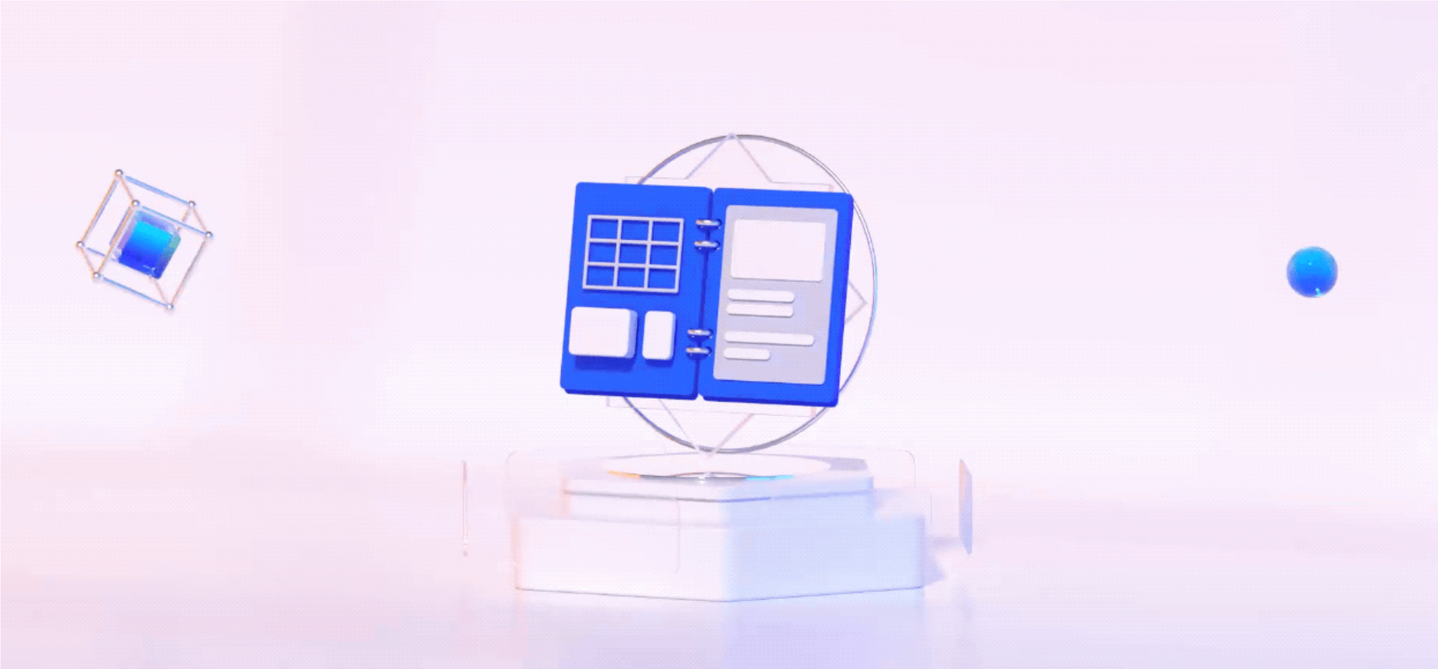


GSA-YOLO-GUI

Project background

> This project includes the powerful Grounded-Segment-Anything and the newest YOLO-NAS. The GUI is made by groundio, which is able to help us understand and use these algorithms easier, while it can be run on the personal PC via public URL.

