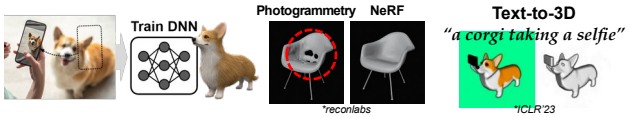
NeuGPU: A Neural Graphics Processing Unit for Instant Modeling and Real-Time Rendering on Mobile AR/VR Devices

Junha Ryu, Hankyul Kwon, Wonhoon Park, Zhiyong Li, Beomseok Kwon, Donghyeon Han, Dongseok Im, Sangyeob Kim, Hyungnam Joo, Minsung Kim, and Hoi-Jun Yoo (KAIST, Daejeon, South Korea)

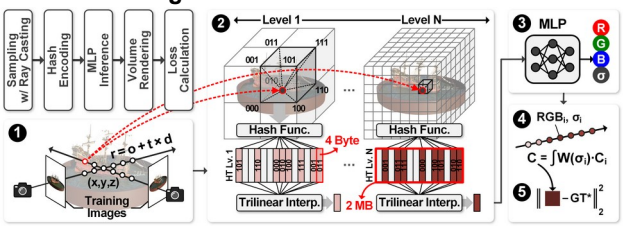
Motivation

- **Why 3D Modeling using Neural Radiance Field?**

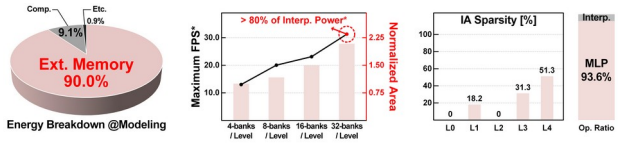


1) **Easy to Make:** Only DNN training w/o high-cost scanner
2) **Photo-realistic:** Robust to transparent & featureless surface
3) **Versatility:** Applied to various down-stream AI tasks

- **3D Modeling Process w/ NeRF**



- **HW Challenges for Real-time Modeling at the Edge**

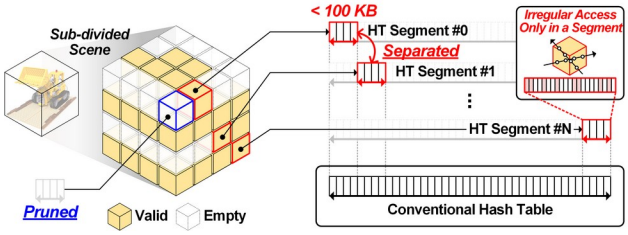


- ***Irregular*** access by hash func. & Large ***MLP Comp.***

➔ ***> 10min.*** modeling & ***< 1-FPS*** rendering on edge GPU

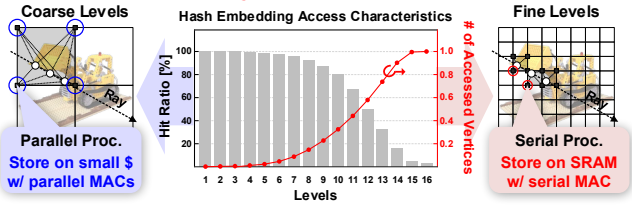
Proposed SW/HW Co-Design

1) Segmented Hashing w/ Spatial Pruning (HT Read)



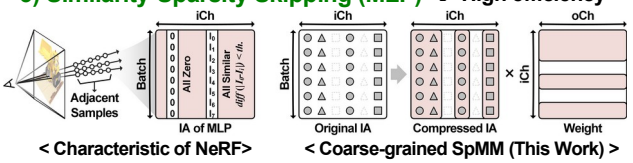
➔ **Eliminates irregular off-chip** access by partitioning HT & Pruning

2) Attention-based Hybrid Interpolation (Trilinear Interp.)



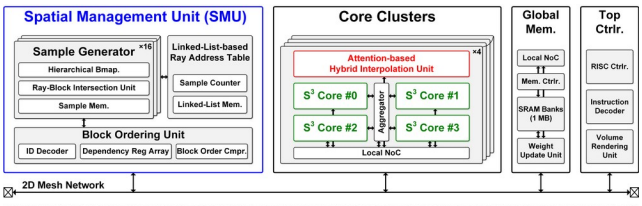
➔ **Power reduction** by mem. access optimization of multi-level HTs

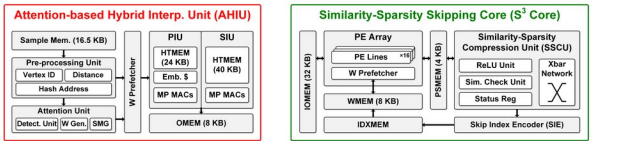
3) Similarity-Sparsity Skipping (MLP) ➔ High efficiency



Key Hardware Blocks

- **Overall Architecture of NeuGPU**





- **Key Building Blocks of NeuGPU**

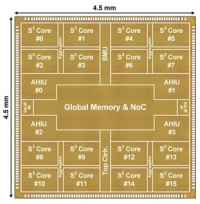
1) Spatial Management Unit (SMU)
- **66%▼** ***off-chip*** memory access supporting SHSP

2) Attention-based Hybrid Interpolation Unit (AHIU)
- **1.51×▲** throughput by skipping far vertices
- **56.4%▼** power by ***hybrid*** Interpolators (PIU & SIU)

3) Similarity-Sparsity Skipping Core (S³ Core)
- **1.61×▲** energy-efficiency & **1.41×▲** throughput by ***coarse-grained*** similarity/sparsity skipping

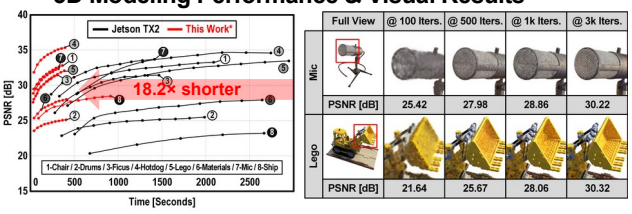
Measurement Results

- **Chip Summary and Performance Comparison**



	Jetson TX2	ISSCC'23	VLSI'23	This Work
NeRF ASIC	X	O	O	O
Modeling Support	△	X	X	O
Technology [nm]	16	28	28	28
Die Area [mm²]	-	20.25	5.07	20.25
Voltage [V]	-	0.6-0.95	1.0	0.68-0.9
Max Frequency [MHz]	1400	250	200	200
Ext. Memory BW [GB/s]	59.7	-	-	1.9
Modeling Performance				
Modeling Speed [Iters/s]	4.1	×8.74▲		35.84
Modeling Time for 30 dB	10m19s	×18.2▼		34s
Modeling Energy [mJ/iter]	4282	×231.4▼		18.5
Modeling Power [mW]	15000	×26.8▼		559.5
Rendering Performance				
Rendering Speed [FPS]	0.48	108.5	30.7	73.5
Rendering Power [mW]	15000	899	129.8	728.4

- **3D Modeling Performance & Visual Results**



- **Demonstration System: 3D Capture of Real World**