IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

LSV-Loc: LiDAR to StreetView Image Cross-modal Localization

Sangmin Lee, Donghyun Choi, and Jee-Hwan Ryu, *Fellow Member, IEEE*

***Abstract*—Accurate global localization remains a fundamental challenge in autonomous vehicle navigation. Traditional methods typically rely on high-definition (HD) maps generated through prior traverses or utilize auxiliary sensors, such as a global positioning system (GPS). However, the above approaches are often limited by high costs, scalability issues, and decreased reliability where GPS is unavailable. Moreover, prior methods require route-specific sensor calibration and impose modality-specific constraints, which restrict generalization across different sensor types. The proposed framework addresses this limitation by leveraging a shared embedding space, learned via a weight-sharing Vision Transformer (ViT) encoder, that aligns heterogeneous sensor modalities, Light Detection and Ranging (LiDAR) images, and geo-tagged StreetView panoramas. The proposed alignment enables reliable cross-modal retrieval and coarse-level localization without HD-map priors or route-specific calibration. Further, to address the heading inconsistency between query LiDAR and StreetView, an equirectangular perspective-n-point (PnP) solver is proposed to refine the relative pose through patch-level feature correspondences. As a result, the framework achieves coarse 3-degree-of-freedom (DoF) localization from a single LiDAR scan and publicly available StreetView imagery, bridging the gap between place recognition and metric localization. Experiments demonstrate that the proposed method achieves high recall and heading accuracy, offering scalability in urban settings covered by public Street View without reliance on HD maps. Our code will be made publicly available at: https://github.com/iismn/IEEE_RA-L_LSV-Loc.**

***Index Terms*—Localization, Autonomous Vehicle Navigation, Place Recognition**

I. Introduction

GLOBAL localization is a critical requirement for autonomous vehicle navigation, as it enables high-level planning and reliable arrival at the intended destination. Without a global estimate of the vehicle position on a map, global route planning and subsequent local trajectory control cannot be executed. Owing to its importance, extensive research has been devoted to global localization in recent years. One major research direction involves localization based on external positioning systems, such as GPS, radio frequency (RF), and ultra-wideband (UWB) sensors [1], [2]. While sensor-based methods are effective in open environments, their accuracy degrades significantly in urban areas with dense structures and GPS-shadowed zones [3]. Moreover, sensor-based systems typically require additional hardware mounted on both the vehicle and the surrounding infrastructure, increasing cost and deployment complexity. To address the challenges above, HD map-based methods have been widely adopted. HD map-based approaches use a mobile mapping system (MMS) to generate detailed maps containing semantic and geometric information like lanes and traffic signs [4]. Localization is then performed by aligning real-time sensor observations with the prior map. However, HD map-based methods require costly pre-driving efforts and repeated data collection, and suffer from limited scalability. Furthermore, localization is restricted to previously mapped areas and is affected by sensor modality mismatches, especially when the inference and mapping sensors differ [5]. Some recent works have attempted to address sensor modality gaps using cross-modal localization techniques [6]–[8]. While

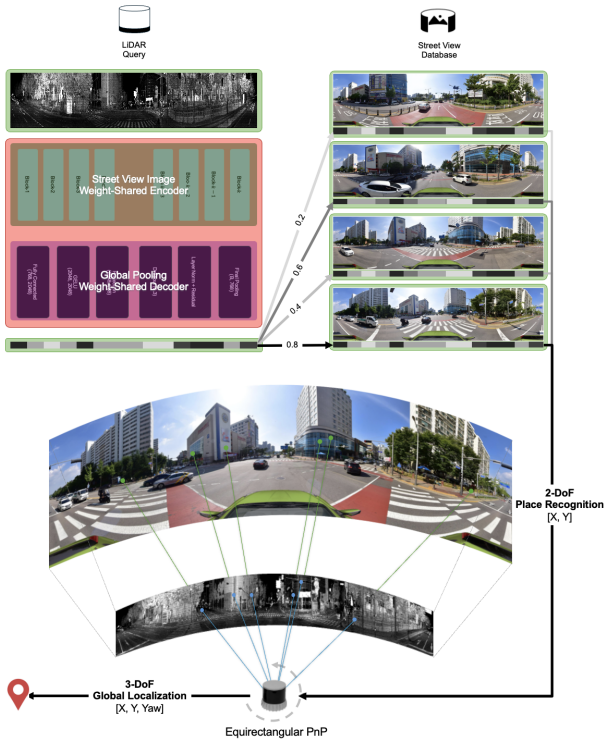


Fig. 1. Overall framework. A multi-modal network extracts global descriptors from LiDAR intensity and StreetView images for place recognition. After retrieval, coarse features from high-attention areas are matched across modalities. An equirectangular PnP solver uses these features to correct heading, achieving coarse localization with accurate heading.

Manuscript received: July 5, 2025; Revised October 2, 2025; Accepted December 15, 2025.

This paper was recommended for publication by Editor Ashis Banerjee upon evaluation of the Associate Editor and Reviewers' comments. This research was supported in part by the Robot Industry Core Technology Development Program (20023294) and in part by the Robot Industry Core Technology Development Program (00423853) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea), and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2025-02213804). (*Corresponding author: Jee-Hwan Ryu.*)

Sangmin Lee and Jee-Hwan Ryu are affiliated with the Department of Civil and Environmental Engineering, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, Korea (e-mail: iismn@kaist.ac.kr; jhryu@kaist.ac.kr)

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

prior methods offer promising results, they still rely on pre-recorded reference data and fail to generalize to unseen environments. Alternatively, several studies have explored public map data as a reference source, such as aerial imagery or satellite maps [9], [10]. However, aerial image-based approaches require an accurate GPS initialization, which is unreliable in obstructed urban settings. Additionally, camera-based localization remains sensitive to lighting conditions and weather, limiting its robustness in real-world deployments.

