COMPUTER VISION PAPER PRESENTATION

By:

Saiteja Juluru

Sai Dinesh Vasireddy

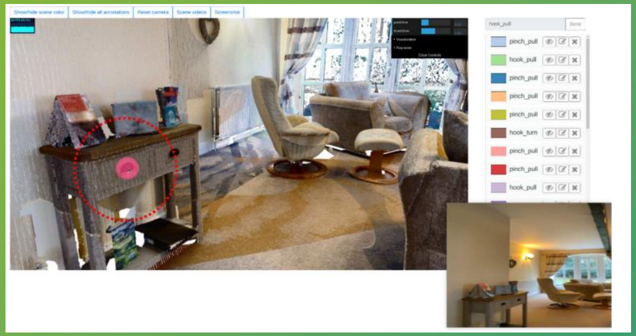
Laxminarayana Vadnala

PAPER TITLE

SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes

Objective of The Research Paper

- To advance the understanding of how machines can interpret and interact with complex 3D environments, by providing a comprehensive dataset and tools for recognizing functional elements and affordances in 3D scenes.
- To Enhance interaction and experience with Virtual Reality/ Augmented Reality



What is Affordance?

- An **affordance** refers to the potential actions or uses that an object or environment offers to an individual, based on its design and properties. It's a concept that originates from psychology and was introduced by James J. Gibson in the 1970s.
- In simpler terms, affordances are the clues or cues that suggest how something can be interacted with. For example:
  - A **door handle** affords **turning or pulling** to open the door.
  - A **button** affords **pressing**.
  - A **chair** affords **sitting**.

DEMO VIDEO

- Demo Video
- Project Information (Git- Repo)

What Does the Research Paper Entail?

This research focuses on **fine-grained functionality and affordance understanding in 3D scenes** through the **SceneFun3D dataset**. The paper presents:

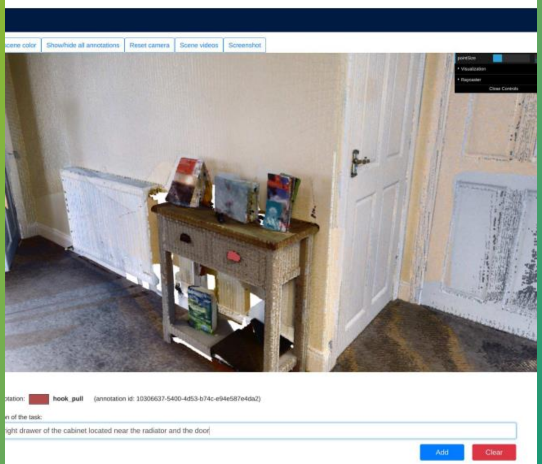
Annotation Framework:
A comprehensive annotation pipeline designed for large, high-resolution 3D point clouds, involving functionality annotations, natural language task descriptions, and motion annotations.

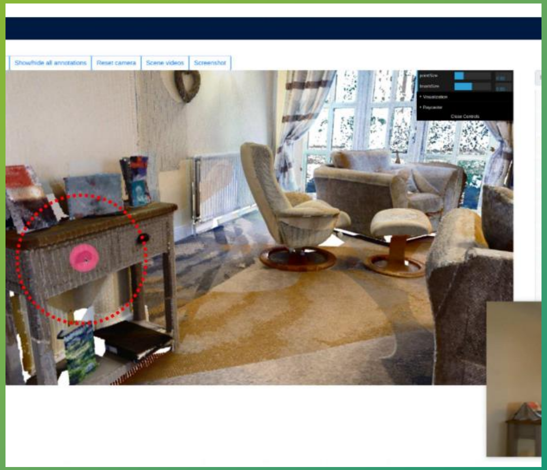
3D Scene Understanding:
It utilizes advanced techniques like **accelerated ray-casting algorithms** for efficient annotation and **semantic understanding** of objects and functional elements in indoor environments.

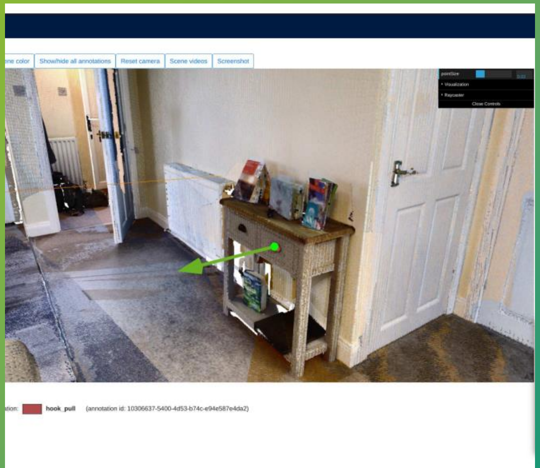
Task-Driven Affordance Grounding:
The paper explores task-specific affordances, using models like **LERF** and **OpenMask3D** for scene interpretation, enabling natural language-based querying and object interaction.

Motion Estimation:
Incorporation of motion parameters into 3D scene understanding, predicting motion types, axes, and origins to model dynamic interactions.

Applications:
The research opens up numerous possibilities for applications in **robotics** (assistive tasks and autonomous interaction) and **augmented reality** (realistic human-scene interactions).







