A Structural-Analysis-Based Information Fusion for Multi-Modal Cross-View Geo-Localization

Xu Yan[1,2†], **Xiaoran Zhang**[1,2†], **Ziwei Shi**[1,2], **Yu Zang**[1,2*], **Weiquan Liu**[3] and **Cheng Wang**[1,2]
[1]Fujian Key Laboratory of Urban Intelligent Sensing and Computing, School of Informatics, Xiamen University, China
[2]Key Laboratory of Multimedia Trusted Perception and Efficient Computing, Xiamen University, China
[3]Jimei University, China
{lemonxwj, zxrfaith}@gmail.com, shizw1995@stu.xmu.edu.cn, zangyu7@126.com, wqliu@jmu.edu.cn, cwang@xmu.edu.cn

Abstract

Cross-view geo-localization (CVGL) aims at localizing a ground-level query by retrieving its corresponding match from a database of geo-tagged satellite images. Existing multi-modal CVGL methods lack a structured design in the fusion stage, limiting their ability to fully exploit the information from multiple modalities. To overcome this limitation, we propose a Structural-Analysis-Based fusion principle that guides the design of network architecture. Following this principle, we present Decoupled Query Fusion (DQF), a novel fusion module that decouples feature interactions through role-specific learnable queries. These learnable queries aggregate features into distinct slots. Specifically, modality-specific queries capture unique information from each modality, while cross-modal queries extract redundant and synergistic information between different modalities. To better discriminate positive samples at varying geographic distances, we further propose Geo-aware Circle Loss, which adaptively weights supervision signal based on the geographical distance between samples and anchors. Extensive experiments on large-scale benchmarks demonstrate that our method significantly outperforms state-of-the-art approaches. Comprehensive ablation studies further validate the effectiveness of each component. Our code and models will be made publicly available.

1 Introduction

Cross-View Geo-Localization (CVGL) aims to retrieve the corresponding location from large-scale aerial or remote sensing reference databases using ground-level observations as queries. Even when the Global Positioning System (GPS) is interfered with, CVGL can still ensure the stable operation of the navigation and positioning system. Therefore, an increasing number of researchers have begun to pay attention to and explore ways to enhance the application effectiveness of CVGL. However, the large viewpoint gap, scale variations, and heterogeneous appearances make cross-domain alignment challenging.

Current studies have primarily concentrated on uni-modal CVGL approaches using either cameras [Shugaev *et al.*, 2024; Mi *et al.*, 2024; Zhu *et al.*, 2022; Shi *et al.*, 2020b; Wang *et al.*, 2022] or LiDAR [Shi *et al.*, 2025]. However, single-modality data often lack sufficient discriminative features to distinguish between correct and incorrect matches. Firstly, the data from the ground-level perspective and the data from the aerial-level perspective have only a small amount of information overlapping. Secondly, there are numerous repetitive elements in a city area, making it difficult to distinguish the remote sensing images of these locations. Thirdly, when the sensor is disturbed, e.g., by sudden changes in light intensity, the performance of the system will be further compromised. To address this, recent work has explored multi-modal fusion of street-view imagery and LiDAR [Wang *et al.*, 2025], showing promise in improving the performance of CVGL. However, current methods mainly rely on the network to implicitly learn the relationships between different modalities during the multi-modal feature fusion stage, lacking structural constraints. This limits the full utilization of multi-modal information, degrading the performance of the model. Additionally, existing methods treat all positive samples uniformly in the supervision stage. However, since positive samples are typically obtained from within a certain neighborhood range of the anchor point, the distances between different positive samples and the anchor vary considerably. As the distance increases, the feature similarity between the anchor and positive samples gradually decreases. If these variations are not distinguished during the training stage, the network may receive conflicting supervision signals, which can degrade the network's performance.

To overcome this limitation, we analyze the fusion process of multi-modal data from the perspective of information theory [Williams and Beer, 2010], and proposed a principle for guiding the design of multi-modal CVGL networks: *In the design of multi-modal fusion networks, structured constraints should be imposed to explicitly decouple the fusion process, thereby avoiding unstructured mixing of information from dif-*

*Corresponding author.

ferent sources. Driven by this principle, we propose Decoupled Query Fusion (DQF). DQF uses role-specific (modality-specific and cross-modal) learnable queries to organize the fusion process of multi-modal features, ensuring that the final fusion representation can more fully retain the information of both. Furthermore, to ensure that the learned metric embedding space better reflects real-world spatial relationships, we propose a Geo-aware Circle Loss. Our loss function will assign adaptive weights to the losses of positive sample pairs based on the geographical distance between them. Specifically, sample pairs with a closer geographical distance will be assigned a larger weight. This operation ensures that the descriptors generated by the network can better reflect the continuity of the geographical space, making the descriptors more discriminative.

The main contributions of this work are:

- We propose a Structural-Analysis-Based information fusion principle to guide the design of the network framework for multi-modal CVGL. This design strategy enables the network to have greater interpretability and allows it to better integrate multi-modal information.
- We present Decoupled Query Fusion, a novel fusion module that disentangles feature interactions through role-specialized learnable queries. Additionally, we propose a Geo-aware Circle Loss, which incorporates geographic priors into supervisory signals, enabling the model to better preserve spatial relationships.
- Extensive experiments demonstrate the superiority of our approach. On LiRSI-KITTI360 and KITTI360-AG, our method achieves **75.4%** and **44.9%** relative improvements over the state-of-the-art methods in Recall@1, respectively. Furthermore, comprehensive ablation studies validate the effectiveness and necessity of each proposed component.