To address limitations, we propose a novel cross-modal localization framework that leverages publicly available, geo-tagged StreetView panoramas. The proposed method deliberately chooses panoramic StreetView over ordinary, limited-FoV images, as its full 360° structure is geometrically critical for resolving heading ambiguities with respect to 360° LiDAR scans. Unlike aerial imagery, StreetView provides ground-level observations that closely align with LiDAR viewpoints and are densely accessible in urban environments, making them an effective and scalable reference source. Inspired by vision-language models that align heterogeneous modalities, the proposed method embeds LiDAR intensity images and StreetView panoramas into a shared representation space using dual ViT encoders with weight sharing. The proposed unified embedding framework enables robust coarse global localization via place recognition and accurate heading estimation without the need for GPS initialization, calibration, or prior mapping. The specific methodology and implementation details of the proposed approach are presented as follows:

- A weight-shared framework is proposed to align common features between LiDAR images and StreetView images, enabling robust cross-modal matching.
- Attention-guided correspondences and an equirectangular PnP formulation are employed to recover heading from aligned features, thereby moving beyond similarity-only contrastive retrieval toward heading-aware, coarse-level 3-DoF localization.
- A localization framework using only a single LiDAR scan and a StreetView image, eliminating the dependence on prior HD maps or expensive positioning sensors.

II. Related Work

This section reviews existing global localization methods based on camera-only, LiDAR-only, and cross-modal sensors, with a focus on their relevance to the proposed framework.

A. Camera-Based Global Localization

Camera-based global localization has been extensively studied due to the low cost and rich semantic content of visual sensors. Early approaches relied on hand-crafted local features like SIFT [11], which were aggregated into global descriptors using Bag-of-Words (BoW) for database retrieval and pose inference. To improve robustness under varying conditions, recent methods have adopted deep feature extractors based on Convolutional Neural Networks (CNNs) [12] and Vision Transformers (ViTs) [13]. Deep feature extractors have been integrated with global feature pooling techniques such as PatchNetVLAD [14] and DELG [15], achieving strong performance in place recognition tasks. However, despite their accuracy in controlled settings, camera-based methods remain vulnerable to performance degradation under low-light conditions or adverse weather [16], and generally assume consistent camera configurations between query and reference images.

B. LiDAR-Based Global Localization

With advances in resolution and sensing range, LiDAR-based methods have been established as reliable alternatives for localization in challenging environments. ScanContext++ [17] exploits structural cues by constructing polar and Cartesian representations for place recognition and loop closure. PointNetVLAD [18] combines point-wise feature aggregation to learn global descriptors directly from raw point clouds.

Recently, transformer-based pipelines have improved large-scale retrieval and registration in the single-modality setting. Xu et al. [19] proposed transformer-driven global descriptors that are robust to viewpoint changes for LiDAR-to-LiDAR place recognition, while Qin et al. [20] formulated a geometric transformer for fast and robust point cloud registration with attention-based correspondence reasoning.

However, LiDAR-based localization typically relies on prior traversals to build detailed reference maps. Without pre-built LiDAR maps, global localization is infeasible. As a result, most prior methods are constrained to loop-closure detection within SLAM rather than zero-shot localization in unseen regions. Moreover, retrieval accuracy is generally guaranteed only within the traversed area and sensor configuration; cross-sensor or cross-dataset generalization can degrade when the sensor type or deployment conditions differ from the reference acquisition.

C. Cross-Modal Localization and Public Map Sources

Recent work has explored cross-modal localization by bridging LiDAR and RGB image domains. Zhao et al. [21] and Lee et al. [7] proposed transforming RGB images into LiDAR-style depth or range maps using monocular estimators. Puligilla et al. [8] adopted CLIP-based vision–language models to embed LiDAR and images in a shared space. However, these approaches typically depend on spatially aligned, calibrated camera–LiDAR pairs with known projection parameters, which constrains applicability to pre-surveyed environments and limits deployment in unseen areas.

Moreover, cross-view retrieval studies have reduced viewpoint gaps through multi-modal fusion or view normalization. Wang et al. [22] integrate ground cameras and LiDAR to retrieve aerial descriptors, advancing multi-modal aerial-ground place recognition. Ye et al. [23] convert StreetView panoramas into a bird-eye-view representation to narrow aerial-ground discrepancies and perform collaborative retrieval across branches. Despite these advances, the methods remain retrieval-oriented and offer limited heading estimation or geometric pose recovery, which limits their applicability to geometric localization and initialization.

