RT-NeRF: Real-Time On-Device Neural Radiance Fields Towards Immersive AR/VR Rendering

Chaojian Li, Sixu Li, Yang (Katie) Zhao, Wenbo Zhu, and Yingyan (Celine) Lin

Georgia Institute of Technology

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

ICCAD'22

EFFICIENT AND INTELLIGENT COMPUTING

NeRF: A Tool to Generate Novel Views

- Neural Radiance Fields (NeRF) can generate **arbitrary new views** of a specific scene **given sparsely sampled scene images**

Inputs: Sparsely sampled views

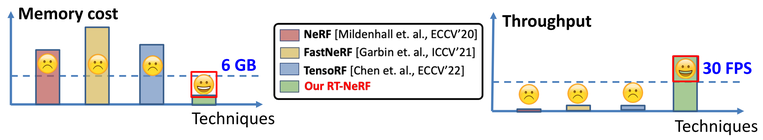
Outputs: Images of any new view

Real-Time NeRF Is Increasingly Demanded

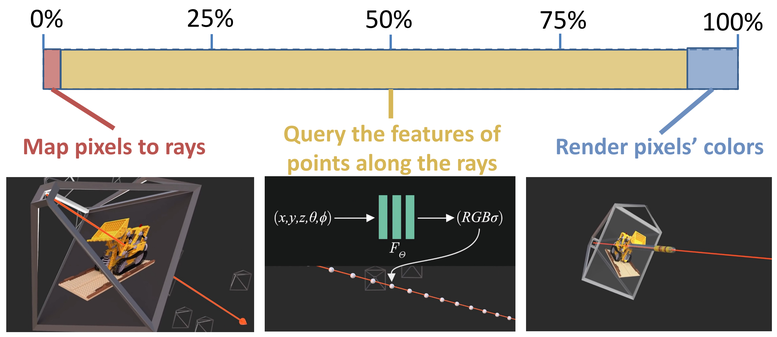
Virtual Meetings | Metaverse | Autonomous Driving Simulation

SOTA Efficient NeRF's Limitations

- **Limitation 1**: **Large memory** requirement
- **Limitation 2**: **Low throughput**



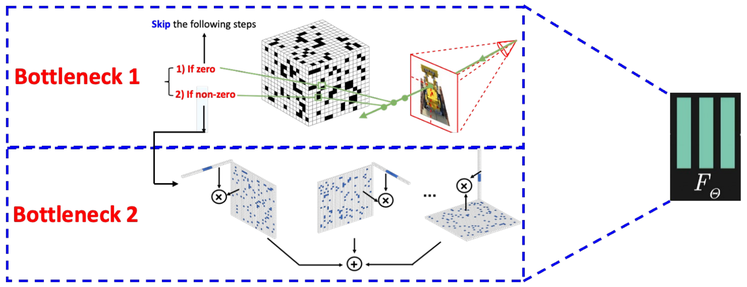
Contribution 1: Analyze the Efficiency Bottleneck



Source: [Mildenhall et. al., ECCV'20]

Contribution 2: Identify Two Key Bottlenecks

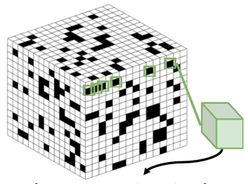
- Dominant step: Query the features of points along the rays
  - **Bottleneck 1 - Locate pre-existing points**
  - **Bottleneck 2 - Compute points' embeddings**



Overview of the Proposed RT-NeRF

To Alleviate Bottleneck 1
(Locate the Pre-Existing Points)

