-city, cross-region, and cross-scene settings. Figure 3 presents retrieval visualizations of SFPR along the driving trajectory on KITTI-360. Overall, these results show that the proposed structural fingerprint substantially improves the accuracy of cross-modal place recognition while maintaining strong robustness.

Table 4: Ablation study on module effectiveness. ✓ and – indicate whether the module is enabled.

FAFA	AGSP	SFEM	FCCL	KITTI-360 00		
				1m	5m	10m
–	–	–	–	14.92	42.82	44.18
✓	–	–	–	18.40	47.52	49.23
✓	✓	–	–	20.29	50.86	52.18
✓	–	✓	–	23.05	60.52	61.71
✓	–	–	✓	20.19	50.44	51.85
✓	✓	–	✓	21.44	52.25	53.30
✓	✓	✓	–	23.89	61.58	62.79
✓	–	✓	✓	24.86	63.72	64.73
✓	✓	✓	✓	**25.60**	**64.34**	**65.19**

Rotation and Offset Robustness. To evaluate the model's robustness to viewpoint variations, we apply random rotations within $[0, 2\pi)$ to the query point clouds. As shown in Tables 2 and 3 (SFPR-Rotate), SFPR remains robust under pose perturbations. Furthermore, to investigate the impact of map data quality, we introduce random offsets within $[0\,\mathrm{m}, 0.5\,\mathrm{m}]$ into the OSM data to simulate localization noise. As shown in Tables 2 and 3 (SFPR-Offset), our method also exhibits consistent robustness against map noise.

Ablation Study. To quantify the contribution of each module, we report ablation results in Table 4. The FAFA module captures the low-frequency global layout and high-frequency local details within each radial ring via FFT-based spectral decomposition, and performs global contextual aggregation through dynamic proposal queries, yielding a rotation-robust global descriptor and significantly improving baseline accuracy. The AGSP module transforms stable anchors into saliency prompts, improving retrieval performance while extracting fingerprint-aware attention. Subsequently, the SFEM module performs geometry-consistent structural fingerprint extraction and matching based on this attention, eliminating false matches that are globally similar but locally geometrically inconsistent. Finally, FCCL serves as a geometric prior during training by feeding fingerprint consistency back into the metric learning process to reshape the embedding space, further enhancing the model's ability to distinguish hard negatives.

6 Conclusion

Inspired by human spatial navigation, we present SFPR, a structural fingerprint-based L2O place recognition framework that moves beyond traditional global descriptor similarity by introducing fine-grained discrimination via fingerprints. Specifically, we propose a fingerprint-aware attention mechanism leveraging sparse but spatially stable anchors to guide the network's focus toward potential fingerprint regions. We then formulate fingerprint extraction and matching as a fingerprint-attention-guided alternating optimization problem and feed the geometric consistency of fingerprints back into training to reshape the embedding space. Experiments demonstrate that SFPR effectively overcomes retrieval bottlenecks in macroscopically homogeneous scenes. Future work will aim to address the stable fingerprint extraction in low-beam point clouds to further enhance the framework's deployment capability in real-world environments.

Acknowledgments and Disclosures

Funding. This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62471415 and 62401225), the Natural Science Foundation of Fujian Province (Grant Nos. 2023J01004, 2025J0141, and 2024J01115), and the Xiamen Municipal Natural Science Foundation (Grant No. 3502Z202472018).

Competing Interests. The authors declare that they have no competing financial interests that could have influenced the work reported in this paper.

References

[1] Yuhao Li, Jianping Li, Zhen Dong, Yuan Wang, and Bisheng Yang. Saliencyi2ploc: Saliency-guided image–point cloud localization using contrastive learning. *Information Fusion*, 118:103015, 2025.

[2] Ethan Griffiths, Maryam Haghighat, Simon Denman, Clinton Fookes, and Milad Ramezani. Hotformerloc: Hierarchical octree transformer for versatile lidar place recognition across ground and aerial views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6648–6658, 2025.

[3] Wen Li, Shangshu Yu, Cheng Wang, Guosheng Hu, Siqi Shen, and Chenglu Wen. Sgloc: Scene geometry encoding for outdoor lidar localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9286–9295, 2023.

[4] Ziwei Shi, Xiaoran Zhang, Wenjing Xu, Yan Xia, Yu Zang, Siqi Shen, and Cheng Wang. L2RSI: Cross-view liDAR-based place recognition for large-scale urban scenes via remote sensing imagery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[5] Shuhao Kang, Youqi Liao, Yan Xia, Olaf Wysocki, Boris Jutzi, and Daniel Cremers. OPAL: Visibility-aware liDAR-to-openstreetmap place recognition via adaptive radial fusion. In *9th Annual Conference on Robot Learning*, 2025.

[6] Xin Kong, Xuemeng Yang, Guangyao Zhai, Xiangrui Zhao, Xianfang Zeng, Mengmeng Wang, Yong Liu, Wanlong Li, and Feng Wen. Semantic graph based place recognition for 3d point clouds. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8216–8223. IEEE, 2020.

[7] Georgi Pramatarov, Daniele De Martini, Matthew Gadd, and Paul Newman. Boxgraph: Semantic place recognition and pose estimation from 3d lidar. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7004–7011. IEEE, 2022.

[8] Sijie Wang, Rui She, Qiyu Kang, Siqi Li, Disheng Li, Tianyu Geng, Shangshu Yu, and Wee Peng Tay. Multi-modal aerial-ground cross-view place recognition with neural odes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 11717–11728, 2025.

[9] Russell A Epstein, Eva Zita Patai, Joshua B Julian, and Hugo J Spiers. The cognitive map in humans: spatial navigation and beyond. *Nature neuroscience*, 20(11):1504–1513, 2017.

[10] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4470–4479, 2018.

[11] Juan Du, Rui Wang, and Daniel Cremers. Dh3d: Deep hierarchical 3d descriptors for robust large-scale 6dof relocalization. In *European Conference on Computer Vision*, pages 744–762. Springer, 2020.

[12] Yan Xia, Yusheng Xu, Shuang Li, Rui Wang, Juan Du, Daniel Cremers, and Uwe Stilla. Soe-net: A self-attention and orientation encoding network for point cloud based place recognition. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11348–11357, 2021.

[13] Jacek Komorowski. Minkloc3d: Point cloud based large-scale place recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1790–1799, 2021.

[14] Lun Luo, Shuhang Zheng, Yixuan Li, Yongzhi Fan, Beinan Yu, Si-Yuan Cao, Junwei Li, and Hui-Liang Shen. Bevplace: Learning lidar-based place recognition using bird's eye view images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8700–8709, 2023.

[15] Junyi Ma, Guangming Xiong, Jingyi Xu, and Xieyuanli Chen. Cvtnet: A cross-view transformer network for lidar-based place recognition in autonomous driving environments. *IEEE Transactions on Industrial Informatics*, 20(3):4039–4048, 2023.

[16] Yan Xia, Mariia Gladkova, Rui Wang, Qianyun Li, Uwe Stilla, Joao F Henriques, and Daniel Cremers. Casspr: Cross attention single scan place recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8461–8472, 2023.

