香港科技大学(广州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

系统枢纽
SYSTEMS HUB

智能交通学域
INTELLIGENT TRANSPORTATION THRUST

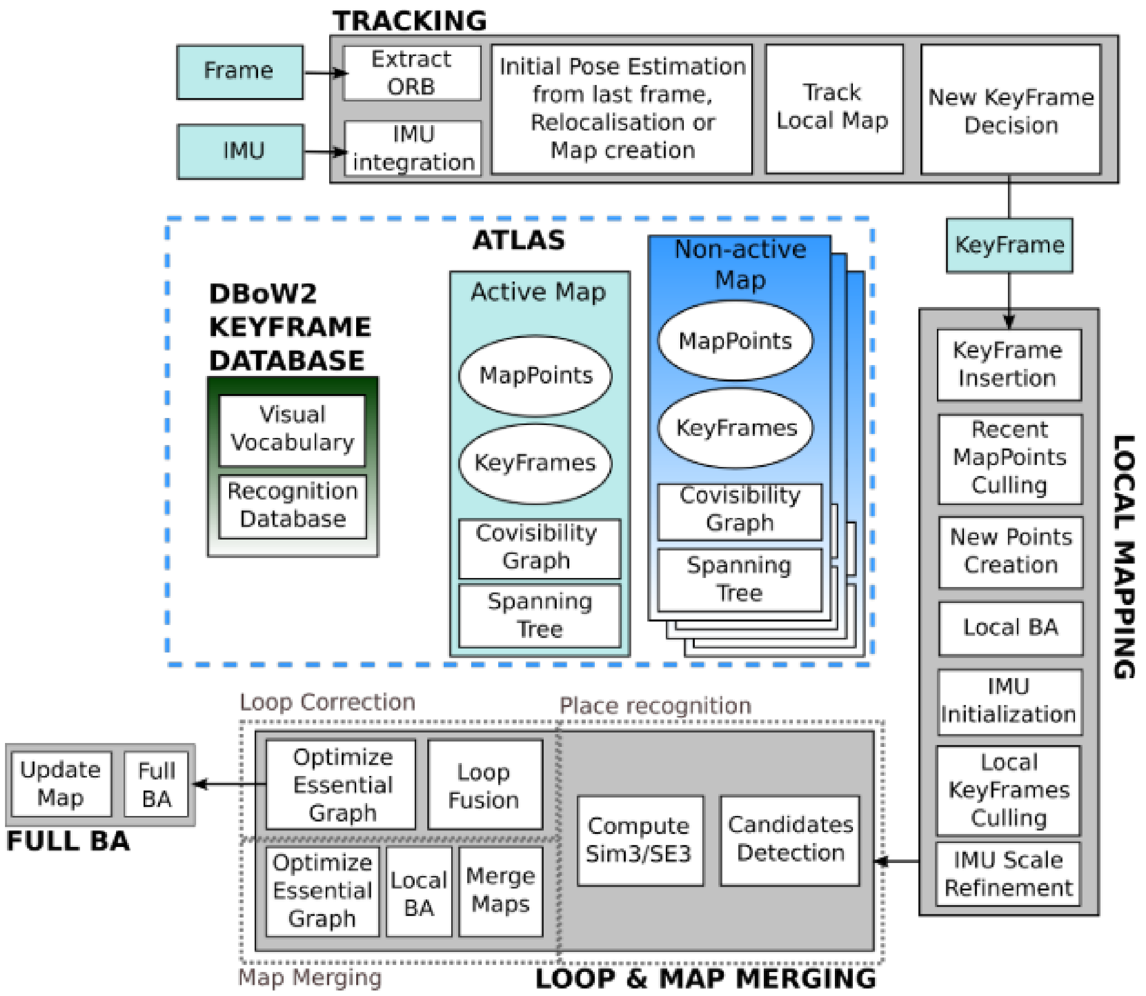
Place Recognition

Graduate Course INTR-6000P

Week 5 - Lecture 10

Changhao Chen
Assistant Professor
HKUST (GZ)

Recap: The SLAM Problem



Visual SLAM
(Simultaneous Localization and Mapping)

Goal: Build a consistent global map of the environment while simultaneously localizing within it.

Focus: Global consistency. Output: A globally consistent map and trajectory.

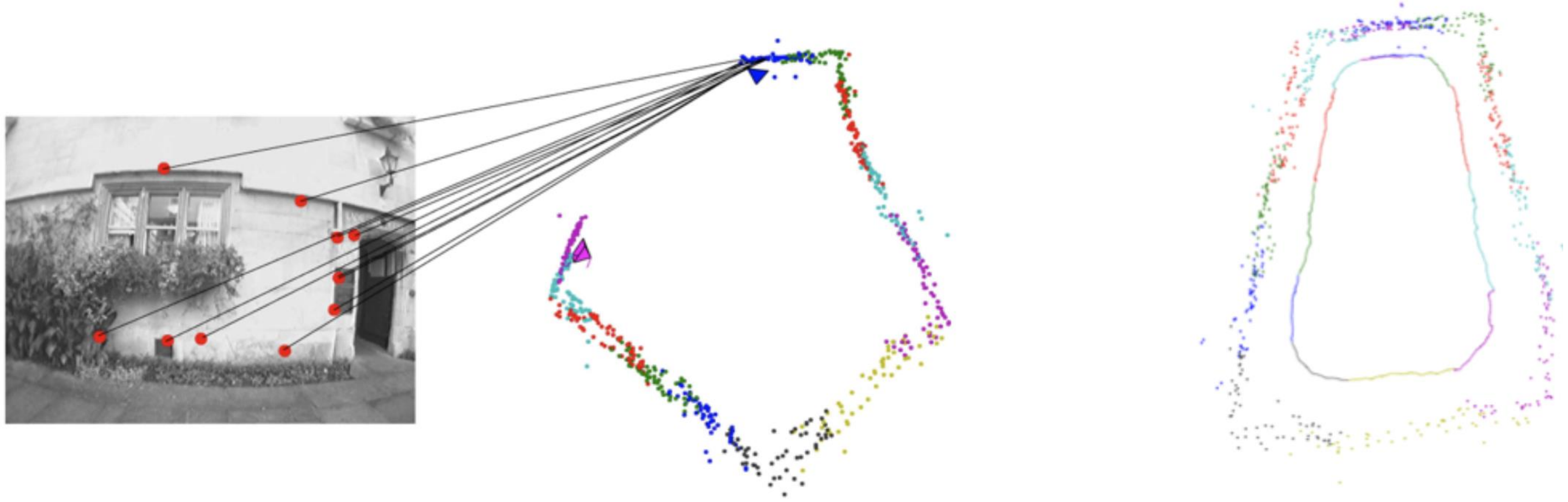
Solution to Drift: Loop Closing - detecting previously visited locations and correcting the entire map.

