

CozmoWander

General Object Recognition & Localization

Zhanxun Liu

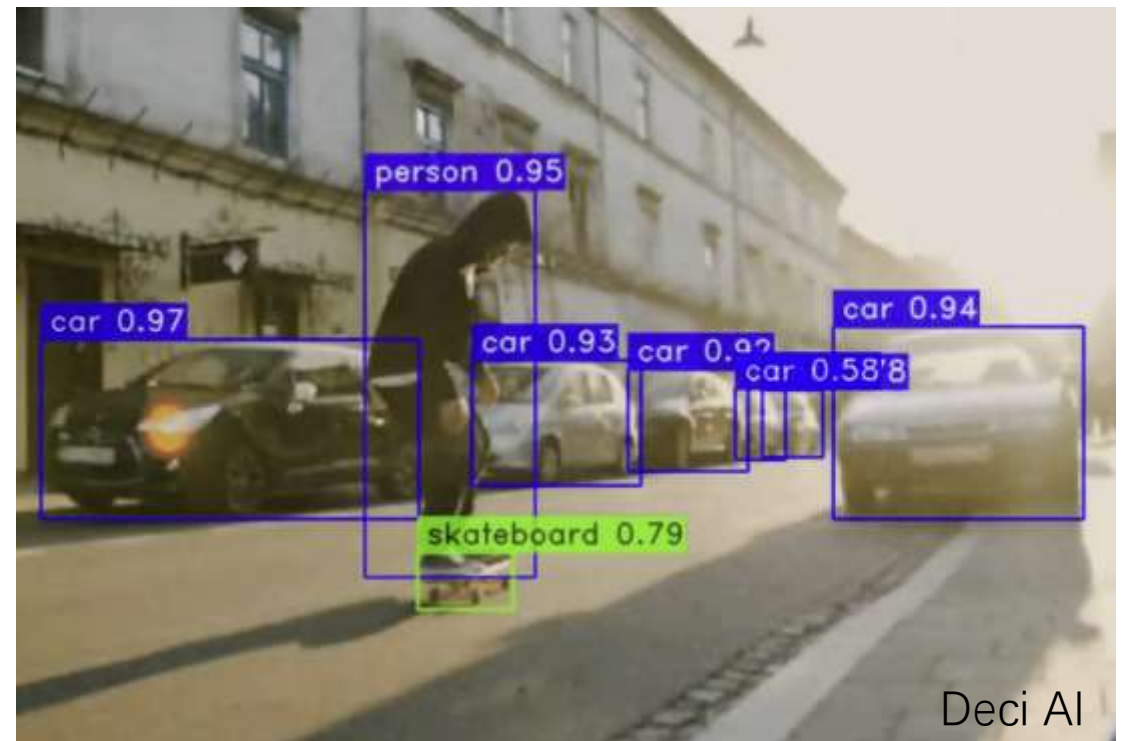
Problem

- Cozmo only knows **cubes** and **Aruco markers**

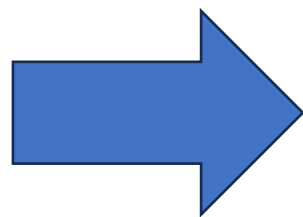


Solution

- ML approach to do **general object recognition and localization**

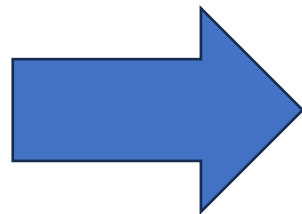


Super-resolution

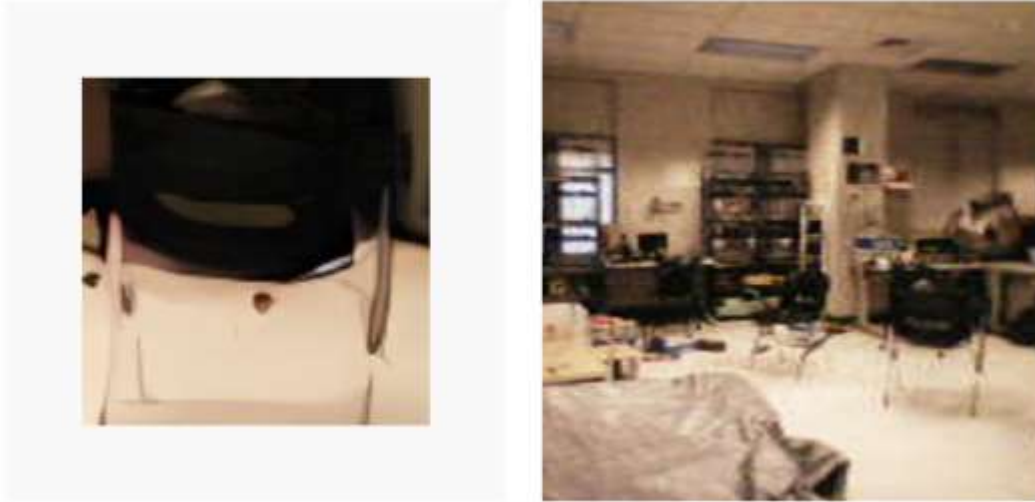


Real-ESRGAN
PRACTICAL RESTORATION

YOLO Object Detection



GPT4-Vision Description Fetch



One image provided is the cropped region of the other image.

