CSE 4/573 Modern CV Topics III:

Visual Place Recognition

Instructor: Zhipeng Zhao

Spatial AI & Robotics Lab

Department of Computer Science and Engineering

SAIR

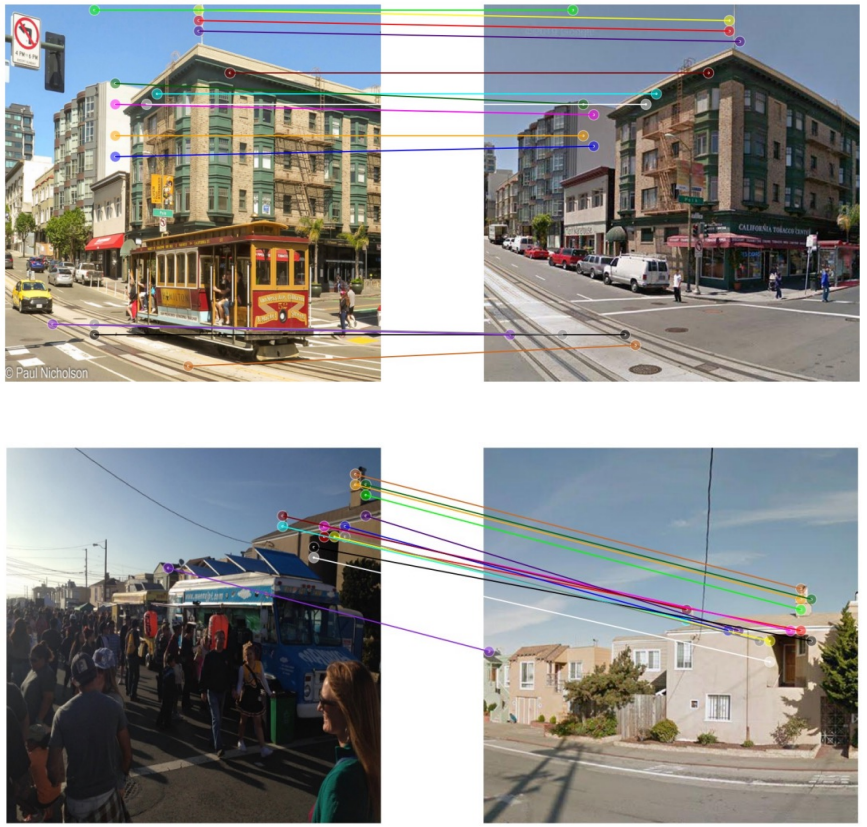
Spatial AI & Robotics Lab

CONTENTS ›

/01 Introduction

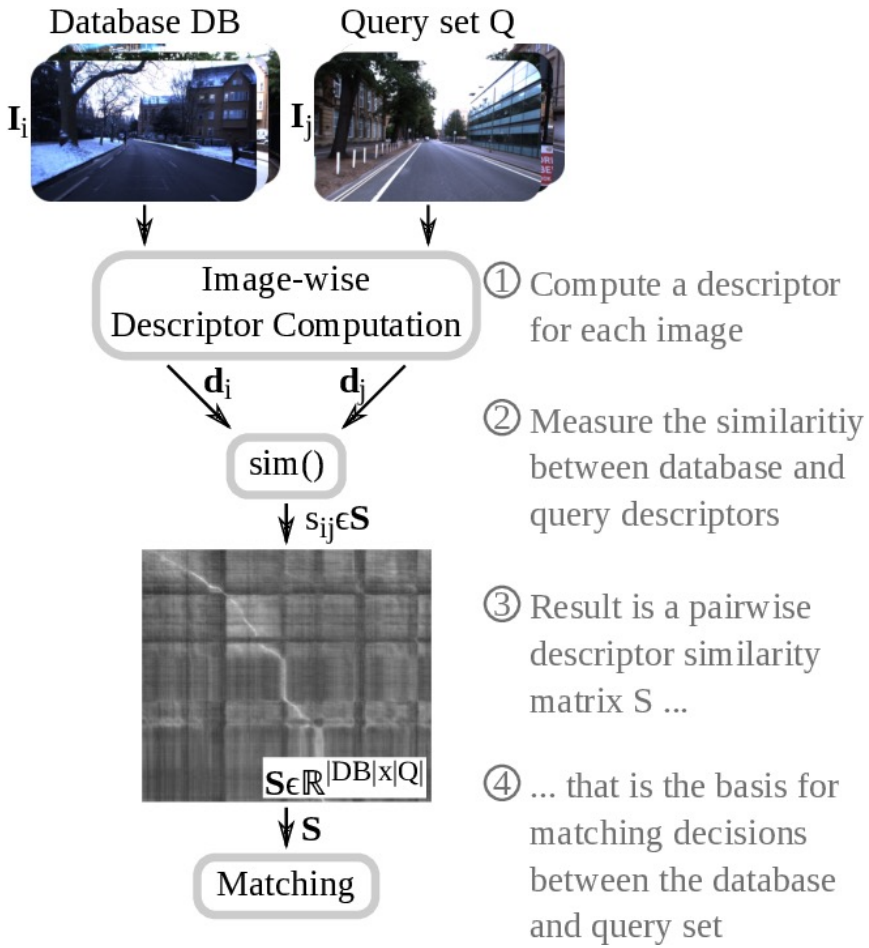
What is Visual Place Recognition(VPR)?

- **Task**: recognize a previously visited location based on visual information (CV).
- **Robotics**: determine whether a robot is in a location it has encountered before.
- **Same place**: visual overlap or location
- **Application**: autonomous vehicles, augmented reality



What is Visual Place Recognition(VPR)?

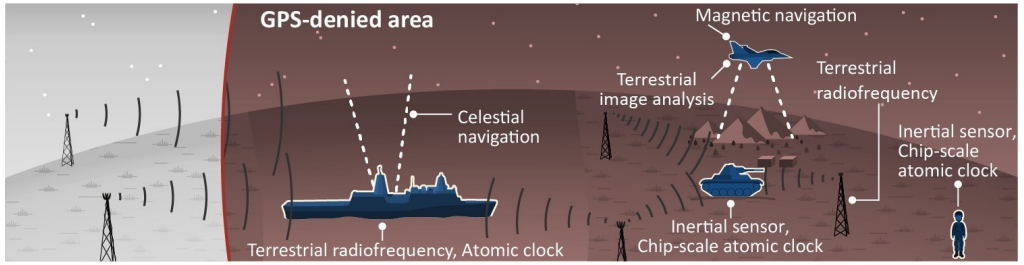
- **Task**: recognize a previously visited location based on visual information (CV).
- **Formulation:** image retrieval
  - **Database:** a reference set composed of images of known places.
  - **Query:** images of unknown places.



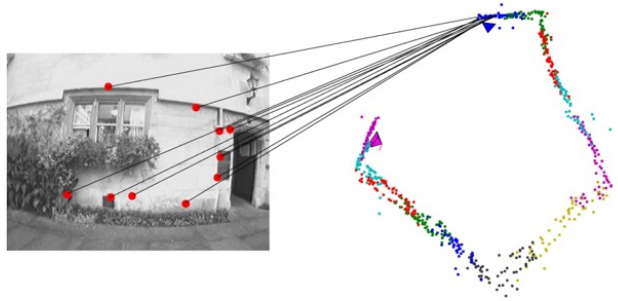
Why we need VPR?

- **Autonomous navigation:** unavailable or weak GNSS signals
  - Occlusions, absorption, and reflections.
  - Urban buildings and indoor space, valleys and caves.
- **Loop closure detection**

  recognize a place when revisiting it in Simultaneous Localization and Mapping (SLAM)
  - **Database:** previously visited places in the map.
  - **Query:** current observation.
- **Visual localization:** initial value for accurate camera pose estimation.

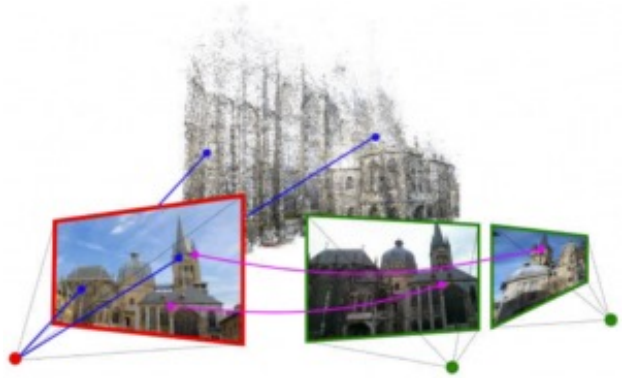


Source: GAO analysis of DOD information. | GAO-21-320SP



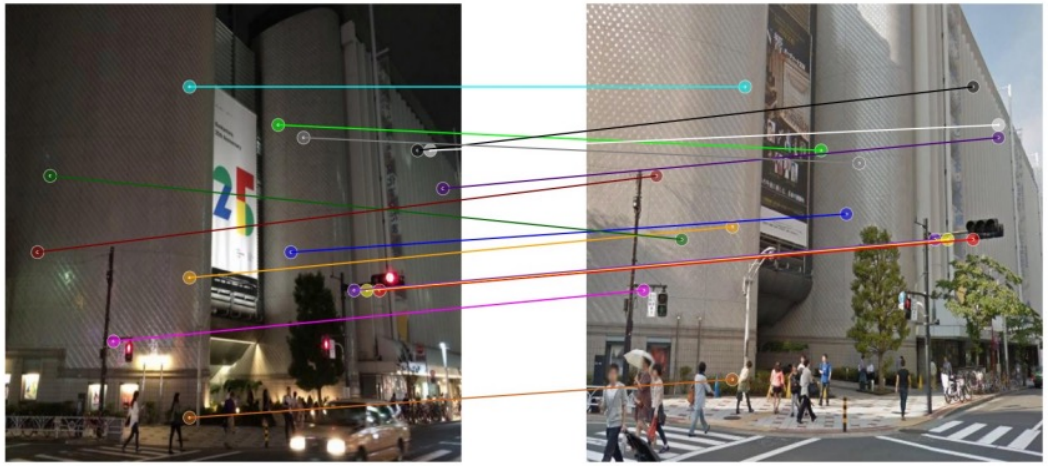
LOOP CLOSURE DETECTION

CLOSED LOOP



Challenge in VPR?

- **Environmental Variability**
  - **Appearance:** varying illumination (day/night), weather (sunny/rainy) and season conditions.
  - **Viewpoint changes**
- **Perceptual aliasing:**
  - Visually similar places (corridors, city streets) leading to false positives.
- **Dynamic objects**
- **Scalability and Generalization**



Approaches for VPR?

- **Traditional handcrafted features**

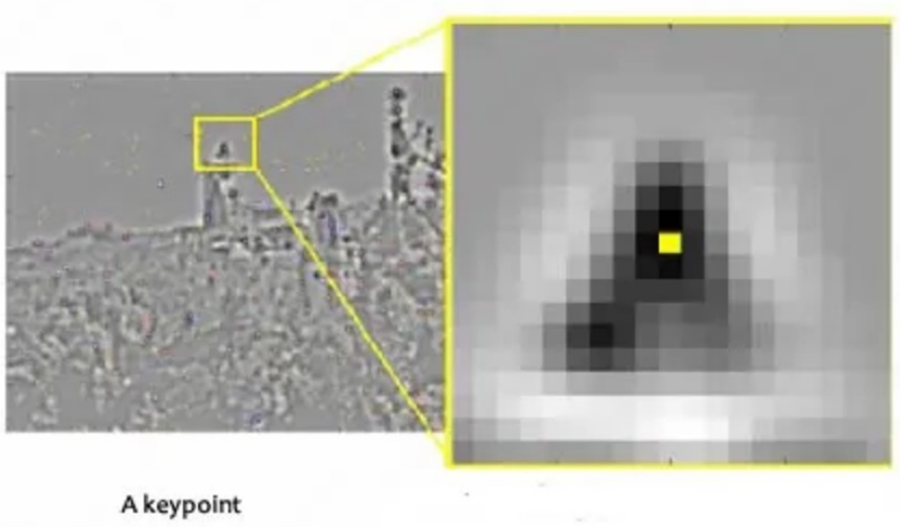
  detect specific visual features like edges or textures.