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

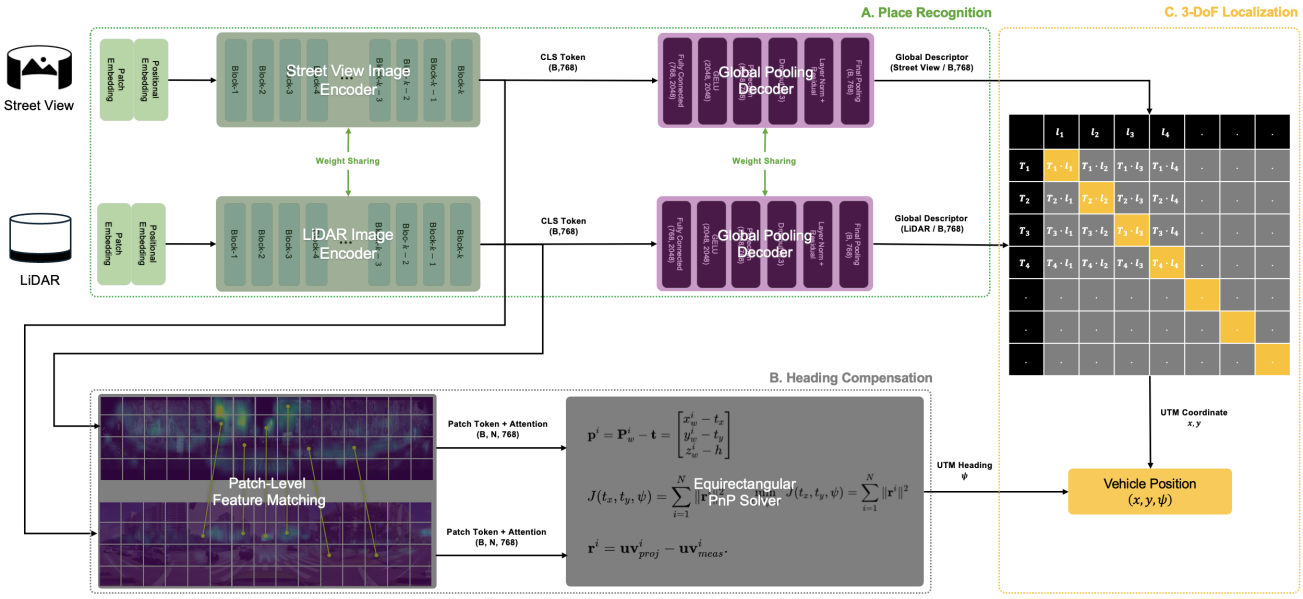


Fig. 2. Overall framework. A weight-sharing ViT encoder extracts features from StreetView and LiDAR images. Global descriptors, derived from the CLS token and a pooling decoder, are matched via cosine similarity for place recognition and trained using InfoNCE loss. Patch-level features are then used in an equirectangular PnP solver to correct heading ambiguity.

To overcome the limitations above, the proposed framework directly matches LiDAR intensity scans with uncalibrated, geo-tagged StreetView panoramas, mitigating the dependency on GPS initialization, HD-map priors, or calibrated data. Unlike map sources requiring prior traversals, public StreetView panoramas are readily available without additional data collection. Building on the geometric consistency provided by StreetView panoramas, retrieval is performed by aligning LiDAR and panoramic features through weight sharing and coarse-level attention matching, followed by equirectangular PnP solving to recover heading and position.

III. Proposed Method

This section presents a LiDAR–StreetView cross-modal localization framework that estimates the 3-degree-of-freedom (3-DoF) pose of an autonomous platform. The proposed method consists of two components: a cross-modal place recognition network and an equirectangular PnP solver. First, the place recognition module extracts global descriptors from both LiDAR intensity and Street View images by projecting them into a shared embedding space. Then, to resolve orientation mismatches, the method applies an equirectangular PnP formulation using local patch-level features. Our two-stage process enables robust 3-DoF global localization without requiring GPS or prior HD maps.

A. Place Recognition

1) Equirectangular image projection: A significant modality gap exists between LiDAR and camera imagery, which poses challenges for learning a unified representation without preprocessing. To mitigate discrepancies, the proposed method projects LiDAR point clouds, containing both depth and intensity information, into an equirectangular image coordinate system. Projecting LiDAR point clouds into an equirectangular image representation aligns their geometry with the 2D structure of StreetView panoramas, both of which follow a consistent latitude–longitude mapping over a full 360° horizontal field of view. This shared projection model reduces the modality gap by preserving spatial correspondences without requiring camera intrinsic parameters or calibration. Unlike aerial or satellite imagery, StreetView panoramas are captured at ground level with viewpoints that naturally align with LiDAR sensors, allowing for spatially consistent feature extraction and effective cross-modal alignment. Given a 3D LiDAR point $\mathbf{P} = [x, y, z]$, the projection onto an equirectangular image is computed as:

$$\theta = \arctan 2(y, x), \tag{1}$$

$$\phi = \arcsin\left(\frac{z}{\sqrt{x^2 + y^2 + z^2}}\right), \tag{2}$$

$$u = \left(\frac{\theta + \pi}{2\pi}\right) \cdot W, \tag{3}$$

$$v = \left(1 - \frac{\phi + \frac{\pi}{2}}{\pi}\right) \cdot H, \tag{4}$$

where W and H denote the horizontal and vertical resolutions, respectively.

Because the sky and nearby-vehicle occlusions lack geometric or semantic information useful for place recognition, utilizing these regions is inefficient. Following panoramic VPR studies that restrict comparisons to the informative FoV [24], [25], the proposed method crops the vertical field of view (FoV) of the equirectangular image from 180° to 90°. For LiDAR images, zero-padding is applied to match the StreetView format, ensuring consistent spatial dimensions and FoV alignment across both modalities. While padding adds non-informative regions, especially for narrow-FoV sensors,

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

our proposed network learns to ignore these tokens, a principle validated in masked-encoding frameworks [26].

2) Place Recognition Network: To address the modality gap between LiDAR and camera imagery, the proposed method adopts a dual-encoder architecture inspired by CLIP, a framework originally developed to align representations between text and images. Analogously, the proposed method projects both LiDAR intensity scans and StreetView panoramas into equirectangular image formats. To extract the same feature from both modalities, the proposed method adapts a ViT backbone for patch-based tokenization and a global self-attention mechanism, which captures structural correspondences for aligning LiDAR intensity images with StreetView panoramas. Unlike CNNs constrained by local receptive fields and inductive biases, the transformer's token-based formulation offers a modality-agnostic embedding space suitable for cross-modal alignment. Furthermore, the self-supervised ViT pretraining network, DINOv2 [27] has been shown to yield geometry-aware mid-level features, improving transferability to heterogeneous modalities. To adapt the ViT for equirectangular geometry, the pretrained positional embeddings were initialized using 2D bilinear interpolation. The learnable embeddings were then fine-tuned on our dataset, allowing the model to effectively learn the new spatial geometry and mitigate potential artifacts arising from the significant change in aspect.

Let $f_{\theta_s}(I_s)$ and $f_{\theta_l}(I_l)$ denote the feature extractors for StreetView and LiDAR images, respectively. Rather than training them independently, the proposed method shares weights $\theta_s = \theta_l$ to enforce a shared embedding space across modalities. Shared embedding space encourages the network to attend to consistent semantic cues, such as building facades or road boundaries, regardless of sensor type.