Loop Closing

Loop Closing - The process of recognizing a previously visited location and correcting the accumulated drift.

Functions:

- **Drift Correction:** Significantly reduces long-term error.
- **Map Consistency:** Produces a globally consistent map.
- **Enables Long-Term Autonomy:** A vehicle can operate for hours/days without getting lost.



Place Recognition

Definition: **The task of determining where an image was taken by matching it against a database of geo-referenced images.**

It's not just image retrieval! It's about *appearance-invariant* recognition.

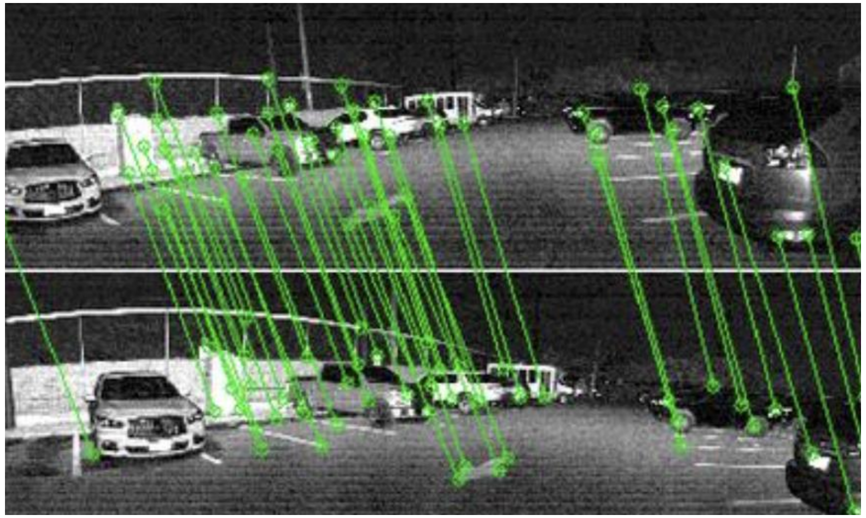
Input: A query image from the vehicle's current view.
Output: A binary decision ("Is this a loop?") and/or a match to a previous location in the map.

(a)

(b)

Why is it Challenging for Intelligent Vehicles?

Viewpoint Change: The same place looks different when approached from a different direction.



Condition Change (Perceptual Aliasing):

- **Time of Day:** Morning vs. Night.
- **Weather:** Sunny vs. Rainy vs. Snowy.
- **Seasons:** Summer vs. Winter.

Dynamic Objects: Cars, pedestrians, which are not part of the "place.”

Structural Changes: Construction, new buildings.

Scale & Speed: Vehicles move faster than robots, requiring efficient algorithms.

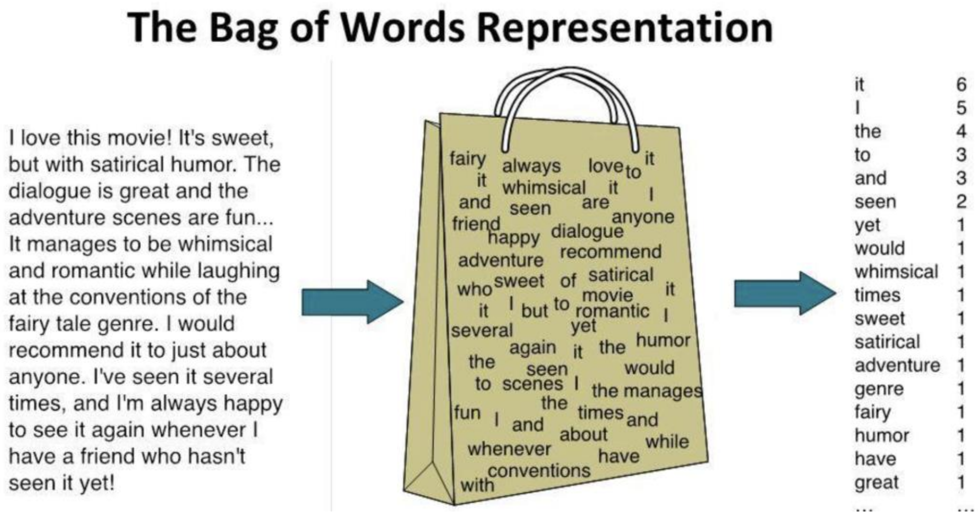
Bag-of-Words (BoW)