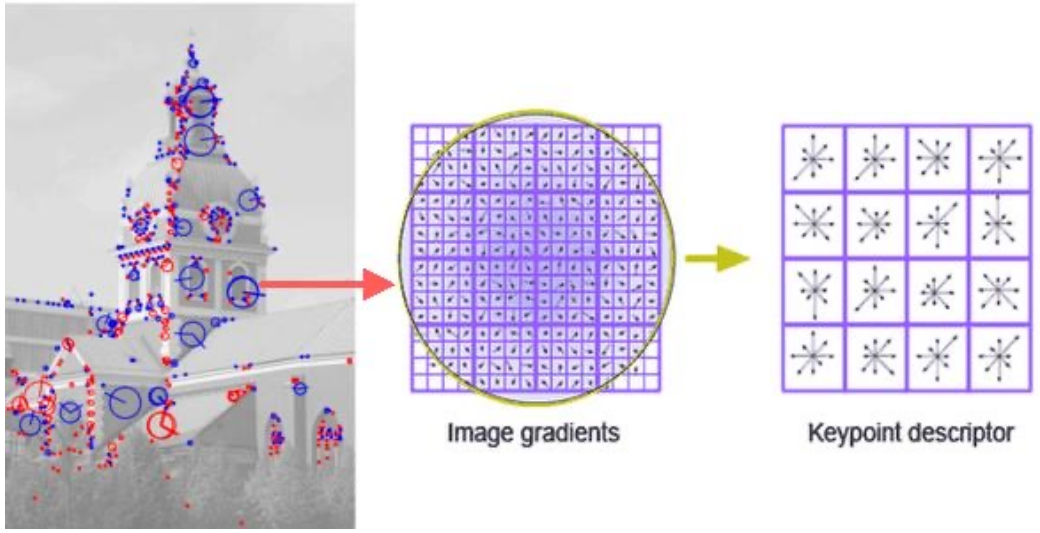
  - Local feature matching.
  - **SIFT, SURF, ORB.**
- **Bag of visual words**
  - Aggregated features.
  - A bag of quantized visual features.
- **Deep Learning-based**
  - Image-based, Point clouds-based, Cross-modal retrieval.

/02 Method

Traditional Methods

- **SIFT** (Scale-Invariant Feature Transform)
  - Detects **keypoints** that are invariant to **scale**, **rotation**, and partially invariant to **illumination**.
  - Extracts **descriptors** that represent the **local appearance** around each **keypoint**.
  - Effective for recognizing places with **distinctive textures or structures**.



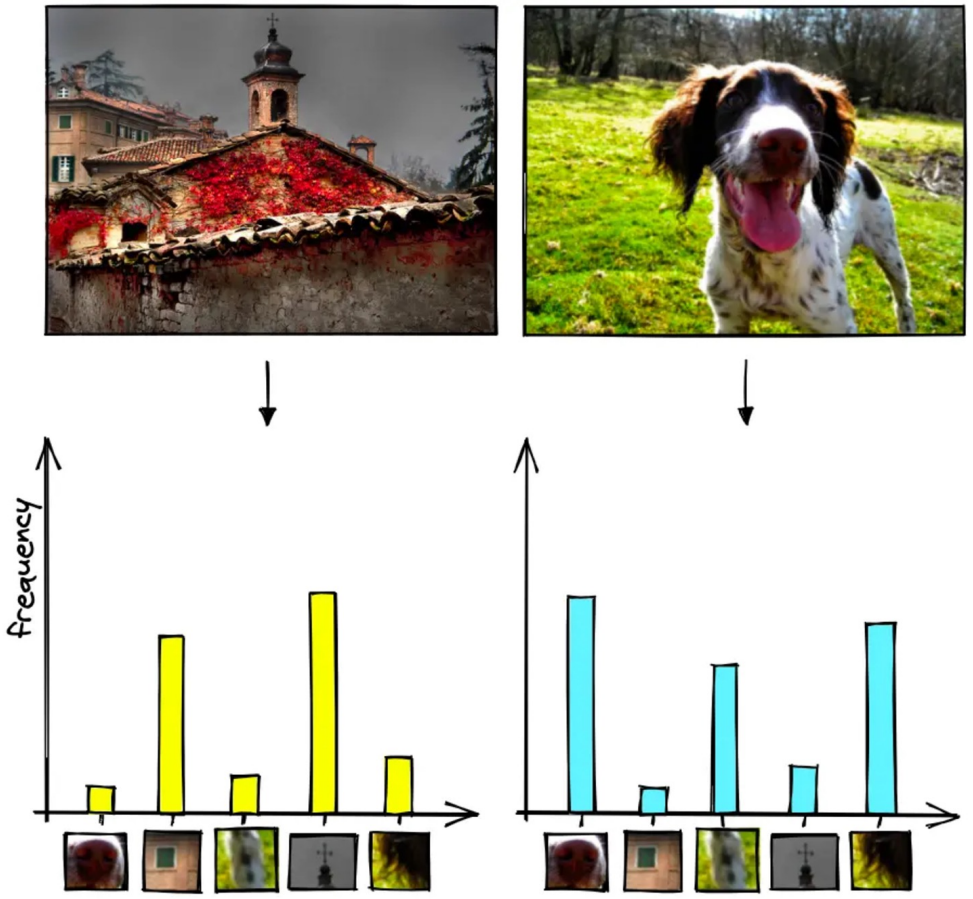


Traditional Methods

- **SURF** (Speeded-Up Robust Features)
  - **Faster** but **less robust** than SIFT for **large viewpoint or scale variation**.
  - Uses **integral images** to speed up the detection of **keypoints** and **descriptors**.
- **ORB** (Oriented FAST and Rotated BRIEF)
  - Combines the **FAST** detector for fast keypoint detection and **BRIEF** descriptor.
  - Widely used for **real-time** and **low-resource** applications with binary descriptors.

Bag of visual words

- **Keypoint** Detection and **Descriptor** Extraction
- **Visual Vocabulary Creation**
  - Apply **k-means clustering** to a large set of feature descriptors.
  - Form a vocabulary of visual words (each cluster).
- Map **each feature descriptor** in an image to the closest visual word
- Represent the image as **a histogram of visual word frequencies**



Deep Learning-based Methods

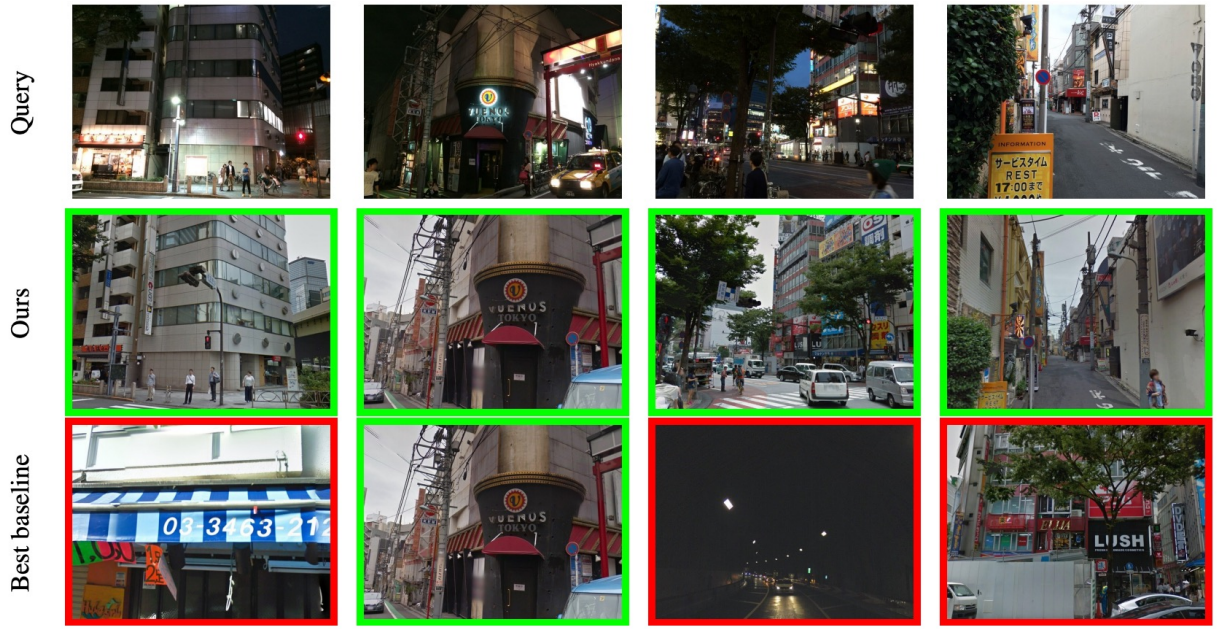
Image-based Retrieval

NetVLAD: CNN architecture for weakly supervised place recognition, CVPR 2016

- An **end-to-end** CNN architecture **directly trainable** for the **place recognition**.
  - NetVLAD: a new generalized VLAD layer pluggable into any CNN.
- A training procedure based on a new weakly supervised ranking loss.

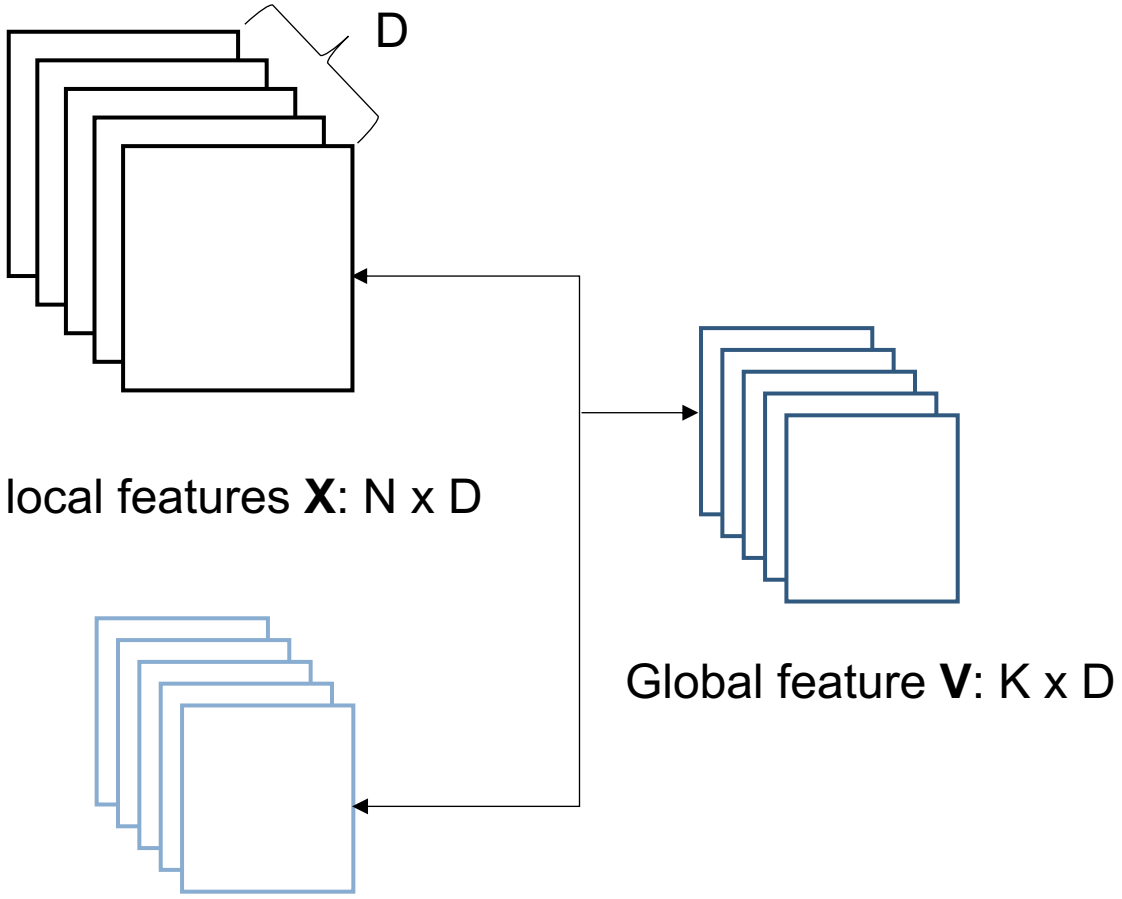
(a) Mobile phone query

(b) Retrieved image of same place



NetVLAD

VLAD(Vector of Local Aggregated Descriptors)