After extracting patch-level features, the [CLS] token from each encoder is used to represent the global scene content. The [CLS] token is passed through a non-linear projection head, which improves the discriminability of the resulting embedding. Following the structure proposed in SimCLR [28], the projection head consists of a linear projection followed by a GELU activation, dropout, and residual connection with layer normalization:

$$z_0 = W_p x_{\text{cls}} + b_p, \tag{5}$$

$$z_1 = \text{Dropout}\left(W_f\,\text{GELU}(z_0) + b_f\right), \tag{6}$$

$$y = \text{LayerNorm}(z_1 + z_0). \tag{7}$$

The resulting global descriptors, y_l for LiDAR and y_s for StreetView, are optimized using the InfoNCE loss [29]:

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp\left(\text{sim}(y_l, y_s)/\tau\right)}{\sum_{k=1}^{N} \exp\left(\text{sim}(y_l, y_k)/\tau\right)}, \tag{8}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is a temperature scaling parameter, and N is the batch size. The InfoNCE contrastive objective drives the model to cluster cross-modal descriptors that originate from the same location while dispersing those from different places, thereby improving the accuracy and robustness of place recognition across heterogeneous sensor inputs.

B. Global Localization via Equirectangular PnP Solver

Heading estimation is essential for route planning, but most place recognition methods provide only 2-DoF position, omitting heading and limiting their utility. To address this limitation, the proposed method jointly estimates position and heading, achieving full 3-DoF global localization by exploiting complementary information between LiDAR and StreetView imagery.

StreetView panoramas offer geo-tagged poses but lack intrinsics and metric depth; LiDAR provides accurate geometry but no global pose. To bridge the gap, the proposed method extracts high-confidence patch-level features from both modalities and solves a perspective-n-point (PnP) problem in the equirectangular image domain. Patch-level features $e_l(v,u)$ and $e_s(v,u)$ are extracted from LiDAR and StreetView images using ViT encoders. To ensure robustness, features are selected from the top $\rho\%$ of self-attention-weighted patches. The attention map is computed as:

$$A = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{9}$$

where Q, K, V are the standard attention components from ViT, and A defines the importance of each spatial patch. For feature matching, Mutual Nearest Neighbor (MNN) filtering is applied. Let $\mathcal{E}_l = \{e_l^i\}$ and $\mathcal{E}_s = \{e_s^j\}$ denote the sets of descriptors from LiDAR and StreetView images, respectively. A match (e_l^i, e_s^j) is retained if:

$$e_s^j = \arg\min_{e\in\mathcal{E}_s} \|e - e_l^i\| \quad \text{and} \quad e_l^i = \arg\min_{e\in\mathcal{E}_l} \|e - e_s^j\|, \tag{10}$$

ensuring that both descriptors are each other's nearest neighbors in Euclidean space. From the matched patch coordinate (u,v) and LiDAR range image $R_{v,u}$, 3D coordinates are reconstructed as:

$$x_{v,u} = R_{v,u}\cos\left(\tfrac{\text{FOV}_v}{2} - v\tfrac{\text{FOV}_v}{H-1}\right)\cos\left(-\pi + u\tfrac{2\pi}{W}\right), \tag{11}$$

$$y_{v,u} = R_{v,u}\cos\left(\tfrac{\text{FOV}_v}{2} - v\tfrac{\text{FOV}_v}{H-1}\right)\sin\left(-\pi + u\tfrac{2\pi}{W}\right), \tag{12}$$

$$z_{v,u} = R_{v,u}\sin\left(\tfrac{\text{FOV}_v}{2} - v\tfrac{\text{FOV}_v}{H-1}\right), \tag{13}$$

where W and H are the horizontal and vertical resolutions, and FOV_v is the LiDAR's vertical FoV.

Extracted 3D points are reprojected onto the equirectangular image plane using:

$$\hat{u} = \frac{\arctan 2(y_{v,u}, x_{v,u}) + \pi}{2\pi} W, \tag{14}$$

$$\hat{v} = \frac{\frac{\text{FOV}_v}{2} - \arcsin(z_{v,u}/\|\mathbf{x}_{v,u}\|)}{\text{FOV}_v} \cdot (H-1), \tag{15}$$

yielding reprojected pixel coordinates $(\hat{u}, \hat{v})$ aligned with StreetView features.

Let $\mathcal{P} = \{\mathbf{X}_i\}$ denote the reconstructed 3D LiDAR points and $\mathcal{Q} = \{\mathbf{q}_i\}$ the matched StreetView coordinates $\mathbf{q}_i = (\hat{u}_i, \hat{v}_i)$. The 3-DoF pose is estimated by solving a robust least-squares optimization with Huber loss [30]. We explicitly

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE I
IMPLEMENTATION DETAILS AND HYPERPARAMETERS FOR NETWORK TRAINING.

Implementation parameter details	
Feature Extractor Backbone (f_θ)	DiNOv2 (ViT-B)
Patch Size (d)	14
Input Image Size	256 × 1024 (FoV: 90° × 360°)
Decision Distance	25 m
Batch Size (B)	10
Epoch	50
Optimizer	AdamW
LR Scheduler	Cosine Annealing
Temperature (τ)	0.07
Loss	InfoNCE
Attention Threshold (ρ)	10%
Augmentation	Horizontal Roll (p=0.5) Photometric (p=0.1→0.8)
Random Seed	Not fixed (stochastic sampling)

constrain the problem to planar 3-DoF, assuming zero roll, pitch, and z-translation. The optimization thus minimizes:

$$\min_{R,t} \sum_{i=1}^{K} \left\| \mathrm{proj}_{\mathrm{eq}}(R\mathbf{X}_i + t) - \mathbf{q}_i \right\|_2^2, \tag{16}$$

where $\mathbf{R}_z(\psi)$ is the 3D rotation matrix for Yaw ψ, $\mathbf{t} = [t_x, t_y, 0]^T$ is the 2D translation vector embedded in $\mathbb{R}^3$, and $\mathrm{proj}_{\mathrm{eq}}(\cdot)$ applies the equirectangular projection. As a result, the proposed equirectangular PnP solver enables reliable heading estimation without prior HD maps or route-specific sensor calibration, achieving accurate global localization in previously unseen areas.