Annotation Types in SceneFun3D

Functionality Annotations

- Identify interactive elements in 3D scenes (e.g., drawers, switches)
- Annotators select affordance category and mask relevant 3D points

Natural Language Task Descriptions

- Describe tasks associated with functional elements (e.g., “Open the cabinet”)
- Collected from human annotators and augmented using ChatGPT

3D Motion Annotations

- Specify how objects move (translation or rotation)
- Include motion type, axis, and origin of motion

Project Goal

Develop a 3D scene understanding system for recognizing **functional affordances** and task-driven actions.

Integrate natural language processing for querying 3D environments and directing task-specific interactions.

Create a system that enables **autonomous robots** or **AR systems** to understand, interpret, and interact with indoor spaces in a **human-like manner**.

Efficient Point Cloud Annotation

High-resolution point clouds require efficient processing for annotation

Accelerated ray-casting algorithm used for fast interaction

Based on **Bounding Volume Hierarchies (BVH)**

3D points grouped into recursive bounding volumes forming a **KD-tree**

Enables **top-down spatial queries**, avoiding naive iteration over all points

Improves annotation speed and lowers computational load

Key Images Mentioned In Research Paper

- **LERF (Language-Embedded Radiance Fields)** combines 3D scene reconstruction (using NeRF) with natural language understanding. It allows machines to interpret and interact with 3D scenes based on text descriptions (e.g., "open the drawer").
- LERF converts text into a language embedding and uses it to adjust the 3D scene, helping robots or virtual agents understand and perform tasks described in natural language.

Figure 10. Robot-scene interaction. The ability to detect functional elements, enables a robot to perform scene interactions such as turning on a light, or opening a drawer.

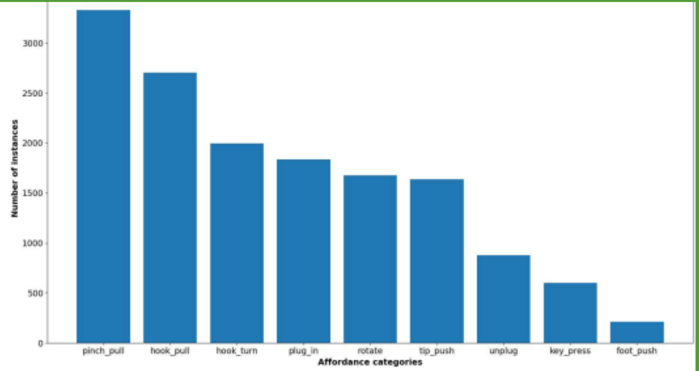


Figure 6. Distribution of affordance categories in the SceneFun3D dataset.

4. Qualitative results on LERF

We also present qualitative results using the LERF model. In Fig. 8 we visualize the response field of LERF on a training frame given a text query.

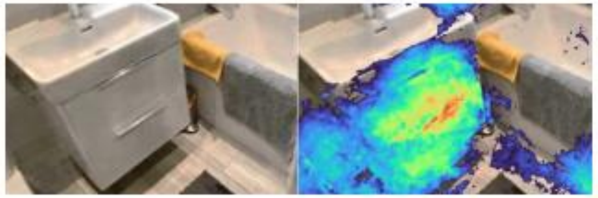


Figure 8. **Response field of LERF** [2]. Query is "Open the bottom drawer." Red means high response, blue low response.

Key Objectives

- **Create 3D Scene Representations from Mobile Photos:**
  - Use **photogrammetry techniques** and depth information from mobile phone cameras to build **3D models** of indoor spaces.
- **Affordance Grounding with Language:**
  - Implement **text-based querying** systems (e.g., CLIP, LERF) to understand affordances in the context of task descriptions.
- **3D Motion Estimation:**
  - Develop **motion models** to predict motion types, directions, and origins within 3D scenes.
- **Task-Specific Interaction:**
  - Enable systems (robots or AR applications) to **understand and act on natural language commands** (e.g., "Open the cabinet door").

What is Point Cloud?

- A point cloud in a 3D image is a collection of data points in a three-dimensional space. Each point represents a specific location in 3D space and may also include attributes like color, intensity, or normal vectors.
- Created using accelerated ray casting and grouped using Bounding Volume Hierarchies (BVH) and enabling with KD tree.

Implementation Steps

Image Capture: Use a mobile phone to capture multiple photos of the scene from different angles.

Photogrammetry with Python: Use libraries like **OpenCV** for feature extraction, camera calibration, and image matching.

Use **React.js and MongoDB** for managing laser scan annotations. Develop 3D interactive interfaces using **Three.js (WebGL-based)**, accessible via a simple web browser.

Scene Understanding: Apply **LERF** (Language-Embedded Radiance Fields) to link 3D models with natural language descriptions, enabling task-driven scene understanding.

Task Execution: Implement functionality to query 3D models based on natural language descriptions (e.g., "open the drawer") using **CLIP** and **OpenCV** for feature matching.

Final Output: 3D scene model with functionality annotations and task descriptions.

Reference

- SceneFun3D: Fine-Grained Functionality and Affordance Understanding in 3D Scenes
- https://scenefun3d.github.io/

THANK YOU AND HAVE A GREAT DAY