[17] 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*, pages 14958–14967, 2024.

[18] Alex Junho Lee, Seungwon Song, Hyungtae Lim, Woojoo Lee, and Hyun Myung. L^2: Lidar-camera loop constraints for cross-modal place recognition. *IEEE Robotics and Automation Letters*, 8(6):3589–3596, 2023.

[19] Sai Shubodh, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–957, 2024.

[20] Huaiyuan Xu, Huaping Liu, Shoudong Huang, and Yuxiang Sun. C2l-pr: Cross-modal camera-to-lidar place recognition via modality alignment and orientation voting. *IEEE Transactions on Intelligent Vehicles*, 2024.

[21] Yun-Jin Li, Mariia Gladkova, Yan Xia, Rui Wang, and Daniel Cremers. Vxp: Voxel-cross-pixel large-scale camera-lidar place recognition. In *2025 International Conference on 3D Vision (3DV)*, pages 1233–1242. IEEE, 2025.

[22] Yan Xia, Zhendong Li, Yun-Jin Li, Letian Shi, Hu Cao, João F Henriques, and Daniel Cremers. Uniloc: Towards universal place recognition using any single modality. *arXiv preprint arXiv:2412.12079*, 2024.

[23] Benjamin Suger and Wolfram Burgard. Global outer-urban navigation with openstreetmap. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1417–1422. IEEE, 2017.

[24] Younghun Cho, Giseop Kim, Sangmin Lee, and Jee-Hwan Ryu. Openstreetmap-based lidar global localization in urban environment without a prior lidar map. *IEEE Robotics and Automation Letters*, 7(2):4999–5006, 2022.

[25] Georgios Floros, Benito Van Der Zander, and Bastian Leibe. Openstreetslam: Global vehicle localization using openstreetmaps. In *2013 IEEE international conference on robotics and automation*, pages 1054–1059. IEEE, 2013.

[26] Philipp Ruchti, Bastian Steder, Michael Ruhnke, and Wolfram Burgard. Localization on openstreetmap data using a 3d laser scanner. In *2015 IEEE international conference on robotics and automation (ICRA)*, pages 5260–5265. IEEE, 2015.

[27] F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *Proceedings 1999 IEEE International Conference on Robotics and Automation (Cat. No.99CH36288C)*, volume 2, pages 1322–1328 vol.2, 1999.

[28] Henning Lategahn, Andreas Geiger, and Bernd Kitt. Visual slam for autonomous ground vehicles. In *2011 IEEE international conference on robotics and automation*, pages 1732–1737. IEEE, 2011.

[29] Fan Yan, Olga Vysotska, and Cyrill Stachniss. Global localization on openstreetmap using 4-bit semantic descriptors. In *2019 European conference on mobile robots (ECMR)*, pages 1–7. IEEE, 2019.

[30] Sangmin Lee and Jee-Hwan Ryu. Autonomous vehicle localization without prior high-definition map. *IEEE Transactions on Robotics*, 40:2888–2906, 2024.

[31] Renaud Dube, Andrei Cramariuc, Daniel Dugas, Hannes Sommer, Marcin Dymczyk, Juan Nieto, Roland Siegwart, and Cesar Cadena. Segmap: Segment-based mapping and localization using data-driven descriptors. *The International Journal of Robotics Research*, 39(2-3):339–355, 2020.

[32] Henrik Stewénius, Steinar H Gunderson, and Julien Pilet. Size matters: exhaustive geometric verification for image retrieval. In *European Conference on Computer Vision*, pages 674–687. Springer, 2012.

[33] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

[34] Kavisha Vidanapathirana, Peyman Moghadam, Sridha Sridharan, and Clinton Fookes. Spectral geometric verification: Re-ranking point cloud retrieval for metric localization. *IEEE Robotics and Automation Letters*, 8(5):2494–2501, 2023.

[35] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Wei Li, Yuexin Ma, Hongsheng Li, Ruigang Yang, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar-based perception. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6807–6822, 2021.

[36] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

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

[38] Yang Zhang, Zixiang Zhou, Philip David, Xiangyu Yue, Zerong Xi, Boqing Gong, and Hassan Foroosh. Polarnet: An improved grid representation for online lidar point clouds semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9601–9610, 2020.

[39] Pierre Duhamel and Martin Vetterli. Fast fourier transforms: a tutorial review and a state of the art. *Signal processing*, 19(4):259–299, 1990.

[40] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[41] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. Boq: A place is worth a bag of learnable queries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17794–17803, 2024.

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

[43] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.

[44] Giseop Kim and Ayoung Kim. Scan context: Egocentric spatial descriptor for place recognition within 3d point cloud map. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4802–4809. IEEE, 2018.

[45] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *International Journal of Robotics Research (IJRR)*, pages 1231–1237, 2013.

[46] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):3292–3310, 2022.

[47] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *2021 IEEE international intelligent transportation systems conference (ITSC)*, pages 3095–3101. IEEE, 2021.

[48] Xin Lai, Yukang Chen, Fanbin Lu, Jianhui Liu, and Jiaya Jia. Spherical transformer for lidar-based 3d recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17545–17555, 2023.

[49] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

A Appendix

A.1 Limitations

While the SFPR framework demonstrates robust retrieval performance, its generalization capability remains constrained when applied to low-beam (e.g., 32-beam) LiDAR data. The limited detection range and extreme spatial sparsity inherent to such sensors make the extraction of stable structural fingerprints and discriminative global features highly challenging. Consequently, direct single-frame cross-modal matching often encounters a severe information bottleneck. To address this, our future work will explore strategies to overcome the place recognition challenges posed by highly sparse point clouds from low-beam LiDARs.

A.2 Impact of Semantic Label Quality

Table 5 demonstrates the impact of semantic label quality on model performance. When employing predicted labels generated by off-the-shelf models like Cylinder3D or SphereFormer, the retrieval performance experiences only a slight degradation compared to using Ground Truth labels. This proves that SFPR does not heavily rely on highly accurate, dense semantic annotations. Its intrinsic fingerprint-aware attention and explicit matching mechanisms effectively concentrate the feature representations on sparse yet highly discriminative structural fingerprint clues. Consequently, even in the presence of global semantic prediction noise, the model maintains high-precision retrieval performance as long as the core structural features are reliably extracted. This fully validates the robustness and immense potential of deploying SFPR in real-world, perception-constrained scenarios.

Table 5: Impact of semantic label quality on place recognition performance.

Semantic Label	KITTI 00		
	1m	5m	10m
Cylinder3D [35]	28.41	88.37	88.72
SphereFormer [48]	28.98	88.83	89.02
Ground Truth [36]	**30.98**	**89.03**	**89.12**

A.3 Effectiveness of SFAA on Fingerprint Extraction and Matching

To validate the necessity of the proposed SFAA guidance strategy, we design SFPR (w/o SFAA Guidance), which discards the SFAA-guided iterative extraction and matching process. Instead, SFPR (w/o SFAA Guidance) exhaustively matches point-cloud fingerprint elements against all fingerprint elements within each candidate OSM tile. As shown in Table 6, exhaustive matching achieves slightly higher recall by exploring a larger correspondence space, but incurs substantially greater computational cost. In contrast, SFAA-guided matching prioritizes high-confidence anchors and progressively expands geometrically consistent correspondences, reducing the matching time from 604.30 to 106.26 ms on KITTI 00 and from 642.64 to 108.53 ms on KITTI 07. These results indicate that the primary benefit of SFAA guidance during SFEM is improved matching efficiency with a modest reduction in recall.