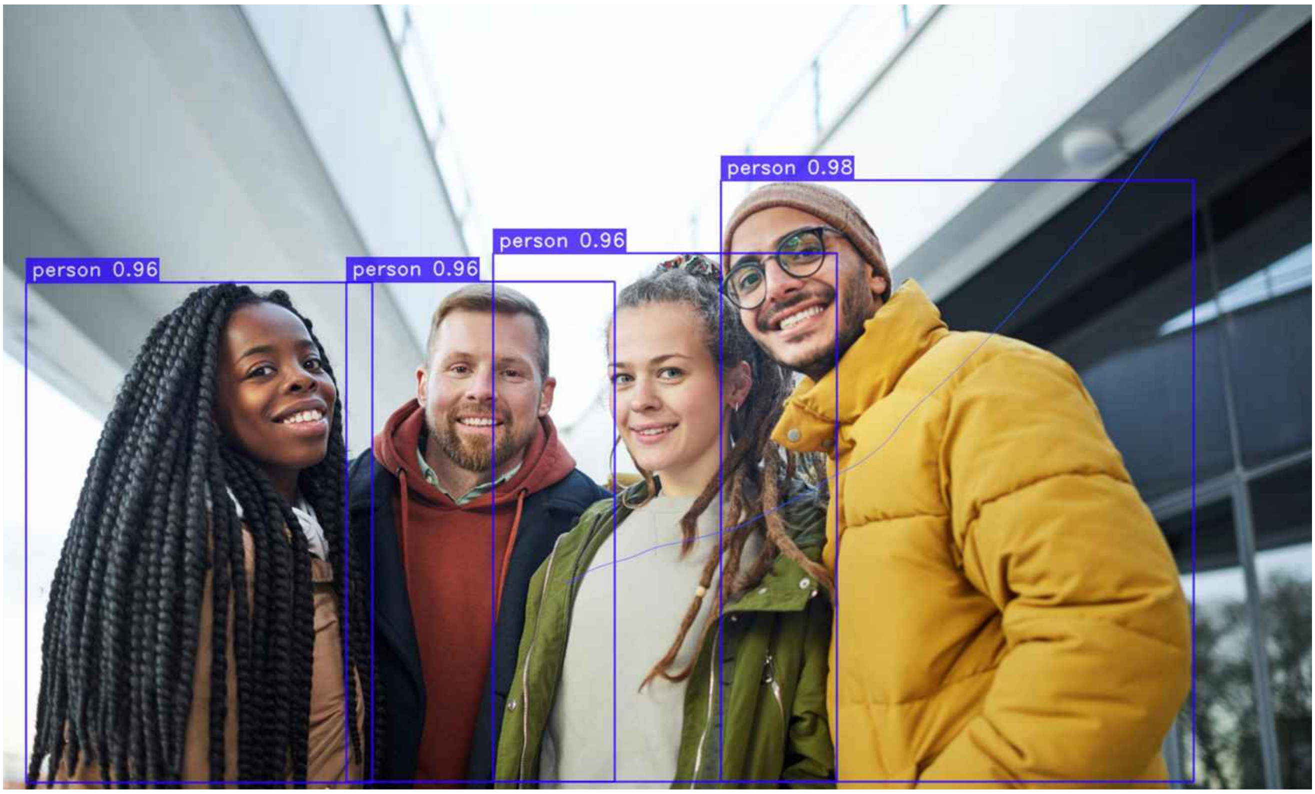
Project intro

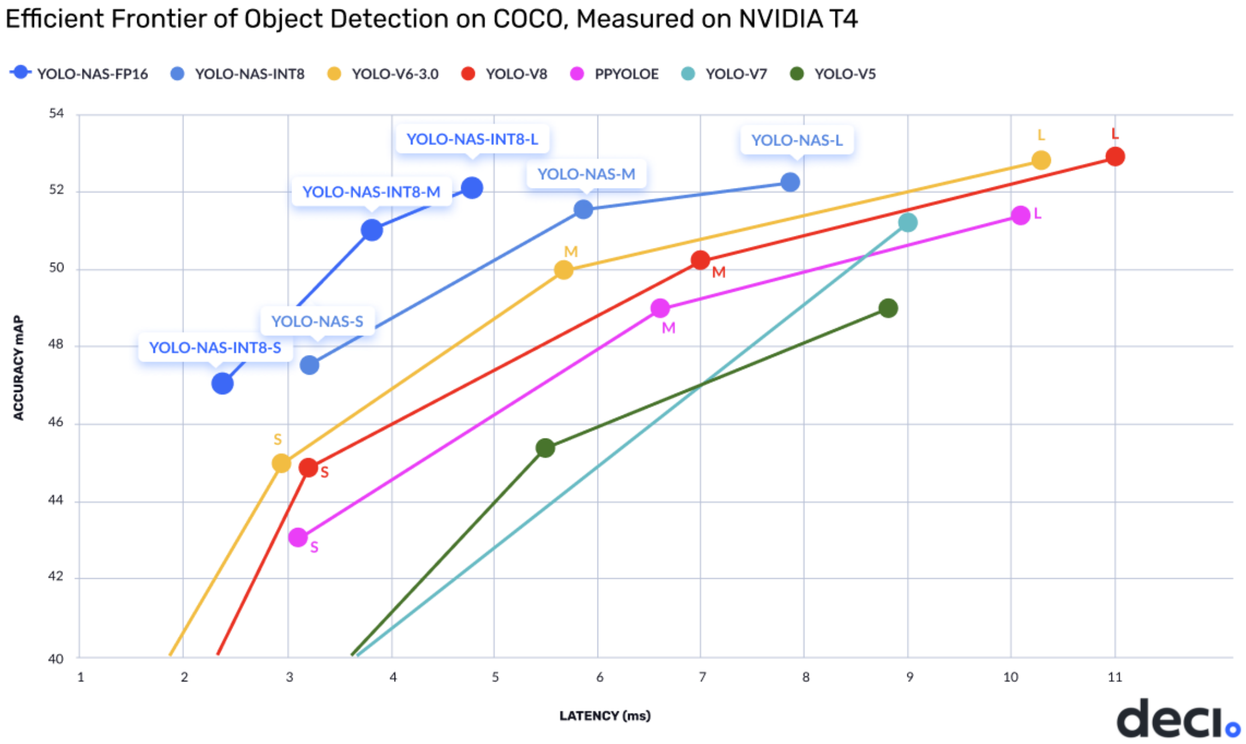
> YOLO --> YOLO-NAS

What is YOLO-NAS ?

> **YOLO** (You Only Look Once) is a real-time object detection algorithm known for its fast and efficient performance. It divides the input image into a grid and predicts bounding boxes and class probabilities for objects within each grid cell. **YOLO-NAS** (You Only Look Once Neural Architecture Search) is a new state-of-the-art real-time object detection model that outperforms previous YOLOv6 and YOLOv8 models in terms of mean average precision (mAP) and inference latency. It is characterized by the incorporation of Neural Architecture Search (NAS) techniques.

More information --> github： YOLO-NAS-github





And then Grounded-Segment-Anything（GSA）？

> Grounded-Segment-Anything (GSA) is a method for image segmentation tasks that aims to associate objects in images with their corresponding textual descriptions. This method combines image segmentation with natural language processing to achieve precise segmentation of objects in images and generate text descriptions related to those objects. The core idea of GSA is to transform the image segmentation task into an end-to-end multimodal

> learning problem. It utilizes two key components: an image segmentation network and a text generation network.