The object in the cropped image is a chair.

Describe the object in the cropped image: color, shape and other properties.

Do not mention anything about size.

The original scene is low-resolution, and the scene and object provided are after super-resolution. The user knows this. Therefore, do not mention anything about resolution.

Just provide the description. No anything else.

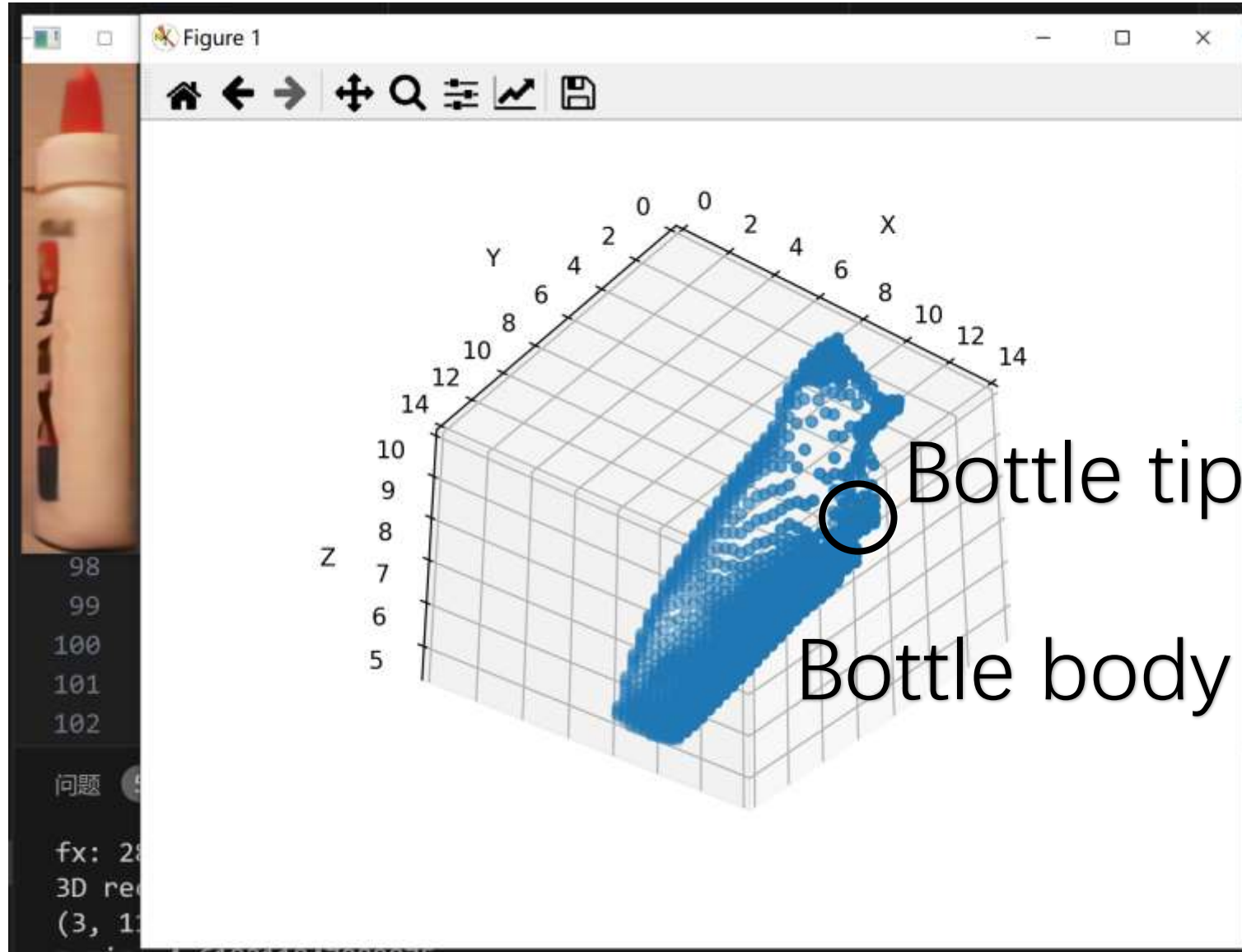
GPT4-Vision Description Fetch



ChatGPT

The object is a chair, black in color. It has a standard structure with four legs supporting it, and both the seat and backrest appear to be made of a soft material. There seems to be a horizontal support at the top of the chair, which could be part of the chair or another object. Behind the chair, there are some blurred lines that may represent details of the chair's back or the silhouette of other objects in the background. The ends of the chair legs are equipped with pads, likely for slip resistance or floor protection. Overall, the chair seems functional and comfortable.