Table 6: Comparison between exhaustive matching and SFAA-guided fingerprint extraction and matching over the Top-15 OSM candidates. SFPR (w/o SFAA Guidance) exhaustively matches all fingerprint elements, whereas SFPR performs progressive extraction and matching guided by SFAA.

Dataset	Method	Recall@Km (%)			Time (ms)
		1m	5m	10m	Top-15 Candidates
KITTI 00	SFPR (w/o SFAA Guidance)	**29.20**	**90.55**	**90.91**	604.30
	SFPR	28.41	88.37	88.72	**106.26**
KITTI 07	SFPR (w/o SFAA Guidance)	**60.04**	**94.37**	**94.55**	642.64
	SFPR	58.40	93.19	94.01	**108.53**

A.4 Precision–Recall Analysis

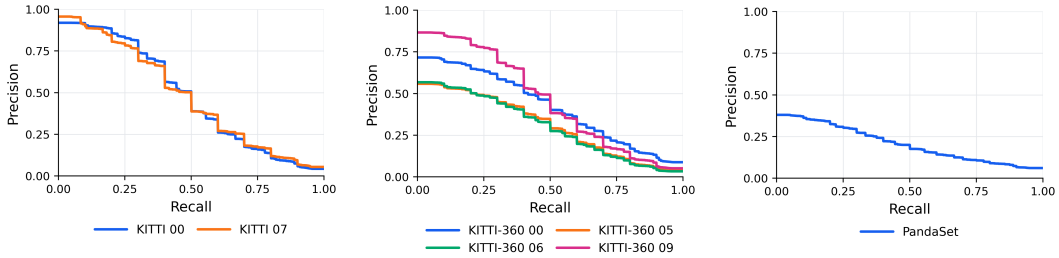

Figure 4: Precision–recall curves of SFPR under a 5 m spatial relevance threshold.

We report the precision–recall curves of SFPR across KITTI, KITTI-360, and PandaSet under a 5 m spatial relevance threshold, as shown in Fig. 4. On KITTI, sequences 00 and 07 exhibit comparable retrieval performance, both maintaining relatively high precision at low recall levels. Under zero-shot evaluation on KITTI-360, SFPR's performance varies across sequences while retaining favorable precision in the low-recall regime. On the more demanding PandaSet, precision gradually decreases as recall increases.

A.5 Robustness to Structural Perturbations

To further evaluate the robustness of SFPR to imperfect OSM data and structural anchor extraction, we conduct additional experiments on KITTI sequence 07 under six types of perturbations: OSM spatial offsets, anchor deletion, anchor semantic noise, anchor coordinate noise, OSM element dropout, and OSM map dropout. For OSM element and map dropout, we evaluate SFPR-Aug, retrained with the corresponding dropout augmentation, whereas the other perturbations use the standard SFPR model without additional perturbation-specific training. These perturbations simulate localization errors, unreliable anchor observations, and incomplete map information. The results are summarized in Table 7.

As shown in Table 7, SFPR exhibits different sensitivities to structural perturbations. Under OSM offsets of up to 5m, Recall@5m remains at 86.10%, although Recall@1m decreases substantially. With up to 40% anchor deletion or semantic noise, Recall@5m remains above 81%. In contrast, larger anchor coordinate errors and incomplete OSM regions lead to more pronounced performance degradation. These results indicate that SFPR maintains useful retrieval performance under moderate structural perturbations, while fine-grained localization remains sensitive to geometric inaccuracies and missing map information.

A.6 Runtime Analysis

We evaluate the inference time on a workstation equipped with an NVIDIA RTX 4090 GPU. As detailed in Table 8, SFPR takes only 6.80 ms for point cloud descriptor generation and 5.24 ms for OSM tile descriptor generation. Although the explicit extraction and geometric matching of structural fingerprints introduce an additional 106.26 ms, SFPR still maintains highly competitive computational efficiency, achieving an optimal trade-off between computational overhead and retrieval performance.

Table 7: Robustness to structural perturbations on KITTI sequence 07.

Perturbation	Level	R@1m	R@5m	R@10m
OSM Offset	0m	58.40	93.19	94.01
	0–5m	26.98	86.10	89.92
	0–10m	14.17	55.59	81.74
Anchor Deletion	0%	58.40	93.19	94.01
	0–20%	49.77	88.01	89.19
	0–40%	46.05	81.20	82.29
Anchor Semantic Noise	0%	58.40	93.19	94.01
	0–20%	53.32	89.83	90.55
	0–40%	44.05	82.74	83.92
Anchor Coordinate Noise	0m	58.40	93.19	94.01
	1m	48.50	85.74	86.56
	2m	37.87	72.66	73.57
OSM Element Dropout	0%	58.40	93.19	94.01
	0–10%	42.33	89.83	90.46
	0–20%	30.79	81.11	81.93
OSM Map Dropout	0%	58.40	93.19	94.01
	0–10%	40.96	85.92	86.74
	0–20%	26.88	74.57	75.84

Table 8: Runtime (ms) of descriptor generation and SFEM.

Method	Point Cloud	OSM Tile	SFEM	Total
SC [44]	31.54	16.68	–	48.22
Building [24]	29.86	54.87	–	84.73
C2L-PR [20]	219.08	316.71	–	535.79
OPAL [5]	1.91	5.14	–	7.05
SFPR	6.80	5.24	106.26	118.3

A.7 Qualitative Retrieval Results

Figure 5 presents a qualitative comparison of Top-1 retrieval results across the KITTI, KITTI-360, and PandaSet datasets under highly challenging query scenarios. As illustrated, baseline methods are easily confused by similar global layouts and mistakenly retrieve places that are globally similar but possess completely different local details. By explicitly leveraging structural fingerprints, the proposed SFPR successfully disambiguates these confusing environments, yielding accurate Top-1 retrievals and demonstrating robustness across diverse scenes.

A.8 Data Processing

Point Cloud Semantic Segmentation. To ensure semantic consistency across point cloud datasets, we unify the semantic labels. For KITTI and KITTI-360, we employ a pre-trained Cylinder3D [35] model to generate 19-class semantic labels following the SemanticKITTI [36] taxonomy. Since PandaSet [47] provides native semantic annotations, we map its original labels to the same SemanticKITTI taxonomy to maintain compatibility. The resulting semantic point clouds are visualized in Figure 6(a).

