Multi-Modal Aerial-Ground Cross-View Place Recognition with Neural ODEs

Sijie Wang*[1] Rui She*[2]
Qiyu Kang[3] Siqi Li[1] Disheng Li[1] Tianyu Geng[1] Shangshu Yu[1] Wee Peng Tay[1]
[1]Nanyang Technological University [2]Beihang University
[3]University of Science and Technology of China
wang1679@e.ntu.edu.sg

Abstract

Place recognition (PR) aims at retrieving the query place from a database and plays a crucial role in various applications, including navigation, autonomous driving, and augmented reality. While previous multi-modal PR works have mainly focused on the same-view scenario in which ground-view descriptors are matched with a database of ground-view descriptors during inference, the multi-modal cross-view scenario, in which ground-view descriptors are matched with aerial-view descriptors in a database, remains underexplored. We propose AGPlace, a model that effectively integrates information from multi-modal ground sensors (cameras and LiDARs) to achieve accurate aerial-ground PR. AGPlace achieves effective aerial-ground cross-view PR by leveraging a manifold-based neural ordinary differential equation (ODE) framework with a multi-domain alignment loss. It outperforms existing state-of-the-art cross-view PR models on large-scale datasets. As most existing PR models are designed for ground-ground PR, we adapt these baselines into our cross-view pipeline. Experiments demonstrate that this direct adaptation performs worse than our overall model architecture AGPlace. AGPlace represents a significant advancement in multi-modal aerial-ground PR, with promising implications for real-world applications.

1. Introduction

Place recognition (PR) serves as a fundamental solution to the problem of localization, treating localization as a retrieval task. It involves maintaining a database of geo-tagged sensor data (e.g., images and point clouds) collected at previously visited places. During inference, PR computes the similarity between the query data and the database data to retrieve the most similar places. PR plays a crucial role in various applications, including simultaneous localization and mapping (SLAM) [4] and augmented reality (AR) [49].

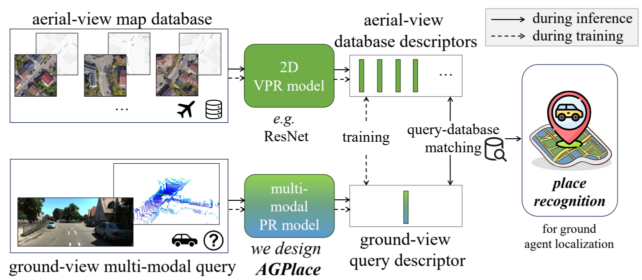

Figure 1. An example pipeline of multi-modal aerial-ground PR. (1) The aerial-view geo-tagged maps (e.g. aerial RGB images, road maps) act as the database. (2) The ground-view multi-modal data (images + point clouds) are the place query to be matched with the database.

PR models are expected to convert the raw sensor data into highly aggregated scene descriptors with much smaller sizes. Traditional techniques leverage hand-crafted features to summarize the query and database data, such as Vector of Locally Aggregated Descriptors (VLAD) [23]. However, these solutions heavily rely on prior human knowledge and may not perform optimally in challenging scenes. Advances in deep learning have facilitated the widespread adoption of learning-based PR works. For example, NetVLAD [3] replaces the manual design of the encoding process used in traditional VLAD with a neural network that learns the representation directly from raw image data.

In recent years, PR has attracted increasing attention from researchers. The initial focus was primarily on developing single-modal solutions, where both query and database come from the same modality, such as images [1, 17, 22, 25, 47, 63, 76] or point clouds [9, 10, 26, 27, 39, 40, 62]. However, with the advancement of multi-sensor processing techniques, there has been a growing interest in multi-modal PR. In this direction, query and database data encompass more than one modality, allowing for richer information fusion. Multi-modal models [28, 29, 74] have demonstrated superior performances to their single-modal counterparts.

*Sijie Wang and Rui She contribute equally. Corresponding author: Rui She.

Apart from models addressing scenarios where both query data and database data come from the same viewpoint, achieving PR in aerial-ground cross-view scenarios [11, 59, 75, 76, 78] presents another challenge. In this case, the database data is collected from aerial views, while the query data is collected from ground views. The aerial-ground PR is particularly valuable for ground robot localization due to extensive aerial RGB or semantic maps offering broad field-of-view (FOV) coverage. Additionally, collecting aerial data is more efficient than ground data, especially in challenging terrains like forests and mountains. These aerial maps provide rich geometric references, enabling ground robots to achieve accurate and effective localization[65, 73]. However, the current state-of-the-art (SOTA) aerial-ground PR models focus on the same-modal setting, i.e., ground *images* are matched to aerial *images*. These models do not consider fusing ground images and point clouds to achieve better cross-view performance.

In this paper, we investigate the challenge of multi-modal PR in the aerial-ground cross-view scenario. More specifically, as shown in Fig. 1, we consider the ground data to be with multiple modalities, consisting of both images and point clouds. On the other hand, aerial data encompasses various formats, including real captured aerial RGB images as well as semantic maps such as road maps.

We introduce **A**erial-**G**round **Place** (**AGPlace**), designed for multi-modal aerial-ground cross-view PR. Our main contributions are summarized as follows:

1. AGPlace utilizes a manifold to fuse multi-modal information by learning manifold chart functions to relate points in 2D or 3D spaces to the manifold. The fusion embedding is constructed using neural ordinary differential equations (ODEs), which produces distinct representations for different ground views, leading to effective PR. Furthermore, by allowing the fusion embedding to evolve on a manifold guided by a multi-domain alignment loss, we allow it to flexibly match the aerial descriptors in the aerial database. The evolved fusion embedding acts as guidance for the extraction of respective modal features to formulate the final scene descriptor.
2. The existing aerial-ground PR datasets like CVUSA [69] and CVACT [36] do not provide the ground LiDAR data, which cannot be used for our multi-modal PR task. Therefore, we establish two benchmark datasets to evaluate AGPlace and related baselines. The datasets are adapted from KITTI360 [35] and nuScenes [8] together with Google Maps to provide multi-sensor ground data and aerial database maps.
3. We conduct extensive numerical experiments demonstrating that AGPlace surpasses previous SOTA multi-modal PR models in aerial-ground cross-view PR. We conduct ablation studies to provide insights into our design.

2. Related Work

2.1. Visual Place Recognition.

In 2D visual PR (VPR), the query and database data are both collected in the 2D image format. Traditionally, handcrafted feature techniques are used to aggregate the whole image, e.g., BoW [13], VLAD [23], and Fisher [45]. These solutions depend heavily on prior human knowledge and may not perform well in challenging scenes. NetVLAD [3] is the first to leverage learnable layers in PR by incorporating soft-weighted feature assignments. NetVLAD has inspired a series of image-based PR works, including general models [1, 7, 17, 22, 25, 63, 77] and aerial-ground models [20, 50, 54, 55, 59, 75, 76].

PointNetVLAD [62] is the first to leverage learning-based methods in point cloud PR. It uses PointNet [46] as the backbone to extract point cloud features followed by a NetVLAD [3] layer to aggregate 3D scene information. With the growing popularity of LiDARs in the field of autonomous driving, there has been a surge in the development of models that leverage point clouds to address the 3D PR problem [9, 27, 39, 72, 79]. In addition, there are also models proposed for aerial-ground 3D PR [16, 24].