2 Related Work

2.1 Uni-Modal Cross-View Geo-Localization

CVM-Net [Hu *et al.*, 2018], as an early exploratory work, adopts a Siamese network architecture and employs NetVLAD [Arandjelovic *et al.*, 2016] for feature aggregation, improving the accuracy of cross-view retrieval through a ranking loss function. SAFA [Shi *et al.*, 2019] reduces the geometric domain gap between ground-level panoramas and satellite images by performing polar coordinate transformation on satellite images, subsequently enhancing the discriminability of feature through spatial attention mechanisms. [Shi *et al.*, 2020a] is the first to jointly train the tasks of position estimation and direction prediction for CVGL. [Regmi and Shah, 2019] employs a conditional generative adversarial network to transform ground-level panoramas into the aerial-level domain, reducing the domain differences with satellite images. [Toker *et al.*, 2021] proposes a multi-task framework to jointly optimize the view synthesis from satellite images to street-view images and the CVGL in an end-to-end manner. TransGeo [Zhu *et al.*, 2022] adopts a Transformer architecture and models cross-domain correspondence through explicit position encoding. EgoTR [Yang *et al.*, 2021] enhances the generalization ability of representations through self-attention and cross-attention mechanisms. DWDR [Wang *et al.*, 2024] reduces the correlation between channels of embedded features through regularization. ConGeo [Mi *et al.*, 2024] leverages contrastive learning to align features from ground-level observations at the same location but with different orientation and fields of view (FoVs), enabling a single model to achieve orientation invariance and FoVs robustness. L2RSI [Shi *et al.*, 2025] significantly reduces the domain gap between LiDAR point clouds and satellite images by aligning them to a unified semantic space, achieving high-precision cross-scene localization through a particle-based estimation method.

2.2 Multi-Modal Cross-View Geo-Localization

Multi-modal CVGL aims to jointly leverage information from multiple ground-level sensors to alleviate the problem of insufficient discriminative information from single-modal observations in complex environments. However, compared to single-modal CVGL, multi-modal CVGL is still in the exploration stage. AGPlace [Wang *et al.*, 2025] fuse ground-level cameras and LiDAR for CVGL, introducing Neural Ordinary Differential Equations to model the continuous cross-modal alignment process, providing pioneering exploration for the research of multi-modal CVGL.

2.3 Supervision for Cross-View Geo-Localization

Early work [Vo and Hays, 2016] explored the application of metric learning in CVGL, arguing that ranking constraints are crucial for the stability of retrieval. Based on this, CVM-Net [Hu *et al.*, 2018] proposed a weighted soft-margin ranking loss to improve the stability of training. ConGeo [Mi *et al.*, 2024] introduced ground-to-ground constraints in addition to ground-to-aerial constraints. Specifically, it employs contrastive learning to align features of ground-level images from the same location but with different orientations and FoVs, thereby enhancing the robustness of model under ground-level viewpoint perturbations. ArcGeo [Shugaev *et al.*, 2024] improved the retrieval accuracy under limited FoVs conditions by introducing cosine margin into the ArcFace loss. In multi-modal CVGL, AGPlace [Wang *et al.*, 2025] proposed a multi-domain alignment loss, enabling the network to learn consistent representations between ground-level fusion features and aerial-level features, as well as between individual modality features and aerial-level features, thereby improving overall model performance.

3 Methodology

3.1 Problem Formulation

Previous research has shown that integrating LiDAR point clouds and remote sensing images into a semantic domain can effectively reduce the modal differences between them, thereby enhancing the generalization ability of the model. Specifically, we follow the semantic data processing procedure in [Shi *et al.*, 2025] to obtain the semantic BEV view S_{bev} of the LiDAR point clouds and the semantic view S_a of the remote sensing image.