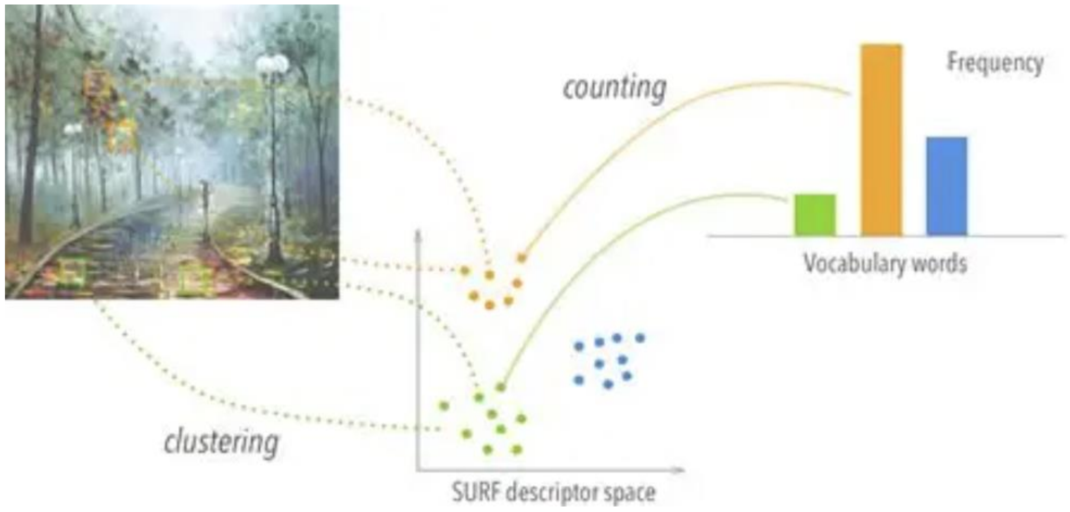
Borrowed from text retrieval. Treat an image as a "bag" of visual words, ignoring their spatial arrangement.

Pipeline:

- **Feature Extraction:** Detect and describe keypoints (e.g., with SIFT).
- **Vocabulary Building:** Cluster all descriptors from the training dataset to create a "visual vocabulary."
- **Quantization:** Assign each new feature to its nearest visual word.
- **Image Representation:** Create a histogram of visual word frequencies for each image.

Matching: Compare histograms using a distance metric (e.g., L1, L2). Fast and scalable!





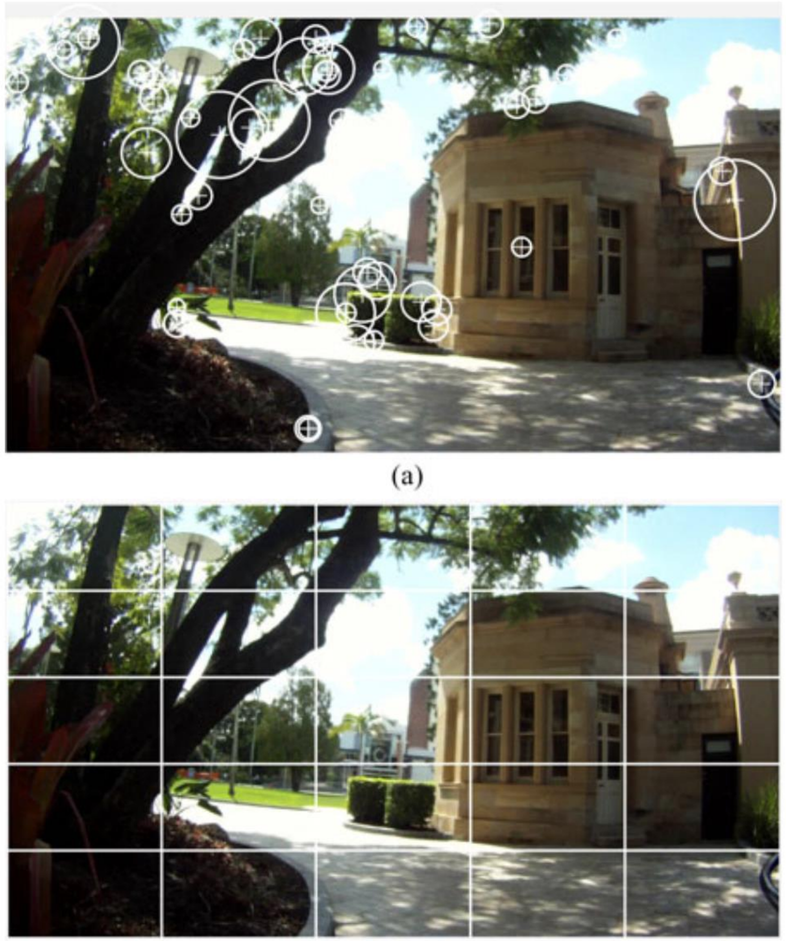
Feature Detectors & Descriptors

•**SIFT (Scale-Invariant Feature Transform):** Robust to scale, rotation, and illumination. Computationally heavy.

•**SURF (Speeded Up Robust Features):** Faster approximation of SIFT.

•**ORB (Oriented FAST and Rotated BRIEF):** Fast, binary descriptor. Good for real-time systems. Less robust than SIFT.

These provide the "words" for the BoW model.