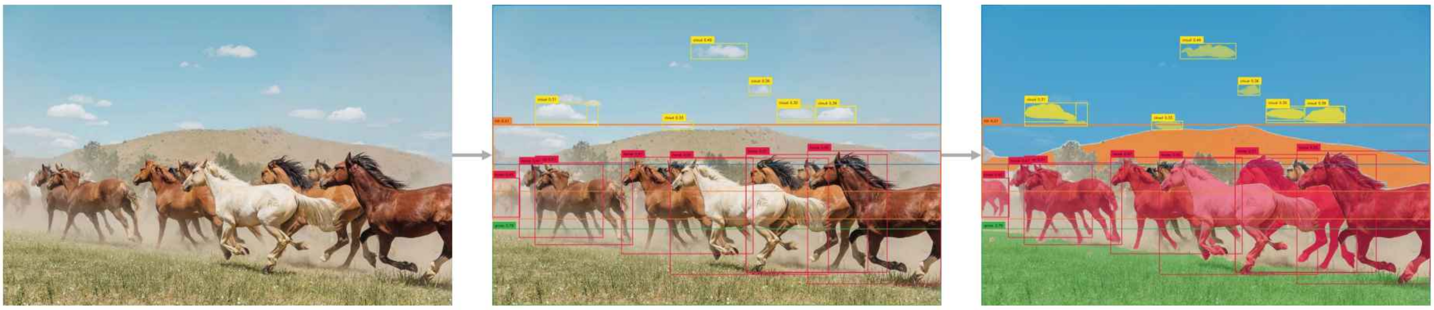
> **The image segmentation network** is responsible for pixel-level segmentation of the input image, separating different object regions in the image. This network can be a traditional deep learning-based segmentation model, such as Fully Convolutional Network (FCN) or Mask R-CNN (Mask Region Convolutional Neural Network).

> **The text generation network**, based on the results of image segmentation, generates textual descriptions associated with each object region. It can be a language model based on recurrent neural networks (RNNs) such as Long Short-Term Memory (LSTM) or Transformer.

> By jointly training these two networks, GSA achieves accurate image segmentation and generation of text descriptions related to the objects. This makes GSA valuable for various tasks such as image analysis, automatic annotation, and image retrieval.

More information --> github：Grounded-Segment-Anything-github

Grounded-SAM



Text Prompt:
"Horse. Clouds. Grasses. Sky. Hill."

Grounding DINO:
Detect Everything

Grounded-SAM:
Detect and Segment Everything

How to use it?

1. Env setup

- (Mini) conda virtual environment
- Environment for deep-learning (pytorch)
- Environment for image handling: numpy, OpenCV, PIL, ...
- OpenCV installation from source
- IDE: Pycharm, ...