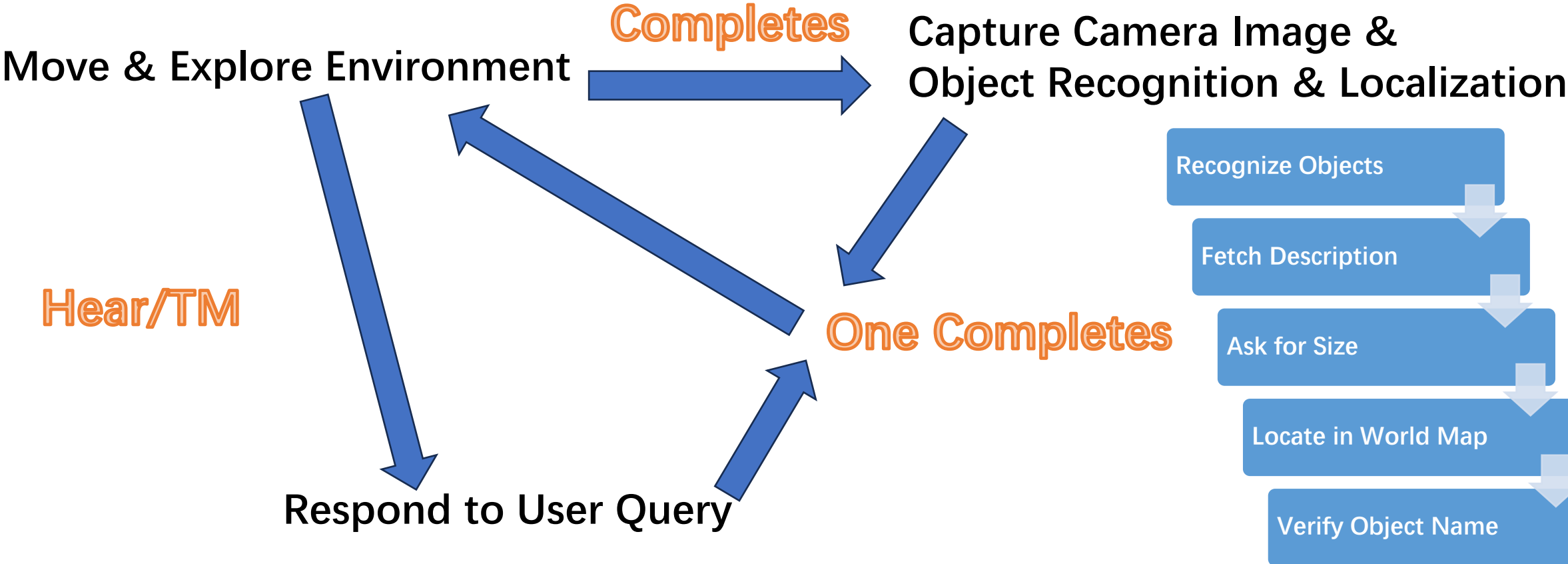
MiDaS Depth Estimation



Camera Space 2 World Space

$$\mathbf{p}_w^h = \begin{bmatrix} R & \mathbf{t} \\ \mathbf{0} & 1 \end{bmatrix} \cdot \mathbf{p}_c^h$$

Integrate into FSM



Human Robot Interaction
with GPT when user says:

What do you know?

The size is 5 centimeters I think.

Object is Iphone.

Tell me a joke.






```
object_detection.py > | ObjectDetection > | get_ops
No. 1928 | | after /bin
from ultralytics import YOLO
import numpy as np
import torch

O recording size...
typing: 'size successfully recorded.'
Checking if all sizes are recorded...
Verifying objects...
3D reconstruction: depth map mean: 0.00599955879519283
3D reconstruction: axial coords x mean: 1150.6
3D reconstruction: axial coords y mean: 775.5
width scale: 1.3112831195028745
fx: 288.87
3D reconstruction: camera space z mean: 1.1244581864405817
Object num: person Position shape: (3,) Depth map shape: (3, 68826)
3D reconstruction: depth map mean: 0.004820111287792398
3D reconstruction: axial coords x mean: 108.5
3D reconstruction: axial coords y mean: 478.6
width scale: 1.1457039190542886
fx: 288.87
3D reconstruction: camera space z mean: 0.1699750660884
Object num: person Position shape: (7,) Depth map shape: (7, 1856786)
O in checking if all names are recorded...
typing description...
Adding to dict...
Count dict length: 1
O to what do
```



Good News: the system of
~**1000 lines** of code works!

Worked

- The system runs smoothly
- Cozmo can recognize chairs, people, phones...
- GPT-vision generates good object description
- The worldmap position of an object is roughly reasonable
- Use multi-threads to process objects concurrently

Not Worked

- Object detection nets can only recognize objects from 80 classes...
No better net found!
- Single-camera depth estimation -> need human support
- Super-resolved image is still of poor quality
- Super-resolution takes long time

Future Look

- Make Cozmo look from different perspectives to enable two-camera depth estimation
- Train object detection model to recognize more objects (in the far future)
- Toy robots of better camera image will help much