My long-term research vision is to build machines that can reconstruct, understand and edit the 2D and 3D visual worlds. To tackle this goal, my recent research has developed ***efficient and editable visual scene representations****, in several frontiers including multi-view stereo, implicit neural representations, 3D reconstruction, and on-device efficient vision.*

1. Visual Scene Editing and Manipulation with Implicit Representations

Stylization for Implicit Representation: While implicit neural representation (INR) reveals multiple advantages (e.g., continuous, compact, etc.) over discrete signals, it is still unknown how we can edit/manipulate these continuous representations. My ECCV'2022 paper [C1] proposes the first unified implicit neural stylization framework. which can edit the appearance of both 2D (SIREN) and 3D (SDF/NeRF) continuous representations. The framework produces faithful stylization results while preserving geometry fidelity by developing a novel self-distillation geometry consistency loss.

Unsupervised Segmentation for NeRF: While existing NeRF-based segmentation methods either require per-view annotations or are confined to synthetic datasets. My latest work, NeRF-SOS [C2], spearheads the progress in self-supervised segmentation in NeRF. By proposing a novel collaborative contrastive loss in both appearance and geometry levels, NeRF-SOS encourages a NeRF backbone to distill compact geometry-aware segmentation clusters from their density fields and the self-supervised 2D visual features. NeRF-SOS obtains convincing segmentation maps for complex real-world indoor and outdoor scenes, without any annotation.

Signal Processing for INR: Existing INR works manipulate their continuous representations via processing on their discretized instance. In my work INSP-Net [C3], we explore directly modifying an INR without explicit decoding. INSP-Net leverages spatial gradients of neural networks, instantiates the signal processing operator as a weighted composition of computational graphs corresponding to the high-order derivatives, to approximate the continuous convolution filters.

Single View NeRF: NeRF is impeded by the requirement of the dense views captured from different angles. My latest work, SinNeRF [C4], trains a NeRF on only one single view. SinNeRF generate pseudo labels in both geometry and appearance levels, yields photo-realistic novel-view synthesis results.

2. Memory and Data-Efficient Visual Scene Modeling

Efficient Multi-view Stereo for High-resolution Images: Recovering the lost dimension from merely 2D images has been the classical goal of multi-view stereo. Recent deep MVS methods construct heavy 3D cost volume to regress the scene depths. However, previous methods are limited when high-resolution inputs are needed since the memory and time costs grow cubically as the volume resolution increases. My CVPR'2020 paper [C5] (oral presentation) introduces a memory and time-efficient cost volume formulation, which is built upon a feature pyramid encoding geometry at gradually finer scales, and narrows the depth search range of each scale by the prediction from the previous stage. This work reduces 50.6% and 59.3% reduction in GPU memory and run-time. It is not only applied to city-scale 3D reconstruction in Alibaba Group but also widely adopted by the subsequent efficient MVS methods.

Efficient Multi-task Learning via Model-Accelerator Co-Design: Multi-task learning (MTL) encapsulates multiple learned tasks (e.g., depth estimation, semantic segmentation) in a single model and often lets those tasks learn better jointly. However, deploying MTL to resource-constrained devices still faces gradient conflict and inefficiency. My work [C6] for the first time adapts mixture-of-experts (MoE) layers into the MTL backbone, along with co-designed hardware innovations. This work significantly reduces the inference FLOPs (88%) and saves more energy ($>$ 9x) than our baselines. It was recognized and awarded by **Qualcomm Innovation Fellowship, 2022**, and **DAC University Demonstration (3rd place).**

Data Efficient Point Cloud Representation via Domain Adaptation: Point clouds are the most straightforward way to preserve 3D spatial information and are close to a number of 3D understanding applications (e.g., autonomous driving, indoor scene parsing). However, most of previous works are supervised learning and therefore require a large amount of annotated data. My paper ECCV'2022 [C7] mitigates the expensive labeling cost by transferring the knowledge from existing labeled source data to unseen unlabeled target data, owing to self-supervision via three-level masked local structure predictions. 11.8% and 5.9% improvements on classification and semantic segmentation are obtained than the baseline method without domain adaptation.

Reference List (due to space limit, please see resume for the more detailed list): [C1] Unified Implicit Neural Stylization (ECCV'2022); [C2] NeRF-SOS: Any-View Self-supervised Object Segmentation on Complex Scenes (preprint'2022); [C3] Signal Processing for Implicit Neural Representations (NeurIPS'2022); [C4] SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image (ECCV'2022); [C5] Cascade Cost Volume for High-Resolution Multi-View Stereo and Stereo Matching (CVPR'2020); [C6] M^{3}ViT: Mixture-of-Experts Vision Transformer for Efficient Multi-task Learning with Model-Accelerator Co-design. (NeurIPS'2022); [C7] Point Cloud Domain Adaptation via Masked Local 3D Structure Prediction (ECCV'2022).