2. Source installation command

```
pip install super-gradients
python -m pip install -e segment_anything
python -m pip install -e GroundingDINO
pip install --upgrade diffusers[torch]
git submodule update --init --recursive
cd Tag2Text && pip install -r requirements.txt
```

3. Move and replace these python files to the root folder

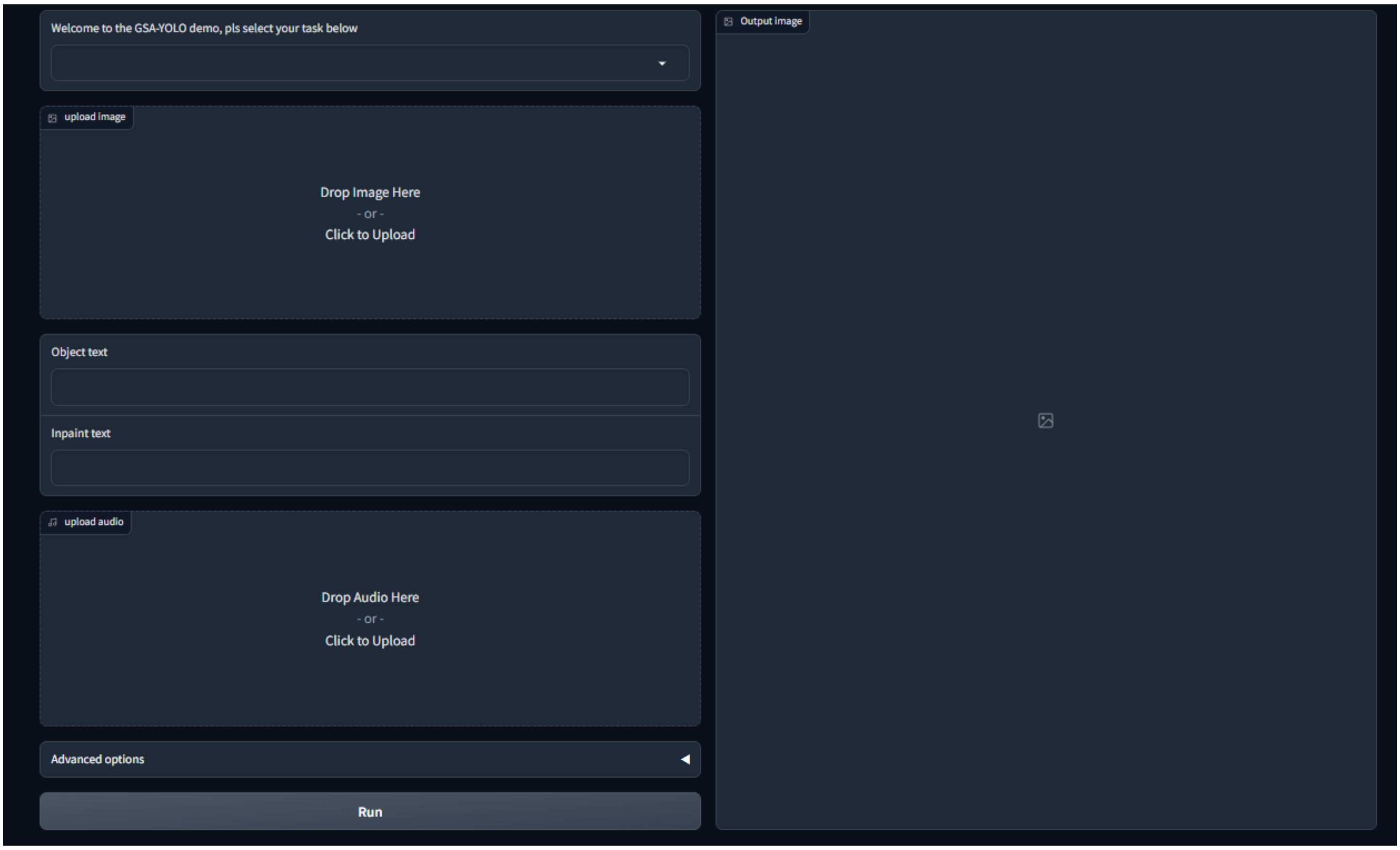
- grounded_sam_demo.py
- grounded_sam_inpainting_demo.py
- grounded_sam_whisper_demo.py
- automatic_label_tag2text_demo.py
- gui.py