(a)

(b)

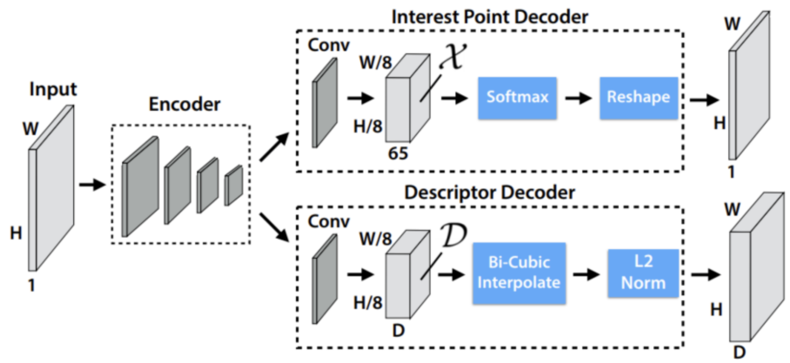
Deep Learning Based Loop-Closing

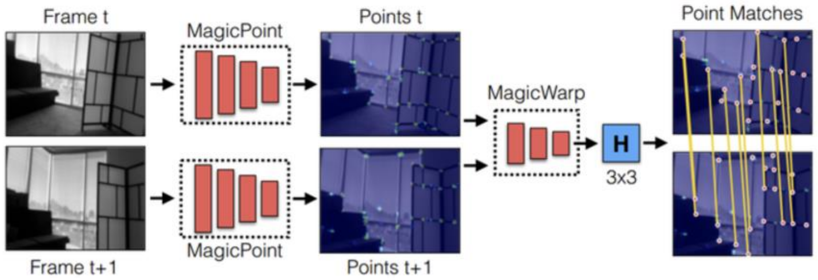
Handcrafted features struggle with severe appearance changes.

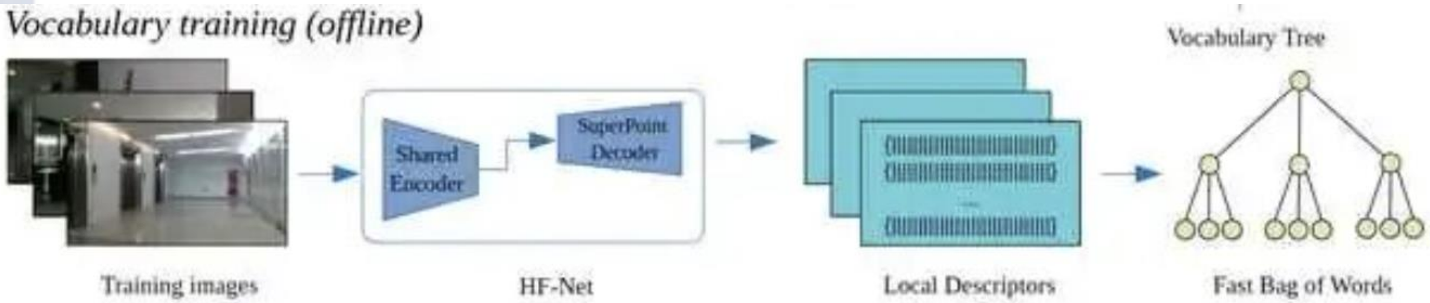
The Power of CNNs: Convolutional Neural Networks can learn powerful, condition-invariant features directly from data.

How? Use a pre-trained CNN (e.g., on ImageNet) as a feature extractor.

- **Global Descriptors:** Use the activations from a fully connected layer as a single vector representing the entire image.
- **Advantage:** More robust to viewpoint and condition changes.







Deep Learning Based Loop-Closing

NetVLAD

VLAD (Vector of Locally Aggregated Descriptors): An improvement over BoW that aggregates the residuals of features with their cluster centers.

•**NetVLAD:** A learnable version of VLAD implemented as a CNN layer.

- **End-to-End Trainable:** The entire network (feature extraction + VLAD aggregation) is trained for the specific task of place recognition.
- **Superior Performance:** Became the new state-of-the-art, significantly outperforming previous methods.

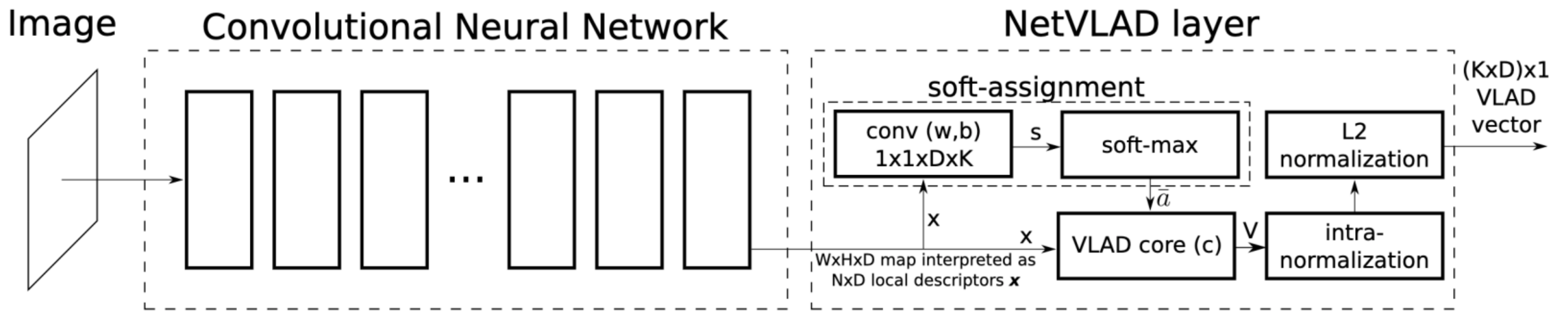


Figure 2. **CNN architecture with the NetVLAD layer.** The layer can be implemented using standard CNN layers (convolutions, softmax, L2-normalization) and one easy-to-implement aggregation layer to perform aggregation in equation (4) ("VLAD core"), joined up in a directed acyclic graph. Parameters are shown in brackets.

