VXP: Voxel-Cross-Pixel Large-Scale Camera-LiDAR Place Recognition

Yun-Jin Li[1,2*], Mariia Gladkova[1,2*], Yan Xia[1,2†], Rui Wang[3], Daniel Cremers[1,2]

[1]Technical University of Munich [2]Munich Center for Machine Learning [3]Microsoft

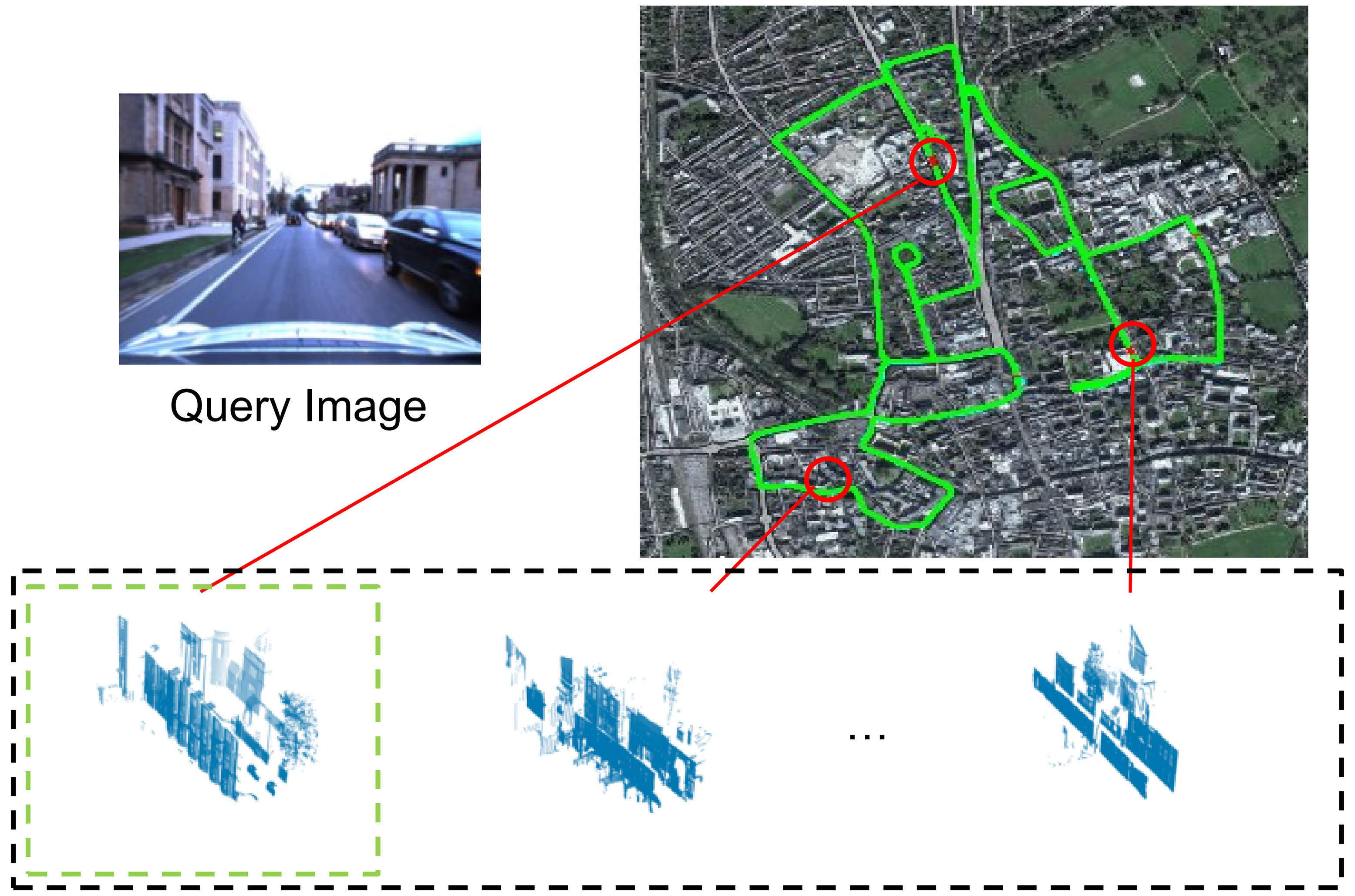
3DV 2025

Motivation

Since GNSS outages are inevitable, other onboard sensors like camera and LiDAR are handy. While fusion-based methods are common, both modalities have limitations in large-scale place recognition in terms of robustness and scalability. Cross-modal frameworks come as a flexible solution to mitigate the problem.

Cross-Modal Place Recognition (PR):

Given a query image or a LiDAR scan, retrieve the closet match of the *other* modality and its corresponding location from the database.



