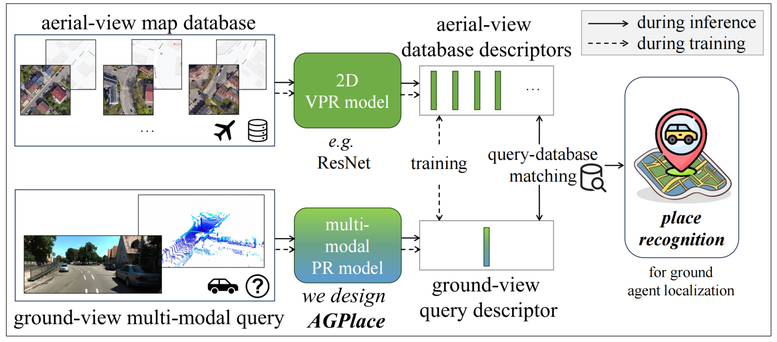
Multi-Modal Aerial-Ground Cross-View Place Recognition with Neural ODEs

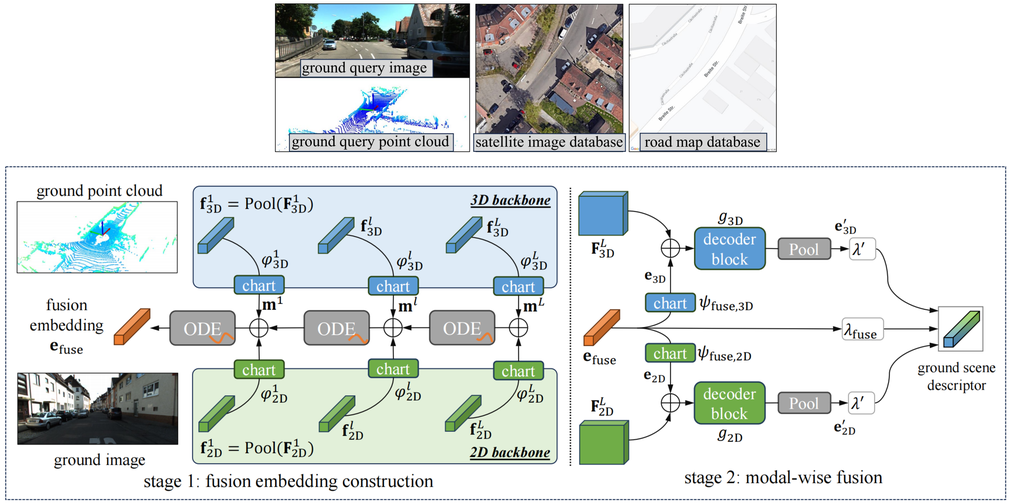
Sijie Wang*[1] Rui She*[2] Qiyu Kang[3] Siqi Li[1] Disheng Li[1] Tianyu Geng[1] Shangshu Yu[1] Wee Peng Tay[1]

[1]Nanyang Technological University [2]Beihang University [3]University of Science and Technology of China

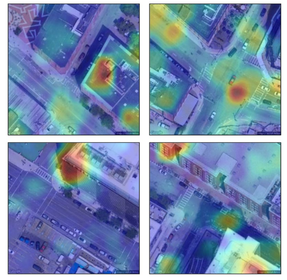
Place recognition (PR) aims at retrieving the query place from a database and plays a crucial role in various applications, including navigation, autonomous driving, and augmented reality. While previous multi-modal PR works have mainly focused on the same-view scenario in which ground-view descriptors 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 under-explored. 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.



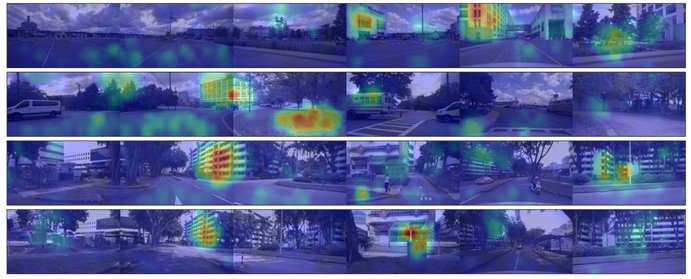
An example pipeline of multi-modal aerial-ground PR. (1) The aerial-view geo-tagged maps (e.g. aerial RGB images, road maps) act as the database. (2) The ground-view multi-modal data (images + point clouds) are the place query to be matched with the database.



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.



aerial view



ground view

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**

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	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**

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.