OSM Tile Rasterization. For OSM maps, we convert the raw data into geo-referenced semantic representations. Specifically, OSM elements are parsed and categorized into three primary classes—Area, Way, and Node—according to the hierarchy defined in Table 9. These elements are subsequently projected onto a local East-North coordinate system and rasterized into a 2D Cartesian grid at a fixed spatial resolution of $\Delta_o = 50$ cm/pixel [5]. The detailed rasterization pipeline is illustrated in Figure 7, and the resulting OSM tile visualization is shown in Figure 8(a).



Figure 5: Qualitative comparison of LiDAR queries and their Top-1 retrieval results within 1m on the KITTI, KITTI-360, and PandaSet datasets. Light blue boxes indicate the query point cloud submaps, black boxes represent the ground truth, green boxes denote correct retrievals, and red boxes indicate incorrect retrievals.

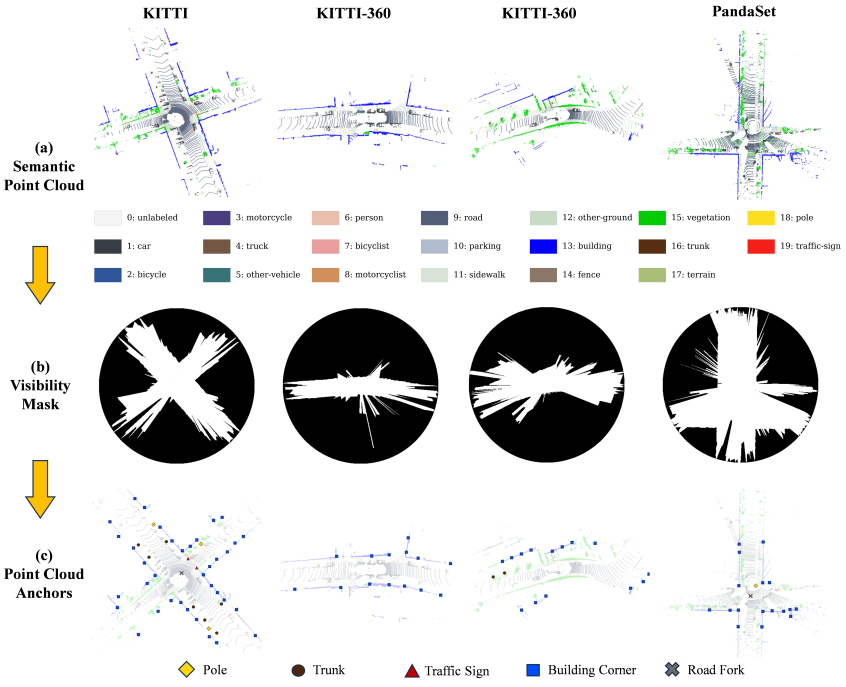

Figure 6: Visualization of the LiDAR point cloud processing pipeline.

Table 9: Details of OSM elements.

Type	Element
Areas	building, parking, playground, grass, park, forest, water
Ways	fence, wall, hedge, kerb, building outline, cycleway, path, road, busway, tree row
Nodes	parking entrance, street lamp, junction, traffic signal, stop sign, give way sign, bus stop, stop area, crossing, gate, bollard, gas station, bicycle parking, charging station, shop, restaurant, bar, vending machine, pharmacy, tree, stone, ATM, toilets, water fountain, bench, waste basket, post box, artwork, recycling station, clock, fire hydrant, pole, street cabinet

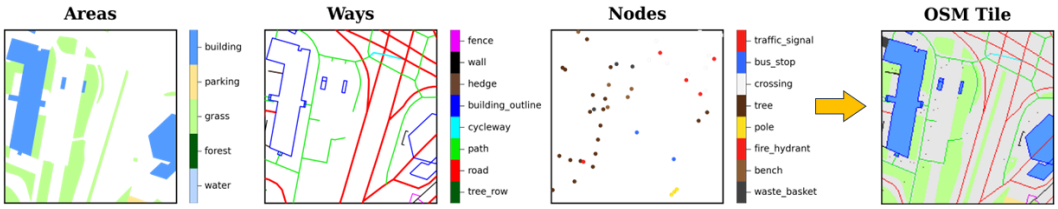

Figure 7: Illustration of area, way, and node channels, together with the full OSM tile.



Figure 8: Visualization of the OSM data processing pipeline.

Visibility Mask Generation. Following OPAL [5], we construct modality-specific visibility masks to alleviate occlusion inconsistencies between LiDAR observations and OSM maps. The masks are estimated on a common polar BEV grid with U radial bins and V angular sectors. Given the valid radial interval $[R_{\min}, R_{\max}]$, the radial and angular resolutions are defined as:

$$\Delta r = \frac{R_{\max} - R_{\min}}{U}, \qquad \Delta\phi = \frac{2\pi}{V}. \tag{22}$$

The center of polar cell (u, v) is given by:

$$r_{u,v} = R_{\min} + \left(u + \frac{1}{2}\right)\Delta r, \qquad \phi_{u,v} = \left(v + \frac{1}{2}\right)\Delta\phi - \pi, \tag{23}$$

where $u \in \{0, \ldots, U-1\}$ and $v \in \{0, \ldots, V-1\}$.

For a LiDAR scan, points are grouped according to their ranges and azimuths. Since regions behind the farthest valid return along a ray are not directly observable, the LiDAR visibility mask $M_P \in \{0,1\}^{U\times V}$ is computed sector-wise as:

$$M_P(u,v) = \begin{cases} 1, & r_{u,v} \le r^P_{\max}(v), \\ 0, & \text{otherwise}, \end{cases} \tag{24}$$

where $r^P_{\max}(v)$ is the maximum observed LiDAR range in angular sector v.

For OSM, map elements are first rasterized in a local Cartesian coordinate frame. To evaluate visibility in the same polar layout, each polar cell is associated with Cartesian sampling coordinates:

$$x_{u,v} = r_{u,v}\cos\phi_{u,v}, \qquad y_{u,v} = r_{u,v}\sin\phi_{u,v}, \tag{25}$$

and the corresponding raster value is obtained by bilinear sampling. Unlike LiDAR, OSM does not provide range returns; therefore, we infer visibility from semantic occluders. In practice, buildings and wall-like structures are used as physical occluders due to their vertical extent and continuity. Let $u^O_{\text{occ}}(v)$ denote the nearest occluder index in sector v. The OSM visibility mask $M_O \in \{0,1\}^{U\times V}$ is defined as:

$$M_O(u,v) = \begin{cases} 0, & u > u^O_{\text{occ}}(v), \quad \text{if an occluder exists in sector } v, \\ 1, & \text{otherwise.} \end{cases} \tag{26}$$

The masks M_P and M_O are concatenated with the corresponding polar BEV features, providing visibility-aware inputs for the two branches.

Figures 6(b) and 8(b) visualize the resulting visibility masks for the point cloud and OSM modalities, respectively.

Visualization of Anchors. Figures 6(c) and 8(c) visualize the anchors extracted from LiDAR scans and OSM tiles, respectively. In particular, pole-like objects, traffic signs, tree trunks, building corners, and road junctions correspond to localized and spatially stable structures in the physical environment. Although the two modalities differ substantially, the extracted anchors exhibit clear cross-modal correspondence. Therefore, they provide a sparse yet informative structural representation of the scene layout. Together with the subsequent fingerprint-aware attention, these anchors form the structural fingerprint elements.