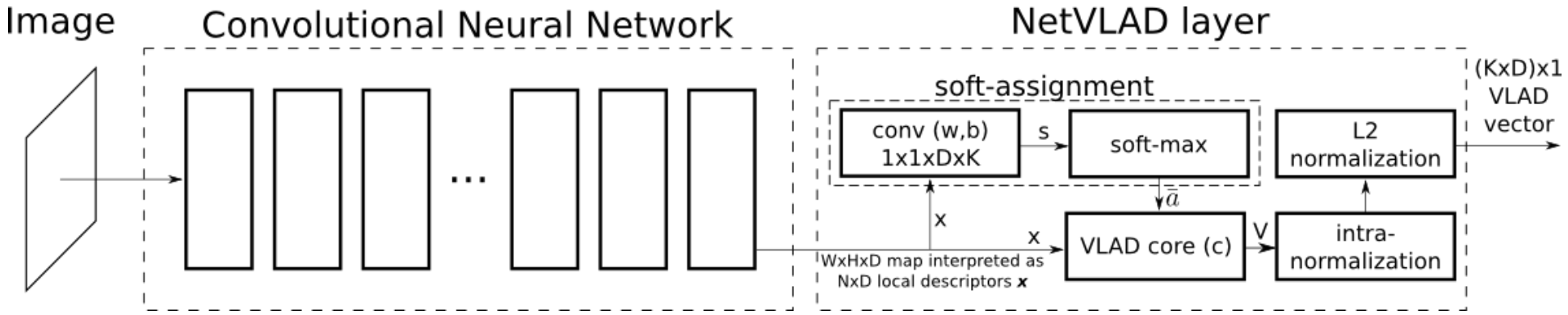


$$V(j,k) = \sum_{i=1}^{N} a_k(x_i)\big(x_i(j) - c_k(j)\big), k \in K, j \in D$$

$$a_k(x_i) = \begin{cases} 1, & if\ x_i\ in\ c_k \\ 0, & else \end{cases}$$

NetVLAD

$$V(j,k) = \sum_{i=1}^{N} \bar{a}_k(x_i)\big(x_i(j) - c_k(j)\big), k \in K, j \in D \qquad \bar{a}_k(x_i) = \frac{e^{-\alpha||x_i - c_k||^2}}{\sum_{k'} e^{-\alpha||x_i - c_{k'}||^2}} = \frac{e^{\omega_k^T x_i + b_k}}{\sum_{k'} e^{\omega_{k'}^T x_i + b_{k'}}}$$



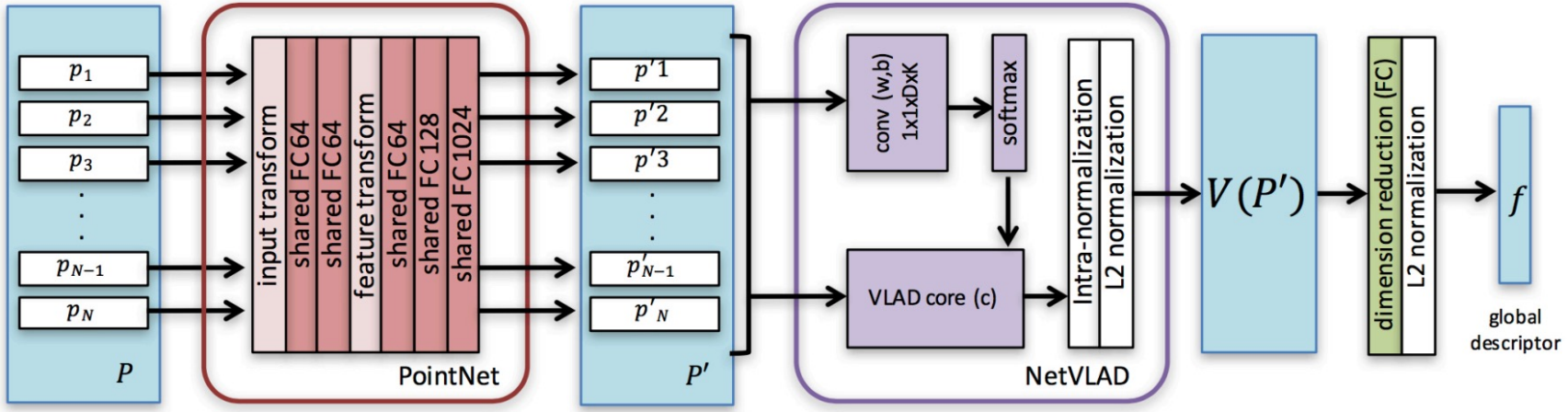
NetVLAD Architecture[2]

Deep Learning-based Methods

Point Clouds-based Retrieval

PointNetVLAD: Deep Point Cloud Based Retrieval for Large-Scale Place Recognition, CVPR 2018

- A combination of the **PointNet** and **NetVLAD** to extract the global descriptor from a 3D **point cloud**.
- A **lazy triplet and quadruplet loss** functions for generalizable global descriptors.



	Average recall
Triplet Loss	71.20
Quadruplet Loss	74.13
Lazy Triplet Loss	**78.99**
Lazy Quadruplet Loss	**80.31**

Deep Learning-based Methods

Cross Modal Retrieval

Attention-Enhanced Cross-modal Localization Between Spherical Images and Point Clouds, IEEE Sensors Journal

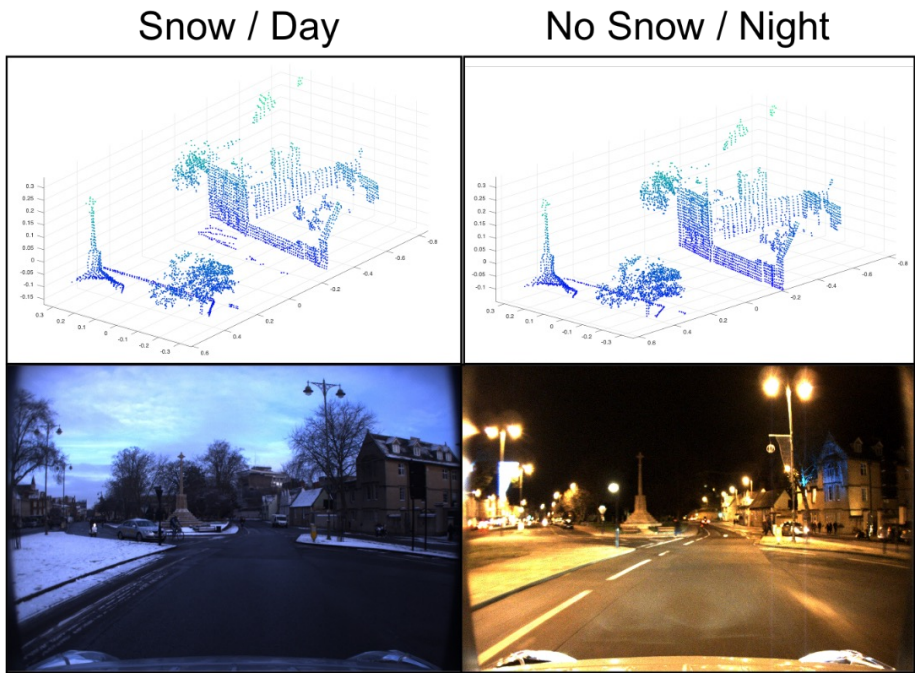
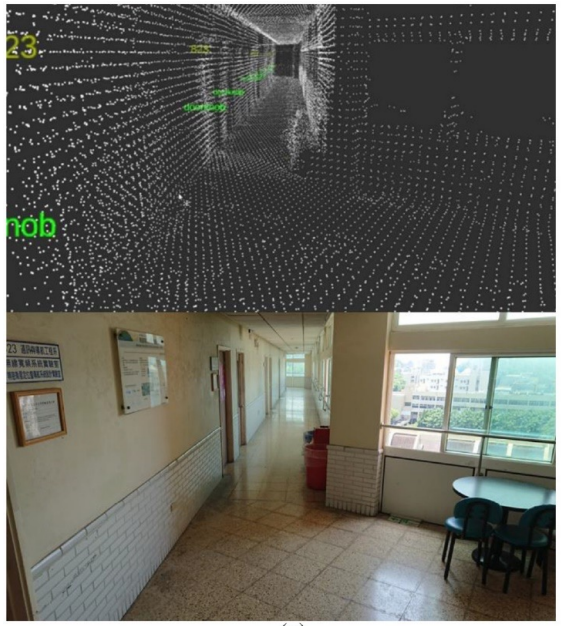
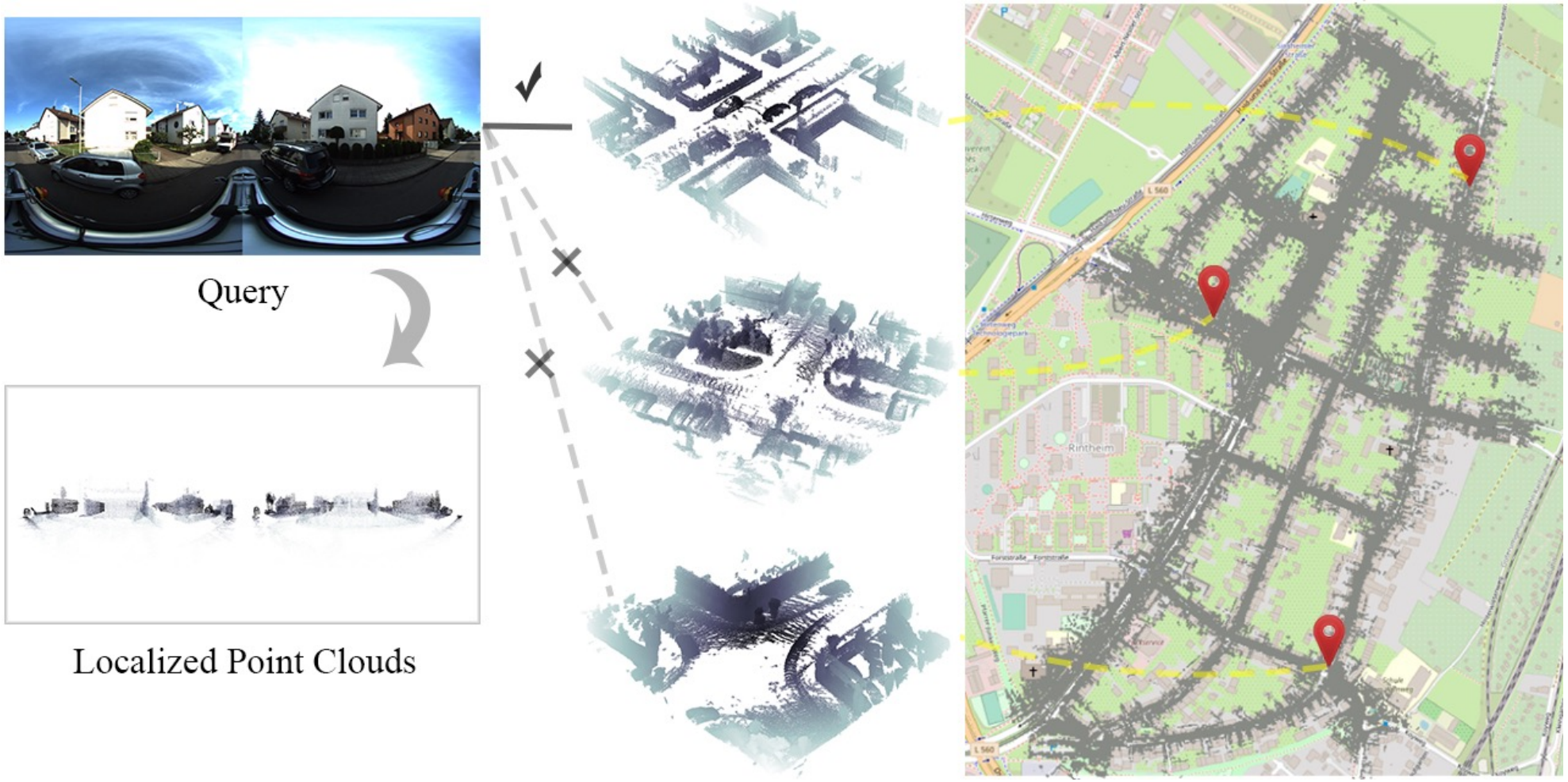


Image-based retrieval is sensitive to illumination and season change



- Point clouds-based retrieval is fragile to geometric degeneracy
- High cost and large volume of LiDAR

Why Cross-Modal Retrieval?