The files above are in the folder: **GSA-YOLO-replace**

4. Running in the terminal

```
Python gui.py
```

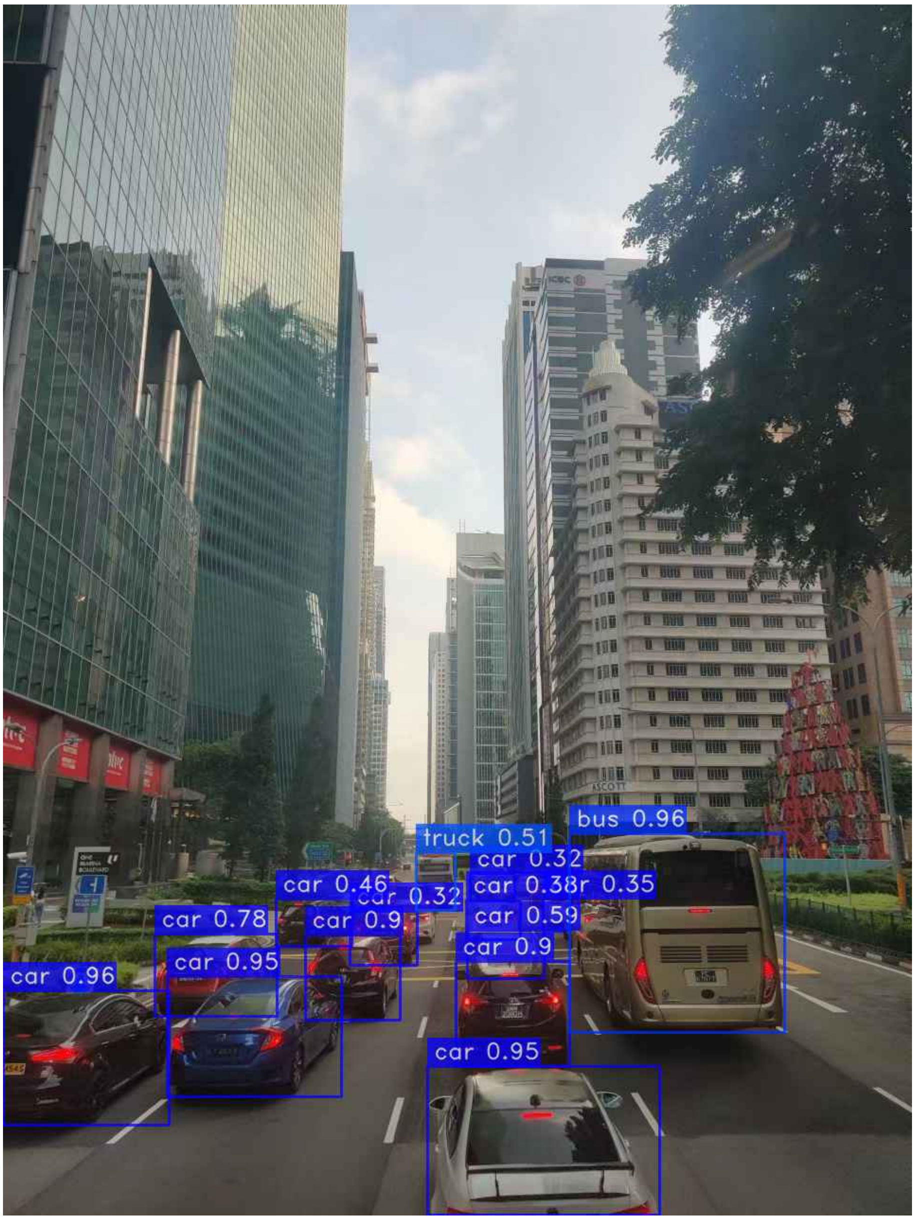
Click on local url or public url in the terminal --> gui



