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

Scene Representation

Graduate Course INTR-6000P

Week 7 - Lecture 14

Changhao Chen
Assistant Professor
HKUST (GZ)

Scene Representation

Definition: An internal model describing the structure, semantics, and dynamics of the environment.

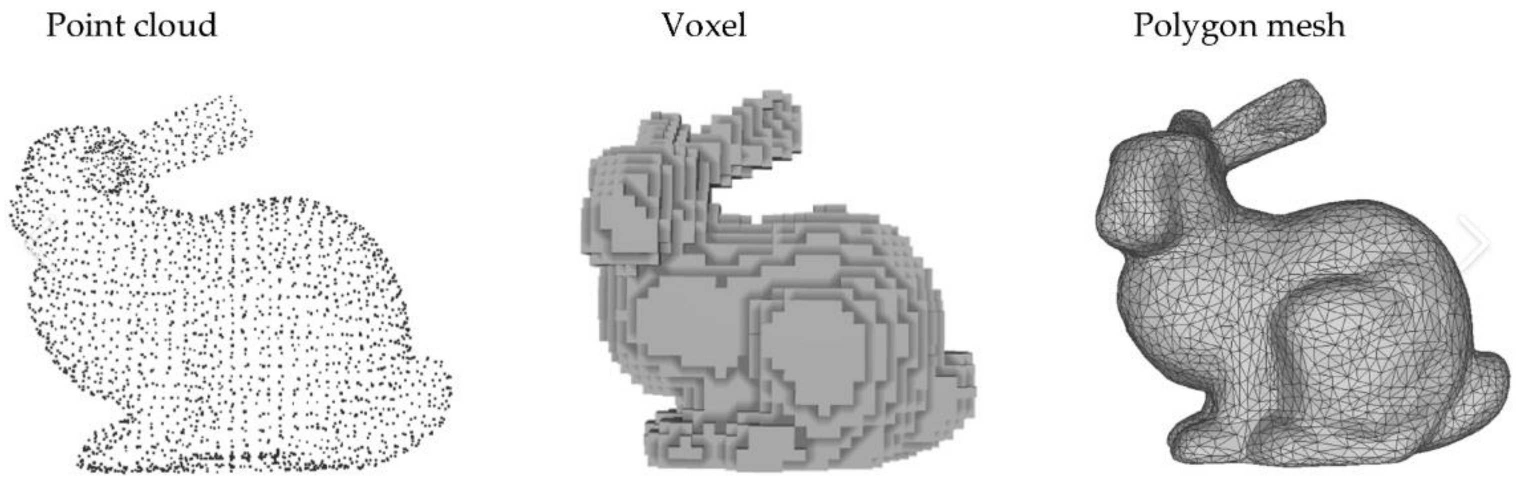
Represents geometry, topology, and meaning.
Must support localization, obstacle avoidance, and decision-making.

Desired Properties:

•Accuracy: Faithful geometric reconstruction
•Compactness: Efficient memory use
•Consistency: Coherent over time
•Interpretability: Supports semantic reasoning
•Real-time update: Suitable for dynamic scenes

Scene Representation

Geometric representations describe the **spatial structure** of the scene.