Query: **more information** at **a lower cost;** Database: **more robust** to illumination changes

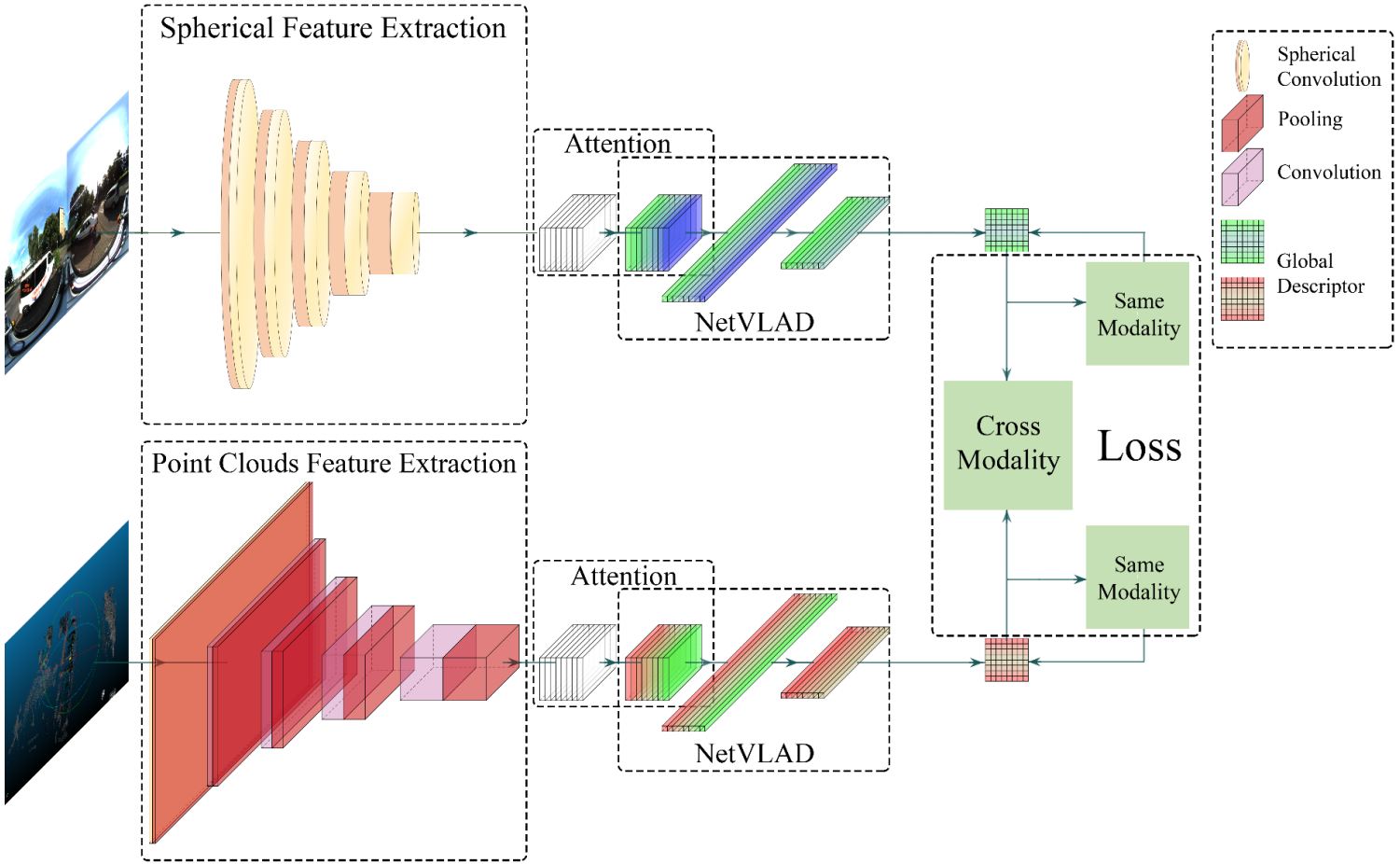
Why Spherical Image?

spherical image

perspective image

A **wider** field-of-view, corresponding to **omni-directional** point clouds

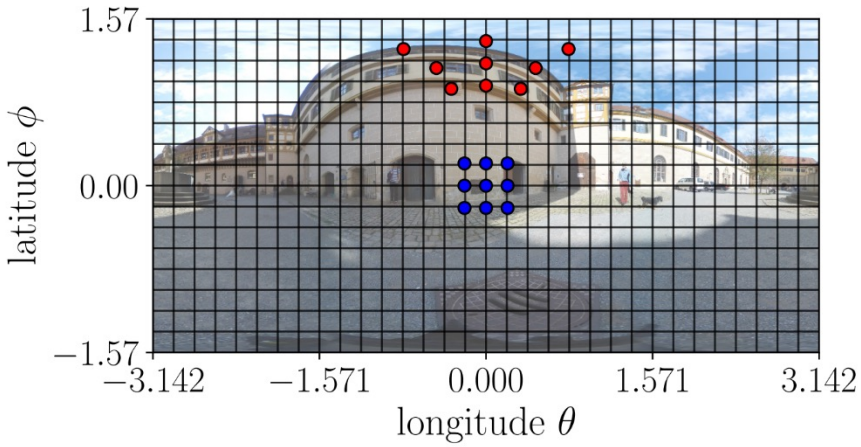
Network Architecture



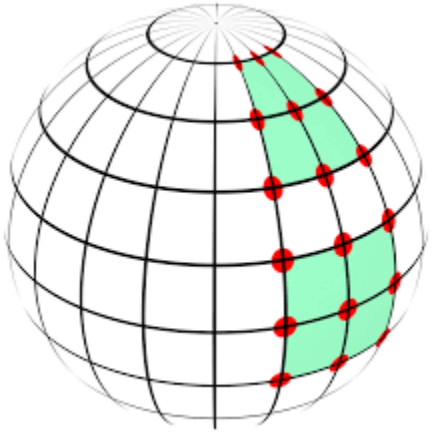
- **Spherical** feature extraction
- Attention enhancement on **specific** features linking two modalities

Spherical Feature Extraction

unwrap

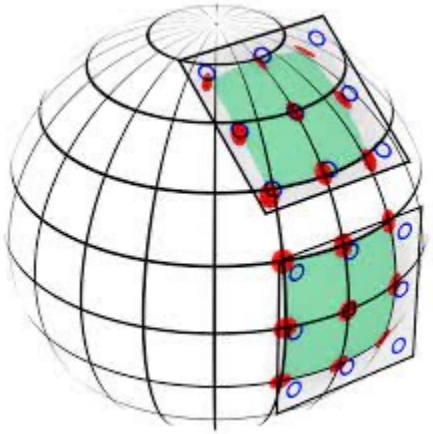


different distortion at different latitude



Regular CNN Filter

ignore the distortion variance

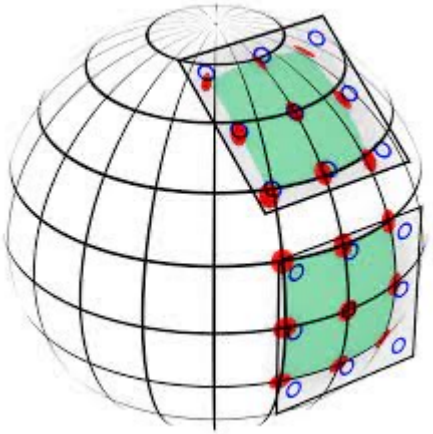


Spherical CNN Filter

adapting the **sample grid**

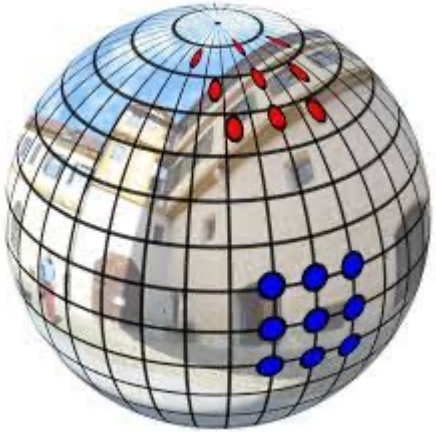
Spherical Feature Extraction

- Sample location defined on the Sphere S^2 surface



Every point $s = (\Phi, \theta) \in S^2, latitude\ \Phi \in \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] and\ longtitude\ \theta \in [-\pi, \pi]$

Let Π denote the tangent plane located at $s_\Pi = (\Phi_\Pi, \theta_\Pi)$



$$s_{(0,0)} = (0,0)$$
$$s_{(\pm1,0)} = (\pm\Delta_\Phi, 0)$$
$$s_{(0,\pm1)} = (0, \pm\Delta_\theta)$$
$$s_{(\pm1,\pm1)} = (\pm\Delta_\Phi, \pm\Delta_\theta)$$

Spherical Feature Extraction

- Projection from the Sphere S^2 to the tangent plane

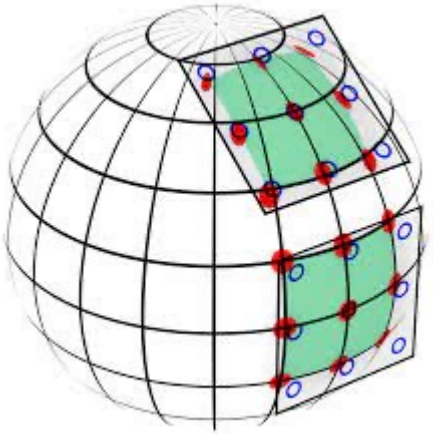
Projection to Π_0, located at $s = (0,0)$

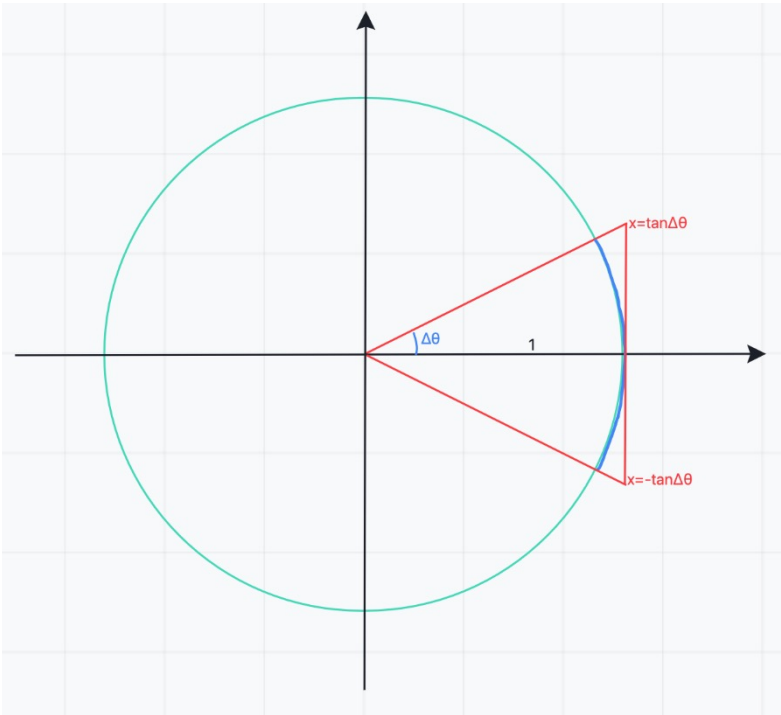
$$x_{(0,0)} = (0,0)$$
$$x_{(\pm1,0)} = (\pm tan\Delta_\theta, 0)$$
$$x_{(0,\pm1)} = (0, \pm tan\Delta_\Phi)$$
$$x_{(\pm1,\pm1)} = (\pm tan\Delta_\theta, \pm sec\Delta_\theta tan\Delta_\Phi)$$

Projection to Π, located at $s_\Pi = (\Phi_\Pi, \theta_\Pi)$

$$x(\emptyset, \theta) = \frac{cos\Phi \sin(\theta - \theta_\Pi)}{sin\Phi_\Pi sin\Phi + cos\Phi_\Pi cos\Phi \cos(\theta - \theta_\Pi)}$$

$$y(\emptyset, \theta) = \frac{cos\Phi_\Pi sin\Phi - sin\Phi_\Pi cos\Phi \cos(\theta - \theta_\Pi)}{sin\Phi_\Pi sin\Phi + cos\Phi_\Pi cos\Phi \cos(\theta - \theta_\Pi)}$$