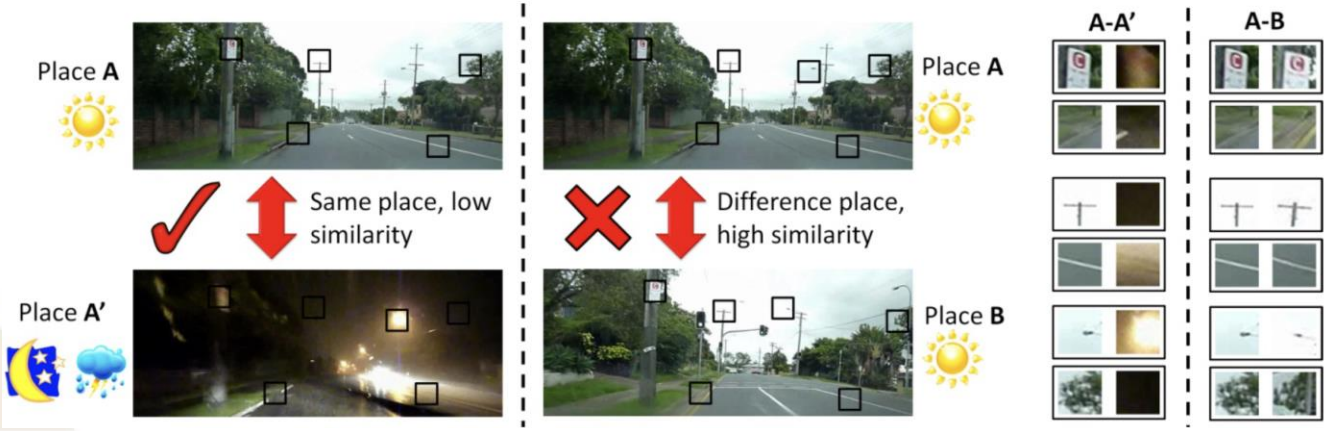
Sequence-Based Methods

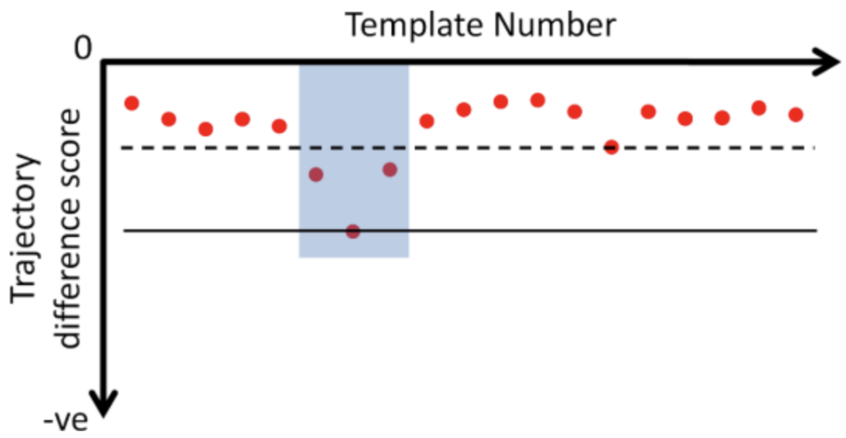
The Problem: A single image might be ambiguous (e.g., two different intersections might look similar).

The Solution: Use a sequence of images (temporal consistency).

Instead of matching one query image, match a short sequence of recent images against sequences in the database.

Typical Algorithms: SeqSLAM, HMM-based methods.





Handcrafted vs. Learning-Based

Handcrafted (BoW + SIFT/ORB):

- *Pros:* Interpretable, doesn't require large training sets, fast (especially ORB).
- *Cons:* Less robust to appearance change, performance plateaus.
- *Best For:* Controlled environments, short-term loops, resource-constrained systems.

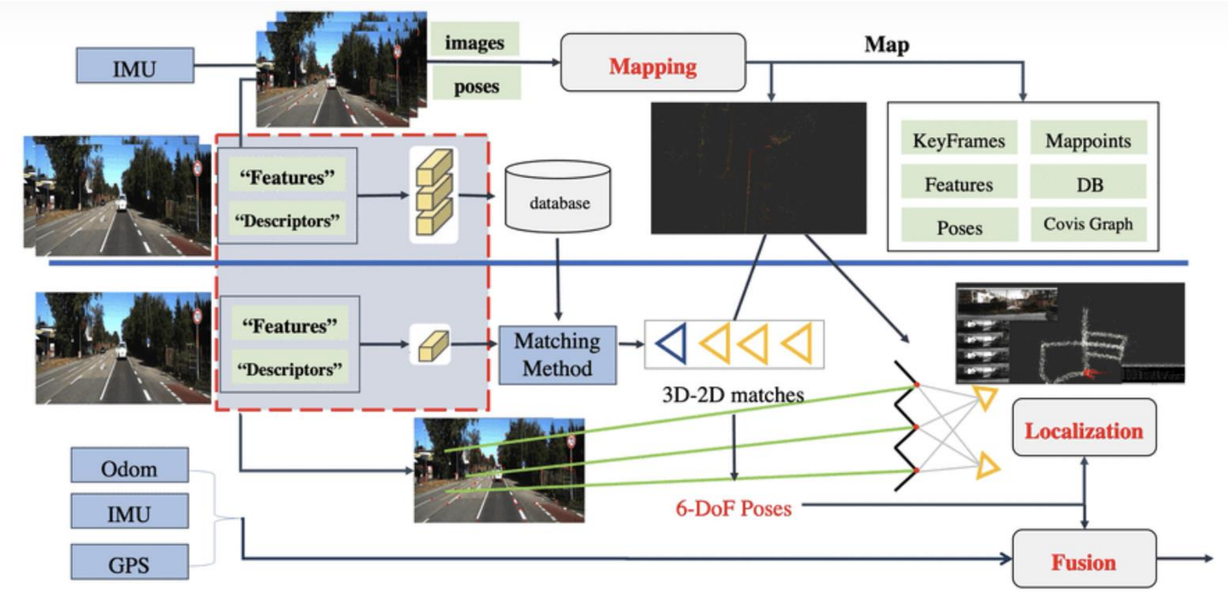
Learning-Based (CNN, NetVLAD):

- *Pros:* Highly robust to appearance change, state-of-the-art performance.
- *Cons:* Requires large datasets for training, "black box," computationally heavier.
- *Best For:* Long-term autonomy, challenging environments (cross-season).