Propose a New Efficient Rendering Pipeline

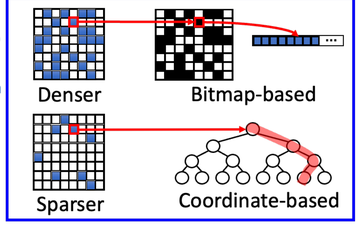


Only query points in the cube

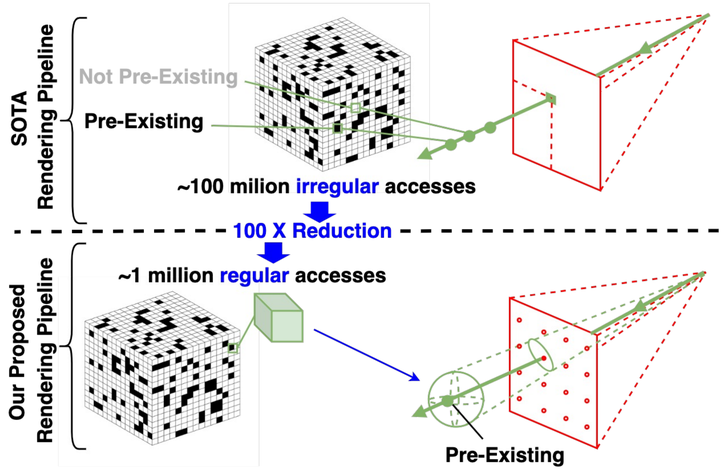
Co-Design

To Alleviate Bottleneck 2
(Compute Points' Embeddings)

Propose a Hybrid Sparse Encoding Scheme & Bi-Direction Trees



Contribution 3: Efficient Rendering Pipeline



Contribution 4: Hybrid Sparse Encoding

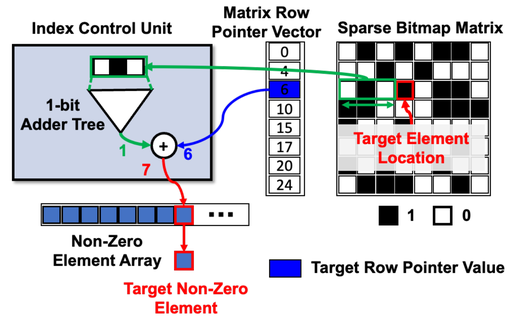
- For **dense** (< 80% sparsity) matrices

Encoding Scheme	Storage Size (↓)	Decoding Throughput (↑)	Resource Utilization (↑)
Bitmap -based	★★★	★	★★★
Our proposed	★★★	★★★	★★★

- For **sparse** (≥ 80% sparsity) matrices

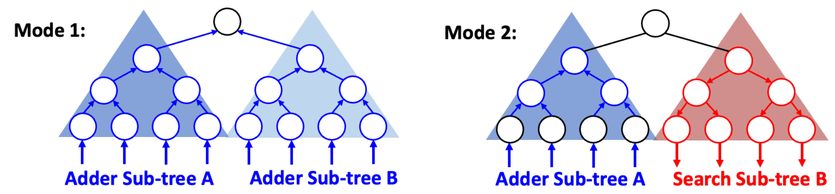
Encoding Scheme	Storage Size (↓)	Decoding Throughput (↑)	Resource Utilization (↑)
Coordinate -based	★★★	★★★	★
Our proposed	★★★	★★★	★★★

Contribution 5: Improved Bitmap-Based Encoding Scheme to Boost Throughput

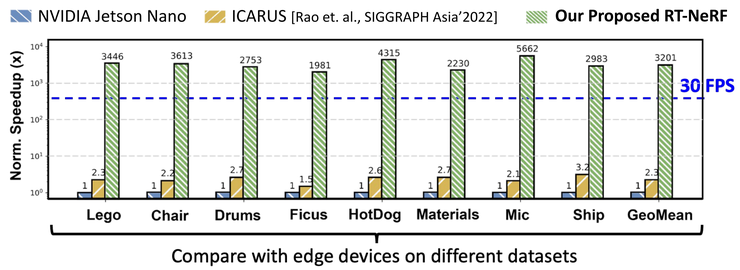


- **Cycle 1:** Check the bitmap matrix element 1 or 0
- **Cycle 2:** Sum up 1-bit bitmap vector and then add the row pointer value
- **Cycle 3:** Fetch the target non-zero element

Contribution 6: Bi-Direction Trees to Boost Utilization



RT-NeRF's Speedup Over Baselines



Our RT-NeRF framework has delivered **the first real-time** neural rendering solution suited for edge applications