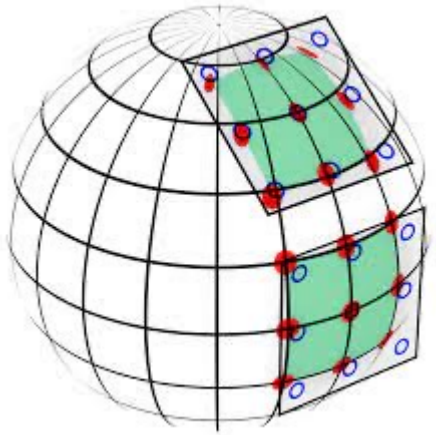
Spherical Feature Extraction

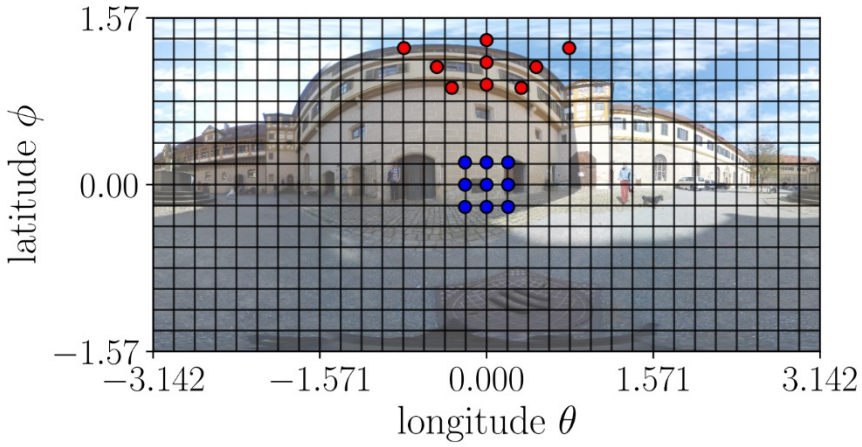
- Projection from the tangent plane to the spherical image

$$\Phi(x,y) = sin^{-1}(cosvsin\Phi_{\Pi} + \frac{ysinvcos\Phi_{\Pi}}{\rho})$$

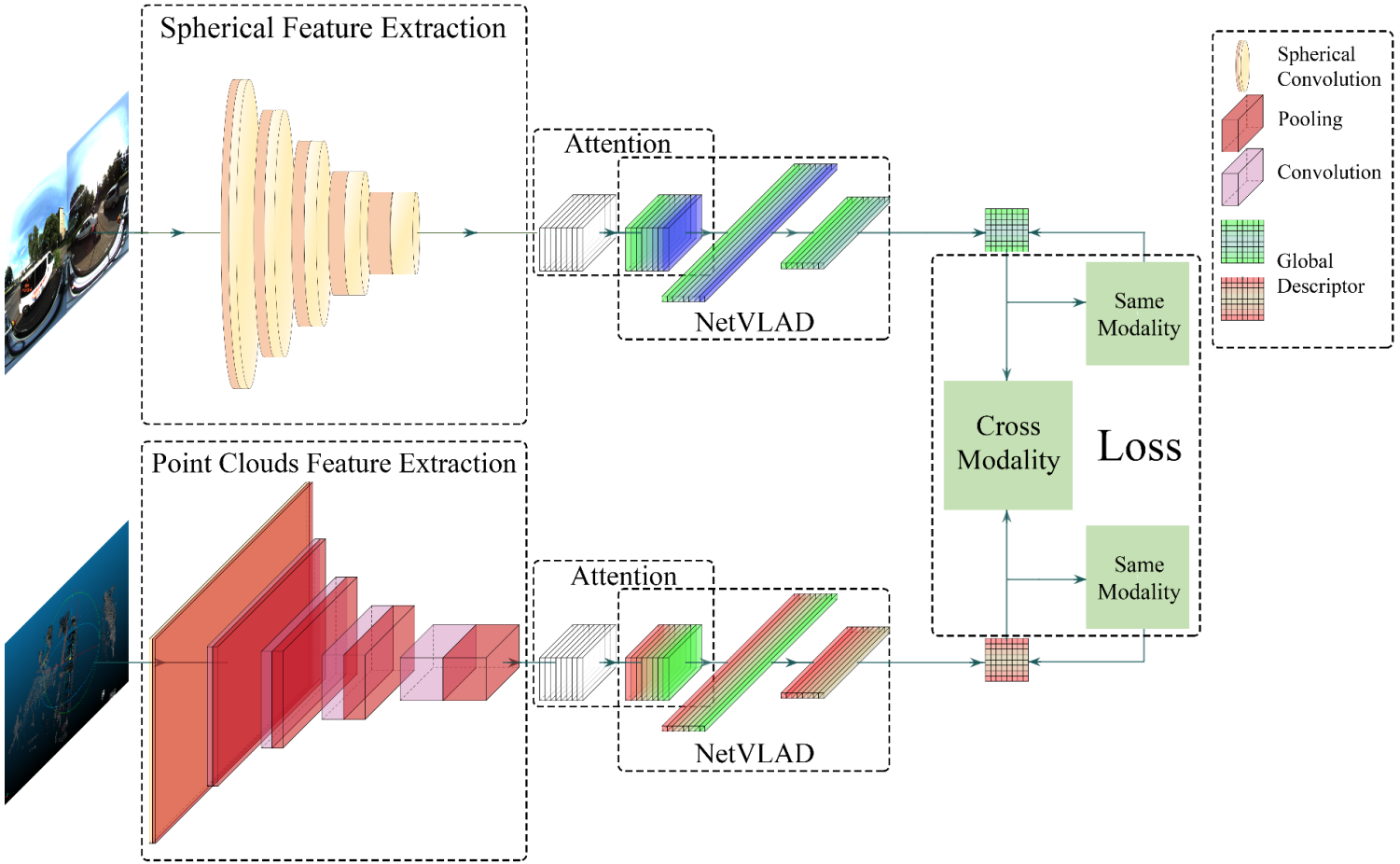
$$\theta(x,y) = \theta_{\Pi} + tan^{-1}(\frac{xsinv}{\rho cos\Phi_{\Pi} cosv - ysin\Phi_{\Pi} sinv})$$

$$where\ \rho = \sqrt{x^2 + y^2}, and\ v = tan^{-1}\rho$$



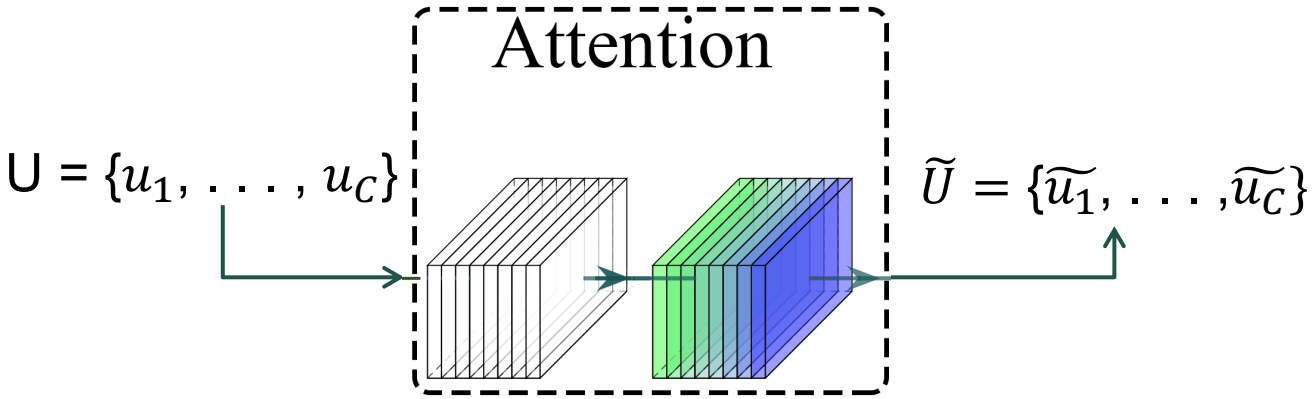


Network Architecture



- **Spherical** feature extraction
- Attention enhancement on **specific** features linking two modalities

Attention Enhancement

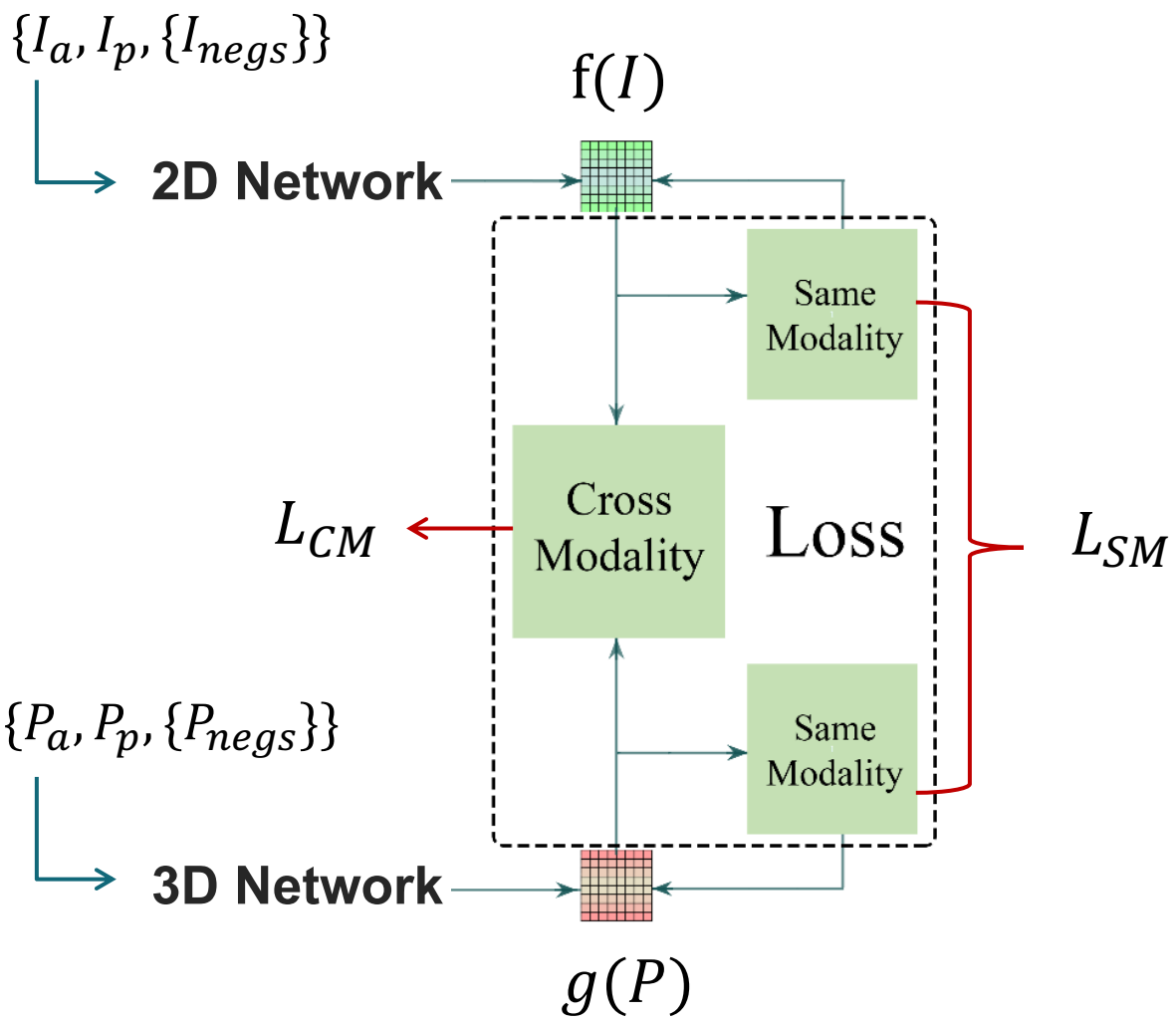


$$z_i = \frac{1}{D_1 \times D_2} \sum_{k=1}^{D_2} \sum_{j=1}^{D_1} u_i(k, j)$$

$$S = \sigma(W_2 \delta(W_1 Z))$$

$$\widetilde{u_i} = s_i u_i,$$

Loss Function



$$L_{anchor} = d(f(I_a), g(P_a))$$

$$L_{ItoP} = [d(f(I_a), g(P_p)) - d(f(I_a), g(P^i_{neg}))]_+$$

$$L_{PtoI} = [d(g(P_a), f(I_p)) - d(g(P_a), f(I^i_{neg}))]_+$$

$$L_{CM} = L_{ItoP} + L_{PtoI}$$

$$L = L_{CM} + 0.1L_{SM} + L_{anchor}$$

Result

ResNet-based Baseline: ResNet-18 + NetVLAD

AE-Spherical Model: Spherical CNN + Attention + NetVLAD

COMPARISON BETWEEN MODELS

model	recall@1	recall@5	recall@1%
ResNet-based Baseline	36.79	52.83	66.98
AE-Spherical Model	**46.23**	**66.04**	**75.47**

