Course Overview

This year's course (Spring 2025) will emphasize applications of GPU acceleration to the latest advances in machine learning-based 3D reconstruction and generation, including topics of neural rendering and neural radiance fields (NeRF), 3D Gaussian Splatting (3DGS), and key models for 3D point cloud understanding.
We will study and consider techniques (building on foundational material we've studied in the previous quarters in GPU Computing and SciVis) for leveraging GPU hardware and architecture to parallelize and accelerate these and other SOTA 3D learning-based methods.

1) CUDA-GL Interop : CUDA – OpenGL Interoperability API
    a) How to use both OpenGL and CUDA and exchange data/memory without making host-device round trips. This will involve learning some more advanced CUDA and OpenGL API techniques:
2) GPU HW Architecture
3) Understanding & Programming Tensor Cores (NVIDIA dedicated GEMM HW)
    a) CUDA Warp matrix API
4) Profiling and analysis of GPU performance and tuning/optimization
    a) Inspect and analyze dedicated GPU hardware counters
    b) Understand interplay of device memory hierarchy, SM warp scheduling, and other factors determining algorithm(s) performance

We will be reviewing example code and applying analysis techniques for several application case studies which illustrate GPU Compute+Graphics interop for massive parallel computation efficiency and interactive visualization

1. In-depth Case Studies include
    1. 3D Point Cloud surface reconstruction, segmentation, and visualization
    2. 3D Gaussian Splatting
    3. Neural rendering
    4. Diffusion-based 3D object generation