C. Implementation Details

All experiments were implemented in PyTorch [31]. The proposed framework adopts the ViT-B backbone from DINOv2 [27] as a shared encoder for both StreetView and LiDAR branches, enabling pretrained camera features to transfer to the LiDAR domain. Input images were resized to 256×1024 while preserving the aspect ratio, and non-informative regions such as sky and foreground vehicles were cropped. To align inputs with the ViT, the LiDAR intensity images were replicated to 3 channels, and all images were subsequently normalized using standard ImageNet statistics. Training and optimization parameters are summarized in Table I.

During training, modality-specific augmentations were applied independently: each modality used a horizontal roll with probability p=0.5 to enhance robustness against misalignments and minor motion distortion, and epoch-scheduled photometric perturbations with p=0.1 to p=0.8. LiDAR inputs additionally underwent nonuniform beam remapping with p=2.0 and FoV/channel simulation for robustness to sensor variation. Global random seeds were not fixed, as the training pipeline employs stochastic data augmentation and epoch-wise sampling.

For pose refinement, an equirectangular PnP solver was used to estimate vehicle heading and planar translation. The optimization minimizes pixel reprojection residuals with a Huber loss (δ=5 px) initialized at (t_x, t_y, ψ)=$(0, 0, 0)$ and a fixed camera height. A valid solution requires $N_{\mathrm{corr}} \geq 3$ correspondences; otherwise, or if convergence fails, the framework reverts to the retrieval pose from the StreetView image.

TABLE II
TRAINING DATASET CONFIGURATION.

Dataset / Sequence	# of Images	Length	LiDAR Channel	LiDAR vFoV
MulRan [32] (DCC01-03, KAIST01)	13472	11.04 km	64	45°,
STheReo [33] (SNU, Valley)	16000	9.94 km	128	45°
HeLiPR [34] (Bridge01, River01, Town01)	29150	37.99 km	128	22.5°
ComplexUrban [35] (01, 02, 13, 15)	16507	20.60 km	64 / 128	45°
MA-LIO [36] (City 01-03)	23858	13.55 km	128	22.5°
Total	98987	93.12 km	# of LiDAR types: 4	

TABLE III
VALIDATION DATASET CONFIGURATION.

Dataset	# of Images	Length	LiDAR Channel	LiDAR vFoV
In-house (DCC 01)	4143	5.42 km	128 / 64 / 32	45°/ 22.5°
In-house (DCC 02)	3079	4.23 km		
In-house (DCC 03)	4041	5.02 km		
In-house (KAIST 01)	8357	9.94 km		
Total	19620	24.61 km	# of LiDAR type: 1	

TABLE IV
TEST DATASET CONFIGURATION.

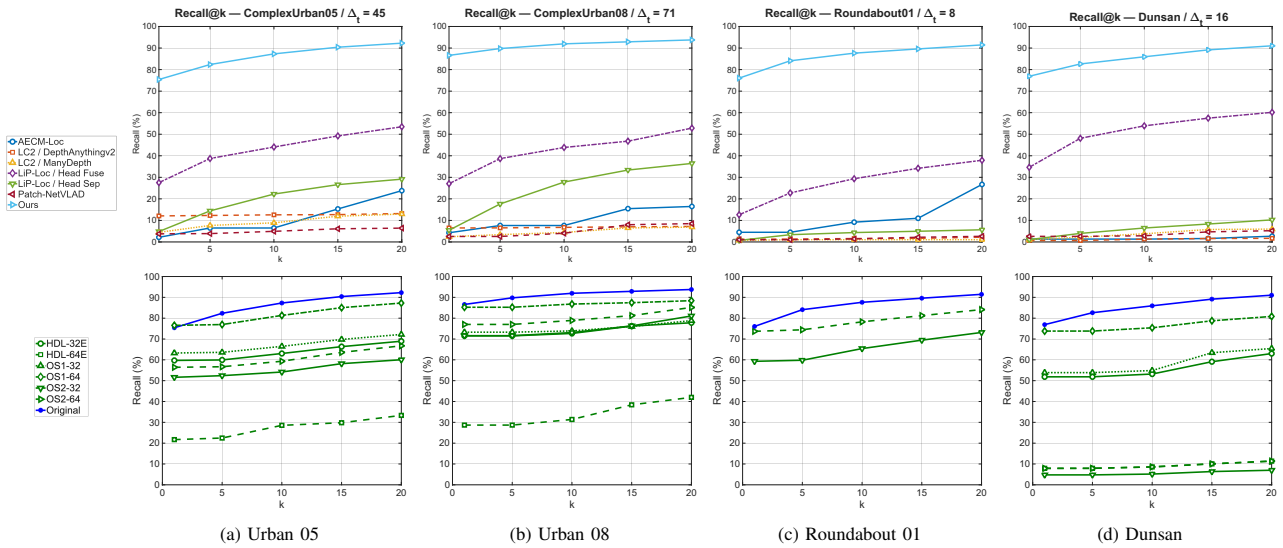
Dataset	# of Images	Length	LiDAR Channel	LiDAR vFoV
ComplexUrban (Seq. 05, 08)	5184	4.43 km	128 / 64 / 32	45° / 22.5°
HeLiPR (Roundabout 01)	6657	9.03 km	128 / 64 / 32	22.5°
In-house (Dunsan)	7081	7.17 km	128 / 64 / 32	45° / 22.5°
Total	18922	20.63	# of LiDAR type: 6	

IV. DATASETS

To train and evaluate the proposed method, a diverse suite of high-resolution LiDAR datasets was used, encompassing a wide range of environments and sensor configurations. Publicly available datasets such as MulRan [32], STheReO [33], HeLiPR [34], and MA-LIO [36] provide raw point clouds from different LiDAR sensor types and serve as the basis for supervised learning and cross-domain evaluation. To simulate alternative sensor configurations, virtual LiDAR scans with 64 and 128 vertical channels were generated from LAS-format point clouds provided by the Complex Urban Dataset [35]. Simulated virtual LiDAR configurations enabled the analysis of the proposed method under varying field-of-view and resolution settings. We also recorded an in-house dataset in new urban regions to assess generalization to unseen sensors and locations. Fine-tuning was performed entirely on the in-house validation set, which was collected along the same routes as MulRan but captured at different times with different LiDAR sensors.