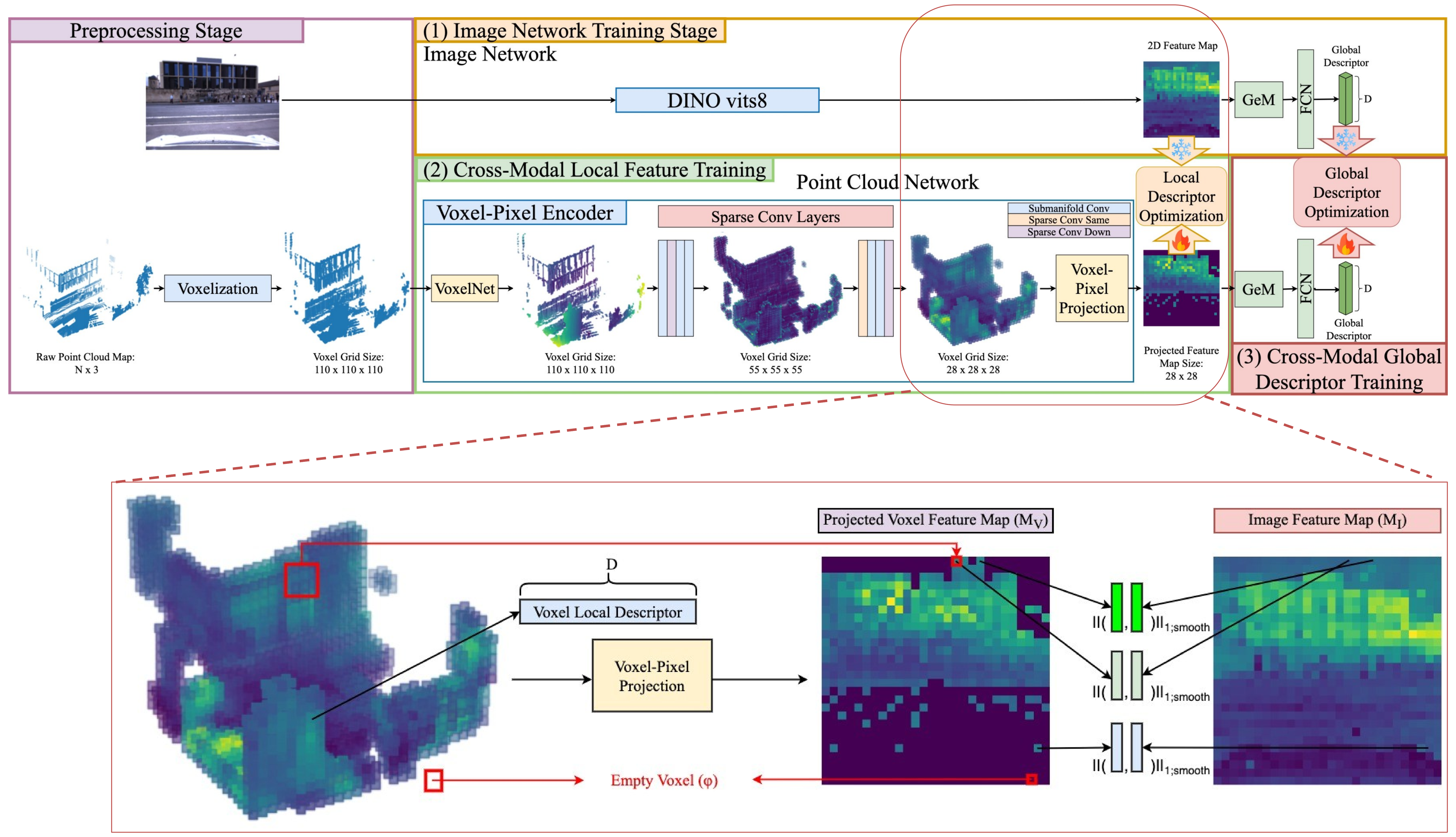
How to effectively design a shared Image-LiDAR latent space to seamlessly switch between two modalities that are completely different?

Contribution

A novel framework for cross-modal place recognition, which bridges the domain gap between images and point clouds by enforcing *local feature similarity* in a fully self-supervised manner.

Our VXP

Figure 1: Our 3-stage pipeline is designed to capture both fine-grained local details (2) and broader global context (3) for successful mapping images and LiDAR point clouds into the shared space.



$$\lambda \mathbf{p} = \mathbf{K}\left(\begin{bmatrix} vx & \cdots & 0 \\ \vdots & vy & \vdots \\ 0 & \cdots & vz \end{bmatrix} \boldsymbol{v} + \frac{1}{2}\begin{bmatrix} vx + 2x_0 \\ vy + 2y_0 \\ vz + 2z_0 \end{bmatrix}\right)$$

Voxel-Pixel projection: given calibration **K** and voxel grid size (vx, vy, xz) we obtain voxel's **v** pixel location **p**

Cross-modal local feature training: We establish the correspondences between the 2D image feature map and 3D voxel feature map using the Voxel-Pixel projection module. Rich local features from the foundation model are distilled to enrich the learned shared space, while the projected voxel features bring geometric consistency.