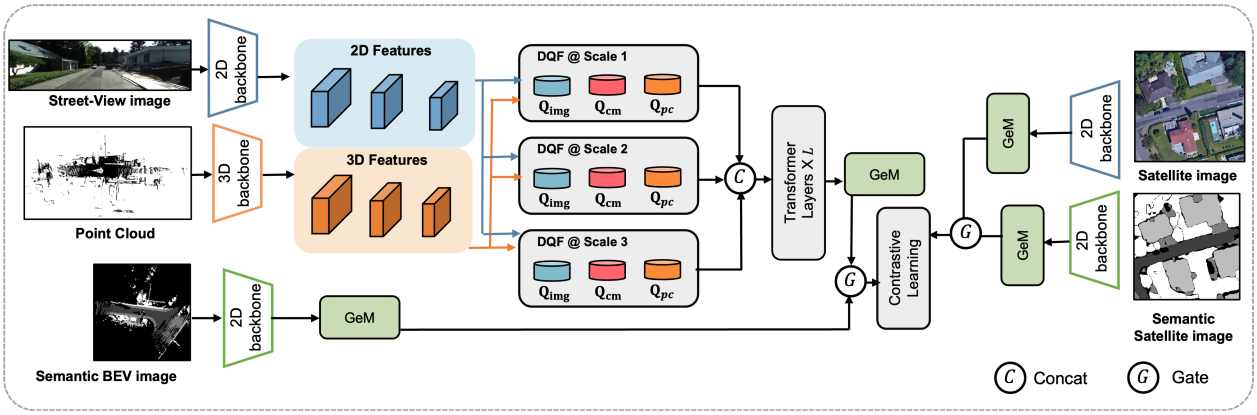


Figure 1: Overview of the proposed framework. The architecture consists of parallel ground-level and aerial-level streams. In the ground-level stream, we employ the DQF module to integrate multi-scale features from street-view images and LiDAR point clouds. In the final stage of the two streams, we employ an adaptive gating module to integrate the global embedding vectors from the original and semantic views, generating a global embedding vector for cross-view retrieval.

Formally, let $\mathcal{G} = \{I_g, P_g, S_{\text{bev}}\}$ denote the multi-modal ground-level input, where I_g and P_g represent the street-view image and raw point clouds, respectively. Let $\mathcal{A} = \{I_a, S_a\}$ represent the aerial-level input consisting of the satellite image and its semantic view. Our objective is to learn two parameterized mapping functions:

$$z_g = f_g(\mathcal{G}; \theta_g), \quad z_a = f_a(\mathcal{A}; \theta_a), \tag{1}$$

where $z_g, z_a \in \mathbb{R}^D$ are D-dimensional global embedding vectors for ground-level input and aerial-level input respectively, and θ_g, θ_a denote the learnable parameters of the ground-level and aerial-level stream networks, respectively.

The learning objective is to optimize the network parameters such that the distance between the global embedding vectors of matching pairs is minimized:

$$(\theta_g^*, \theta_a^*) = \arg\min_{\theta_g, \theta_a} \mathbb{E}_{(\mathcal{G},\mathcal{A})\sim\mathcal{D}^+} \left[d(f_g(\mathcal{G}; \theta_g), f_a(\mathcal{A}; \theta_a)) \right], \tag{2}$$

where $\mathcal{D}^+$ represents the set of positive ground-aerial pairs, $d(\cdot, \cdot)$ is a distance metric.

3.2 Architecture Overview

As shown in Figure 1, our network framework takes ground-level data (street-view images, LiDAR point clouds, semantic BEV views) and aerial-level data (remote sensing images and their semantic views) as inputs, and learns discriminative global embedding vectors through contrastive learning to achieve robust cross-view retrieval.

Ground-Level Stream. In the ground-level stream, we employ different backbones to extract multi-scale features from street-view images and raw point clouds, respectively. To better integrate multi-modal information, we employ the DQF module to fuse features from both modalities at each scale. Subsequently, these multi-scale fused features are aggregated through L transformer layers and pooled via Generalized Mean (GeM) pooling to obtain the global fusion embedding vectors z_{go} for the street-view images and the raw point clouds. Concurrently, semantic features are extracted through a separate backbone and pooled using GeM pooling, resulting in the global embedding vector z_{gs} of the semantic BEV view. Finally, we fuse z_{go} and z_{gs} through an adaptive gated module to obtain the ground-level global embedding vector z_g.

Aerial-Level Stream. The aerial-level stream follows a symmetric encoding and integration pipeline. Specifically, we employ different backbones with GeM pooling to process the original view I_a and semantic view S_a of the remote sensing images independently, yielding the global embedding vectors z_{ao} and z_{as}, respectively. Similarly, we use an adaptive gated module to fuse z_{ao} and z_{as} to obtain the aerial-level global embedding vector z_a.

3.3 Multi-Branch Feature Encoding

Street-View Image Encoder. For the street-view image I_g, we employ a 2D backbone $\mathcal{B}_{img}$ to extract multi-scale features:

$$\{F_{img}^i\}_{i=1}^3 = \mathcal{B}_{img}(I_g), \tag{3}$$

where F_{img}^i denotes the feature map of street-view image at scale i.

LiDAR Point Clouds Encoder. For the raw point clouds P_g, we employ a 3D backbone $\mathcal{B}_{pc}$ to extract multi-scale features:

$$\{F_{pc}^i\}_{i=1}^3 = \mathcal{B}_{pc}(P_g), \tag{4}$$

where F_{pc}^i denotes the feature vectors of LiDAR point clouds at scale i.