Function demo：

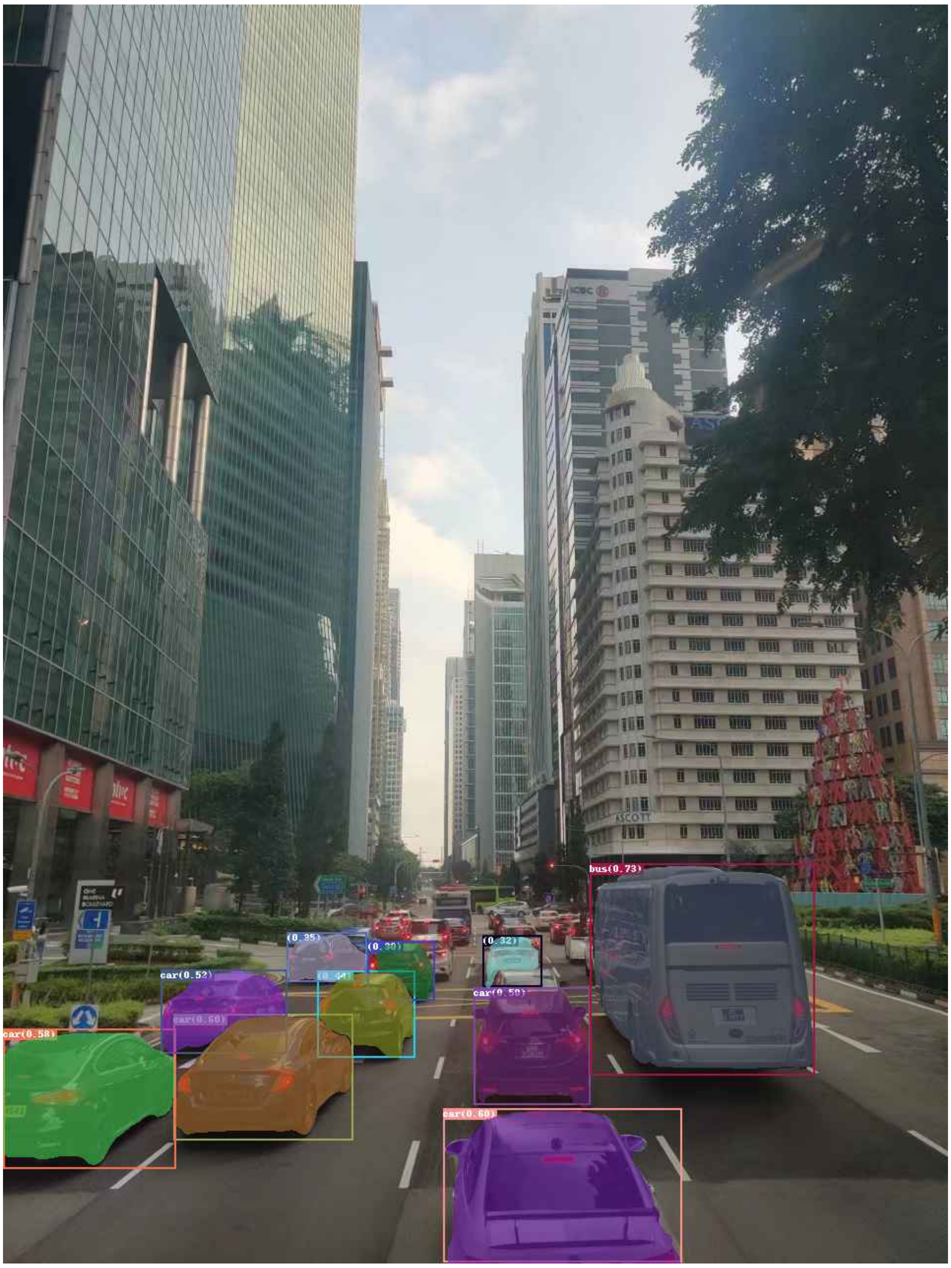
1. **YOLO-NAS:**

   Input: upload image (do not input others)