2.2. Aerial-Ground Place Recognition.

Aerial-ground PR has become a popular task in localization. This cross-view PR task involves using ground data as the query and aerial data as the database. Unlike general PR models, aerial-ground PR models must address the inherent view differences between aerial and ground data.

SAFA [53] employs polar transformation as prior knowledge to align aerial and ground panoramic images. CDTE [61] enhances this approach by training a generative model to better apply the polar transformation. However, the effectiveness of the polar transformation heavily depends on the geometric correspondence between the two views. Subsequent approaches have focused on directly aligning scene descriptors. TransGeo [76] utilizes shared backbone weights to extract both aerial and ground descriptors. Sample4Geo [11] introduces hard negative mining strategies and employs the InfoNCE loss [48] for symmetric discrimination between aerial and ground features. UCVLG [33] proposes an unsupervised framework to guide cross-view PR models for improved retrieval and re-ranking. ArcGeo [59] introduces the angular loss to better explore cosine similarity between different samples.

Despite these advancements, existing aerial-ground models primarily focus on same-modal inputs. The challenge of effectively fusing ground image and point cloud inputs for aerial-ground PR remains unresolved.

2.3. Multi-Modal Place Recognition.

Recently, multi-modal fusion using information from both cameras and LiDARs has emerged as a promising research topic, demonstrating superior performance in various tasks [37, 66–68]. In the field of PR, MinkLoc++ [28] pioneered the use of separate branches for 2D image and 3D point cloud feature extraction, which are then combined to form the final scene descriptor. Due to their superior performance compared to single-sensor solutions, there has been a growing number of multi-modal PR approaches [15, 29, 41, 57, 71, 74].

However, these multi-modal PR works focus only on the ground-ground same-view situation, where the construction of a ground-view database requires tedious collection of ground data. In contrast, utilizing aerial maps as the database offers a potential solution to address the aforementioned challenges, yet it remains an under-explored area [14].

2.4. Cross-View Metric Localization.

Another important application of cross-view PR is metric localization, which is usually deployed as the second stage in fine-grained localization tasks. After retrieving aerial maps or patches through cross-view PR models, metric localization models can estimate a more accurate sensor pose based on these map references [12, 31, 56].

3. Proposed Approach

In this section, we first present the problem formulation of multi-modal aerial-ground PR, followed by detailed descriptions of our proposed AGPlace approach.

3.1. Problem Formulation

In this work, we address the challenge of aerial-ground multi-modal PR as shown in Fig. 1. We focus on the scenario where the aerial-view 2D data (satellite RGB images and road maps) serve as the database, while the ground-view multi-modal data (2D RGB images + 3D LiDAR point clouds) are used as queries. This can be formulated as follows:

$$\arg\max_{f_{\mathrm{G}}, f_{\mathrm{A}}} \sum_{i,j} \mathbf{\Gamma}(\arg\min_{j}(\mathcal{L}(f_{\mathrm{G}}(\mathbf{I}_{\mathrm{G}}^{i}, \mathbf{P}_{\mathrm{G}}^{i}), f_{\mathrm{A}}(\mathbf{I}_{\mathrm{A}}^{j}))) \in \mathcal{R}^{i}), \tag{1}$$

where $f_{\mathrm{G}}(\cdot,\cdot)$ denotes the ground multi-modal query model, $f_{\mathrm{A}}(\cdot)$ is the aerial database model. $\mathcal{L}(\cdot,\cdot)$ is a distance function. $\mathbf{I}_{\mathrm{G}}$ and $\mathbf{P}_{\mathrm{G}}$ are ground query images and point clouds respectively, $\mathbf{I}_{\mathrm{A}}$ is the aerial database maps, $\mathcal{R}^{i}$ is the positive neighborhood set of query i, $\mathbf{\Gamma}(\cdot)$ is the indicator function.

3.2. Modal-Specific Feature Extraction

The overall pipeline is shown in Fig. 2. We follow the previous cross-view VPR solution[32, 76] that uses two different networks (i.e., not sharing weights) to handle aerial and ground inputs respectively. For the aerial part $f_{\mathrm{A}}(\cdot)$, since VPR for the 2D modality is well-established, we can use off-the-shelf 2D VPR methods (e.g., ResNet-GeM[47]) to extract aerial descriptors. And our design is mainly focusing on the ground part $f_{\mathrm{G}}(\cdot,\cdot)$ that takes multi-modal query inputs. Given 2D image and 3D point cloud query inputs, we first use separate 2D and 3D backbones to extract the respective basic features in each modality. Using hierarchical information is a common technique in various tasks, and we follow this manner to extract multiple feature maps. We denote the feature maps at the l-th backbone block as $\mathbf{F}_{\mathrm{2D}}^{l}$ and $\mathbf{F}_{\mathrm{3D}}^{l}$, and all the gathered feature maps are denoted respectively as:

$$\mathbf{F}_{\mathrm{2D}} = \left\{ \mathbf{F}_{\mathrm{2D}}^{l} \in \mathbb{R}^{N^{l} \times C} \,:\, l \in [L] \right\}, \tag{2}$$

$$\mathbf{F}_{\mathrm{3D}} = \left\{ \mathbf{F}_{\mathrm{3D}}^{l} \in \mathbb{R}^{N^{l} \times C} \,:\, l \in [L] \right\}, \tag{3}$$

where N^{l} is the number of feature vectors at block l, C is the number of corresponding feature channels,[1] and L is the number of backbone blocks.[2]

For images, we use the standard 2D format. For point clouds, we use the 3D voxels format, which is a common format in 3D place recognition[28]. 3D point clouds can also be projected onto 2D planes and extracted by 2D backbones, e.g. spherical projection and birds'-eye-view (BEV) projection. However, these projection methods would either lose information or change the neighboring connections, which would result in insufficient feature extraction.

3.3. Stage 1: Fusion Embedding Construction

To achieve effective information fusion between different modalities in multi-modal PR, we employ a two-stage fusion strategy as depicted in Fig. 2. In this first stage, in addition to the vanilla 2D and 3D feature branches, we introduce a third branch as a surrogate manifold for fusing feature information from the 2D and 3D branches and constructing the fusion embedding.

On the other hand, in the problem of aerial-ground PR, the significant domain gap between aerial and ground data poses a challenge for domain alignment. Introducing a third branch can act as a platform to help bridge this gap. The visualization of descriptor distances, shown in Fig. 3, demonstrates that the inclusion of the fusion manifold aids in better aligning the two domains, resulting in the aerial-ground scene descriptor distance that more closely corresponds to the actual geometric distance.

At the fusion embedding construction stage, we leverage the extracted feature maps $\mathbf{F}_{\mathrm{2D}}$ and $\mathbf{F}_{\mathrm{3D}}$ from each modal backbone to build the fusion embedding. Inspired by the concept of space point motion on manifolds in differential

[1] For illustration simplicity, we assume 2D and 3D backbones have the same number of feature vectors and channels.

[2] We denote $[N] = \{1, \ldots, N\}$.