6-DoF Pose Estimation

- **Loop Closing is a two-step process:**
  - **Place Recognition:** "I've been here before." (This lecture's focus)
  - **Geometric Verification:** "Where exactly am I relative to before?”

- **Geometric Verification:**
  - Use feature matching (e.g., with RANSAC) to compute the relative pose (6-DoF: x, y, z, roll, pitch, yaw) between the current view and the matched place.
  - This relative pose is the "constraint" fed into the pose graph optimizer.



6-DoF Pose Estimation

PoseNet: A Convolutional Network for Real-Time 6-DOF Camera Relocalization (2015)

$$loss(I) = \|\hat{\mathbf{x}} - \mathbf{x}\|_2 + \beta \left\| \hat{\mathbf{q}} - \frac{\mathbf{q}}{\|\mathbf{q}\|} \right\|_2$$

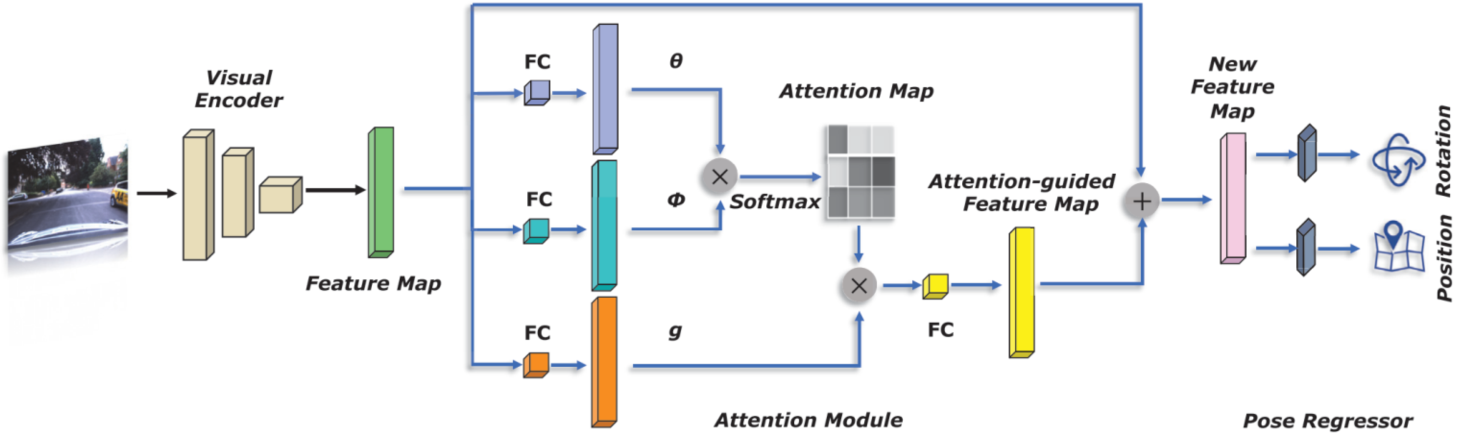
$$\mathbf{p} = [\mathbf{x}, \mathbf{q}]$$

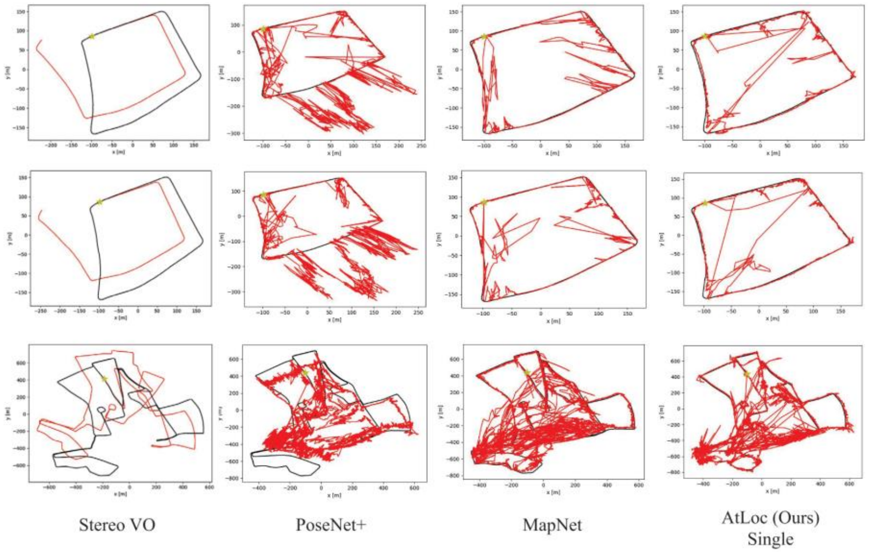
Figure 5: **7 Scenes dataset** example images from left to right; Chess, Fire, Heads, Office, Pumpkin, Red Kitchen and Stairs.