Semantic BEV Encoder. We extract the features of the semantic BEV view S_{bev} using a 2D backbone, and then obtain its global embedding vector z_{gs} through GeM pooling:

$$z_{gs} = \text{GeM}(\mathcal{B}_{bev}(S_{\text{bev}})). \tag{5}$$

Aerial-Level Encoders. The aerial-level stream encodes I_a and S_a through dual backbones and GeM pooling:

$$z_{ao} = \text{GeM}\left(\mathcal{B}_a^{rgb}(I_a)\right), \quad z_{as} = \text{GeM}(\mathcal{B}_a^{sem}(S_a)). \tag{6}$$

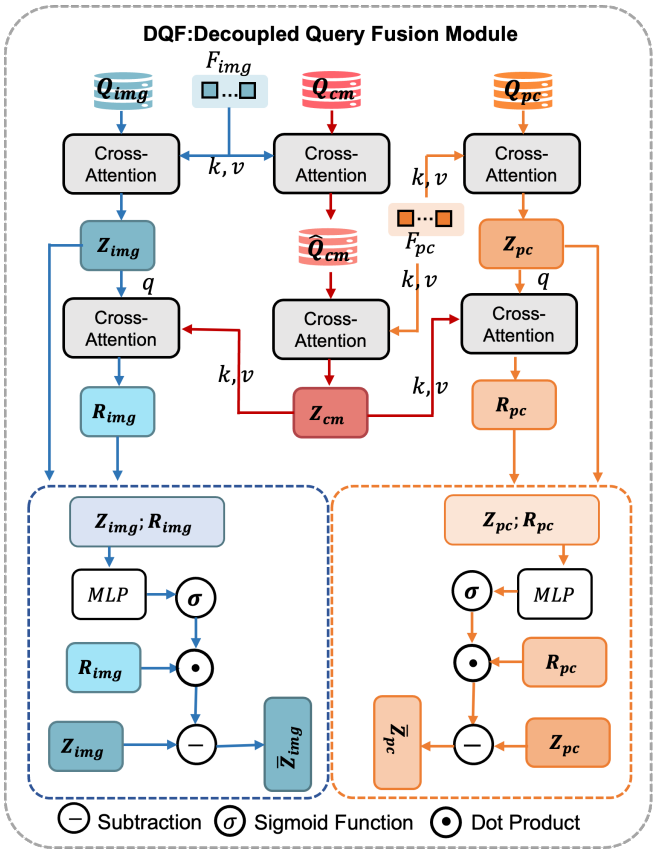


Figure 2: Illustration of the Decoupled Query Fusion module.

3.4 Decoupled Query Fusion

Structural-Analysis-Based Information Fusion Principle

From an information-theoretic perspective, the information contained in the global embedding vectors generated by multi-modal data at the ground-level can be decomposed into three parts:

- *Unique information*: Information that exists only in street-view images or LiDAR point clouds, e.g., color information in images and 3D geometric structure of point clouds.
- *Redundant information*: Information that exists simultaneously in street-view images and LiDAR point clouds, such as the outlines of buildings and the boundaries of roads.
- *Synergistic information*: Information that can only be obtained through the interaction of street-view images and LiDAR point clouds, e.g., complete environment understanding in obstructed scenarios.

Obviously, in order to fully utilize the information of multi-modal data, the final fused representation should retain the above three parts of information as much as possible. However, if the network framework lacks structural constraints, the model is prone to confuse the information of each part during the training process. This will cause the network to overly rely on one part of the information, resulting in the loss of other information. Therefore, we propose a Structural-Analysis-Based information fusion principle: *In the design of multi-modal fusion networks, structured constraints should be imposed to explicitly decouple the fusion process, thereby avoiding unstructured mixing of information from different sources.*

Decoupled Query Fusion Module

Motivated by the Structural-Analysis-Based information fusion principle, we propose the Decoupled Query Fusion module, as shown in Figure 2. DQF separately extracts modality-specific and cross-modal information through role-specific learnable queries.

For the i-th scale, the encoded token sequences of street-view images and point clouds are defined as $X^i_{img} \in \mathbb{R}^{N_{img} \times d}$ and $X^i_{pc} \in \mathbb{R}^{N_{pc} \times d}$, respectively. DQF maintains three learnable query sets:

$$Q^i_{img} \in \mathbb{R}^{K_u \times d}, \quad Q^i_{pc} \in \mathbb{R}^{K_u \times d}, \quad Q^i_{cm} \in \mathbb{R}^{K_c \times d}, \tag{7}$$

where K_u and K_c denote the numbers of modality-specific and cross-modal queries.

Modality-Specific Slot. Modality-specific queries extract information from their corresponding modalities through intra-modal interactions:

$$Z^i_{img} = \mathrm{CA}(Q^i_{img}, X^i_{img}, X^i_{img}), \tag{8}$$

$$Z^i_{pc} = \mathrm{CA}(Q^i_{pc}, X^i_{pc}, X^i_{pc}), \tag{9}$$

where $\mathrm{CA}(\cdot)$ denotes the multi-head cross-attention taking (q, k, v) as query, key and value, respectively.