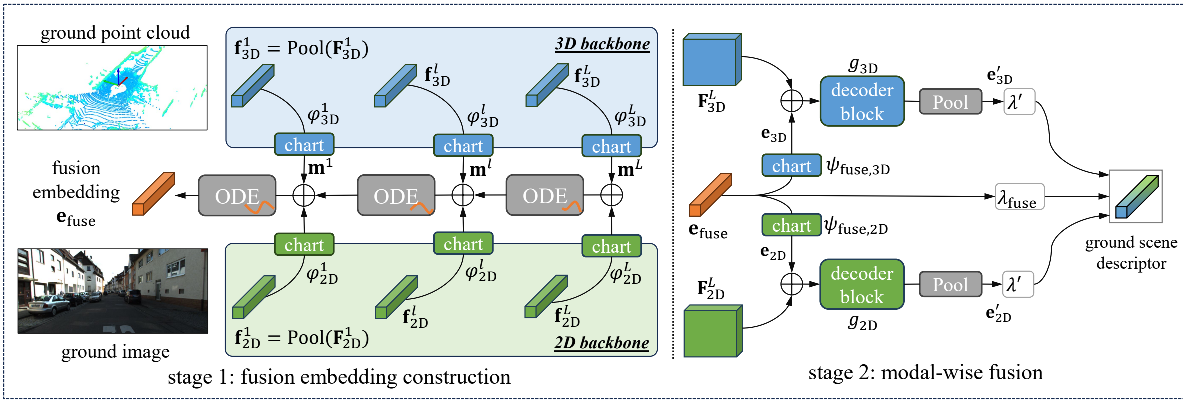

Figure 2. The pipeline overview. In the first stage, the ground image and point cloud are processed with separate backbone branches, the output features of which are used to build the fusion embedding. In the second stage, the constructed fusion embedding is mapped into the respective modal spaces to achieve further modal-wise feature extraction.

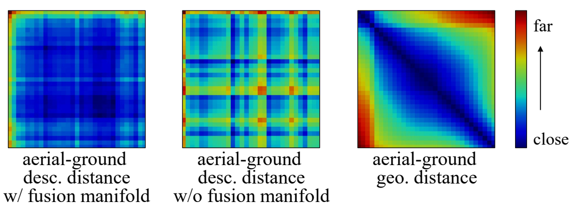

Figure 3. Aerial-ground (database-query) scene descriptor distance visualization. Symmetrization is applied for better visualization.

geometry [30], we model the fusion embedding construction as a multi-block state evolution process (shown in Fig. 4). This process starts from the last block and progresses towards the first block (i.e., $L \to 1$), allowing the embedding state to incorporate more global-level information at later blocks that is beneficial for the PR task. At each block, the evolution consists of two sub-processes: (1) state initialization and (2) state updating. By allowing the fusion embedding to evolve according to a multi-domain alignment loss function (see Sec. 3.5), we enable it to flexibly match the aerial descriptors in the aerial database.

State Initialization.

At each backbone block l, we denote the initial fusion state as $\gamma^l(t=0) \in \mathcal{M}^C$ on the C-dimensional manifold $\mathcal{M}^C$, and define it as:

$$\gamma^l(0) = \begin{cases} \mathbf{m}^l & \text{if } l = L, \\ \mathbf{m}^l + \gamma^{l+1}(T) & \text{otherwise}, \end{cases} \tag{4}$$

where T is the state updating end time, and $\mathbf{m}^l \in \{\mathbf{m}^l\}_{l=1}^L$ is the fusion state momentum. Specifically, $\mathbf{m}^l$ is constructed using feature vectors from the respective l-th backbone block:

$$\mathbf{m}^l = \varphi^l_{2D}(\mathbf{f}^l_{2D}) + \varphi^l_{3D}(\mathbf{f}^l_{3D}). \tag{5}$$

Here, $\varphi^l_{2D}(\cdot), \varphi^l_{3D}(\cdot) : \mathbb{R}^C \to \mathcal{M}^C$ are learnable manifold chart functions representing connections between the 2D/3D space and the fusion manifold. A chart for a manifold $\mathcal{M}$ provides a local coordinate system, capturing local structures of $\mathcal{M}$. $\mathbf{f}^l_{2D}, \mathbf{f}^l_{3D} \in \mathbb{R}^C$ in (5) are the pooled 2D and 3D features from the respective modal space:

$$\mathbf{f}^l_{2D} = \text{Pool}(\mathbf{F}^l_{2D}), \quad \mathbf{f}^l_{3D} = \text{Pool}(\mathbf{F}^l_{3D}), \tag{6}$$

where $\text{Pool}(\cdot)$ is the global average pooling function. The pooling function can highly summarize the global information of the whole scene, which is beneficial for constructing a fusion embedding with rich global multi-modal domain representations and thus benefits the PR task. This differs from dense prediction tasks (e.g. depth estimation, segmentation) that more dig into local patterns.

State Updating. In Riemannian geometry, the movement of a point on a manifold can be described by a parametric curve [30]. Inspired by this, we model the process of updating the fusion embedding state as the movement of a point in the manifold.

ODEs serve as a fundamental tool for describing dynamical systems[52, 60, 64]. They express how a system state evolves over time, capturing relationships between the system variables and the rates of change. Based on this, the movement dynamics on a manifold can be described by an ODE:

$$\frac{d\gamma(t)}{dt} = f(t, \gamma(t)), \tag{7}$$

where $\gamma(t)$ denotes the state trajectory. The following non-intersection property of ODE solutions is well-known.

Theorem 1. *(Non-intersection of ODE solutions [21]) Given the ODE in (7), where f is continuous in t and globally Lipschitz continuous in γ. Let $\gamma_1(t)$ and $\gamma_2(t)$ be two solutions of the ODE in (7). If the initial conditions $\gamma_1(0) \neq \gamma_2(0)$, then it holds that $\gamma_1(t) \neq \gamma_2(t)$ for all $t \in [0, \infty)$.*

The non-intersection property of ODE solutions in Theorem 1 indicates that different ODE inputs guarantee different

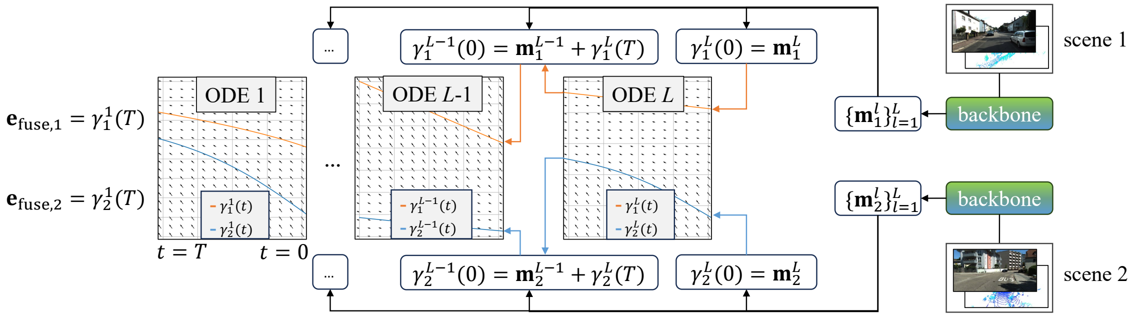

Figure 4. Fusion embedding evolution process starting from the last block and ending at the first block (i.e. $L \to 1$). The trajectories contain the movement of fusion embedding and are depicted by neural ODEs. In each ODE block, different inputs at time 0 guarantee different outputs at time T.