Scene	# Frames Train	# Frames Test	Spatial Extent (m)	SCoRe Forest (Uses RGB-D)	Dist. to Conv. Nearest Neighbour	PoseNet	Dense PoseNet
King's College	1220	343	140 x 40m	N/A	3.34m, 2.96°	1.92m, 2.70°	1.66m, 2.43°
Street	3015	2923	500 x 100m	N/A	1.95m, 4.51°	3.67m, 3.25°	2.96m, 3.00°
Old Hospital	895	182	50 x 40m	N/A	5.38m, 4.51°	2.31m, 2.69°	2.62m, 2.45°
Shop Façade	231	103	35 x 25m	N/A	2.10m, 5.20°	1.46m, 4.04°	1.41m, 3.59°
St Mary's Church	1487	530	80 x 60m	N/A	4.48m, 5.65°	2.65m, 4.24°	2.45m, 3.98°
Chess	4000	2000	3 x 2 x 1m	0.03m, 0.66°	0.41m, 5.60°	0.32m, 4.06°	0.32m, 3.30°
Fire	2000	2000	2.5 x 1 x 1m	0.05m, 1.50°	0.54m, 7.77°	0.47m, 7.33°	0.47m, 7.02°
Heads	1000	1000	2 x 0.5 x 1m	0.06m, 5.50°	0.28m, 7.00°	0.29m, 6.00°	0.30m, 6.09°
Office	6000	4000	2.5 x 2 x 1.5m	0.04m, 0.78°	0.49m, 6.02°	0.48m, 3.84°	0.48m, 3.62°
Pumpkin	4000	2000	2.5 x 2 x 1m	0.04m, 0.68°	0.58m, 6.08°	0.47m, 4.21°	0.49m, 4.06°
Red Kitchen	7000	5000	4 x 3 x 1.5m	0.04m, 0.76°	0.58m, 5.65°	0.59m, 4.32°	0.58m, 4.17°
Stairs	2000	1000	2.5 x 2 x 1.5m	0.32m, 1.32°	0.56m, 7.71°	0.47m, 6.93°	0.48m, 6.54°

6-DoF Pose Estimation

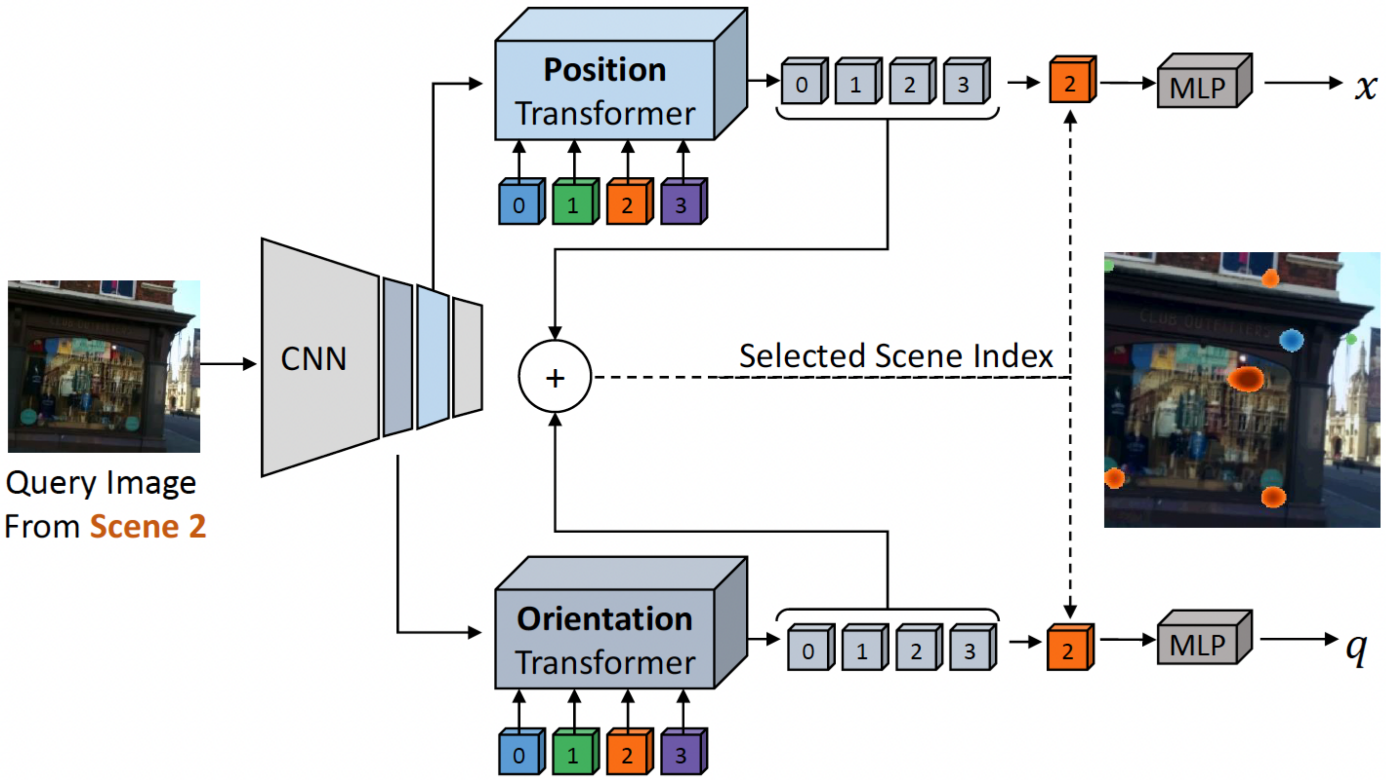
AtLoc: Attention guided camera localization (2020)