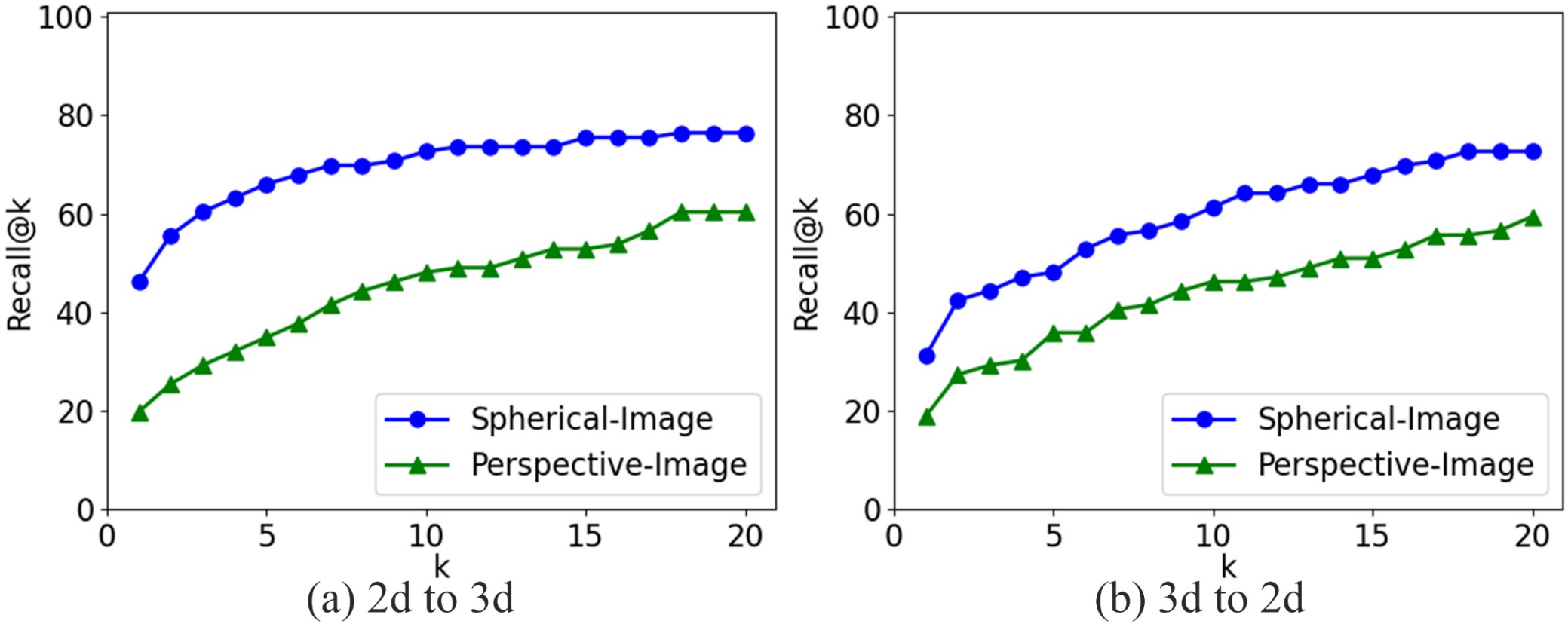
ABLATION STUDY

			recall@1		recall@1%	
Base	*SCNN*	*Attention*	2D to 3D	3D to 2D	2D to 3D	3D to 2D
✓	✓	✗	33.02	31.13	71.70	61.32
✓	✗	✓	37.74	30.19	70.75	**68.87**
✓	✓	✓	**46.23**	**31.13**	**75.47**	67.92

COMPARISON BETWEEN LOSSES

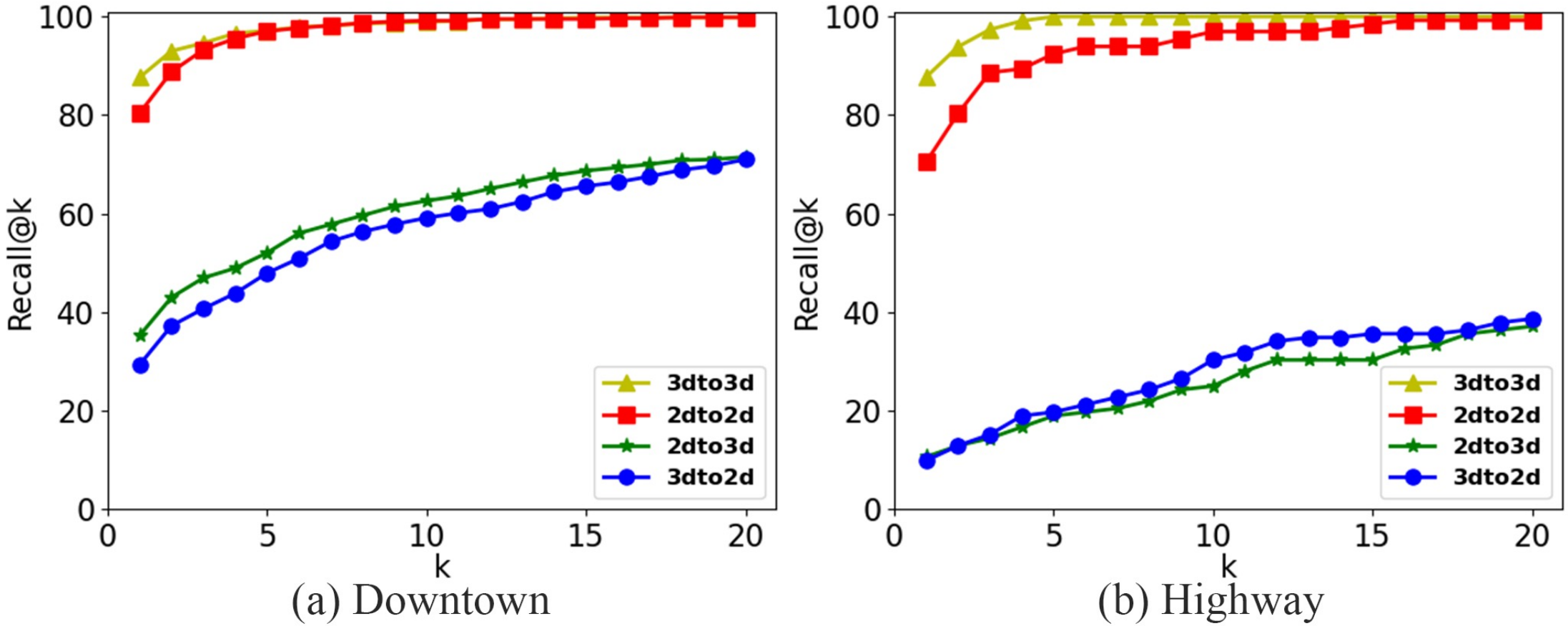
loss	recall@1	recall@5	recall@10	recall@20
$0.1\mathcal{L}_{SM} + \mathcal{L}_{anchor}$	16.04	33.96	41.51	57.55
$0.1\mathcal{L}_{SM} + \mathcal{L}_{CM}$	28.30	42.45	57.55	67.92
$\mathcal{L}_{sum}$	**37.74**	**53.77**	**59.43**	**70.75**

Result



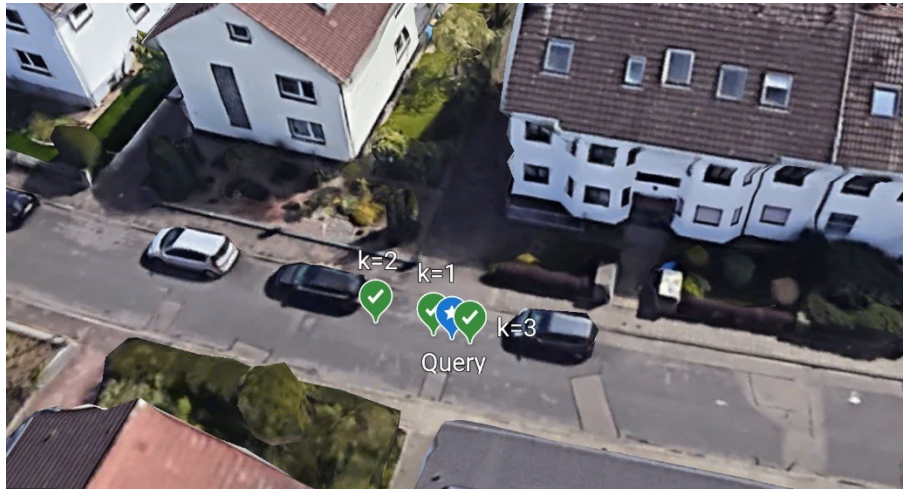
Comparison between spherical image and perspective image

Result



Average Recall@k of AE-Spherical Model on (a) Downtown and (b) Highway scenarios

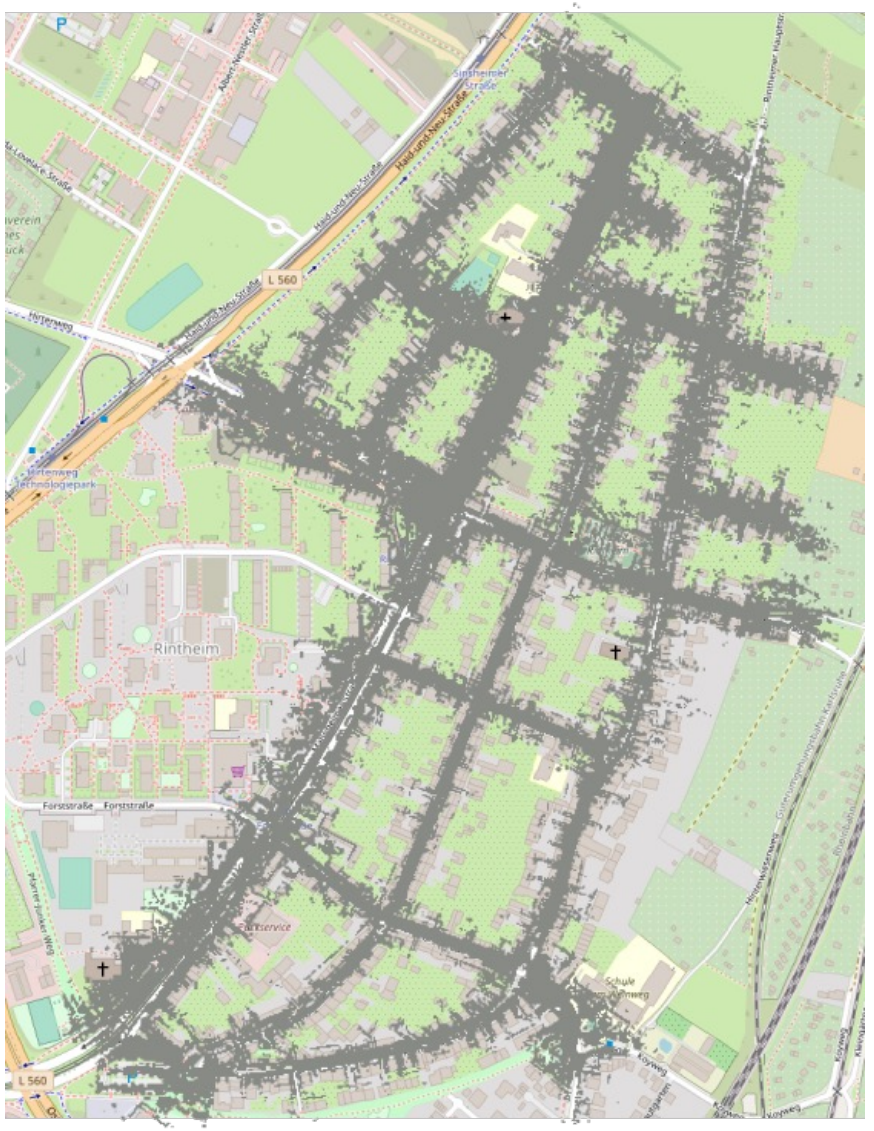
Query spherical Image



Retrieval

Top 3 Point Clouds

Point Clouds Database



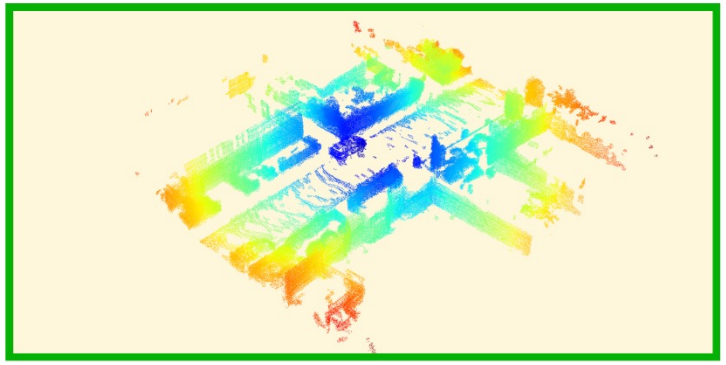
Top k Point Clouds

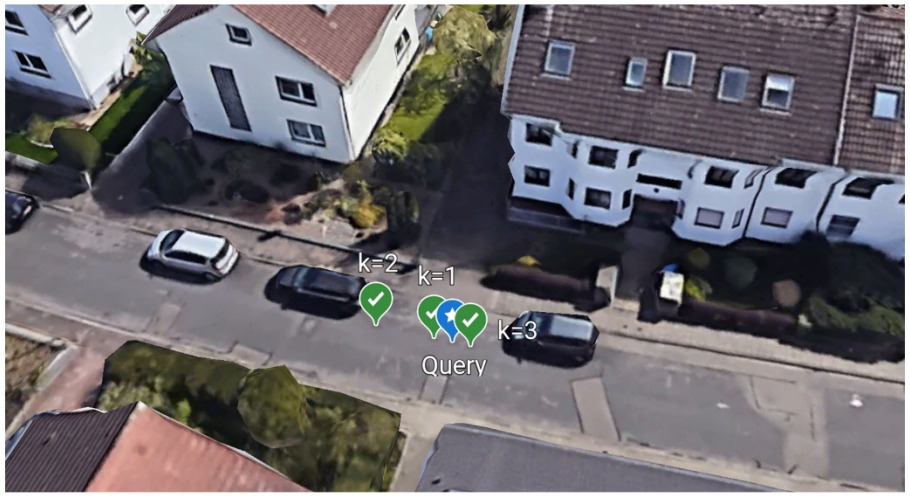
Corresponding Images

(For visualization purpose)