ODE outputs. This property helps to construct a more effective embedding that benefits PR. In our pipeline in (4), given that the initial fusion states at block l are different for different places, neural ODEs can update feature representations while theoretically maintaining the feature differences at this block, which lead to *distinct scene descriptors yielding more robust PR* (cf. Fig. 4).

Specifically, we update the fusion embedding from its initial state using parameterized neural ODEs at each block l, which can be regarded as point movements on the manifold:

$$\frac{d\gamma^l(t)}{dt} = f_{\theta^l}(\gamma^l(t)), \tag{8}$$

where $\gamma^l(t)$ is the fusion embedding state trajectory at block l, $f_{\theta^l}(\cdot)$ is a non-linear neural block with learnable parameters θ^l that describes the trajectory dynamics. By solving (8), we can obtain the updated state $\gamma^l(T)$ at the end time T. At the next block $l-1$, the start state $\gamma^{l-1}(0)$ is initialized based on momentum $\mathbf{m}^{l-1}$ and the last end state $\gamma^l(T)$, as shown in Fig. 4 and (4). After progressive updating, finally, we obtain the end state at the first block $\gamma^1(T)$ as the final constructed fusion embedding $\mathbf{e}_{\text{fuse}} = \gamma^1(T) \in \mathcal{M}^C$. The following result follows immediately from Theorem 1.

Corollary 1. *(Distinguished fusion states.) In each neural ODE block l, given ODE inputs $\gamma_1^l(0) \neq \gamma_2^l(0)$ from two different scenes that are different, then the two ODE outputs at this ODE block are different, i.e., $\gamma_1^l(T) \neq \gamma_2^l(T)$.*

As stated in Corollary 1, the neural ODE process can produce different fusion states, which would help construct more distinguished representations for different places and thus contribute to more effective PR performance. The visualized illustration of Corollary 1 is shown in Fig. 4.

3.4. Stage 2: Modal-Wise Fusion

The fusion embedding $\mathbf{e}_{\text{fuse}}$ captures rich multi-modal fusion information, which in turn serves as effective guidance for individual fine-grained local-level modal feature learning.

With the surrogate fusion embedding injected into the respective 2D/3D spaces, the tight coupling connection of the 2D and 3D embeddings could be softened, where the perturbation of a modality could be mitigated by the surrogate fusion information. As shown in Tab. 2, with the help of included fusion embedding in this stage 2, the final scene descriptor could be more robust against sensor failing.

Specifically, as illustrated in Fig. 2, we project the fusion embedding into the corresponding modal space to obtain 2D and 3D embeddings:

$$\mathbf{e}_{\text{2D}} = \psi_{\text{fuse,2D}}(\mathbf{e}_{\text{fuse}}), \quad \mathbf{e}_{\text{3D}} = \psi_{\text{fuse,3D}}(\mathbf{e}_{\text{fuse}}), \tag{9}$$

where $\psi_{\text{fuse,2D}}(\cdot), \psi_{\text{fuse,3D}}(\cdot) : \mathcal{M}^C \to \mathbb{R}^C$ are chart functions that establish connections from the fusion manifold to the respective 2D and 3D spaces. These 2D and 3D embeddings capture rich multi-modal information, which are further combined with the original modal feature maps $\mathbf{F}^L_{\text{2D}}$ and $\mathbf{F}^L_{\text{3D}}$ from the last block L to construct more effective representations through decoder blocks $g_{\text{2D}}(\cdot), g_{\text{3D}}(\cdot)$:

$$\mathbf{e}'_{\text{2D}} = \text{Pool}(g_{\text{2D}}(\mathbf{e}_{\text{2D}} \oplus \mathbf{F}^L_{\text{2D}})), \tag{10}$$

$$\mathbf{e}'_{\text{3D}} = \text{Pool}(g_{\text{3D}}(\mathbf{e}_{\text{3D}} \oplus \mathbf{F}^L_{\text{3D}})), \tag{11}$$

where $\oplus$ denotes broadcast addition. By injecting the highly summarized fusion embedding, the respective local modal feature maps can capture more global information and achieve more robust representations. The decoded embeddings $\mathbf{e}'_{\text{2D}}$ and $\mathbf{e}'_{\text{3D}}$ are used to construct the final scene descriptor in the sequel.

Our final ground scene descriptor consists of the decoded modal embeddings, as well as the fusion embedding. It is obtained as follows:

$$\mathbf{e}_{\text{G}} = \lambda'(\mathbf{e}'_{\text{2D}} + \mathbf{e}'_{\text{3D}}) + \lambda_{\text{fuse}}\mathbf{e}_{\text{fuse}}, \tag{12}$$

where $\lambda', \lambda_{\text{fuse}}$ are weights to balance different components.

3.5. Multi-View Multi-Modal Loss Function

In the multi-modal cross-view PR task, the features are interacting not only in different modalities (2D and 3D) but also in different views (aerial and ground). Thus, sufficient feature alignment across various domains is critical for this PR task. To achieve this goal, we propose the multi-view multi-modal (MVMM) loss function. The MVMM loss treats PR as a binary classification problem, where positive pairs should have smaller feature distances and are assigned ground-truth labels “0”, while negative pairs have ground-truth labels “1”.

Given two sets of embeddings $\mathcal{D}_1 = \{e_{1i}\}_{i\in[N_1]}, \mathcal{D}_2 = \{e_{2j}\}_{j\in[N_2]}$ from two domains, we first define the basic domain distance loss as:

$$\mathcal{L}(\mathcal{D}_1, \mathcal{D}_2) = \frac{1}{N_1 N_2}\sum_{ij}\Big(y_{ij}\log(\sigma(\|e_{1i}-e_{2j}\|_2)) + (1-y_{ij})\log(1-\sigma(\|e_{1i}-e_{2j}\|_2))\Big), \tag{13}$$

where $y_{ij} \in \{0, 1\}$ indicates positive and negative pair labels, and $\sigma(\cdot)$ is the sigmoid function. We then further define the MVMM loss as:

$$\ell_{\mathrm{MVMM}} = \mathcal{L}(\mathcal{D}_{\mathrm{A}}, \mathcal{D}_{\mathrm{A}}) + \mathcal{L}(\mathcal{D}_{\mathrm{G}}, \mathcal{D}_{\mathrm{A}} \cup \mathcal{D}_{\mathrm{G}}) + \mathcal{L}(\mathcal{D}_{\mathrm{G2D}}, \mathcal{D}_{\mathrm{A}} \cup \mathcal{D}_{\mathrm{G2D}}) + \mathcal{L}(\mathcal{D}_{\mathrm{G3D}}, \mathcal{D}_{\mathrm{A}} \cup \mathcal{D}_{\mathrm{G3D}}), \tag{14}$$

where $\mathcal{D}_{\mathrm{A}} = \{e_{\mathrm{A}i}\}_{i\in M}$ is a set of aerial descriptors. $\mathcal{D}_{\mathrm{G}} = \{\mathbf{e}_{\mathbf{G}j}\}_{j\in N}$, $\mathcal{D}_{\mathrm{G2D}} = \{\mathbf{f}^{L}_{2\mathrm{D}j}\}_{j\in N}$, $\mathcal{D}_{\mathrm{G3D}} = \{\mathbf{f}^{L}_{3\mathrm{D}j}\}_{j\in N}$ are sets of corresponding ground final, 2D, and 3D embeddings as in (6) and (12). The MVMM loss aligns features not only in different domains but also within the same domains. That is, the MVMM loss also optimizes the ground-ground and aerial-aerial feature relationships, which could better help cross-view PR. As shown in Fig. 7, the roof that can only be perceived by aerial views is emphasized, which indicates the aerial-aerial same-view discrimination is implicitly hidden in the whole cross-view discrimination process. In addition, the MVMM loss also directly aligns ground 2D and 3D features to the aerial features, which would implicitly contribute to more effective aerial-ground alignment for cross-view PR.