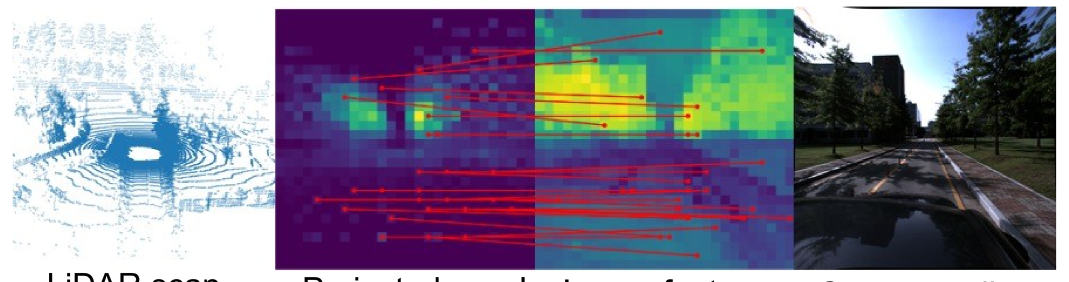


Figure 2: Local feature correspondences in local feature space after the local feature training

$$\mathcal{L}_{local} = \sum_{p \in \mathcal{M}_v} \| d_i * \mathcal{M}_v(\boldsymbol{p}) - \mathcal{M}_I(\boldsymbol{p}) \|$$

Local feature loss: For every projected voxel location **p** we enforce cross-modal consistency between voxel-based $\mathcal{M}_v$ and image-based $\mathcal{M}_I$ feature maps.

Results

Dataset	Oxford RobotCar [4]		ViViD++ [5]		KITTI [6]	
AR@1%	2D-3D	3D-2D	2D-3D	3D-2D	2D-3D	3D-2D
Cattaneo [1]	77.3	70.4	**99.6**	98.6	23.4	28.7
LC² [2]	81.2	73.8	96.0	94.6	- -	- -
LIP-Loc [3]	77.8	73.6	98.4	93.0	**40.9**	29.3
VXP (Ours)	**84.4**	**76.9**	**99.6**	**99.8**	38.6	**38.3**

Table 1: Cross-modal evaluation. Our model achieves SOTA performance across 3 large-scale datasets.