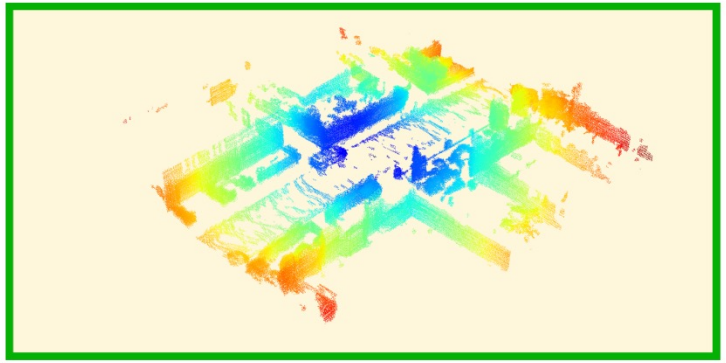
Query

k=1

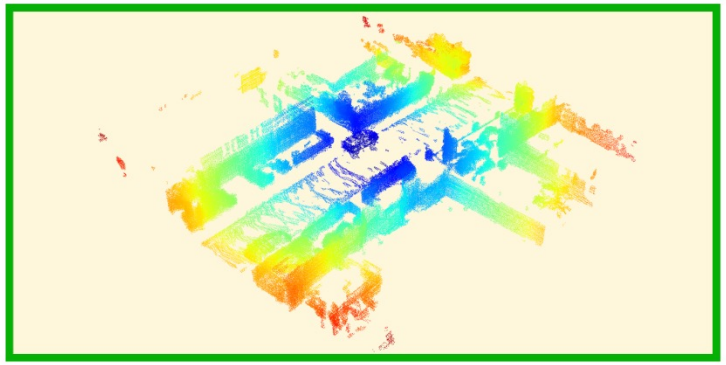




k=2



k=3

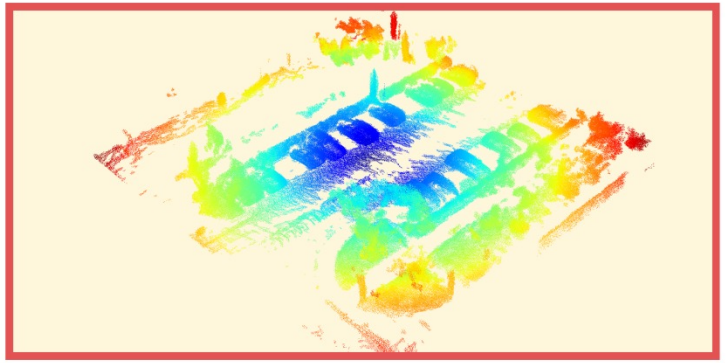


Top k Point Clouds

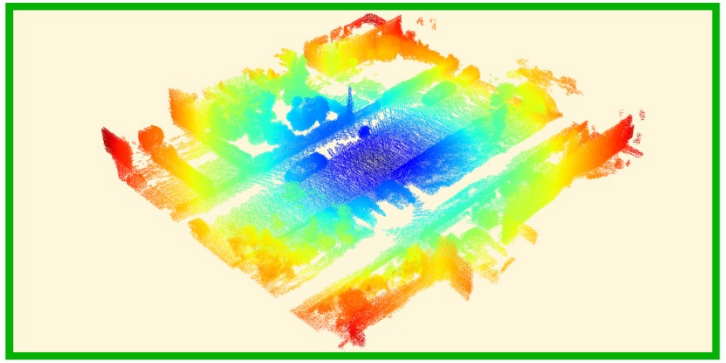
Corresponding Images
(For visualization purpose)

Query

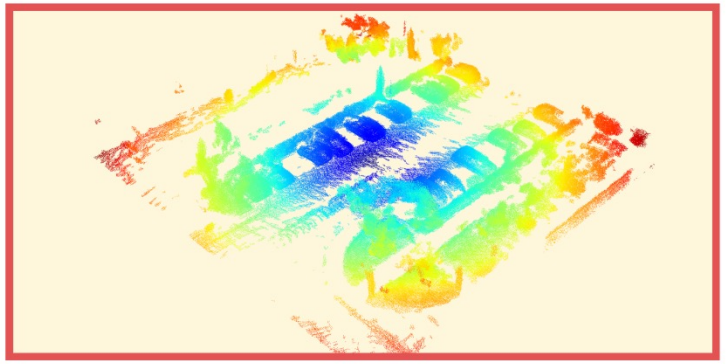
k=1



k=2



k=3



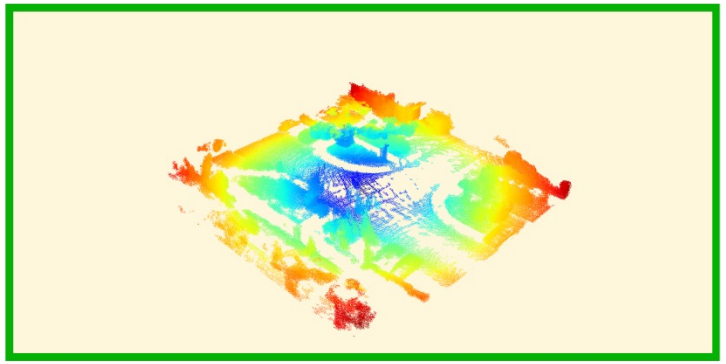
Top k Point Clouds

Corresponding Images

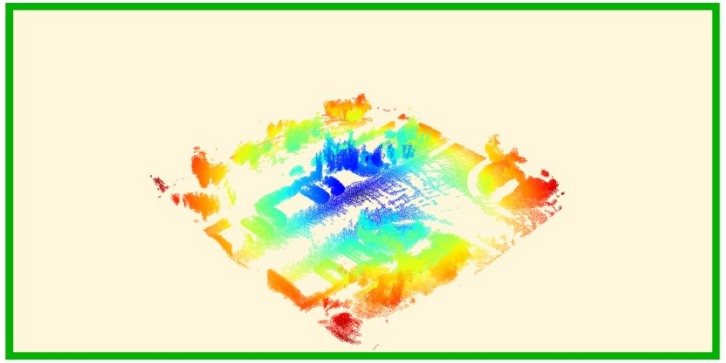
(For visualization purpose)

Query

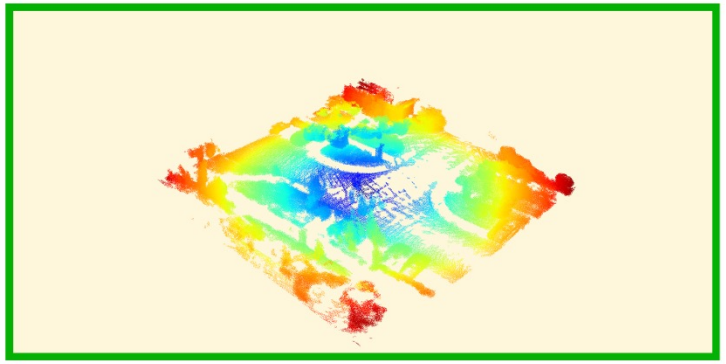
k=1

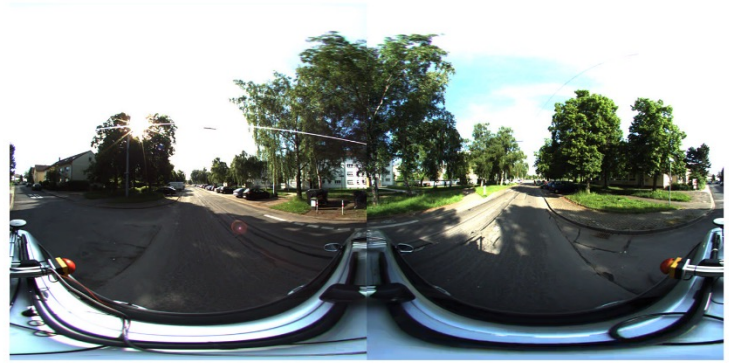


k=2



k=3



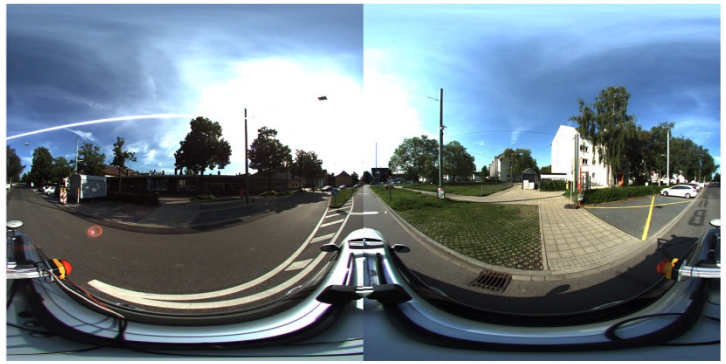


Top k Point Clouds

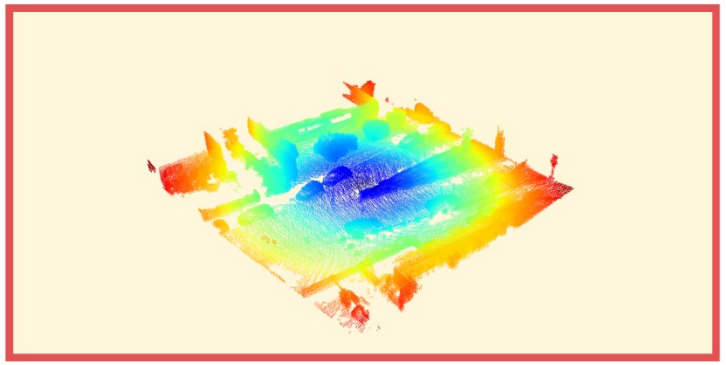
Corresponding Images

(For visualization purpose)

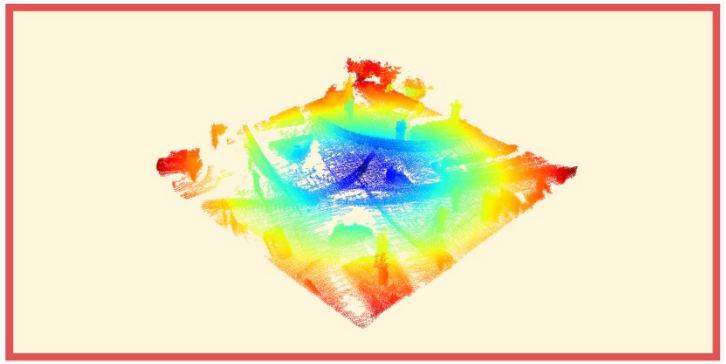
Query



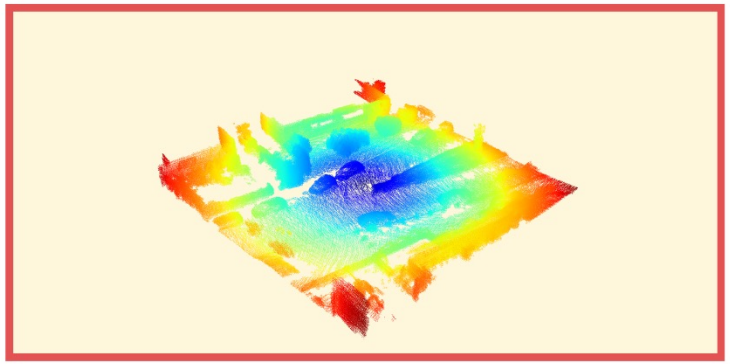
k=1



k=2



k=3



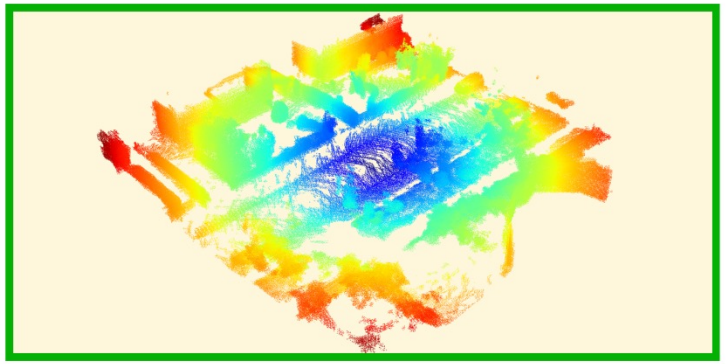
Top k Point Clouds

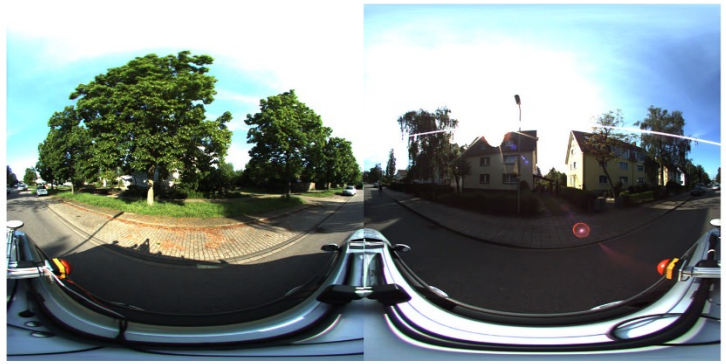
Corresponding Images

(For visualization purpose)