The above MVMM loss can be regarded as the direct guidance that forces pair distances close to some target values. On the other hand, we can also force negative pairs to have greater distances than positive pairs, which can be regarded as the relative guidance. Specifically, we use the traditional triplet loss[19] to achieve the relative guidance:

$$\ell_{\mathrm{tri}} = \frac{1}{N}\sum_{j}\max(\|\mathbf{e}_{\mathrm{G}i} - \mathbf{e}^{\mathrm{p}}_{\mathrm{A}}\|_2 - \|\mathbf{e}_{\mathrm{G}i} - \mathbf{e}^{\mathrm{n}}_{\mathrm{A}}\|_2 + m, 0), \tag{15}$$

where m is the margin. $\mathbf{e}^{\mathrm{p}}_{\mathrm{A}}, \mathbf{e}^{\mathrm{n}}_{\mathrm{A}}$ are positive and negative aerial descriptors sampled from $\mathcal{D}_{\mathrm{A}}$. Following previous works [6, 11, 59], hard negative mining is adopted during training. Our final loss is a combination of the directly-guided MVMM loss and the relatively-guided triplet loss with the weight α:

$$\ell = \alpha\ell_{\mathbf{MVMM}} + \ell_{\mathrm{tri}}. \tag{16}$$

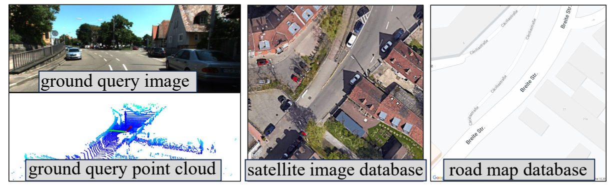

Figure 5. Data visualization from the KITTI360-AG dataset.

4. Experiments

4.1. Datasets and Implementation Details

To evaluate the performance of multi-modal aerial-ground PR models, we create benchmark datasets that consist of comprehensive ground sensor data and corresponding aerial database maps. We specifically construct two datasets, namely *KITTI360-AG* and *nuScenes-AG*. Examples are shown in Fig. 5. In addition, we also test our models on the Oxford RobotCar benchmark datasets used in previous multi-model ground-ground PR evaluations.

KITTI360-AG. For the ground dataset, we use the vanilla KITTI360 [35] dataset as the base to provide ground-view images, point clouds, and GNSS coordinates. We use a total of 7 sequences, in which the first 85% frames are for training and the last 15% frames are for testing, such that the test frames consist of seen and unseen areas for both fitting and generalization evaluation. The aerial data is generated based on the GNSS coordinates of the ground data frames. We use the Google Maps Static API[3] to download the corresponding aerial images, including satellite images, and road maps. We set the Google Maps Static API parameters as follows: scale=1, zoom=20, size=640×640. Each aerial image covers an approximately 75×75 m^2 area.

nuScenes-AG. The nuScenes [8] dataset provides a rich collection of sensor data collected from real urban driving scenarios, which also contains multi-view cameras. We download the aerial data based on GNSS coordinates with similar configurations mentioned in KITTI360-AG. The official train/test split is used in our setting, such that test scenes contain seen and unseen areas for both fitting and generalization evaluation.

Oxford RobotCar Benchmark. The Oxford RobotCar benchmark dataset is a public benchmark used for multi-modal ground-ground PR. There are two different splits used by MinkLoc++ (Oxford-Mink+) and AdaFusion (Oxford-Ada) respectively. And we compare our model under both settings.

Implementation. Our selected baselines mainly consist of four types, including ground-view 2D/3D/multi-modal base-

[3]https://developers.google.com/maps/documentation/maps-static/overview

lines and aerial-ground 2D baselines. We use ResNet-18-Avg [18] as the aerial network for all PR works except for AnyLoc [25] for a fair comparison (with a head to align descriptor dimensions). We select open-sourced SOTA multi-modal models as ground network baselines for query descriptor construction, mainly including MinkLoc++ [28], AdaFusion [29], LCPR [74], MSSPlace [41], and UMF [15], VSGP-PR [14] does not provide open-source codes and EINet [71] only provides data splits without model codes, which hinders us from comparing with them. Our backbones are constructed based on MinkLoc++ [28]. We use recall@K ($K = 1, 5, 10$)(R@K), average recall@1%(AR@%1), and average recall@1(AR@1) as the metrics. The positive retrieval threshold is set as 25 m. All experiments are conducted on a Tesla A100 GPU. More dataset, baseline, and implementation details are provided in the supplement.

4.2. Main Results

KITTI360-AG (Satellite and Roadmap Databases). We conduct an evaluation of AGPlace on the KITTI360-AG dataset as demonstrated in Tab. 1 (Best is in **bold** and second best is underlined). AGPlace outperforms all other baselines in both satellite and road map settings. On the other hand, when using the road map aerial database, most models can already show considerable recall performance. This suggests that semantic maps, which are easily accessible and created without the need for satellites, drones, or airplanes, can already provide sufficient information for practical PR tasks. This finding opens up interesting possibilities for future research in the field [51, 70]. We also test the 2D VPR methods with the strong foundation backbone DINOV2[43], some of which can even surpass 3D counterparts.

nuScenes-AG (Ground Sensor Failing). We next experiment on the nuScenes-AG dataset that consists of a LiDAR and multiple cameras, where AGPlace can also achieve significant performance. With the fusion embedding injected into the respective 2D/3D spaces in stage 2, the final scene descriptor demonstrates stronger robustness both camera and LiDAR failing. This demonstrates that our model can be more applicable in real deployment conditions.

Oxford Benchmark Dataset. To investigate the inherent power of our ground-view model for PR, we perform a comparison with previous SOTA baselines using the *ground-ground PR* task. As in Tab. 3, our model (without using extra data, re-ranking, or multiple cameras) achieves better performance than previous SOTA counterparts, which further verifies the effectiveness of our design.

4.3. Ablation Studies

Module Ablation. We verify the effectiveness by ablating each module in our network. As shown in Tab. 4, the two fusion stages in Fig. 2 and the MVMM loss in Sec. 3.5 can contribute to the better final performance, confirming the efficacy of our approach.