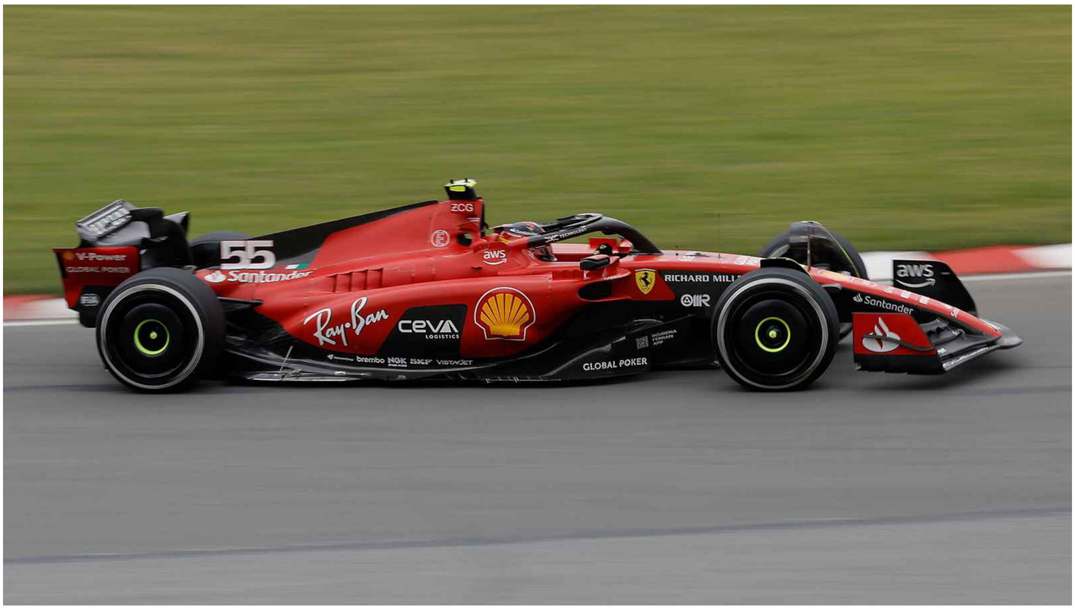
2. **Grounded-SAM:**

   Input: upload image, Object text: **car, bus**

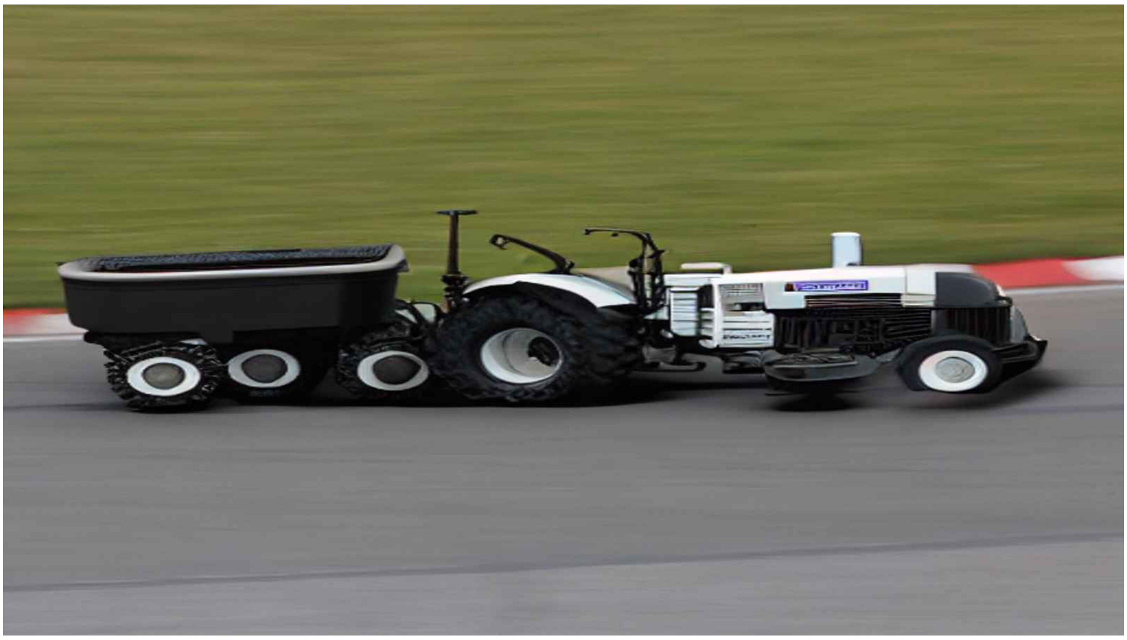


3. **Grounded-SAM with Inpainting:**

   Input: upload image, Object text: **F1 car**, Impaint text: **tractor**

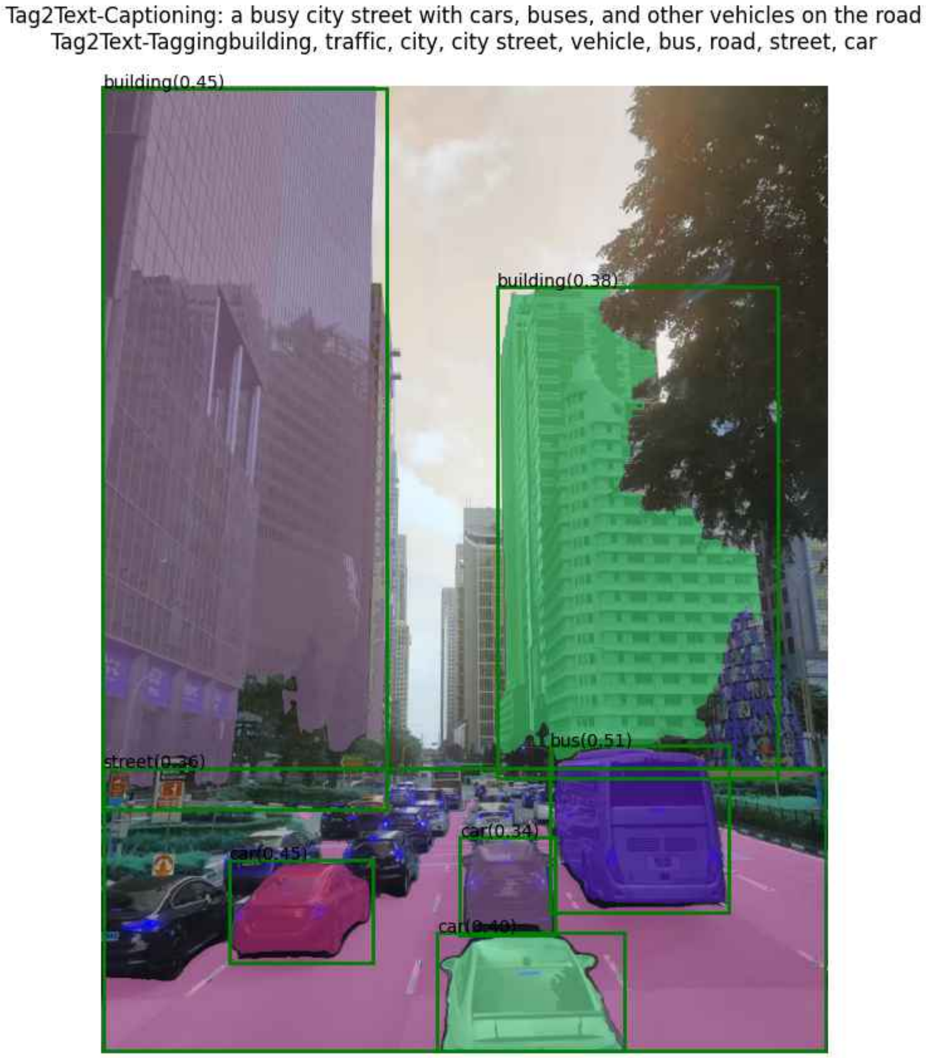


Output:



4. **Grounded-SAM for Automatic Labeling:**

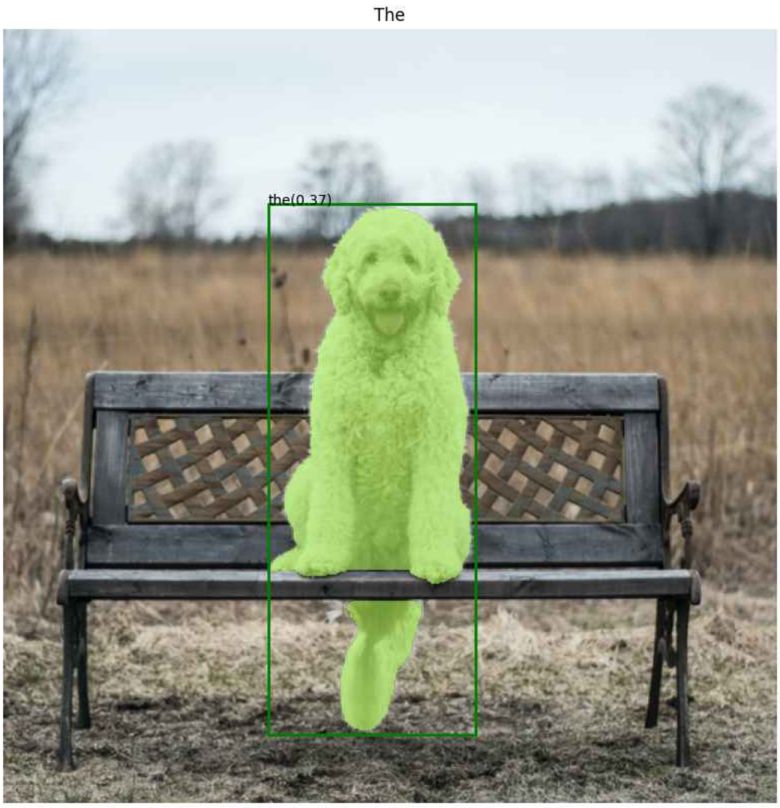
   Input: upload image (do not input others)



Tag2Text-Captioning: a busy city street with cars, buses, and other vehicles on the road
Tag2Text-Taggingbuilding, traffic, city, city street, vehicle, bus, road, street, car

5. **Grounded-SAM with Whisper:**

   Input: upload image, upload audio： **“The dog”**



Future Tasks

- Add other functions to the gui (OSX, VISAM, ...)
- Show a video and live stream in the output block.
- TBC...

Related information

> First edition by Hongzhao Xiao 13/7/2023