A.9 Implementation Details

To ensure strict spatial separation between the training and testing sets, we use KITTI sequences 01, 02, 04, 05, 06, and 08 for training, while sequences 00 and 07 are used for testing. Sequence 03 is omitted due to the absence of ground-truth poses. For zero-shot generalization evaluation, we utilize sequences 00, 05, 06, and 09 from KITTI-360, along with the San Francisco sequences from PandaSet.

During data processing, the point clouds are cropped to retain points within a radial range from 3m to 50m. The corresponding OSM tiles are configured with a spatial scale of 100m $\times$ 100m. During training, OSM tiles are extracted exactly at the ground-truth poses. For evaluation, they are uniformly sampled along the testing trajectories at 1m intervals. The polar BEV representation is discretized into $R = 480$ radial rings and $P = 360$ angular sectors.

Table 10: Hyperparameter sensitivity on KITTI sequence 07. Top-1 Recall@Km (%) is reported. Each parameter is varied independently while others remain fixed.

Parameter	Value	R@1m	R@5m	R@10m
α	0.3	56.68	90.10	90.92
	0.7	58.40	93.19	94.01
	1.1	58.07	92.73	93.46
λ_{fp}	0	57.95	91.82	92.46
	0.025	60.76	94.19	95.10
	0.05	58.40	93.19	94.01
m_{fp}	0	58.67	94.37	95.10
	0.1	58.40	93.19	94.01
	0.2	60.40	94.94	95.78

The proposed SFPR is implemented in PyTorch [49]. The model is trained using the AdamW optimizer [50] with a batch size of 20. The learning rate is initialized at 1×10^{-3}, with a linear warm-up phase followed by a cosine annealing schedule that gradually decays the learning rate to 1×10^{-5}. All experiments are conducted on an NVIDIA GeForce RTX 4090 GPU with 24GB of memory.

Hyperparameter Sensitivity. We evaluate the sensitivity of three key hyperparameters, including α, λ_{fp}, and m_{fp}, on KITTI sequence 07. As shown in Table 10, SFPR maintains relatively stable performance across moderate parameter variations. Moreover, the same default configuration is used across all datasets, suggesting that SFPR does not require extensive dataset-specific hyperparameter tuning.

A.10 Reformulation of the Structural Fingerprint Extraction and Matching Problem

For each candidate fingerprint element pair $(i, j) \in \mathcal{U}$, we define the squared geometric residual under a rigid transformation $T \in SE(2)$ as

$$u_{ij}(T) = \left\| T\mathbf{x}_i^p - \mathbf{x}_j^o \right\|_2^2, \tag{27}$$

where $\mathbf{x}_i^p$ and $\mathbf{x}_j^o$ denote the 2D coordinates of the corresponding point-cloud and OSM fingerprint elements, respectively.

Using the fingerprint element geometric residual in Eq. (27), the objective in Eq. (11) can be rewritten as

$$(\mathcal{S}^*, T^*) = \arg \max_{\mathcal{S} \subseteq \mathcal{U},\, T \in SE(2)} \sum_{(i,j) \in \mathcal{S}} w_{ij} \left(1 - \lambda u_{ij}(T)\right). \tag{28}$$

Here, $\lambda > 0$ controls the tolerance to fingerprint element geometric residuals.

A.10.1 Introduction of the Truncated Quadratic Robust Kernel

In practical cross-modal structural fingerprint matching, the candidate fingerprint element correspondence set $\mathcal{U}$ inevitably contains false matches. If all candidates were treated equally, fingerprint element correspondences with large geometric residuals could dominate the fingerprint matching objective and bias the pose update. The role of the factor λ in Eq. (28) is therefore not merely to rescale the geometric residual, but to induce a residual-dependent fingerprint element selection threshold.

For a fixed transformation T, each candidate fingerprint element correspondence contributes $w_{ij}(1 - \lambda u_{ij}(T))$ to the objective if it is selected into $\mathcal{S}$. Since $w_{ij} \geq 0$, selecting a candidate correspondence is useful only when its fingerprint element geometric residual is small enough to produce a non-negative contribution:

$$1 - \lambda u_{ij}(T) \geq 0 \quad \Longleftrightarrow \quad u_{ij}(T) \leq \tau^2, \qquad \tau^2 = \frac{1}{\lambda}. \tag{29}$$

Thus, λ defines a geometric selection radius τ, such that candidate fingerprint element correspondences with geometric residuals below this threshold can improve the matching score, whereas those beyond this threshold are rejected by structural fingerprint element selection. Substituting this residual-thresholded selection rule into Eq. (28) gives

$$\max_{T\in SE(2)} \sum_{(i,j)\in\mathcal{U}} w_{ij} \max\big(1-\lambda u_{ij}(T),0\big) = \max_{T\in SE(2)} \sum_{(i,j)\in\mathcal{U}} w_{ij} \left(1-\frac{1}{\tau^2}\min\{u_{ij}(T),\tau^2\}\right). \tag{30}$$

Since $\sum_{(i,j)\in\mathcal{U}} w_{ij}$ is independent of T, maximizing Eq. (30) is equivalent to minimizing the truncated least-squares objective

$$E(T) \triangleq \frac{1}{2} \sum_{(i,j)\in\mathcal{U}} w_{ij}\, \phi_\tau(u_{ij}(T)), \qquad \phi_\tau(u) = \min\{u,\tau^2\}. \tag{31}$$

Equivalently,

$$\phi_\tau(u) = \begin{cases} u, & u \le \tau^2, \\ \tau^2, & u > \tau^2. \end{cases} \tag{32}$$

This truncated kernel assigns a constant cost to candidate fingerprint element correspondences with large geometric residuals. When $u_{ij}(T) > \tau^2$, the corresponding term no longer affects the fingerprint pose-update direction:

$$\frac{\partial}{\partial\theta}\left(\frac{1}{2} w_{ij}\phi_\tau\left(u_{ij}(T(\theta))\right)\right) = \frac{\partial}{\partial\theta}\left(\frac{1}{2} w_{ij}\tau^2\right) = 0. \tag{33}$$

Therefore, false fingerprint element matches with large geometric residuals are prevented from dominating the fingerprint matching process.

A.10.2 Discrete Binary-Gated Reformulation

Although the truncated objective in Eq. (31) provides robustness, it also introduces a non-smooth transition at $u = \tau^2$. Moreover, the rotation component of $T \in SE(2)$ makes the fingerprint pose estimation problem non-convex. Rather than optimizing the truncated objective directly, we introduce binary gates to make structural fingerprint element selection explicit and obtain a more stable alternating optimization procedure.