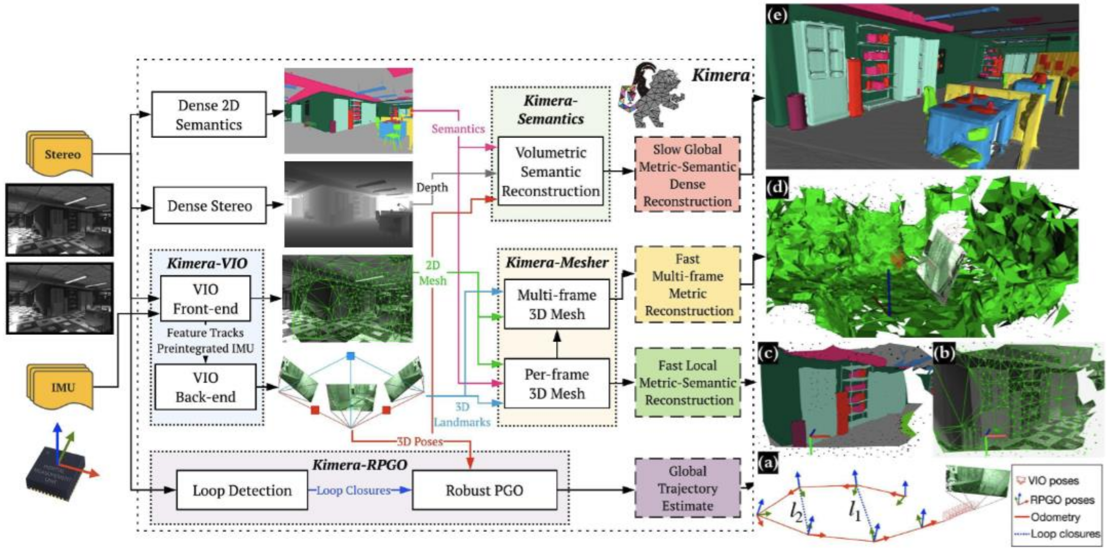
Common forms:
1.Point clouds
2.Meshes
3.Voxel grids
4.Signed distance fields (SDFs)
5.Neural Scene Representations

Point Clouds

- Collection of 3D points sampled from scene surfaces.
- Directly obtained from LiDAR or stereo triangulation.
- Pros: Simple, flexible, accurate.
- Cons: Unstructured, memory-intensive, lacks topology.

Mesh-Based Representations

•Define surfaces using **vertices**, **edges**, and **faces**.
•Generated from point clouds via **Delaunay triangulation** or **Poisson reconstruction**.
•Pros: Compact, visually interpretable.
•Cons: Difficult to update dynamically; limited scalability.



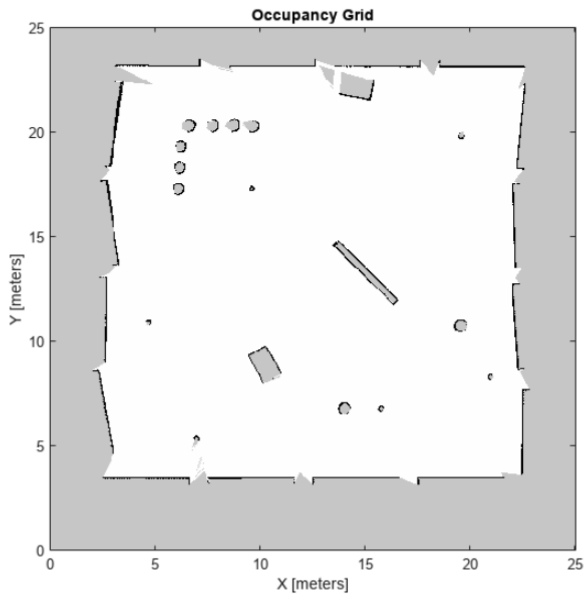
Volumetric Representations

•Environment represented as voxel grid or **Signed Distance Field (SDF)**.
•Used in **KinectFusion**, **ElasticFusion**, **TSDF-SLAM**.
•Each voxel stores occupancy or distance to nearest surface.
•Enables continuous surface extraction via zero-crossing.

The Mapping Problem

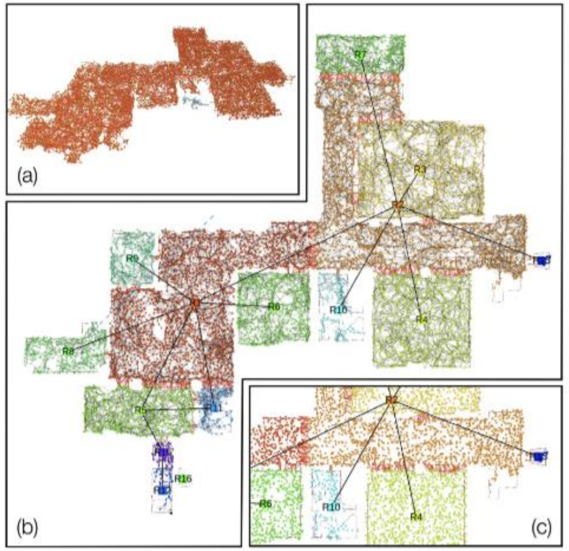
Occupancy Grids

- Divide space into discrete cells; each cell stores occupancy probability.
- Key in 2D/3D mapping and path planning.
- Update rule: Bayesian fusion of sensor measurements.
- Used in autonomous driving for free-space estimation.