Cross-Modal Slot. Cross-modal queries capture redundant and synergistic information by sequentially interacting with both modalities:

$$\tilde{Q}^i_{cm} = \mathrm{CA}(Q^i_{cm}, X^i_{img}, X^i_{img}), \tag{10}$$

$$Z^i_{cm} = \mathrm{CA}(\tilde{Q}^i_{cm}, X^i_{pc}, X^i_{pc}). \tag{11}$$

Redundancy-Aware Gated Suppression. Currently, redundant information exists simultaneously in cross-modal queries and modality-specific queries. This would consume the limited representation capacity and prevent the final fusion representation from fully leveraging the information of multi-modal data. Therefore, we introduce the Redundancy-aware Gated Suppression mechanism to reduce the redundant information in modality-specific queries, thereby ensuring that the redundant information is concentrated as much as possible in cross-modal queries:

$$R^i_{img} = \mathrm{CA}(Z^i_{img}, Z^i_{cm}, Z^i_{cm}), \tag{12}$$

$$\bar{Z}^i_{img} = Z^i_{img} - \sigma\big(\mathrm{MLP}(\mathrm{Concat}(Z^i_{img}, R^i_{img}))\big) \odot R^i_{img}, \tag{13}$$

$$R^i_{pc} = \mathrm{CA}(Z^i_{pc}, Z^i_{cm}, Z^i_{cm}), \tag{14}$$

$$\bar{Z}^i_{pc} = Z^i_{pc} - \sigma\big(\mathrm{MLP}(\mathrm{Concat}(Z^i_{pc}, R^i_{pc}))\big) \odot R^i_{pc}, \tag{15}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathrm{MLP}(\cdot)$ is a Multi-Layer Perceptron and $\odot$ denotes element-wise multiplication.

Hybrid Slot Aggregation. The fused representation at scale i is obtained by concatenating the refined modality-specific queries and the cross-modal queries:

$$U^i = \mathrm{Concat}(\bar{Z}^i_{img},\, Z^i_{cm},\, \bar{Z}^i_{pc}). \tag{16}$$

3.5 Multi-Scale Feature Aggregation

The multi-scale fusion features obtained from Sec.3.4 will be concatenated together. Subsequently, we use the L-layer multi-head self-attention (MHSA) module to integrate the high-resolution detailed features and the low-resolution semantic features:

$$H = \mathrm{MHSA}^{(L)}(\mathrm{Concat}(U^1, U^2, U^3)). \tag{17}$$

Subsequently, we aggregate the features into a global embedding vector through GeM pooling:

$$z_{go} = \mathrm{GeM}(H). \tag{18}$$

3.6 Semantic-Aware Gated Integration

To integrate the global embedding vectors of the original view and the semantic view, we employ an adaptive gated module. This mechanism enables the model to dynamically adjust the weight ratios between the two views according to the feature distributions of different samples, thereby achieving more flexible feature representation.

Ground-Level View. Given the z_{go} and z_{gs}, we compute a channel-wise gating mechanism to obtain the fused global embedding vectors for the ground-level views:

$$g_g = \sigma(\mathrm{MLP}_g(\mathrm{Concat}(z_{go},\, z_{gs}))), \tag{19}$$

$$z_g = g_g \odot z_{go} + (1 - g_g) \odot z_{gs}. \tag{20}$$

Aerial-Level View. Similarly, given the z_{ao} and z_{as}, we apply a channel-wise gating mechanism to obtain the fused aerial-level global embedding vector:

$$g_a = \sigma(\mathrm{MLP}_a(\mathrm{Concat}(z_{ao},\, z_{as}))), \tag{21}$$

$$z_a = g_a \odot z_{ao} + (1 - g_a) \odot z_{as}. \tag{22}$$

3.7 Loss Function

To ensure robust cross-view alignment and effective feature decoupling, we propose a composite objective function consisting of a geometry-aware contrastive loss and a multi-scale regularization term.

Geo-Aware Circle Loss

Standard metric learning losses treat all positive pairs uniformly. In the CVGL task, the positive samples are usually selected from the specific-sized spatial neighborhood around the anchor point. Therefore, the distances between different positive samples and the anchor point are different. If these differences are not distinguished during the training process, the network may receive conflicting supervisory signals, which will lead to a decline in its performance. To address this, we propose a Geo-aware Circle Loss, which calibrates the supervisory signal based on the spatial distance between the positive sample pairs.

We use a Gaussian kernel function to assign corresponding weights to positive samples of different distances:

$$w_{ij} = \exp\left(-\frac{\|p^i_g - p^j_a\|^2}{2h^2}\right), \tag{23}$$

where $p^i_g, p^j_a \in \mathbb{R}^2$ represent the Universal Transverse Mercator (UTM) coordinates of the ground-level and aerial-level samples, and h denotes the bandwidth parameter that controls the decay rate.

We adapt the Circle Loss [Sun *et al.*, 2020] by integrating w_{ij} into the gradient re-weighting mechanism. The geometry-aware coefficients for positive pairs $\hat{\alpha}^j_p$ and negative pairs α^k_n are computed as:

$$\hat{\alpha}^j_p = w_{ij} \cdot \mathrm{ReLU}(O_p - s_{ij}), \quad \alpha^k_n = \mathrm{ReLU}(s_{ik} - O_n), \tag{24}$$

where s represents the similarity score between sample pairs, O_p and O_n denote the target boundaries for positive and negative scores, respectively. Consequently, the final geometry-aware circle loss is formulated as:

$$\mathcal{L}_{geo} = \log\left[1 + \sum_{k\in\mathcal{N}} e^{\gamma\alpha^k_n(s_{ik}-\Delta_n)} \sum_{j\in\mathcal{P}} e^{-\gamma\hat{\alpha}^j_p(s_{ij}-\Delta_p)}\right], \tag{25}$$

where $\mathcal{P}$ and $\mathcal{N}$ denote the sets of positive and negative samples for the anchor sample i, respectively, γ is the scaling factor that controls the penalty strength, and Δ_p, Δ_n are the margins. By incorporating the geometric weight w_{ij} into $\hat{\alpha}^j_p$, the loss function implements differentiated supervision on positive samples of different spatial distances. Specifically, the contribution of distant matches is weakened, while the contribution of close matches is enhanced. This mechanism guides the model to learn more fine-grained spatial correspondences.

