ODIN:

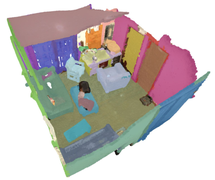
A Single Model for 2D and 3D Segmentation

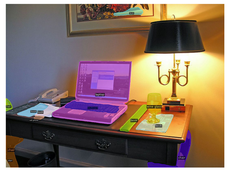
https://odin-seg.github.io
https://github.com/ayushjain1144/odin

Ayush Jain, Pushkal Katara, Nikolaos Gkanatsios, Adam W. Harley,
Gabriel Sarch, Kriti Aggarwal, Vishrav Chaudhary, Katerina Fragkiadaki

CVPR SEATTLE, WA JUNE 17-21, 2024

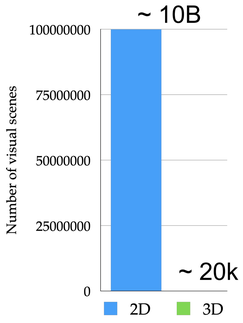
Goal: Given a single RGB image or 3D scene, segment and classify all objects present in the scene





3D and 2D Instance Segmentation with a single model

Unified 2D-3D Architecture is a promising way to scale 3D Learning



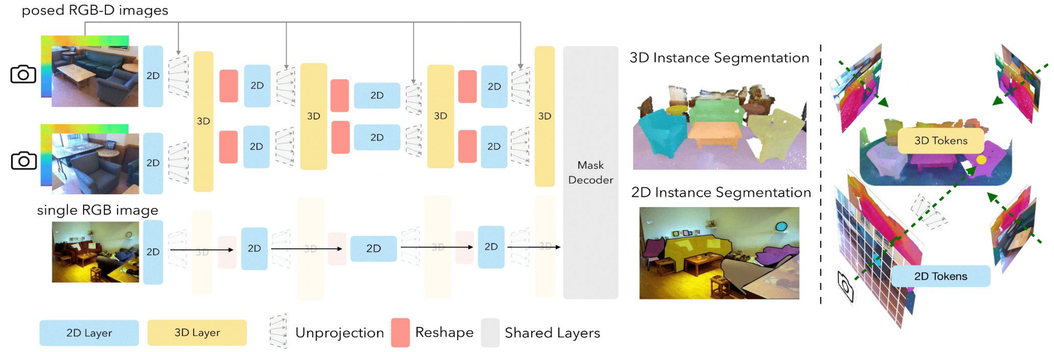
Lack of 3D data is a key reason behind not having a 3D Foundation model yet

Key Ideas:

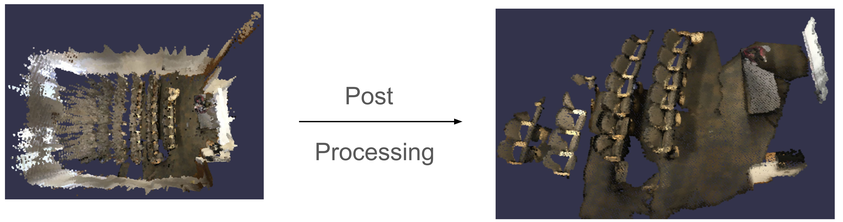
- Make minimal changes in 2D architecture to maximally load pre-trained 2D weights
- Use RGB-D as the 3D representation to make 2D and 3D inputs similar
- Interleave 2D layers with 3D layers for effectively using 3D information
- Use segmentation as a unified 2D-3D output space

ODIN can utilize 2D pre-trained weights and both 2D and 3D data to help 3D learning

ODIN: (Left) Given a posed RGB-D sequence as input, ODIN alternates between a within-view 2D fusion and a cross-view 3D fusion. When the input is a single RGB image, the 3D fusion layers are skipped.
(Right) At each 2D-to-3D transition, ODIN unprojects 2D feature tokens to their 3D locations using sensed depth and camera intrinsics and extrinsics.



Sensor Point Cloud vs Mesh Point Cloud



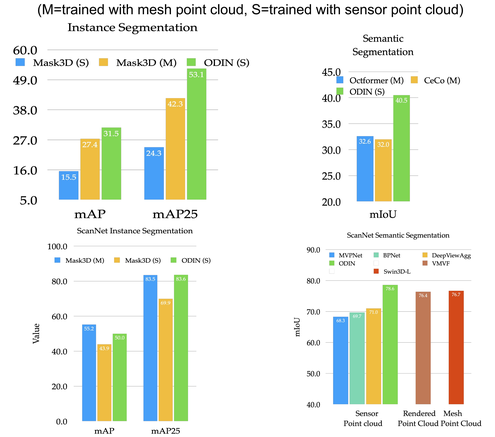
Sensor RGB-D Point Cloud

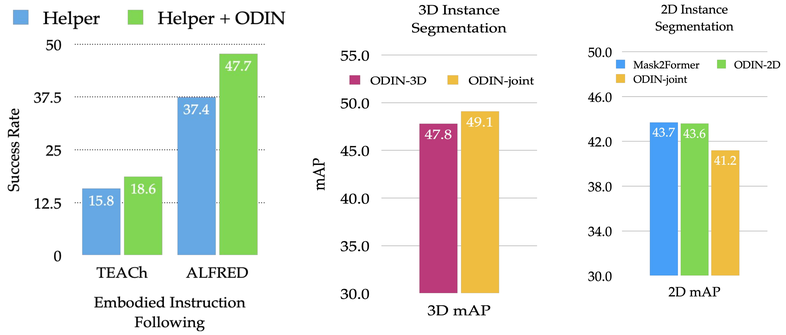
Mesh Point Cloud

Fine-grained misalignments in sensor and mesh point clouds creates major discrepancy in 2D-3D evaluation setups which we systematically quantify

Results: We outperform prior work, especially on long-tailed ScanNet200 benchmark

(M=trained with mesh point cloud, S=trained with sensor point cloud)





Multiview processing of ODIN helps embodied agents

Joint 2D-3D training helps 3D performance

Future Work: 2D-3D cross-domain generalization, scaling up