Topological Maps

- Represent scene as a **graph** of nodes (places) and edges (connectivity).
- Compact representation suitable for large-scale navigation.
- Example: Pose graph in SLAM, lane graph in HD maps.

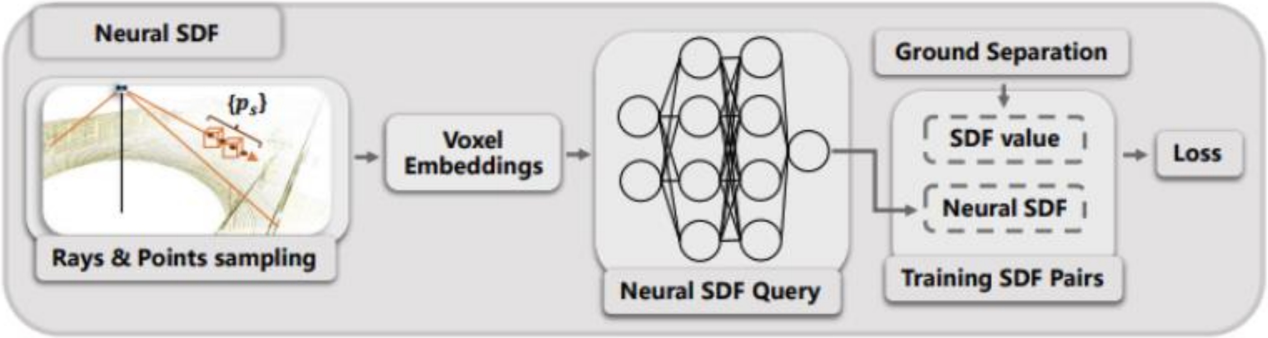


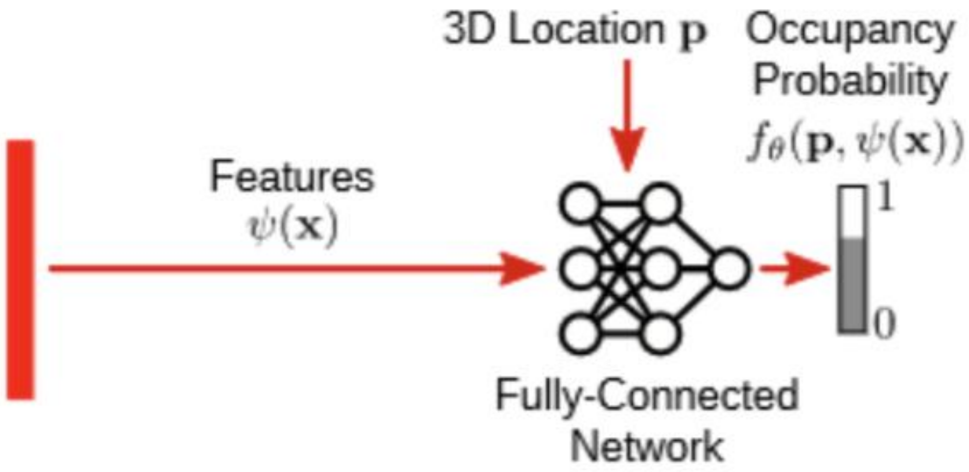
Neural Implicit Representations

Encode 3D scenes as continuous neural fields.
Implicit function

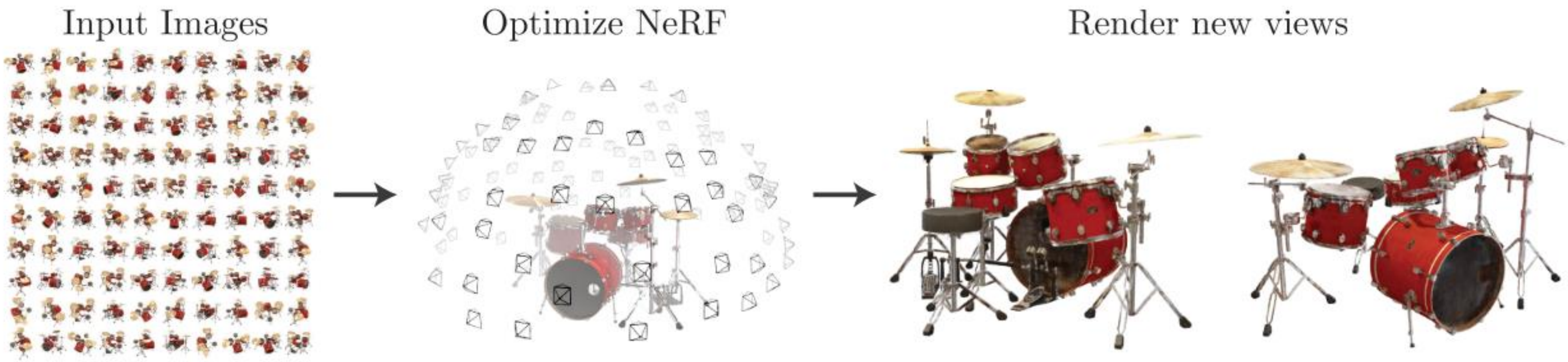
$$f_\theta(\mathbf{x}) \rightarrow$$ color, density, occupancy.

Smooth, memory-efficient, and differentiable.





Neural Rendering



NeRF: Represent scenes as continuous radiance and density fields.
Learn by minimizing photometric loss across views.
Enable high-fidelity rendering from new viewpoints.

$$F_\theta : (\mathbf{x}, \mathbf{d}) \to (\mathbf{c}, \sigma)$$