For each candidate fingerprint element correspondence, we define

$$\gamma_{ij} \in \{0,1\}, \qquad (i,j) \in \mathcal{U}. \tag{34}$$

Here, $\gamma_{ij} = 1$ indicates that the candidate fingerprint element correspondence is selected as a geometrically consistent fingerprint element match, whereas $\gamma_{ij} = 0$ indicates that it is excluded from the selected structural fingerprint and its cost is saturated at the truncation level. The binary-gated objective is defined as

$$\widetilde{E}(T,\gamma) \triangleq \frac{1}{2} \sum_{(i,j)\in\mathcal{U}} w_{ij} \left[\gamma_{ij} u_{ij}(T) + (1-\gamma_{ij})\tau^2\right], \qquad \gamma \in \{0,1\}^{|\mathcal{U}|}. \tag{35}$$

This formulation is equivalent to the truncated kernel because, for any scalar $u \ge 0$,

$$\min_{\gamma\in\{0,1\}} \left[\gamma u + (1-\gamma)\tau^2\right] = \min\{u,\tau^2\} = \phi_\tau(u). \tag{36}$$

Applying Eq. (36) to each fingerprint element geometric residual $u_{ij}(T)$ gives

$$E(T) = \min_{\gamma\in\{0,1\}^{|\mathcal{U}|}} \widetilde{E}(T,\gamma). \tag{37}$$

The resulting optimization can be handled by alternating between two simple steps:

$$\min_{T\in SE(2)} E(T) = \min_{T\in SE(2)} \min_{\gamma\in\{0,1\}^{|\mathcal{U}|}} \widetilde{E}(T,\gamma). \tag{38}$$

Given T, the gate update performs residual-thresholded structural fingerprint element selection. Given γ, the pose update reduces to a weighted rigid matching problem over the selected fingerprint element correspondences.

A.11 Solving the Alternating Optimization Problem

Given the joint discrete-continuous optimization problem in Eq. (38), we solve it by alternating minimization. At iteration k, the two updates are

$$\gamma^{(k+1)} \in \arg\min_{\gamma\in\{0,1\}^{|\mathcal{U}|}} \widetilde{E}\Big(T^{(k)}, \gamma\Big), \tag{39}$$

$$T^{(k+1)} \in \arg\min_{T\in SE(2)} \widetilde{E}\Big(T, \gamma^{(k+1)}\Big). \tag{40}$$

Fingerprint Extraction. We first fix the current fingerprint pose $T^{(k)}$ and update the binary gates. Since $\widetilde{E}(T^{(k)}, \gamma)$ is separable over candidate fingerprint element correspondences, each gate can be optimized independently:

$$\gamma_{ij}^{(k+1)} \in \arg\min_{\gamma\in\{0,1\}} \Big[\gamma u_{ij}\Big(T^{(k)}\Big) + (1-\gamma)\tau^2\Big]. \tag{41}$$

This update follows directly from the minimization objective. For a fixed candidate fingerprint element correspondence (i, j), choosing $\gamma = 1$ assigns its fingerprint element geometric residual cost $u_{ij}(T^{(k)})$, whereas choosing $\gamma = 0$ assigns the saturated cost τ^2. Therefore, the optimal gate selects the smaller of these two values:

$$\gamma_{ij}^{(k+1)} = \begin{cases} 1, & u_{ij}\big(T^{(k)}\big) \le \tau^2, \\ 0, & u_{ij}\big(T^{(k)}\big) > \tau^2. \end{cases} \tag{42}$$

Equivalently,

$$\gamma_{ij}^{(k+1)} = \mathbb{I}\Big[u_{ij}\Big(T^{(k)}\Big) \le \tau^2\Big], \tag{43}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function. Thus, the indicator is not an additional heuristic; it is the closed-form minimizer of the binary gate subproblem. The selected structural fingerprint element correspondence set at iteration $k+1$ is then

$$\mathcal{C}^{(k+1)} = \Big\{(i,j) \in \mathcal{U} \;\Big|\; \gamma_{ij}^{(k+1)} = 1\Big\}. \tag{44}$$

Fingerprint Matching. We next fix $\gamma^{(k+1)}$. Since the terms with $\gamma_{ij} = 0$ contribute only the constant penalty τ^2, they do not affect the minimizer with respect to T. Therefore, updating the transformation reduces to estimating the rigid transformation that best matches the selected fingerprint element correspondences $\mathcal{C}^{(k+1)}$:

$$T^{(k+1)} = \arg\min_{T\in SE(2)} \sum_{(i,j)\in\mathcal{C}^{(k+1)}} w_{ij} \left\|T\mathbf{x}_i^p - \mathbf{x}_j^o\right\|_2^2. \tag{45}$$

Let $T = (\mathbf{R}, \mathbf{t})$, where $\mathbf{R} \in SO(2)$ and $\mathbf{t} \in \mathbb{R}^2$. For a fixed selected fingerprint element correspondence set $\mathcal{C}$, we write the fingerprint matching objective as

$$F_{\mathcal{C}}(\mathbf{R}, \mathbf{t}) \triangleq \frac{1}{2} \sum_{(i,j)\in\mathcal{C}} w_{ij} \left\|\mathbf{R}\mathbf{x}_i^p + \mathbf{t} - \mathbf{x}_j^o\right\|_2^2. \tag{46}$$

For a fixed rotation $\mathbf{R}$, the objective is a quadratic function of the translation $\mathbf{t}$. Thus, the optimal translation can be obtained by setting the derivative with respect to $\mathbf{t}$ to zero:

$$\frac{\partial F_{\mathcal{C}}(\mathbf{R}, \mathbf{t})}{\partial \mathbf{t}} = \sum_{(i,j)\in\mathcal{C}} w_{ij} \left(\mathbf{R}\mathbf{x}_i^p + \mathbf{t} - \mathbf{x}_j^o\right) = \mathbf{0}. \tag{47}$$

Rearranging Eq. (47) gives

$$\left(\sum_{(i,j)\in\mathcal{C}} w_{ij}\right)\mathbf{t} = \sum_{(i,j)\in\mathcal{C}} w_{ij}\mathbf{x}_j^o - \mathbf{R}\sum_{(i,j)\in\mathcal{C}} w_{ij}\mathbf{x}_i^p. \tag{48}$$

Define the total weight and the weighted centroids of the selected fingerprint elements as

$$W \triangleq \sum_{(i,j)\in\mathcal{C}} w_{ij}, \qquad \bar{\mathbf{x}}^p \triangleq \frac{1}{W}\sum_{(i,j)\in\mathcal{C}} w_{ij}\mathbf{x}_i^p, \qquad \bar{\mathbf{x}}^o \triangleq \frac{1}{W}\sum_{(i,j)\in\mathcal{C}} w_{ij}\mathbf{x}_j^o. \tag{49}$$

Then the optimal translation for a fixed rotation is

$$\mathbf{t}^\star(\mathbf{R}) = \bar{\mathbf{x}}^o - \mathbf{R}\bar{\mathbf{x}}^p. \tag{50}$$

This shows why the weighted centroids appear naturally: the optimal translation brings the rotated point-cloud fingerprint element centroid to the OSM fingerprint element centroid.