ODE Updating and Fusion Direction. During the state

Model	Type	Satellite R@1/5/10	Road Map R@1/5/10
ConvAP (ResNet-18) [2]	2D	25.2 / 40.8 / 48.6	20.3 / 32.0 / 39.7
CosPlace (ResNet-18) [5]	2D	22.8 / 38.9 / 47.4	17.4 / 31.1 / 38.8
MixVPR (ResNet-18) [1]	2D	19.4 / 30.5 / 37.3	19.3 / 32.0 / 39.1
AnyLoc* (DINOV2-s) [25]	2D	4.2 / 8.7 / 13.0	5.4 / 5.5 / 5.9
SALAD (DINOV2-s) [22]	2D	26.1 / 34.3 / 39.3	17.0 / 27.4 / 33.5
TransGeo (DINOV2-s) [76]	2D+C	22.9 / 32.9 / 39.2	22.4 / 31.9 / 36.9
Sample4Geo (ConvNeXt-t) [11]	2D+C	27.0 / 41.7 / 49.0	22.7 / 41.0 / 47.6
ArcGeo (ConvNeXt-t) [59]	2D+C	28.2 / 41.3 / 48.5	24.4 / 39.6 / 46.5
MinkLoc3DV2 [27]	3D	26.5 / 39.1 / 46.0	23.3 / 36.0 / 43.8
BEVPlace [39]	3D	22.3 / 33.9 / 40.8	23.6 / 35.5 / 41.4
VXP [34]	3D+C	20.7 / 34.5 / 41.0	22.4 / 34.9 / 42.0
Lip-Loc [58]	3D+C	29.9 / 42.2 / 49.0	24.5 / 35.6 / 42.4
MinkLoc++ [28]	MM	28.9 / 39.3 / 44.9	26.5 / 40.8 / 48.8
AdaFusion [29]	MM	26.5 / 39.6 / 47.8	27.2 / 41.6 / 49.3
MSSPlace [41]	MM	25.5 / 37.4 / 45.0	26.7 / 41.2 / 49.0
LCPR [74]	MM	27.7 / 44.4 / 50.6	24.5 / 39.0 / 47.7
UMF [15]	MM	27.1 / 42.6 / 49.2	25.6 / 40.4 / 49.7
AGPlace (ours)	MM+C	**32.0 / 47.6 / 54.9**	**28.2 / 43.3 / 52.0**

Table 1. Aerial-ground PR results on the KITTI-360-AG dataset using satellite or road map aerial sources. "*" denotes the model is frozen and purely relies on pre-trained weights. "C" denotes the model is designed for cross-view/cross-modal PR. Hard negative mining is applied to all models.

Model	cams + LiDAR R@1	cams + LiDAR R@5	cams fail R@1	cams fail R@5	LiDAR fail R@1	LiDAR fail R@5
MinkLoc++ [28]	70.4	82.1	16.9	27.3	2.6	7.3
AdaFusion [29]	71.9	82.3	9.0	17.8	3.5	6.7
MSSPlace [41]	71.3	82.2	17.4	29.8	5.0	8.7
LCPR [74]	57.7	74.2	3.2	6.2	8.6	17.1
UMF [15]	69.4	82.5	2.7	7.3	0.8	2.9
AGPlace w/o stg. 2	73.9	84.2	19.2	30.1	6.9	16.7
AGPlace (ours)	**75.6**	**87.2**	**22.8**	**33.1**	**12.9**	**23.0**

Table 2. Aerial-ground PR results on the nuScenes-AG dataset using satellite image database. "fail" denotes dropping the modality input during testing. All models are trained with both modalities.

Model	Extra train data	Oxford-Mink+ AR@1%	Oxford-Mink+ AR@1	Oxford-Ada AR@1%	Oxford-Ada AR@1
CORAL [44]		96.1	-	-	-
PIC-Net [38]		98.2	-	-	-
Cues-Net [42]		-	-	-	98.0
MinkLoc++ [28]		99.1	96.7	-	-
AdaFusion [29]		-	-	99.2	98.2
HMPR [57]		-	-	99.6	96.9
UMF [15]	✓	99.1	97.9	-	-
MSSPlace [41]		99.1	97.6	-	-
HMPR + *re-rank*		-	-	99.7	99.0
UMF + *re-rank*	✓	99.3	**98.3**	-	-
MSSPlace + *mul-cam*		99.5	98.2	-	-
AGPlace (ours)		**99.7**	**98.3**	**99.9**	**99.3**

Table 3. Ground-ground PR Results on the Oxford benchmark datasets. "-" denotes the result is not provided in the paper.

Model	KITTI360-AG (Sat.) R@1	R@5	R@10
full	**32.0**	**47.6**	**54.9**
full w/o stage 1	28.5	43.9	51.0
full w/o stage 2	30.6	45.3	53.0
full w/o ℓ_{MVMM}	30.9	45.2	51.7

Table 4. Module ablation.

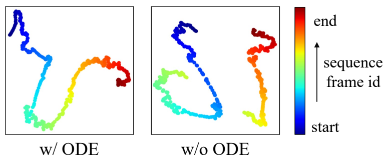

Figure 6. t-SNE plots of ground scene descriptors (from a consecutive frame sequence). ODEs can help build more consistent descriptors that align with the consecutive geometry.

evolution process, we parameterize the fusion embedding state trajectories as learnable neural ODEs. Notably in Tab. 5, ODEs perform better than multi-layer perceptron (MLP) and attention, both of which cannot guarantee outputs' differences. In addition, the t-SNE visualization in Fig. 6 shows ODEs can build more consecutive scene descriptors, underscoring the necessity of the ODE-based fusion state updating process for PR.

Next, we evaluate the performance of different fusion directions, as outlined in Tab. 5. The direction "block $L \rightarrow$ block 1", which transfers information from high-level summaries to detailed low-level representations, emerges as the preferred choice for fusion embedding construction.

State Updating Method	KITTI360-AG (Sate.) R@1	R@5	R@10
w/ MLP	30.0	45.2	51.9
w/ attention	30.4	45.4	52.8
w/ ODE	**32.0**	**47.6**	**54.9**
Fuse Direction			
block 1 $\rightarrow$ block L	31.7	46.8	54.0
block 1 $\leftarrow$ block L	**32.0**	**47.6**	**54.9**

Table 5. State updating method and fusion direction.

Aerial Modality and Salience Visualization. We also explore fusing satellite images and road maps for more representative aerial descriptors. As in Tab. 6, the fusion on both ground and aerial sides gives even better performance, which indicates an interesting research topic for aerial fusion in the future. To investigate which part of ground and aerial inputs are focused in the PR network, we visualize the salience maps as shown in Fig. 7. The geometric landmarks are emphasized (e.g. buildings and roads), which verifies the geo-feature extraction effectiveness of our model architecture and the alignment effectiveness of our loss function.

Runtime Performance. Finally, we assess the runtime speed and memory consumption of our model. AGPlace demonstrates acceptable real-time performance, satisfying basic deployment and application requirements. All compared models exhibit comparable memory usage, with no significant discrepancies observed in this metric.