where

- $\mathbf{x}$: 3D location
- $\mathbf{d}$: viewing direction
- $\mathbf{c}$: RGB color
- σ: volume density

Neural Rendering

1. Volume Rendering Principle

Each pixel's color is computed by integrating the emitted radiance and transmittance along the camera ray:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\,\sigma(\mathbf{r}(t))\,c(\mathbf{r}(t),\mathbf{d})\,dt$$

2. Training Objective

•Rendered color is compared with ground truth image pixel color

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$$

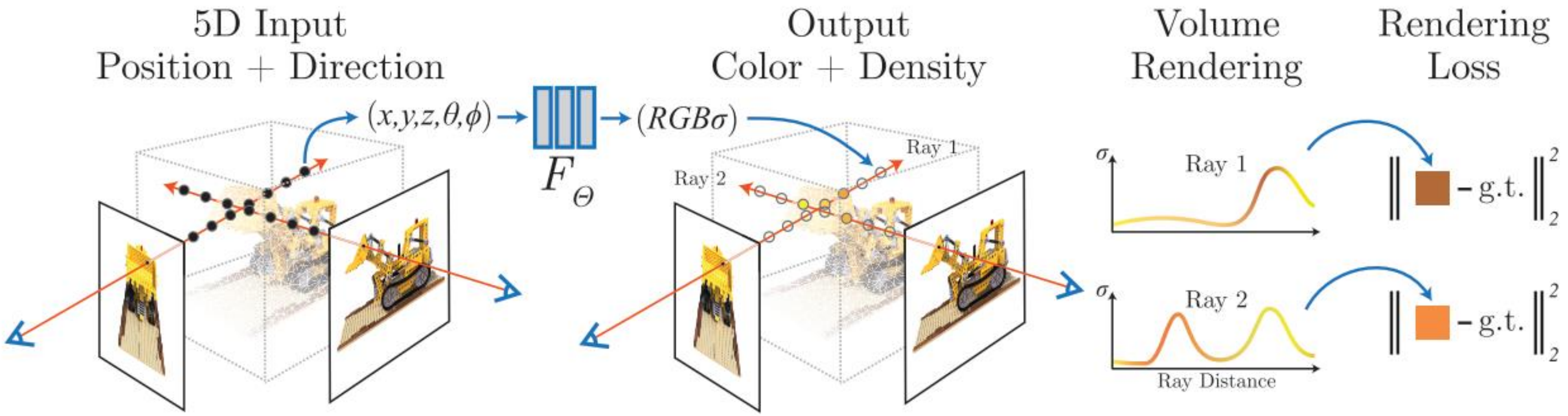
•**Loss:**

$$\mathcal{L} = \sum_{\mathbf{r}} \left\|C(\mathbf{r}) - C^*(\mathbf{r})\right\|^2$$

3. Optimization

•Train MLP parameters θ using multiple images with known camera poses.