Multi-Scale Decoupling Regularization

To ensure the structural independence between modality-specific and cross-modal features during multi-modal fusion, we impose an orthogonality constraint. Specifically, for the i-th scale,

$$\mathcal{L}^i_{orth} = \left\|\bar{Z}^i_{img}(Z^i_{cm})^\top\right\|^2_F + \left\|\bar{Z}^i_{pc}(Z^i_{cm})^\top\right\|^2_F, \tag{26}$$

where $\|\cdot\|_F$ represents the Frobenius norm. The multi-scale regularization loss is formulated as:

$$\mathcal{L}_{DQF} = \sum_{i=1}^{3} \mathcal{L}^i_{orth}. \tag{27}$$

Overall Loss Function

The complete training objective is formulated as a weighted combination of the geometry-aware contrastive loss and the decoupling regularization:

$$\mathcal{L}_{total} = \mathcal{L}_{geo} + \lambda\mathcal{L}_{DQF}, \tag{28}$$

where the hyperparameter λ controls the trade-off between cross-view alignment accuracy and feature disentanglement quality.

Figure 3: Visualization of the extended LiRSI-KITTI360 dataset. Left: Vehicle trajectories of all sequences overlaid on satellite imagery, with training sequences shown in blue and the test sequence in red. Right: The satellite image covering the test area, illustrating the geographic region used for evaluation.

4 Experiment

4.1 Datasets

We evaluate our proposed approach on LiRSI-KITTI360 [Shi *et al.*, 2025] and KITTI360-AG [Wang *et al.*, 2025] benchmark datasets.

KITTI360-AG. KITTI360-AG is a multi-modal cross-view geo-localization dataset built upon KITTI-360 [Liao *et al.*, 2022]. It correlates the ground-level modality with spatially aligned aerial imagery. Following [Wang *et al.*, 2025], we conduct experiments on 7 sequences (00, 02, 03, 04, 05, 06, 09), using the first 85% of frames for training and the remaining 15% for testing. Following the protocol in LiRSI [Shi *et al.*, 2025], we construct dense point cloud submaps at 5-meter intervals along the vehicle trajectory.

LiRSI-KITTI360. LiRSI-KITTI360 is a large-scale dataset derived from KITTI-360, designed for LiDAR-based place recognition using remote sensing imagery. To adapt this dataset for our research, we expand each LiDAR frame with corresponding street-view image. To evaluate the model's generalization capability, we set sequence 02 as the test set while using the remaining 8 sequences (00, 03, 04, 05, 06, 07, 09, 10) for training, as shown in Figure 3. To prevent data leakage, we remove the training samples located in the test remote sensing area.

4.2 Implementation Details

Our method is implemented in PyTorch. For feature extraction, we adopt MinkLoc3D [Komorowski, 2021] with a voxel size of 0.5 m as the 3D backbone and an ImageNet-pretrained ResNet18 [He *et al.*, 2016] as the 2D backbone. The network is trained using the Adam optimizer for 100 epochs with a batch size of 16. We set the initial learning rates of 1×10^{-4} for the 3D branch and 1×10^{-3} for the 2D branch, both scheduled with CosineAnnealingWarmRestarts including a 5-epoch linear warmup. Positive pairs are defined using a spatial threshold of 30 m. The circle loss adopts a scale factor $\gamma = 80$ and a margin $m = 0.4$. For the proposed module, we set $L = 2$, $K_u = 64$, $K_c = 32$, and $h = 20$.

Method	R@1	R@5	R@10
Sample4Geo [Deuser *et al.*, 2023]	27.6	50.1	58.5
L2RSI [Shi *et al.*, 2025]	<u>43.0</u>	<u>64.2</u>	<u>71.8</u>
MinkLoc++ [Komorowski *et al.*, 2021]	26.8	48.2	55.1
AdaFusion [Lai *et al.*, 2022]	27.9	50.6	60.4
MSSPlace [Melekhin *et al.*, 2025]	25.2	47.3	54.7
LCPR [Zhou *et al.*, 2023]	28.1	51.9	60.2
UMF [García-Hernández *et al.*, 2024]	27.4	50.8	61.5
AGPlace [Wang *et al.*, 2025]	29.9	53.0	63.4
Ours	**62.3**	**77.3**	**86.0**

Table 1: Quantitative comparison on the KITTI360-AG dataset. We report Recall@K (%) under the orientation-free setting. Best results are highlighted in **bold** and the second-best results are <u>underlined</u>.

4.3 Experimental Results

Evaluation Metrics

Following [Wang *et al.*, 2025], we adopt $Recall@K$ to evaluate the retrieval performance, where $K \in \{1, 5, 10\}$. $Recall@K$ is defined as the proportion of queries that contain at least one ground-truth match among the top-K retrieved results. Specifically, for each ground-level query, we rank all candidates in the remote sensing image database by their feature similarity to the query. A retrieval is deemed successful if any of the top-K candidates lies within a spatial distance threshold from the query location. This metric directly reflects the usability of the model in practical application scenarios.