Ground Modality	Aerial Modality	KITTI360-AG R@1	R@5	R@10
image + point cloud	satellite	32.0	47.6	54.9
image + point cloud	roadmap	28.2	43.3	52.0
image + point cloud	satellite + roadmap	**34.7**	**48.7**	**55.9**

Table 6. Ground and aerial modality comparison.

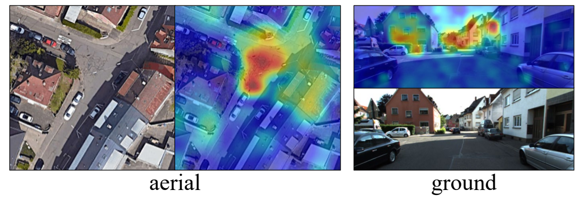

Figure 7. Salience map visualization.

Model	Speed (FPS)	GPU Memory Usage (GB)
MinkLoc++ [28]	80	**0.58**
AdaFusion [29]	72	0.59
MSSPlace [41]	75	0.59
LCPR [74]	**125**	0.72
UMF [15]	70	0.60
AGPlace (ours)	62	0.61

Table 7. Runtime performance comparison on a Tesla A100.

5. Conclusion and Limitations

In this work, we have proposed AGPlace, a multi-modal model for addressing the aerial-ground PR problem. AGPlace leverages neural ODEs rooted in differential geometry to achieve effective multi-modal feature interaction. Our experiments demonstrate that AGPlace outperforms previous baselines and performs well in large-scale scenes, making it suitable for deployment in real-world environments.

However, this work primarily focuses on ground query data collected in urban environments, which may limit its generalizability to other realistic settings. In future work, we plan to expand our dataset and model to include more diverse and challenging scenarios. This will enable us to establish a more comprehensive and generalized multi-modal cross-view PR pipeline.

6. Acknowledgment

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority under its Future Communications Research & Development Programme.

References

[1] Amar Ali-Bey, Brahim Chaib-Draa, and Philippe Giguere. MixVPR: Feature mixing for visual place recognition. In

Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pages 2998–3007, 2023. 1, 2, 7

[2] Amar Alibey, Brahim Chaibdraa, and Philippe Giguere. GSV-Cities: Toward appropriate supervised visual place recognition. *Neurocomputing*, 513:194–203, 2022. 7

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016. 1, 2

[4] Saba Arshad and Gon-Woo Kim. Role of deep learning in loop closure detection for visual and LiDAR slam: A survey. *Sensors*, 21(4):1243, 2021. 1

[5] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4878–4888, 2022. 7

[6] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5396–5407, 2022. 6

[7] Gabriele Berton, Gabriele Trivigno, Barbara Caputo, and Carlo Masone. EigenPlaces: Training viewpoint robust models for visual place recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11080–11090, 2023. 2

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 2, 6

[9] Lineng Chen, Huan Wang, Hui Kong, Wankou Yang, and Mingwu Ren. PTC-Net: Point-wise Transformer with sparse convolution network for place recognition. *IEEE Robotics and Automation Letters*, 2023. 1, 2

[10] Xieyuanli Chen, Thomas Läbe, Andres Milioto, Timo Röhling, Olga Vysotska, Alexandre Haag, Jens Behley, and Cyrill Stachniss. OverlapNet: Loop closing for LiDAR-based SLAM. *arXiv preprint arXiv:2105.11344*, 2021. 1

[11] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 2, 6, 7

[12] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21621–21631, 2023. 3

[13] Dorian Gálvez-López and Juan D Tardos. Bags of binary words for fast place recognition in image sequences. *IEEE Transactions on Robotics*, 28(5):1188–1197, 2012. 2

[14] Peng Gao, Jing Liang, Yu Shen, Sanghyun Son, and Ming C Lin. Visual, spatial, geometric-preserved place recognition for cross-view and cross-modal collaborative perception. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 11079–11086, 2023. 3, 7

[15] Alberto García-Hernández, Riccardo Giubilato, Klaus H Strobl, Javier Civera, and Rudolph Triebel. Unifying local and global multimodal features for place recognition in aliased and low-texture environments. *arXiv preprint arXiv:2403.13395*, 2024. 3, 7, 8

[16] Tianrui Guan, Aswath Muthuselvam, Montana Hoover, Xijun Wang, Jing Liang, Adarsh Jagan Sathyamoorthy, Damon Conover, and Dinesh Manocha. CrossLoc3D: Aerial-ground cross-source 3D place recognition. *arXiv preprint arXiv:2303.17778*, 2023. 2

[17] Stephen Hausler, Sourav Garg, Ming Xu, Michael Milford, and Tobias Fischer. Patch-NetVLAD: Multi-scale fusion of locally-global descriptors for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14141–14152, 2021. 1, 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 7

[19] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *Proceedings of the International Conference on Learning Representations*, pages 84–92, 2015. 6

[20] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. 2

[21] Edward L Ince. *Ordinary differential equations*. Courier Corporation, 1956. 4

[22] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17658–17668, 2024. 1, 2, 7

[23] Hervé Jégou, Florent Perronnin, Matthijs Douze, Jorge Sánchez, Patrick Pérez, and Cordelia Schmid. Aggregating local image descriptors into compact codes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(9): 1704–1716, 2011. 1, 2

[24] Yingrui Jie, Yilin Zhu, and Hui Cheng. Heterogeneous deep metric learning for ground and aerial point cloud-Based place recognition. *IEEE Robotics and Automation Letters*, 2023. 2

[25] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 2023. 1, 2, 7

[26] 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. 1

[27] Jacek Komorowski. Improving point cloud based place recognition with ranking-based loss and large batch training. In *Proceedings of the International Conference on Pattern Recognition*, pages 3699–3705, 2022. 1, 2, 7

[28] Jacek Komorowski, Monika Wysoczańska, and Tomasz Trzcinski. MinkLoc++: LiDAR and monocular image fusion for place recognition. In *Proceedings of the International Joint Conference on Neural Networks*, pages 1–8, 2021. 1, 3, 7, 8

[29] Haowen Lai, Peng Yin, and Sebastian Scherer. AdaFusion: Visual-LiDAR fusion with adaptive weights for place recognition. *IEEE Robotics and Automation Letters*, 7(4):12038–12045, 2022. 1, 3, 7, 8

[30] John M Lee and John M Lee. *Smooth manifolds*. Springer, 2012. 4

[31] Ted Lentsch, Zimin Xia, Holger Caesar, and Julian FP Kooij. SliceMatch: Geometry-guided aggregation for cross-view pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17225–17234, 2023. 3

[32] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16719–16729, 2024. 3

[33] Guopeng Li, Ming Qian, and Gui-Song Xia. Unleashing unlabeled data: A paradigm for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16719–16729, 2024. 2

[34] Yun-Jin Li, Mariia Gladkova, Yan Xia, Rui Wang, and Daniel Cremers. Vxp: Voxel-cross-pixel large-scale image-lidar place recognition. *arXiv preprint arXiv:2403.14594*, 2024. 7

[35] 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. 2, 6

[36] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 2

[37] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *Proceedings of the IEEE international conference on robotics and automation*, pages 2774–2781, 2023. 3

[38] Yuheng Lu, Fan Yang, Fangping Chen, and Don Xie. PIC-Net: Point cloud and image collaboration network for large-scale place recognition. *arXiv preprint arXiv:2008.00658*, 2020. 7