We then define the centered fingerprint element coordinates

$$\widetilde{\mathbf{x}}_i^p \triangleq \mathbf{x}_i^p - \bar{\mathbf{x}}^p, \qquad \widetilde{\mathbf{x}}_j^o \triangleq \mathbf{x}_j^o - \bar{\mathbf{x}}^o. \tag{51}$$

Substituting Eq. (50) into Eq. (46) removes the translation variable and yields the rotation-only fingerprint matching problem:

$$\min_{\mathbf{R}\in SO(2)} \frac{1}{2}\sum_{(i,j)\in\mathcal{C}} w_{ij}\left\|\mathbf{R}\widetilde{\mathbf{x}}_i^p - \widetilde{\mathbf{x}}_j^o\right\|_2^2. \tag{52}$$

We now expand the squared norm:

$$\left\|\mathbf{R}\widetilde{\mathbf{x}}_i^p - \widetilde{\mathbf{x}}_j^o\right\|_2^2 = \|\mathbf{R}\widetilde{\mathbf{x}}_i^p\|_2^2 + \left\|\widetilde{\mathbf{x}}_j^o\right\|_2^2 - 2\left(\widetilde{\mathbf{x}}_j^o\right)^\top \mathbf{R}\widetilde{\mathbf{x}}_i^p. \tag{53}$$

Since $\mathbf{R}\in SO(2)$ is orthogonal, $\|\mathbf{R}\widetilde{\mathbf{x}}_i^p\|_2^2 = \|\widetilde{\mathbf{x}}_i^p\|_2^2$. Therefore, the first two terms in Eq. (53) are independent of $\mathbf{R}$, and minimizing Eq. (52) is equivalent to maximizing the weighted cross term:

$$\max_{\mathbf{R}\in SO(2)} \sum_{(i,j)\in\mathcal{C}} w_{ij}\left(\widetilde{\mathbf{x}}_j^o\right)^\top \mathbf{R}\widetilde{\mathbf{x}}_i^p. \tag{54}$$

Using the cyclic invariance of the trace operator, the scalar cross term can be written as

$$\left(\widetilde{\mathbf{x}}_j^o\right)^\top \mathbf{R}\widetilde{\mathbf{x}}_i^p = \mathrm{tr}\Big(\left(\widetilde{\mathbf{x}}_j^o\right)^\top \mathbf{R}\widetilde{\mathbf{x}}_i^p\Big) = \mathrm{tr}\Big(\mathbf{R}\widetilde{\mathbf{x}}_i^p\left(\widetilde{\mathbf{x}}_j^o\right)^\top\Big). \tag{55}$$

Hence,

$$\begin{aligned}\sum_{(i,j)\in\mathcal{C}} w_{ij}\left(\widetilde{\mathbf{x}}_j^o\right)^\top \mathbf{R}\widetilde{\mathbf{x}}_i^p &= \sum_{(i,j)\in\mathcal{C}} w_{ij}\,\mathrm{tr}\Big(\mathbf{R}\widetilde{\mathbf{x}}_i^p\left(\widetilde{\mathbf{x}}_j^o\right)^\top\Big) \\ &= \mathrm{tr}\left(\mathbf{R}\sum_{(i,j)\in\mathcal{C}} w_{ij}\widetilde{\mathbf{x}}_i^p\left(\widetilde{\mathbf{x}}_j^o\right)^\top\right).\end{aligned} \tag{56}$$

Define the weighted cross-covariance matrix of the selected fingerprint elements as

$$\mathbf{C} \triangleq \sum_{(i,j)\in\mathcal{C}} w_{ij}\widetilde{\mathbf{x}}_i^p\left(\widetilde{\mathbf{x}}_j^o\right)^\top \in \mathbb{R}^{2\times 2}. \tag{57}$$

The rotation subproblem is therefore equivalent to

$$\max_{\mathbf{R}\in SO(2)} \mathrm{tr}\left(\mathbf{R}\mathbf{C}\right). \tag{58}$$

Let the singular value decomposition of $\mathbf{C}$ be

$$\mathbf{C} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top, \qquad \mathbf{\Sigma} = \mathrm{diag}(\sigma_1,\sigma_2), \quad \sigma_1 \geq \sigma_2 \geq 0. \tag{59}$$

Substituting this decomposition into the trace objective gives

$$\mathrm{tr}(\mathbf{R}\mathbf{C}) = \mathrm{tr}\big(\mathbf{R}\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\big) = \mathrm{tr}\big(\mathbf{V}^\top\mathbf{R}\mathbf{U}\mathbf{\Sigma}\big), \tag{60}$$

where the second equality again uses cyclic invariance of the trace. Let

$$\mathbf{X} \triangleq \mathbf{V}^\top\mathbf{R}\mathbf{U}. \tag{61}$$

Because $\mathbf{R}$, $\mathbf{U}$, and $\mathbf{V}$ are orthogonal matrices, $\mathbf{X}$ is also orthogonal. Thus,

$$\operatorname{tr}(\mathbf{RC}) = \operatorname{tr}(\mathbf{X\Sigma}) = \sigma_1 x_{11} + \sigma_2 x_{22}. \tag{62}$$

Since the diagonal entries of an orthogonal matrix are bounded by $[-1, 1]$, the trace is maximized by making the diagonal entries as close to 1 as possible. If the determinant constraint were ignored, this would yield $\mathbf{X} = \mathbf{I}$ and hence $\mathbf{R} = \mathbf{VU}^\top$. However, $\mathbf{VU}^\top$ may have determinant -1, which corresponds to a reflection rather than a valid rotation in $SO(2)$.

To exclude reflection solutions, we introduce the diagonal correction matrix

$$\mathbf{D} = \operatorname{diag}\left(1, \det\left(\mathbf{VU}^\top\right)\right). \tag{63}$$

The optimal fingerprint rotation is then given by the weighted Kabsch solution

$$\mathbf{R}^\star = \mathbf{VDU}^\top. \tag{64}$$

This construction enforces the rotation constraint:

$$\det(\mathbf{R}^\star) = \det(\mathbf{V}) \det(\mathbf{D}) \det(\mathbf{U}^\top) = \det(\mathbf{VU}^\top) \det(\mathbf{D}) = \left(\det(\mathbf{VU}^\top)\right)^2 = 1. \tag{65}$$

Therefore, $\mathbf{R}^\star \in SO(2)$ and mirror-reflection solutions are removed.

Finally, the corresponding optimal fingerprint translation is

$$\mathbf{t}^\star = \bar{\mathbf{x}}^o - \mathbf{R}^\star \bar{\mathbf{x}}^p. \tag{66}$$

Thus, for a fixed selected fingerprint element correspondence set $\mathcal{C}$, the optimal fingerprint rigid transformation is

$$T^\star = (\mathbf{R}^\star, \mathbf{t}^\star). \tag{67}$$

Overall Procedure. Combining the two subproblems gives the alternating optimization procedure:

1. **Initialization.** Initialize $T^{(0)}$ from high-confidence candidate fingerprint element correspondences induced by the fingerprint-aware attention mechanism.
2. **Fingerprint Extraction.** Update the binary gates using Eq. (43), and construct the selected structural fingerprint element correspondence set $\mathcal{C}^{(k+1)}$ using Eq. (44).
3. **Fingerprint Matching.** Estimate the rigid transformation that best matches the selected fingerprint element correspondences $\mathcal{C}^{(k+1)}$ using Eqs. (64) and (66), and update $T^{(k+1)}$.
4. **Iteration.** Repeat the Fingerprint Extraction and Fingerprint Matching steps until convergence or until the maximum number of iterations is reached.