For each LiDAR location, the corresponding equirectangular panorama was retrieved using publicly accessible Google StreetView [37] and NAVER Maps [38] APIs. Since StreetView imagery is captured at discrete spatial intervals, each LiDAR frame was paired with the closest available panorama based on geographic proximity. All LiDAR data were collected

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.



(a) Urban 05 (b) Urban 08 (c) Roundabout 01 (d) Dunsan

Fig. 3. Comparison of Recall@N across scenes. Top row: experimental result between prior method; bottom row: zero-shot recall of multiple LiDAR configurations. For the Roundabout01 sequence, zero-shot experiments were conducted by varying the number of vertical channels, as the LiDAR used in this sequence had a vertical FoV of only 22.5°, making simulations with wider FoVs infeasible. Δt denotes time differences between the query and database captured date in months.

independently of the StreetView imagery, with different acquisition times and environmental conditions, resulting in inherent temporal gaps between the two modalities. Details regarding the training, validation, and test dataset configurations are summarized in Table II, III, and Table IV.

V. Experimental Results and Evaluation

This section presents the quantitative evaluation of the proposed method, focusing on recall accuracy and 3-DoF coarse localization performance across multiple public datasets and LiDAR configurations.

A. Recall Accuracy Evaluation

To validate the accuracy of the proposed cross-modal place recognition framework, recall accuracy is evaluated on multiple datasets by comparing LiDAR scans against geo-tagged StreetView images. A retrieved image is considered a correct match if it is within 25 m of the LiDAR ground truth position, a conventional threshold [15], [39], [40] for evaluating the coarse initialization required for 3-DoF localization. The proposed method is evaluated against recent cross-modal localization approaches: LiP-Loc [8], LC^2 [7], and AECM-Loc [6]. For a fair comparison, LiP-Loc[1] and AECM-Loc[2] were retrained on our dataset using their respective default hyperparameters. Since LC^2 does not provide training code, we fine-tuned the pretrained model and additionally evaluated using Depth Anything v2 [41] to replace the original monocular depth estimator. Additionally, we extend LiP-Loc with our proposed projection pooling and shared embedding space to assess its performance gain under our framework. To further examine robustness across sensor variations, synthetic LiDAR configurations were derived by modifying the FoV and vertical resolution of existing high-resolution sensors. Specifically, FoV was reduced from 45° to 22.5°, and vertical resolution was downsampled from 128 channels to 64 and 32 channels, respectively. In addition to the cross-modal baselines, a strong vision-only method, Patch-NetVLAD [14], was evaluated. Patch-NetVLAD was trained on LiDAR intensity images and StreetView panoramas using the public implementation and released hyperparameters.[3]

As shown in Fig. 3, the proposed method demonstrates better recall performance across all sequences by successfully matching LiDAR scans with their corresponding StreetView images. Our method achieves accuracy despite large heading/position discrepancies and no camera intrinsics. In contrast, existing methods degrade under unseen conditions, likely due to their reliance on aligned viewpoints and known calibration. Our method enables reliable place recognition using only public StreetView imagery, without prior traversals or sensor calibration. Furthermore, across different LiDAR configurations, the proposed method consistently outperforms existing approaches in zero-shot settings. However, performance degrades in certain urban environments when tested with sensors offering narrower vertical FoVs or oriented toward the ground plane.

B. Localization Accuracy Evaluation

As shown in Table V and illustrated in Fig. 4, the proposed method substantially reduces the heading error ϕ_{cp} relative to raw heading output ϕ of the place recognition step, even under large inter-modal viewpoint differences. Qualitative results validate the proposed method to recover heading error from disparate sensor observations without access to camera intrinsics or pre-surveyed maps, though the position error $\mathbf{P}, \mathbf{P}_{cp}$ remains largely unaffected due to the coarse spatial resolution of the matched features.

We further analyze the impact of attention-based feature filtering. When only a small subset of features with extremely high attention scores (e.g., $\rho < 10\%$) is selected, the system

[1]https://github.com/Shubodh/liploc
[2]https://github.com/Zhaozhpe/AE-CrossModal
[3]https://github.com/QVPR/Patch-NetVLAD

IEEE Robotics and Automation Letters (RA-L) paper, presented at ICRA 2026, Vienna, Austria. Cite as RA-L paper.

TABLE V
EQUIRECTANGULAR PNP COMPENSATION RESULT WITH DIFFERENT CONFIDENCE LEVEL.

Sequence	Complex Urban 05				Complex Urban 08				HeLiPR Roundabout 01				Inhouse Dunsan			
	ϕ	ϕ_{cp}	**P**	$\mathbf{P}_{cp}$	ϕ	ϕ_{cp}	**P**	$\mathbf{P}_{cp}$	ϕ	ϕ_{cp}	**P**	$\mathbf{P}_{cp}$	ϕ	ϕ_{cp}	**P**	$\mathbf{P}_{cp}$
$\rho = 100\%$	23.44° (98.69°)	28.74° (99.21°)	5.65 m (3.96 m)	5.30 m (3.40 m)	57.88° (127.38°)	48.02° (111.83°)	5.49 m (4.43 m)	4.69 m (3.63 m)	7.71° (101.00°)	22.89° (101.85°)	11.09 m (5.57 m)	10.88 m (5.27 m)	20.44° (102.83°)	15.46° (100.93°)	7.01 m (4.61 m)	6.15 m (4.00 m)
$\rho = 50\%$		3.24° (101.30°)		5.01 m (3.14 m)		21.08° (124.47°)		4.48 m (3.56 m)		12.62° (100.68°)		10.86 m (5.36 m)		13.10° (104.39°)		6.21 m (4.05 m)
$\rho = 10\%$		8.02° (77.71°)		5.26 m (3.20 m)		12.18° (103.63°)		4.45 m (3.47 m)		2.12° (65.48°)		10.75 m (5.10 m)		3.65° (90.38°)		6.05 m (3.89 m)
$\rho = 5\%$		10.66° (68.40°)		5.24 m (3.41 m)		8.00° (85.36°)		4.51m (3.92 m)		0.77° (54.44°)		10.98 m (5.37 m)		0.10° (78.19°)		6.23 m (4.18 m)
$\rho = 3\%$		6.87° (64.52°)		5.42 m (3.78 m)		5.98° (75.51°)		4.74 m (5.01 m)		1.23° (51.55°)		11.40 m (6.32 m)		3.20° (69.01°)		6.50 m (4.99 m)

* ϕ denotes the heading error and **P** denotes the position error. Standard deviation is annotated under the mean value.