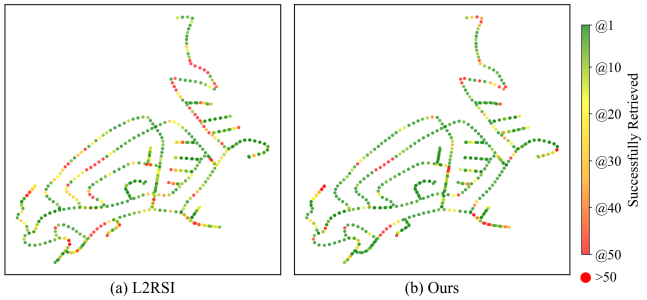
Main Results

We mainly compare our method with multi-modal methods, including MinkLoc++ [Komorowski *et al.*, 2021], AdaFusion [Lai *et al.*, 2022], MSSPlace [Melekhin *et al.*, 2025], LCPR [Zhou *et al.*, 2023], UMF [García-Hernández *et al.*, 2024], and AGPlace [Wang *et al.*, 2025]. Additionally, among the uni-modal CVGL methods, we also select the representative Sample4Geo [Deuser *et al.*, 2023] and L2RSI [Shi *et al.*, 2025] for comparison. For the experiments with Sample4Geo, we use street-view images as the ground-level input.

Results on KITTI360-AG. We validate the fitting and generalization capability of our approach on KITTI360-AG benchmark. On this dataset, we follow the experimental protocol from [Wang *et al.*, 2025] and operate without orientation information. As shown in Table 1, our method outperforms existing state-of-the-art approaches across all evaluation metrics. Notably, our method achieves 62.3% on Recall@1, achieving a 44.9% relative improvement over L2RSI. This substantial improvement demonstrates our method can learn more discriminative global feature representations, highlighting its strong potential for real-world deployment.

Results on LiRSI-KITTI360. To further validate the generalization capability of our method, we conduct additional cross-scene experiments on the LiRSI-KITTI360 dataset. Specifically, we evaluate two experimental settings:

1. **Orientation-Free**: The model does not utilize any orientation prior information, relying solely on visual and geometric features for matching.



(a) L2RSI (b) Ours

Figure 4: Visualization of retrieval performance along the testing trajectory. Points closer to green represent locations where the correct match appears in the top-ranked retrieval results, requiring fewer candidates. Conversely, points closer to red indicate locations where more retrieval results must be examined to find the correct match.

Method	Orientation-Free			Orientation-Aware ($\pm 15°$)		
	R@1	R@5	R@10	R@1	R@5	R@10
Sample4Geo [Deuser *et al.*, 2023]	2.6	5.9	10.2	11.1	20.7	28.0
L2RSI [Shi *et al.*, 2025]	<u>7.6</u>	<u>18.9</u>	<u>25.2</u>	<u>23.2</u>	<u>42.8</u>	<u>54.7</u>
MinkLoc++ [Komorowski *et al.*, 2021]	1.8	3.8	8.3	8.2	15.2	23.5
AdaFusion [Lai *et al.*, 2022]	2.3	5.2	9.4	10.6	18.5	27.2
MSSPlace [Melekhin *et al.*, 2025]	1.5	3.7	7.8	6.5	13.4	20.1
LCPR [Zhou *et al.*, 2023]	1.8	4.1	8.0	8.9	15.8	23.0
UMF [García-Hernández *et al.*, 2024]	2.1	4.5	8.6	9.8	17.3	24.4
AGPlace [Wang *et al.*, 2025]	3.1	8.7	13.9	12.6	25.1	33.8
Ours	**16.6**	**31.1**	**41.6**	**40.7**	**65.6**	**76.2**

Table 2: Quantitative comparison on the LiRSI-KITTI360 dataset. We report Recall@K (%) under both orientation-free and orientation-aware ($\pm 15°$) settings. Best results are highlighted in **bold** and the second-best results are <u>underlined</u>.

2. **Orientation-Aware**: The model incorporates orientation information to guide the retrieval process. To simulate the measurement errors of sensors (e.g., magnetometer) in real-world scenarios, we inject random perturbations within $\pm 15°$ into the ground-truth orientation data.

As shown in Table 2, our method significantly outperforms all baseline methods in the orientation-free setting, demonstrating its superior capability in extracting rotation-invariant features. In the orientation-aware setting, our method also maintains a significant advantage. These results validate the strong cross-scene generalization ability of our approach. It is worth noting that after introducing directional information, Recall@1 and Recall@10 have increased from 16.6% and 41.6% to 40.7% and 76.2%, respectively. This highlights the critical role of orientation information in cross-scene retrieval task. To provide intuitive insights into the retrieval performance, we visualize the per-query results along the test trajectory in Figure 4.

4.4 Ablation Studies

To validate the effectiveness of individual components in our framework, we conduct ablation studies on the LiRSI-KITTI360 dataset. All experiments are performed under the orientation-aware configuration. We investigate the contributions of two key components: the DQF module and the Geo-aware Circle Loss. The quantitative results are presented in Table 3.