•The network learns to encode scene geometry (via density) and appearance (via color).



Neural Rendering

Extensions:

NeRF-SLAM (2022): Jointly optimizes camera poses and radiance field → implicit dense mapping.
iMAP, NICE-SLAM: Real-time incremental NeRF-based SLAM systems.
NeRF++ / Instant-NGP: Improve rendering efficiency with spatial decomposition and hash encoding.
Dynamic NeRF (D-NeRF): Models temporal variations for moving scenes.

Mapping and Tracking Framework

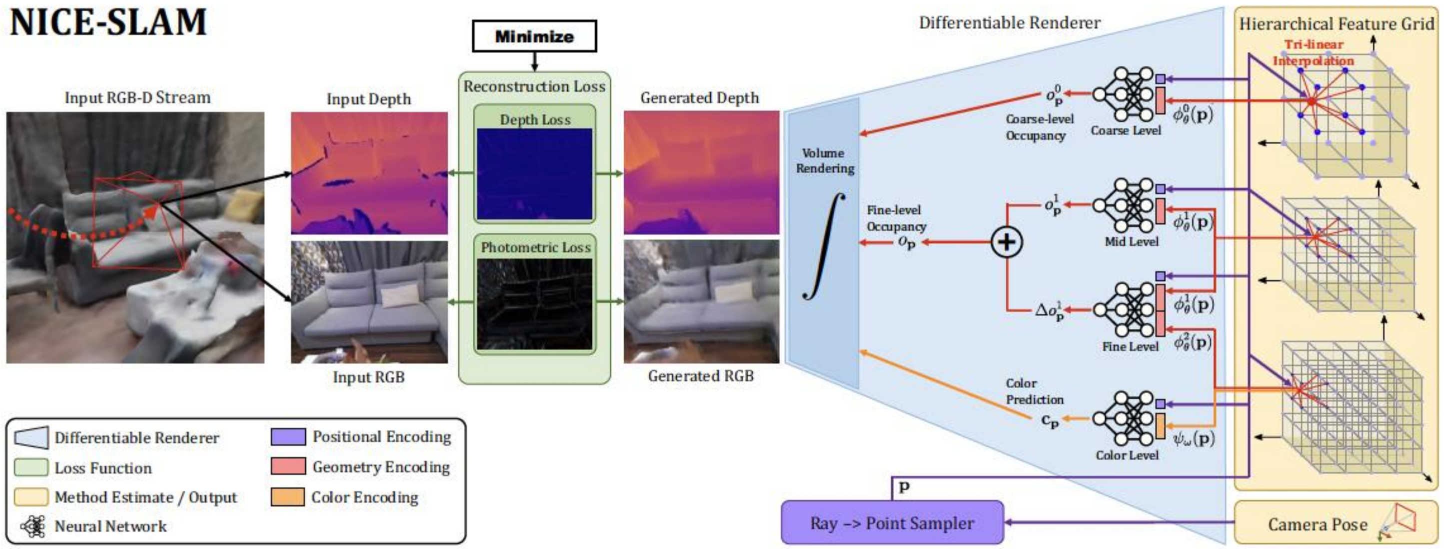
The SLAM system alternates between **mapping** and **tracking**:

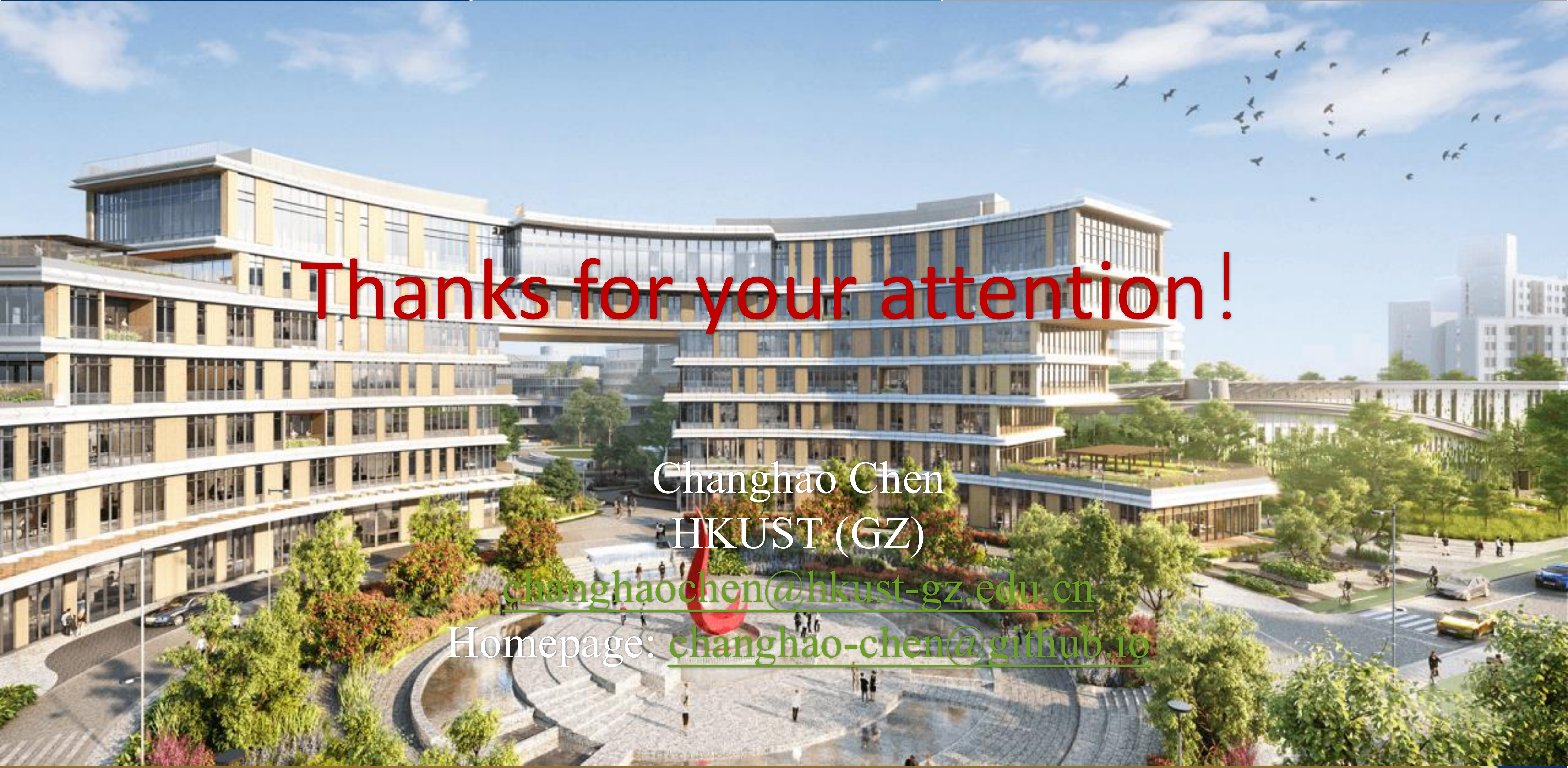
•**Mapping Phase**

- Uses the current RGB and depth images as *ground truth*.
- Predicts RGB and depth from the current feature grids via rendering.
- Constructs **photometric** and **geometric loss functions** between predicted and measured data.
- Jointly optimizes **camera poses** and **feature grid values**.

•**Tracking Phase**

- Given the current feature grids and an estimated pose, renders predicted RGB and depth.
- Compares them with the actual camera observations.
- Optimizes **camera pose** by minimizing rendering errors.





Thanks for your attention！

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