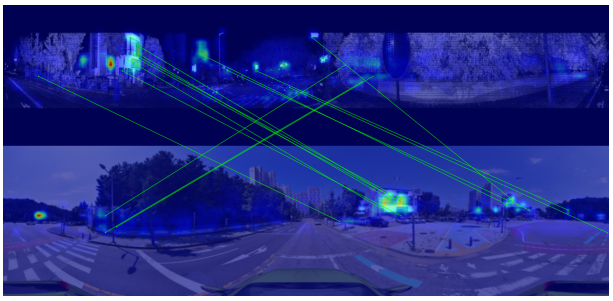


Fig. 4. Feature matching result between LiDAR and StreetView image, overlaid with attention map. Through local feature matching with attention-based filtering, the proposed method can solve the equirectangular PnP problem even when a large heading discrepancy exists.

TABLE VI
ABLATION STUDY OF EMBEDDING SHARING WITH RECALL PERFORMANCE.

Ablation Study				
Embedding Shared Space		Recall@1	Recall@5	Recall@10
Encoder	Pooling			
Fuse	Fuse	74.65 %	81.42 %	85.12 %
Sep	Fuse	52.03 %	58.10 %	62.58 %
Sep	Sep	48.64 %	56.39 %	62.08 %

TABLE VII
COMPUTATIONAL RUNTIME ANALYSIS.

Computational Runtime	
Global Descriptor Matching	13.31 ms
Mutual Nearest Neighbor	77.11 ms
Coarse Feature Matching	1.41 ms
Equirectangular PnP Solving	10.16 ms

becomes more sensitive to noise due to insufficient correspondence support, leading to degraded accuracy. In contrast, using the top $\rho = 10\%$ of attentive patches strikes a balance between confidence and redundancy, resulting in more robust and accurate heading estimates. While the coarse-level patch features are sufficient for heading compensation, their limited spatial granularity restricts improvements in positional accuracy. Experimental results suggest that finer-grained geometric alignment may be necessary to further enhance full 3-DoF localization performance.

C. Ablation Study and Limitation

1) Ablation Study: To assess the impact of key architectural choices, ablation studies are conducted on projection pooling and encoder weight sharing. As shown in Table VI, a fused projection head is beneficial. Furthermore, the ablation study provides the core empirical justification for the proposed shared representation over modality-specific alternatives. Replacing the shared-weight architecture with separate encoders drastically degrades performance, with Recall@1 dropping from 74.65% to 48.64%. This result confirms that a unified encoder is crucial for learning modality-invariant geometric features from both 2D projections, validating the proposed approach. These experiments highlight the importance of tightly coupled representations for achieving robust coarse-level 3-DoF localization through cross-modal place recognition.

2) Limitation: Although the proposed method enables 3-DoF coarse-level localization, including heading estimation through coarse-level equirectangular PnP, accurate metric localization in the x and y dimensions remains limited due to the spatial granularity of patch-level features. Furthermore, performance significantly deteriorates when applied to low-resolution LiDAR, as the reduced vertical FoV and point density limit scene understanding. The method also assumes a rotating LiDAR configuration and is not directly compatible with MEMS or flash LiDAR architectures. Furthermore, we evaluated the runtime on the Complex Urban 05 sequence. While the framework can operate at 9-10 Hz, as shown in Table VII, we identified a significant bottleneck in the Mutual Nearest Neighbor (MNN) search, which is required for robust feature matching.

VI. CONCLUSION

This paper presented LSV-Loc, a novel cross-modal localization framework that enables 3-DoF coarse-level localization in unseen environments using public StreetView and onboard LiDAR. Unlike HD map-based methods, it aligns heterogeneous modalities in a shared embedding space via a weight-sharing ViT encoder. Furthermore, an equirectangular PnP solver estimates heading by matching coarse patch-level features without camera calibration.

Extensive experiments show competitive accuracy in cross-modal coarse-level localization and accurate heading estimation across diverse sensor configurations. We evaluated zero-shot generalization across different LiDAR types with StreetView images. Results validate the potential of the proposed framework for real-world deployment, eliminating the need for high-cost mapping or pre-surveyed environments.

Future work will explore robustness on challenging configurations, like low-resolution or ground-facing LiDAR. Additional data from MEMS LiDAR or modality-specific augmentations may extend generalization. Finally, combining coarse

matching with fine-grained modules could enable centimeter-level accuracy without camera calibration, even in complex scenes.

REFERENCES

[1] N. M. Drawil, H. M. Amar, and O. A. Basir, "Gps localization accuracy classification: A context-based approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 14, no. 1, pp. 262–273, 2012.

[2] H. Wymeersch, S. Maranò, W. M. Gifford, and M. Z. Win, "A machine learning approach to ranging error mitigation for uwb localization," *IEEE transactions on communications*, vol. 60, no. 6, pp. 1719–1728, 2012.

[3] P. D. Groves, "Shadow matching: A new gnss positioning technique for urban canyons," *The journal of Navigation*, vol. 64, no. 3, pp. 417–430, 2011.

[4] Y. Gong, X. Zhang, J. Feng, X. He, and D. Zhang, "Lidar-based hd map localization using semantic generalized icp with road marking detection," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3379–3386, IEEE, 2024.

[5] R. Liu, J. Wang, and B. Zhang, "High definition map for automated driving: Overview and analysis," *The Journal of Navigation*, vol. 73, no. 2, pp. 324–341, 2020.