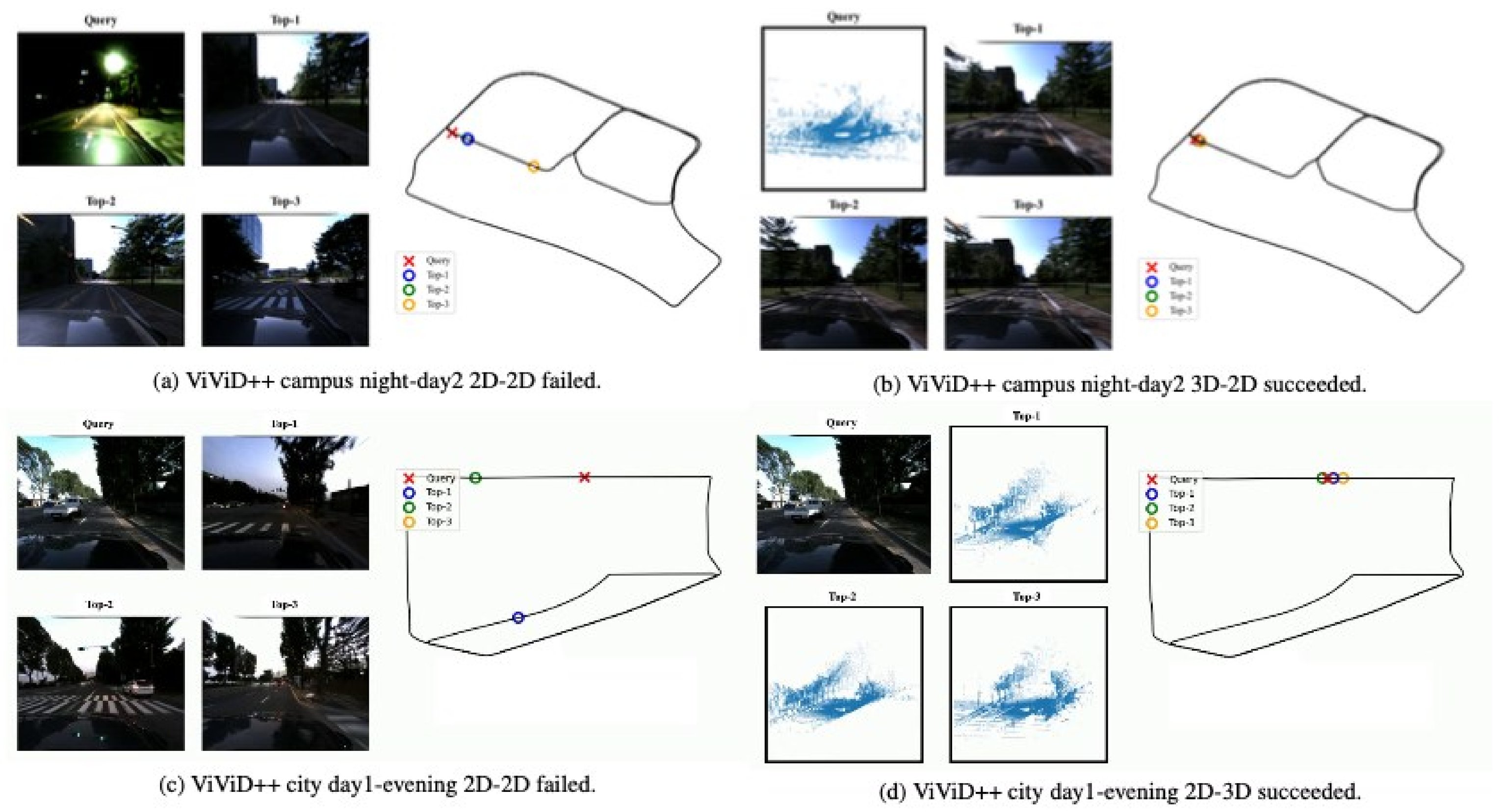
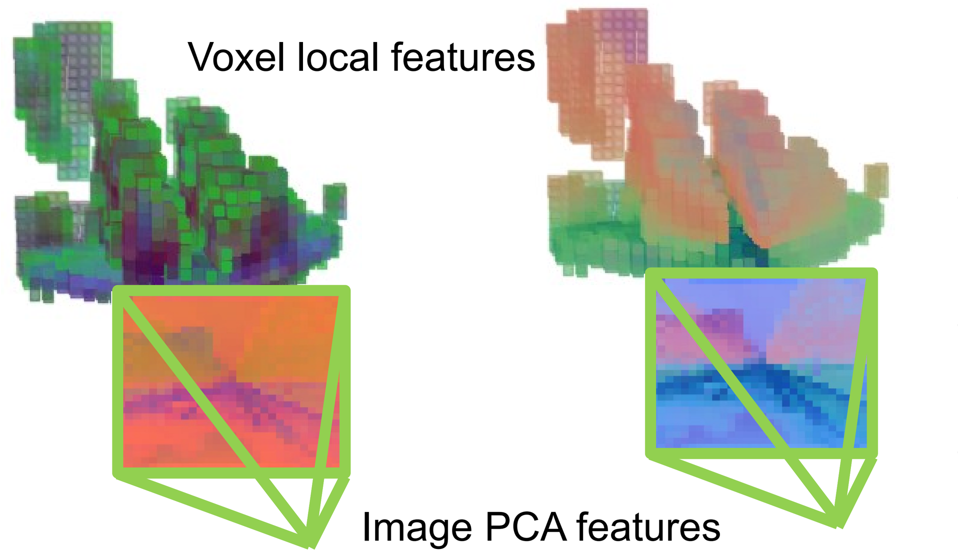


Figure 3: While uni-modal methods suffer from inherent data limitations (low lighting, repetitive geometrical structures), our cross-modal method can utilize the stronger modality.

Local Correspondences are beneficial for PR



AR@1%	2D-3D	3D-2D
Global-only	81.5	74.7
Local + Global	**84.4**	**76.9**

More

Yun-Jin Li

Mariia

References [1] Cattaneo et al., Global visual localization in lidar-maps through shared 2d-3d embedding space, ICRA 2020; [2] Lee et al., Lc2: Lidar-camera loop constraints for cross-modal place recognition, RA-L 2023; [3] Shubodh et al., Lip-loc: Lidar image pre-training for cross-modal localization, WACV-W 2024; [4] Maddern et al., 1 year, 1000 km: The oxford robotcar dataset, IJRR 2017; [5] Lee et al., Vivid++: Vision for visibility dataset, RA-L 2022; [6] Geiger et al., Are we ready for autonomous driving? the kitti vision benchmark suite, CVPR 2012