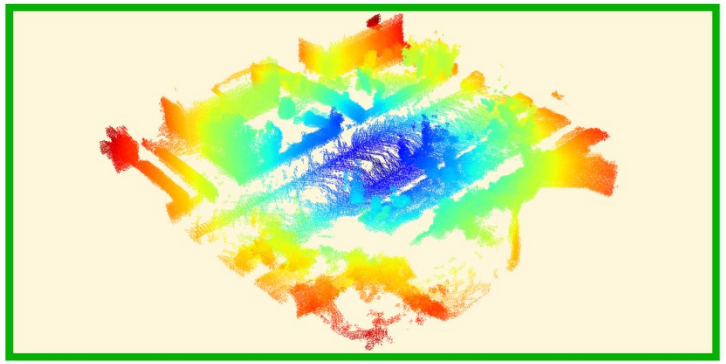
Query

k=1

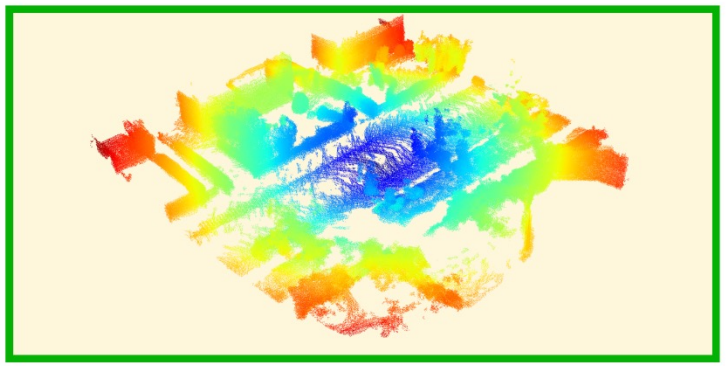




k=2



k=3

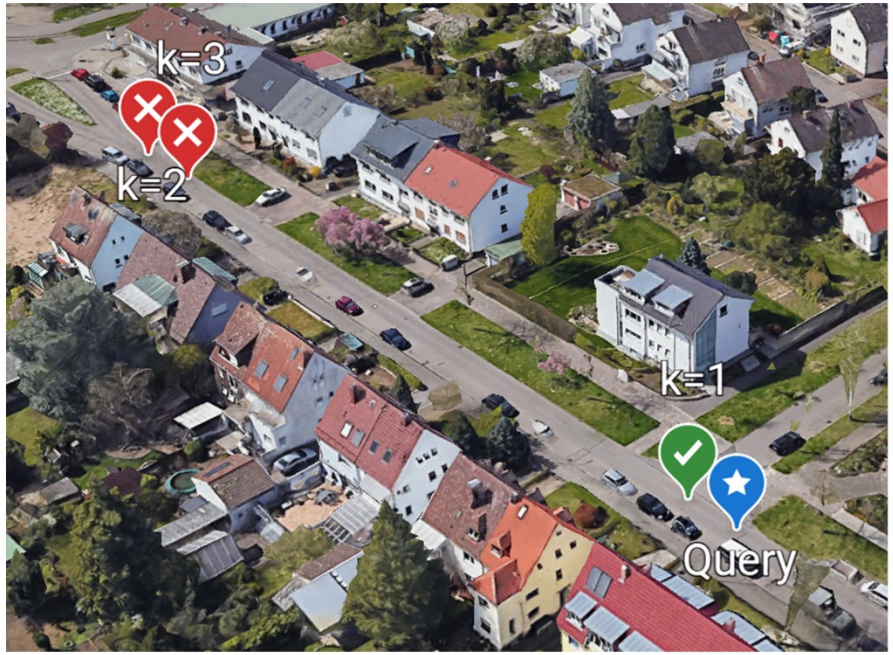


Top k Point Clouds

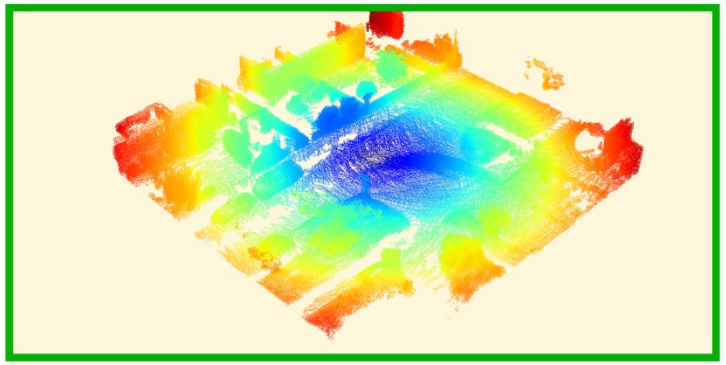
Corresponding Images

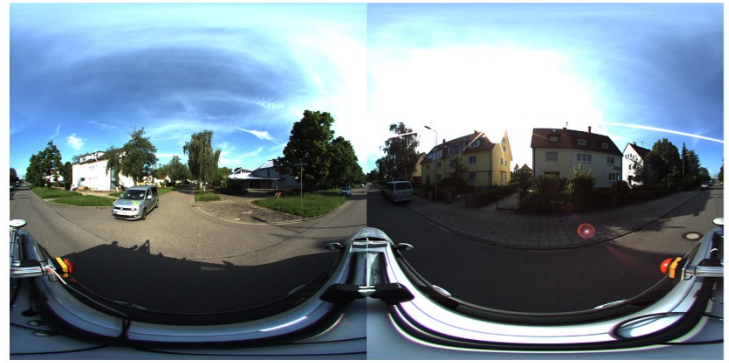
(For visualization purpose)

Query

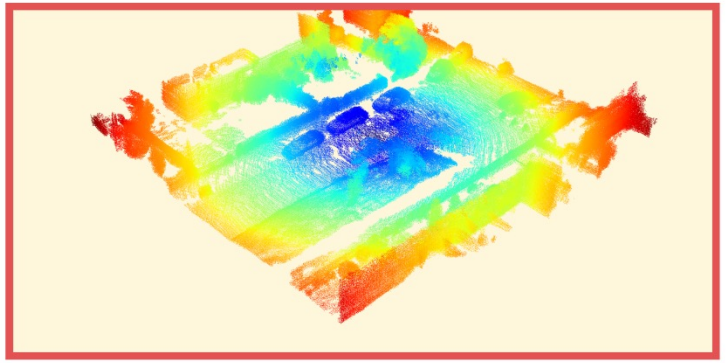


k=1

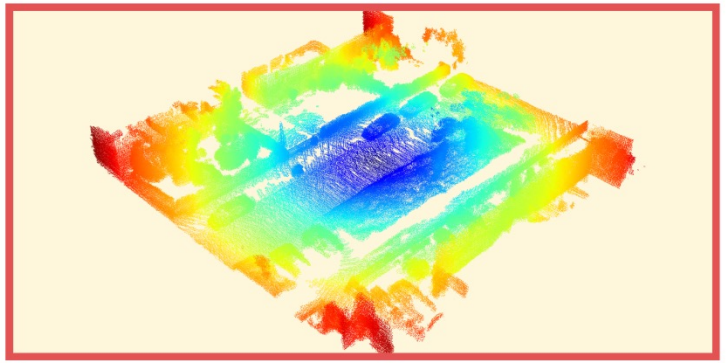




k=2



k=3

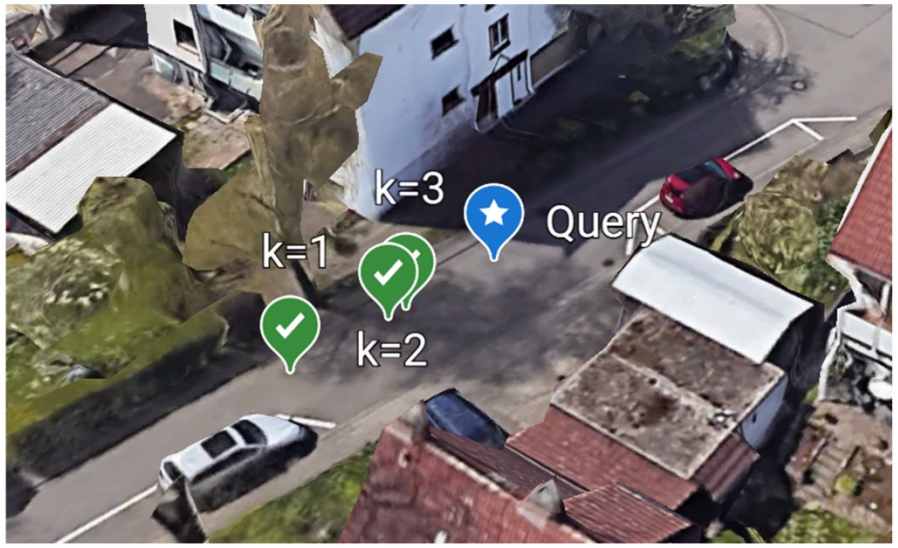


Top k Point Clouds

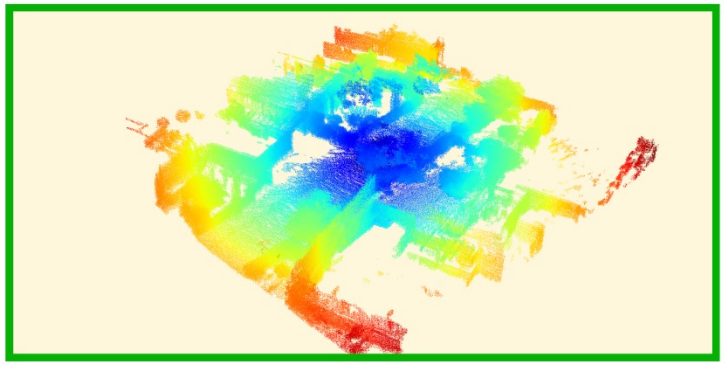
Corresponding Images

(For visualization purpose)

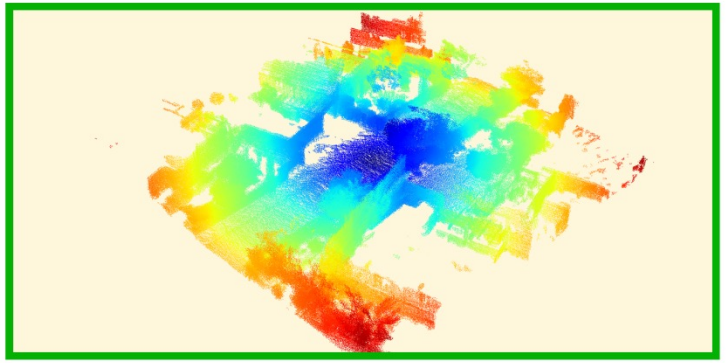
Query



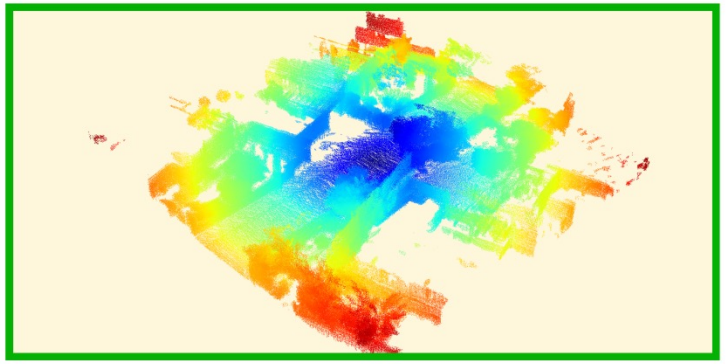
k=1



k=2



k=3



What are the categories of deep learning-based visual place recognition based on data modality?