[6] Z. Zhao, H. Yu, C. Lyu, W. Yang, and S. Scherer, "Attention-enhanced cross-modal localization between spherical images and point clouds," *IEEE Sensors Journal*, vol. 23, no. 19, pp. 23836–23845, 2023.

[7] A. J. Lee, S. Song, H. Lim, W. Lee, and H. Myung, "$(LC)^2$: Lidar-camera loop constraints for cross-modal place recognition," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3589–3596, 2023.

[8] S. S. Puligilla, M. Omama, H. Zaidi, U. S. Parihar, and M. Krishna, "LIP-Loc: Lidar image pretraining for cross-modal localization," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pp. 948–957, IEEE, 2024.

[9] F. Fervers, S. Bullinger, C. Bodensteiner, M. Arens, and R. Stiefelhagen, "Continuous self-localization on aerial images using visual and lidar sensors," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7028–7035, IEEE, 2022.

[10] T. Y. Tang, D. De Martini, S. Wu, and P. Newman, "Self-supervised learning for using overhead imagery as maps in outdoor range sensor localization," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1488–1509, 2021.

[11] D. G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[13] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*, 2021.

[14] S. Hausler, S. Garg, M. Xu, M. Milford, and T. Fischer, "Patch-netvlad: Multi-scale fusion of locally-global descriptors for place recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14141–14152, 2021.

[15] A. Araujo, J. Y. Zhang, A. Matsukawa, and G. Toderici, "Unifying deep local and global features for image search," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 726–743, 2020.

[16] Y. Zhang, A. Carballo, and H. Yang, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 196, pp. 146–177, 02 2023.

[17] G. Kim, S. Choi, and A. Kim, "Scan Context++: Structural place recognition robust to rotation and lateral variations in urban environments," *IEEE Transactions on Robotics*, vol. 38, no. 3, pp. 1856–1874, 2022.

[18] M. A. Uy and G. H. Lee, "PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4470–4479, IEEE, 2018.

[19] Z. Qiao, Z. Yu, H. Yin, and S. Shen, "Transloc3d: Point cloud based large-scale place recognition using adaptive receptive fields," *Pattern Recognition*, vol. 151, p. 110393, 2024.

[20] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, S. Ilic, D. Hu, and K. Xu, "Geotransformer: Fast and robust point cloud registration with geometric transformer," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 9806–9821, 2023.

[21] Z. Zhao, H. Yu, C. Lyu, W. Yang, and S. Scherer, "Attention-enhanced cross-modal localization between spherical images and point clouds," *IEEE Sensors Journal*, vol. 23, no. 19, pp. 23836–23845, 2023.

[22] S. Wang, R. She, Q. Kang, S. Li, D. Li, T. Geng, S. Yu, and W. P. Tay, "Multi-modal aerial-ground cross-view place recognition with neural odes," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 11717–11728, 2025.

[23] J. Ye, Z. Lv, W. Li, J. Yu, H. Yang, H. Zhong, and C. He, "Cross-view image geo-localization with panorama-bev co-retrieval network," in *European Conference on Computer Vision*, pp. 74–90, Springer, 2024.

[24] Z. Shi, H. Shi, K. Yang, Z. Yin, Y. Lin, and K. Wang, "Panovpr: Towards unified perspective-to-equirectangular visual place recognition via sliding windows across the panoramic view," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*, pp. 1333–1340, IEEE, 2023.

[25] S. Orhan and Y. Baştanlar, "Efficient search in a panoramic image database for long-term visual localization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1727–1734, 2021.

[26] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

[27] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "DINOv2: Learning robust visual features without supervision," *Transactions on Machine Learning Research*, 2024. Accepted at TMLR.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607, PMLR, 2020.

[29] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.

[30] P. J. Huber, "Robust estimation of a location parameter," *Annals of Mathematical Statistics*, vol. 35, no. 1, pp. 73–101, 1964.

[31] A. Paszke, "Pytorch: An imperative style, high-performance deep learning library," *arXiv preprint arXiv:1912.01703*, 2019.

[32] G. Kim, Y. S. Park, Y. Cho, J. Jeong, and A. Kim, "MulRan: Multimodal range dataset for urban place recognition," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, (Paris, France), May 2020.

[33] S. Yun, M. Jung, J. Kim, S. Jung, Y. Cho, M. Jeon, and A. Kim, "STHeReO: Stereo thermal dataset for research in odometry and mapping," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, (Kyoto, Japan), pp. 3857–3864, IEEE, 2022.

[34] M. Jung, W. Yang, D. Lee, H. Gil, G. Kim, and A. Kim, "HeLiPR: Heterogeneous lidar dataset for inter-lidar place recognition under spatiotemporal variations," *The International Journal of Robotics Research*, vol. 43, no. 12, pp. 1867–1883, 2024.

[35] J. Jeong, Y. Cho, Y.-S. Shin, H. Roh, and A. Kim, "Complex urban dataset with multi-level sensors from highly diverse urban environments," *The International Journal of Robotics Research*, vol. 38, no. 6, pp. 642–657, 2019.

[36] M. Jung, S. Jung, and A. Kim, "Asynchronous multiple lidar-inertial odometry using point-wise inter-lidar uncertainty propagation," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 4211–4218, 2023.

[37] Google, "Street View Static API." [Online]. Available: http://googleusercontent.com/maps.google.com/streetview/overview, 2025. Accessed: May 26, 2025.

[38] Naver Corp., "NAVER Maps API v3." [Online]. Available: https://navermaps.github.io/maps.js.en/docs/, 2025. Accessed: May 26, 2025.

[39] R. Arandjelović, P. Gronát, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5297–5307, 2016.

[40] J. Komorowski, M. Wysoczańska, and T. Trzcinski, "Minkloc++: lidar and monocular image fusion for place recognition," in *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2021.

[41] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *Advances in Neural Information Processing Systems*, vol. 37, pp. 21875–21911, 2024.