6-DoF Pose Estimation

Learning Multi-Scene Absolute Pose Regression with Transformers (2022)



Method	Average [m/deg]	Ranks
Single-scene APRs		
PoseNet [17]	0.44/10.4	10/11
BayesianPN [15]	0.47/9.81	11/8
LSTM-PN [35]	0.31/9.86	8/9
GPoseNet [8]	0.31/9.95	8/8
PoseNet-Learnable [16]	0.24/7.87	7/4
GeoPoseNet [16]	0.23/8.12	5/5
MapNet [7]	0.21/7.78	4/3
IRPNet [29]	0.23/8.49	5/7
AttLoc [36]	0.20/7.56	2/2
Multi-scene APRs		
MSPN [3]	0.20/8.41	2/6
MS-Transformer (Ours)	**0.18/ 7.28**	**1/1**

Future Trends & Research Directions

1) **Semantic Place Recognition:** Use object detectors (e.g., for buildings, trees, traffic signs) to create a semantic description of a place, which is more invariant to weather/season.

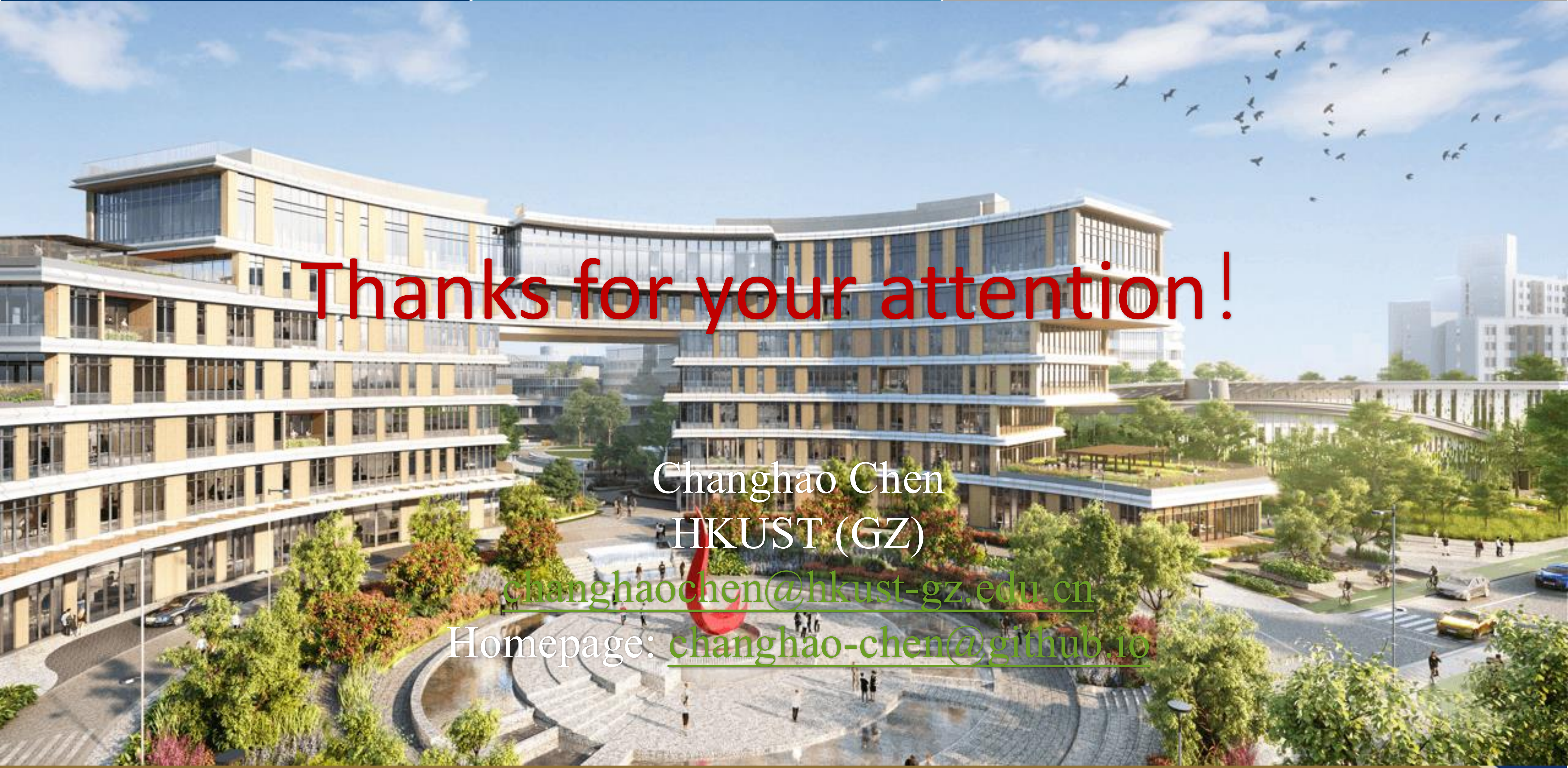
2) **Multi-Modal Fusion:** Combine cameras with LiDAR or Radar. LiDAR's 3D structure is largely invariant to lighting.

3) **Dynamic Object Removal:** Use CNNs to segment out dynamic objects before generating the place descriptor.

4) **Lifelong Learning:** Updating the map and place recognition database over time to account for permanent changes.

5) **Extreme Efficiency:** Making NetVLAD-like models run on embedded vehicle hardware.

6) **Cross-Modal Retrieval:** "Find this place based on a satellite image" or "based on a textual description."



Thanks for your attention！

Changhao Chen
HKUST (GZ)
changhaochen@hkust-gz.edu.cn
Homepage: changhao-chen@github.io