After convergence, the final structural fingerprint is defined as the selected fingerprint element correspondence set:

$$\mathcal{S}^\star = \left\{(i, j) \in \mathcal{U} \;\middle|\; \gamma_{ij}^\star = 1\right\}, \tag{68}$$

and its associated fingerprint pose estimate is given by the final transformation $T^\star$. The matching score computed from the selected structural fingerprint is used as the final fingerprint similarity score for place recognition.

A.12 Monotonic Descent of the Alternating Updates

We next show that the alternating updates used for structural fingerprint element selection and fingerprint matching monotonically decrease the joint binary-gated energy. Specifically, each substep does not increase $\widetilde{E}(T, \gamma)$, and thus the energy sequence generated by the alternating procedure is monotonically non-increasing.

Fingerprint extraction. We first fix the current fingerprint pose $T^{(k)}$ and update the binary gates:

$$\gamma^{(k+1)} \in \arg\min_{\gamma \in \{0,1\}^{|\mathcal{U}|}} \widetilde{E}(T^{(k)}, \gamma). \tag{69}$$

Using the definition of the binary-gated objective, we have

$$\widetilde{E}(T^{(k)}, \gamma) = \frac{1}{2} \sum_{(i,j)\in\mathcal{U}} w_{ij} \left[\gamma_{ij} u_{ij}(T^{(k)}) + (1-\gamma_{ij})\tau^2\right]. \tag{70}$$

This objective is separable over candidate fingerprint element correspondences. Therefore, each gate can be updated independently by solving

$$\gamma_{ij}^{(k+1)} \in \arg\min_{\gamma\in\{0,1\}} \left[\gamma\, u_{ij}(T^{(k)}) + (1-\gamma)\tau^2\right] = \begin{cases} 1, & u_{ij}(T^{(k)}) \le \tau^2, \\ 0, & u_{ij}(T^{(k)}) > \tau^2. \end{cases} \tag{71}$$

Thus, the gate update selects candidate fingerprint element correspondences whose geometric residuals are no larger than the truncation threshold τ^2. Equivalently, $\gamma^{(k+1)}$ is a global minimizer of $\widetilde{E}(T^{(k)}, \gamma)$ over all feasible binary gates. Hence,

$$\widetilde{E}(T^{(k)}, \gamma^{(k+1)}) \le \widetilde{E}(T^{(k)}, \gamma^{(k)}). \tag{72}$$

Fingerprint matching. We next fix the selected gates $\gamma^{(k+1)}$ and update the fingerprint pose:

$$T^{(k+1)} \in \arg\min_{T\in SE(2)} \widetilde{E}(T, \gamma^{(k+1)}). \tag{73}$$

The selected fingerprint element correspondence set induced by $\gamma^{(k+1)}$ is

$$\mathcal{C}^{(k+1)} = \left\{(i,j) \in \mathcal{U} \;\middle|\; \gamma_{ij}^{(k+1)} = 1\right\}. \tag{74}$$

With this notation, the joint energy can be decomposed as

$$\widetilde{E}(T, \gamma^{(k+1)}) = \frac{1}{2} \sum_{(i,j)\in\mathcal{C}^{(k+1)}} w_{ij} u_{ij}(T) + \frac{1}{2} \sum_{(i,j)\in\mathcal{U}\setminus\mathcal{C}^{(k+1)}} w_{ij}\tau^2. \tag{75}$$

The second term is independent of T. Therefore, minimizing $\widetilde{E}(T, \gamma^{(k+1)})$ with respect to T is equivalent to solving a fingerprint matching problem over the selected fingerprint element correspondences:

$$T^{(k+1)} = \arg\min_{T\in SE(2)} \sum_{(i,j)\in\mathcal{C}^{(k+1)}} w_{ij} \left\|T\mathbf{x}_i^p - \mathbf{x}_j^o\right\|_2^2. \tag{76}$$

As shown in Appendix A.11, this subproblem admits a closed-form weighted Kabsch solution. Since $T^{(k+1)}$ minimizes the pose subproblem for the fixed selected fingerprint element correspondence set $\mathcal{C}^{(k+1)}$, we have

$$\widetilde{E}(T^{(k+1)}, \gamma^{(k+1)}) \le \widetilde{E}(T^{(k)}, \gamma^{(k+1)}). \tag{77}$$

Monotonic descent of the full iteration. Combining the non-increase property of the fingerprint extraction step in Eq. (72) with that of the fingerprint matching step in Eq. (77), we obtain

$$\widetilde{E}(T^{(k+1)}, \gamma^{(k+1)}) \le \widetilde{E}(T^{(k)}, \gamma^{(k+1)}) \le \widetilde{E}(T^{(k)}, \gamma^{(k)}). \tag{78}$$

Therefore, the energy sequence

$$\left\{\widetilde{E}(T^{(k)}, \gamma^{(k)})\right\}_{k\ge 0} \tag{79}$$

is monotonically non-increasing.

Moreover, since $w_{ij} \ge 0$, $u_{ij}(T) \ge 0$, and $\tau^2 > 0$, the joint energy $\widetilde{E}(T, \gamma)$ is lower bounded by zero. Hence, the alternating procedure admits monotonic descent of the binary-gated objective value.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction faithfully represent the primary contributions and overall scope of this study.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of our approach in Sec. 6 and provide a more comprehensive analysis in Appendix A.1.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Complete mathematical derivations and rigorous proofs for all theoretical formulations are provided in Appendices A.10, A.11, and A.12.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide comprehensive descriptions of the experimental setup and dataset processing pipeline in Sec. 5, A.8, and A.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Please visit our project homepage, which is publicly available at `https://github.com/Wiakoel/SFPR`.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All necessary training and evaluation details are comprehensively specified in Appendices A.8 and A.9.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We conducted five independent training and evaluation runs using different random seeds. On KITTI sequences 00 and 07, Recall@1m was $28.66 \pm 0.74\%$ and $58.61 \pm 0.81\%$ (mean $\pm$ standard deviation), respectively. The main comparison tables report results obtained with the fixed random seed 42, consistent with the evaluation protocol used for comparison with OPAL.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As stated in the supplementary materials, all experiments were conducted using a single NVIDIA RTX 4090 GPU with 24GB of memory.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, our research fully adheres to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: Our work is technology oriented, with the main goal of improving the positioning accuracy of robots, as well as the efficiency and reliability of tasks such as autonomous driving. It does not have a direct social impact, but we certainly hope that it can indirectly have a positive impact on the industry, economy, and society.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The data sources in this work are all legally collected.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Yes, all external assets (code, data, models) used in this paper are properly cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not introduce or release any new datasets or assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.