OpenScene

3D Scene Understanding with Open Vocabularies

Songyou Peng
ETH Zurich and Max Planck Institute for Intelligent Systems

Arxivores @ Stability.ai
June 15, 2023
Who Am I?

• 4th Year PhD Student
  • Marc Pollefeys
  • Andreas Geiger

• Internships during PhD
  • 2021: Michael Zollhoefer
  • 2022: Tom Funkhouser

• Graduate this fall 😊
My PhD Topics: Neural Scene Representations
for 3D reconstruction and 3D scene understanding

- Convolutional Occupancy Nets
  ECCV 2020 (Spotlight)

- Shape As Points
  NeurIPS 2021 (Oral)

- KiloNeRF
  ICCV 2021

- UNISURF
  ICCV 2021 (Oral)

- NICE-SLAM
  CVPR 2022

- OpenScene
  CVPR 2023

- MonoSDF
  NeurIPS 2022

- NICER-SLAM
  arXiv 2023
Input Images

3D Reconstruction
Input 3D Geometry

Traditional Semantic Segmentation

Only train and test on a few common classes
Input 3D Geometry

3D Scene Understanding Tasks w/o Labels

- Affordance prediction
- Material identification
- Physical property estimation
- Rare object retrieval
- Activity site prediction
- Fine-grained semantic segmentation
- Many more...
OpenScene
3D Scene Understanding with Open Vocabularies
CVPR 2023

Songyou Peng  Kyle Genova  Chiyu "Max" Jiang  Andrea Tagliasacchi  Marc Pollefeys  Tom Funkhouser
**Key Idea:** Co-embed 3D features with CLIP features

**CLIP:** Contrastive Language-Image Pre-Training

Radford et al.: Learning Transferable Visual Models From Natural Language Supervision. ICML 2021
Key Idea: Co-embed 3D features with CLIP features
Key Idea: Co-embed 3D features with CLIP features

Note: bold word embeddings are approximate
How to Learn Such Text-Image-3D Co-Embeddings?
Step 1: Multi-view Feature Fusion

3D Geometry

Per-pixel Features
(visualize with PCA)

RGB Images

$f^{2D}$

3D Geometry

RGB Images

$f^{2D}$


Step 2: 3D Distillation

\[ \mathcal{L} = 1 - \cos(f^{2D} - f^{3D}) \]
Step 3: 2D-3D Ensemble

3D Geometry

2D-3D Ensemble Features

Choose the feature with the highest max score among all prompts

2D-3D Ensemble Features (visualize with PCA)

\[
\begin{align*}
    s_{n}^{2D} &= \cos(f_2^{2D}, t_n) \\
    s_{n}^{3D} &= \cos(f_3^{3D}, t_n)
\end{align*}
\]

Select \( f_2^{2D} \) or \( f_3^{3D} \) based on their max score.
Open-Vocabulary, Zero-shot

3D Semantic Segmentation
Results on nuScenes

Input Lidar Points

MinkowskiNet (Fully supervised)

Ours (Zero-shot)

GT Label
Our Zero-shot 3D Segmentation
(20 classes)
Our Zero-shot 3D Segmentation
(160 classes)
Comparison

Most Common Classes

Matterport3D Top 160 Classes

(ranked by number of instances in training set)
Comparison

(mAcc (%)

Matterport3D Top 160 Classes (ranked by number of instances in training set)

- Fully supervised
- Ours

Rarest Classes
Ablation

![Chart showing mIoU (%) for ScanNet and Matterport3D with different fusion methods: 2D Fusion, 3D Distillation, and 2D-3D Ensemble. The results are 41.4%, 46.0%, 47.5%, 41.3%, and 42.6% respectively.](chart.png)
Image-based 3D Scene Query
Our Segmentation

Input 3D Geometry

Image Queries

Our Segmentation

Input 3D Geometry

Given 3D Geometry
Interactive Demo

Open-vocabulary 3D Scene Exploration
Take-home Message

• We enable a **wide range of applications** by open-vocabulary queries

• This can hopefully influence how people train 3D scene understanding systems in the future

• Our real-time demo already shows the **possibility to directly apply to AR/VR**
OpenScene

3D Scene Understanding with Open Vocabularies

- Input 3D Point Cloud
- "fan" - Object
- "made of metal" - Material
- "kitchen" - Room Type
- Zero-shot Semantic Segmentation
- "anything soft" - Property
- "where to sit" - Affordance
- "work" - Activity

pengsongyou.github.io/openscene