[39] 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. 1, 2, 7

[40] Junyi Ma, Jun Zhang, Jintao Xu, Rui Ai, Weihao Gu, and Xieyuanli Chen. OverlapTransformer: An efficient and yaw-angle-invariant Transformer network for LiDAR-based place recognition. *IEEE Robotics and Automation Letters*, 7(3): 6958–6965, 2022. 1

[41] Alexander Melekhin, Dmitry Yudin, Ilia Petryashin, and Vitaly Bezuglyj. MSSPlace: Multi-sensor place recognition with visual and text semantics. *arXiv preprint arXiv:2407.15663*, 2024. 3, 7, 8

[42] Amadeus Oertel, Titus Cieslewski, and Davide Scaramuzza. Augmenting visual place recognition with structural cues. *IEEE Robotics and Automation Letters*, 5(4):5534–5541, 2020. 7

[43] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOv2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[44] Yiyuan Pan, Xuecheng Xu, Weijie Li, Yunxiang Cui, Yue Wang, and Rong Xiong. CORAL: Colored structural representation for bi-modal place recognition. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2084–2091, 2021. 7

[45] Florent Perronnin, Yan Liu, Jorge Sánchez, and Hervé Poirier. Large-scale image retrieval with compressed fisher vectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3384–3391, 2010. 2

[46] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 2

[47] 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. 1, 3

[48] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 2

[49] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. LaMAR: Benchmarking localization and mapping for augmented reality. In *Proceedings of the European Conference on Computer Vision*, pages 686–704, 2022. 1

[50] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, 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. 2

[51] Paul-Edouard Sarlin, Eduard Trulls, Marc Pollefeys, Jan Hosang, and Simon Lynen. SNAP: self-supervised neural maps for visual positioning and semantic understanding. *arXiv preprint arXiv:2306.05407*, 2023. 7

[52] Rui She, Qiyu Kang, Sijie Wang, Yuan-Rui Yang, Kai Zhao, Yang Song, and Wee Peng Tay. Robustmat: Neural diffusion for street landmark patch matching under challenging environments. *IEEE Transactions on Image Processing*, 32: 5550–5563, 2023. 4

[53] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32, 2019. 2

[54] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am I looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020. 2

[55] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11990–11997, 2020. 2

[56] Yujiao Shi, Fei Wu, Akhil Perincherry, Ankit Vora, and Hongdong Li. Boosting 3-DoF ground-to-satellite camera localization accuracy via geometry-guided cross-View Transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21516–21526, 2023. 3

[57] Chengfu Shu and Yutao Luo. A hierarchical and multi-modal framework for place recognition with a learnable metric. *IEEE Transactions on Intelligent Vehicles*, 2024. 3, 7

[58] Sai Shubodh, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pre-training for cross-modal localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 948–957, 2024. 7

[59] Maxim Shugaev, Ilya Semenov, Kyle Ashley, Michael Klaczynski, Naresh Cuntoor, Mun Wai Lee, and Nathan Jacobs. ArcGeo: Localizing limited field-of-view images using cross-view matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 209–218, 2024. 2, 6, 7

[60] Yang Song, Qiyu Kang, Sijie Wang, Kai Zhao, and Wee Peng Tay. On the robustness of graph neural diffusion to topology perturbations. *Advances in Neural Information Processing Systems*, 35:6384–6396, 2022. 4

[61] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 2

[62] Mikaela Angelina Uy and Gim Hee Lee. PointNetVLAD: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4470–4479, 2018. 1, 2

[63] Ruotong Wang, Yanqing Shen, Weiliang Zuo, Sanping Zhou, and Nanning Zheng. TransVPR: Transformer-based place recognition with multi-level attention aggregation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13648–13657, 2022. 1, 2

[64] Sijie Wang, Qiyu Kang, Rui She, Wee Peng Tay, Andreas Hartmannsgruber, and Diego Navarro Navarro. RobustLoc: Robust camera pose regression in challenging driving environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6209–6216, 2023. 4

[65] Shan Wang, Chuong Nguyen, Jiawei Liu, Yanhao Zhang, Sundaram Muthu, Fahira Afzal Maken, Kaihao Zhang, and Hongdong Li. View from above: Orthogonal-view aware cross-view localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14843–14852, 2024. 2

[66] Sijie Wang, Rui She, Qiyu Kang, Xingchao Jian, Kai Zhao, Yang Song, and Wee Peng Tay. DistilVPR: Cross-modal knowledge distillation for visual place recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10377–10385, 2024. 3

[67] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. OpenOccupancy: A large scale benchmark for surrounding semantic occupancy perception. *arXiv preprint arXiv:2303.03991*, 2023.

[68] Zeyu Wang, Dingwen Li, Chenxu Luo, Cihang Xie, and Xiaodong Yang. DistillBEV: Boosting multi-camera 3D object detection with cross-modal knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8637–8646, 2023. 3

[69] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 2

[70] Hang Wu, Zhenghao Zhang, Siyuan Lin, Xiangru Mu, Qiang Zhao, Ming Yang, and Tong Qin. MapLocNet: Coarse-to-fine feature registration for visual re-localization in vavigation maps. *arXiv preprint arXiv:2407.08561*, 2024. 7

[71] Jingyi Xu, Junyi Ma, Qi Wu, Zijie Zhou, Yue Wang, Xieyuanli Chen, and Ling Pei. Explicit interaction for fusion-based place recognition. *arXiv preprint arXiv:2402.17264*, 2024. 3, 7

[72] Tian-Xing Xu, Yuan-Chen Guo, Zhiqiang Li, Ge Yu, Yu-Kun Lai, and Song-Hai Zhang. TransLoc3D: point cloud based large-scale place recognition using adaptive receptive fields. *Communications in Information and Systems*, 23(1):57–83, 2023. 2

[73] Yanhao Zhang, Yujiao Shi, Shan Wang, Ankit Vora, Akhil Perincherry, Yongbo Chen, and Hongdong Li. Increasing SLAM pose accuracy by ground-to-satellite image registration. *arXiv preprint arXiv:2404.09169*, 2024. 2

[74] Zijie Zhou, Jingyi Xu, Guangming Xiong, and Junyi Ma. LCPR: A multi-scale attention-based LiDAR-camera fusion network for place recognition. *IEEE Robotics and Automation Letters*, 2024. 1, 3, 7, 8

[75] Sijie Zhu, Taojiannan Yang, and Chen Chen. VIGOR: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 2

[76] Sijie Zhu, Mubarak Shah, and Chen Chen. TransGeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022. 1, 2, 3, 7

[77] Sijie Zhu, Linjie Yang, Chen Chen, Mubarak Shah, Xiaohui Shen, and Heng Wang. R2Former: Unified retrieval and reranking Transformer for place recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19370–19380, 2023. 2

[78] Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023. 2

[79] Kamil Żywanowski, Adam Banaszczyk, Michał R Nowicki, and Jacek Komorowski. MinkLoc3D-SI: 3D LiDAR place recognition with sparse convolutions, spherical coordinates, and intensity. *IEEE Robotics and Automation Letters*, 7(2): 1079–1086, 2021. 2