Model	R@1	R@5	R@10
full	**40.7**	**65.6**	**76.2**
full w/o DQF	34.1	58.9	70.6
full w/o Geo-Constraint	33.9	59.1	71.0

Table 3: Ablation study of key components on LiRSI-KITTI360 under the orientation-aware setting. "full" represents our proposed approach. "w/o Geo-Constraint" denotes training with standard Circle Loss instead of our proposed Geo-aware Circle Loss.

DQF. To evaluate the effectiveness of the DQF module, we remove it from the complete framework. The results are shown in the second row of Table 3. Without the DQF module, Recall@1 drops from 40.7% to 34.1%. This significant decrease demonstrates the module's crucial role in enhancing feature discriminability. This also provides an additional verification of the effectiveness of our proposed Structural-Analysis-Based information fusion principle.

Geo-Aware Circle Loss. To evaluate the contribution of Geo-aware Circle Loss, we replace it with the standard Circle Loss while keeping all other components unchanged. The results are presented in the third row of Table 3. After removing the supervision weighting mechanism based on geographical distance, Recall@1 decreased by 6.8 percentage points. Compared to the standard Circle Loss which treats all positive sample pairs equally, Geo-aware Circle Loss provides more fine-grained supervision signals based on the geographical distance between samples. This enables the network to better distinguish between positive samples from different geographical locations and prevents the inappropriate alignment of samples with large spatial separations, leading to more discriminative feature representations.

5 Conclusion

In this work, we investigate multi-modal cross-view geo-localization. Specifically, we propose a Structural-Analysis-Based information fusion principle that guides the design of our DQF module. Furthermore, we introduce a Geo-aware Circle Loss that provides fine-grained supervision signals based on the geographical distance between samples. Extensive experiments demonstrate that our method significantly outperforms prior state-of-the-art approaches.

However, in the task of cross-scene generalization, our method relies heavily on orientation information to achieve strong retrieval performance, which constitutes a notable limitation. Future work will focus on extracting rotation-invariant features, aiming to maintain robust performance even in the absence of directional cues.

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China (No. 62471415, 62401225); the Natural Science Foundation of Fujian Province of China (No. 2023J01004, 2025J09046, 2024J01115); the Natural Science Foundation of Xiamen of China (No. 3502Z202472018); the Jimei University Scientific Research Start-up Funding Project (No.ZQ2024034).

Contribution Statement

† denotes that Xu Yan and Xiaoran Zhang are co-first authors and contribute equally to this work.

References

[Arandjelovic *et al.*, 2016] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[Deuser *et al.*, 2023] 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.

[García-Hernández *et al.*, 2024] 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. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3991–3998. IEEE, 2024.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[Hu *et al.*, 2018] 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 Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.

[Komorowski *et al.*, 2021] Jacek Komorowski, Monika Wysoczańska, and Tomasz Trzcinski. Minkloc++: lidar and monocular image fusion for place recognition. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[Komorowski, 2021] 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.

[Lai *et al.*, 2022] 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.

[Liao *et al.*, 2022] 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.

[Melekhin *et al.*, 2025] Alexander Melekhin, Dmitry Yudin, Ilia Petryashin, and Vitaly Bezuglyj. Mssplace: Multi-sensor place recognition with visual and text semantics. *IEEE Access*, 2025.

[Mi *et al.*, 2024] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. Congeo: Robust cross-view geo-localization across ground view variations. In *European Conference on Computer Vision*, pages 214–230. Springer, 2024.

[Regmi and Shah, 2019] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019.

[Shi *et al.*, 2019] 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.

[Shi *et al.*, 2020a] 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.

[Shi *et al.*, 2020b] 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*, volume 34, pages 11990–11997, 2020.

[Shi *et al.*, 2025] Ziwei Shi, Xiaoran Zhang, Wenjing Xu, Yan Xia, Yu Zang, Siqi Shen, and Cheng Wang. L2rsi: Cross-view lidar-based place recognition for large-scale urban scenes via remote sensing imagery. *arXiv preprint arXiv:2503.11245*, 2025.

[Shugaev *et al.*, 2024] 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.

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

[Toker *et al.*, 2021] 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.

[Vo and Hays, 2016] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*, pages 494–509. Springer, 2016.

[Wang *et al.*, 2022] Teng Wang, Shujuan Fan, Daikun Liu, and Changyin Sun. Transformer-guided convolutional neural network for cross-view geolocalization. *arXiv preprint arXiv:2204.09967*, 2022.

[Wang *et al.*, 2024] Tingyu Wang, Zhedong Zheng, Zunjie Zhu, Yaoqi Sun, Chenggang Yan, and Yi Yang. Learning cross-view geo-localization embeddings via dynamic

weighted decorrelation regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–12, 2024.

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

[Williams and Beer, 2010] Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010.

[Yang *et al.*, 2021] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with evolving transformer. *arXiv preprint arXiv:2107.00842*, 2021.

[Zhou *et al.*, 2023] 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*, 9(2):1342–1349, 2023.

[Zhu *et al.*, 2022] 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.