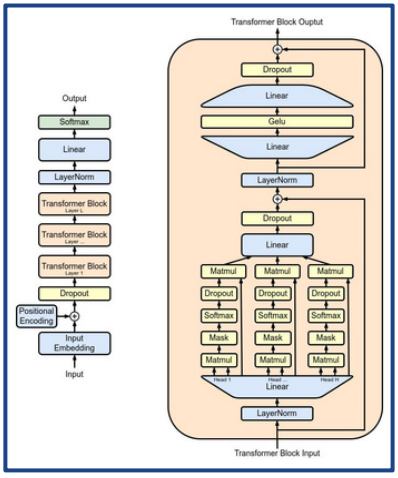
Prompt Engineering for Vision Models

What is a Prompt?

"A photorealistic image of an astronaut riding a horse on the moon."



[0.24, -0.18, 0.14, 0.07, -0.03, ..., 0.23]

What is Visual Prompting?

Visual prompting is a method of interacting with a pre-trained model to accomplish a specific task that it might not necessarily have been explicitly trained to do.

This often involves passing a set of instructions to the model, describing what you'd like it to do.

"Highlight the dog on the left."

Prompt vs. Input

Input (Data)

Prompt (Instructions)

"Segment the dog on the left."

Traditional ML Workflows

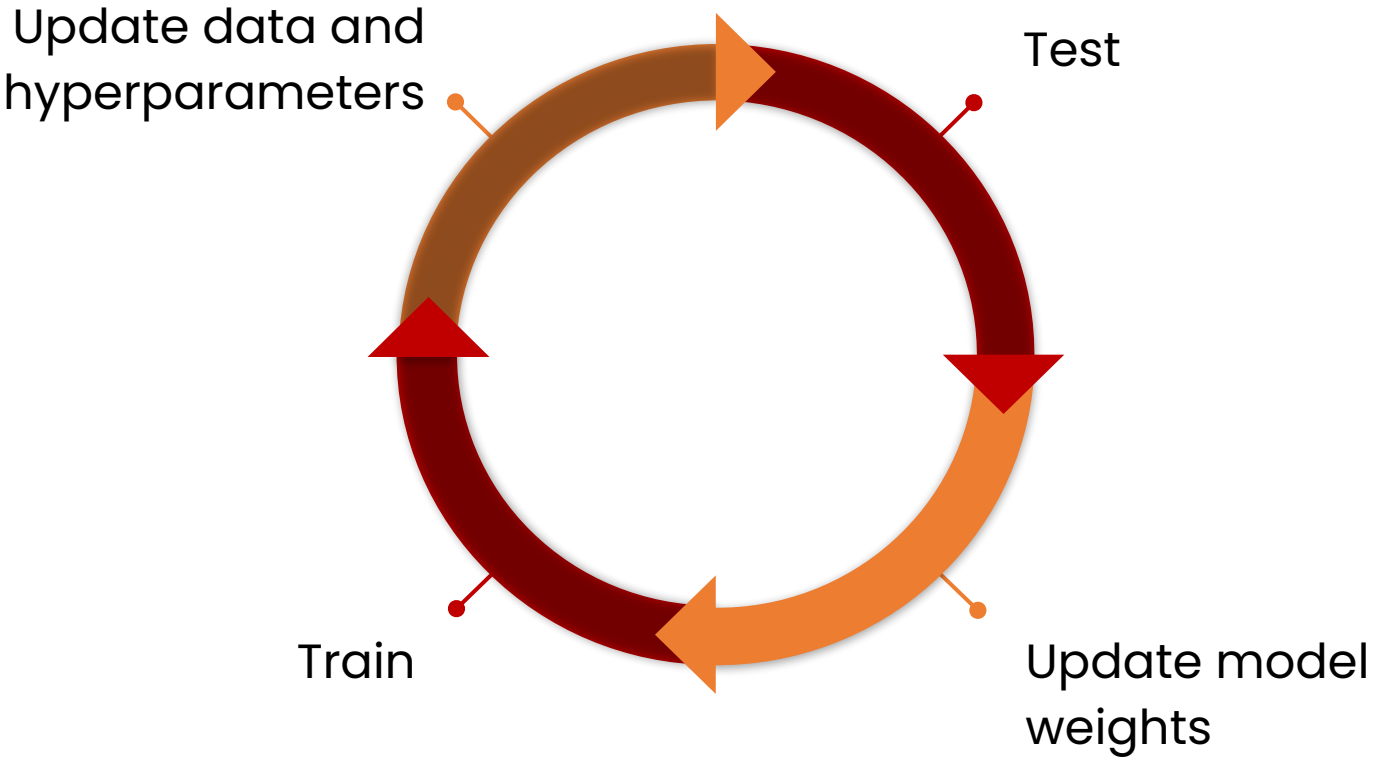


Image segmentation

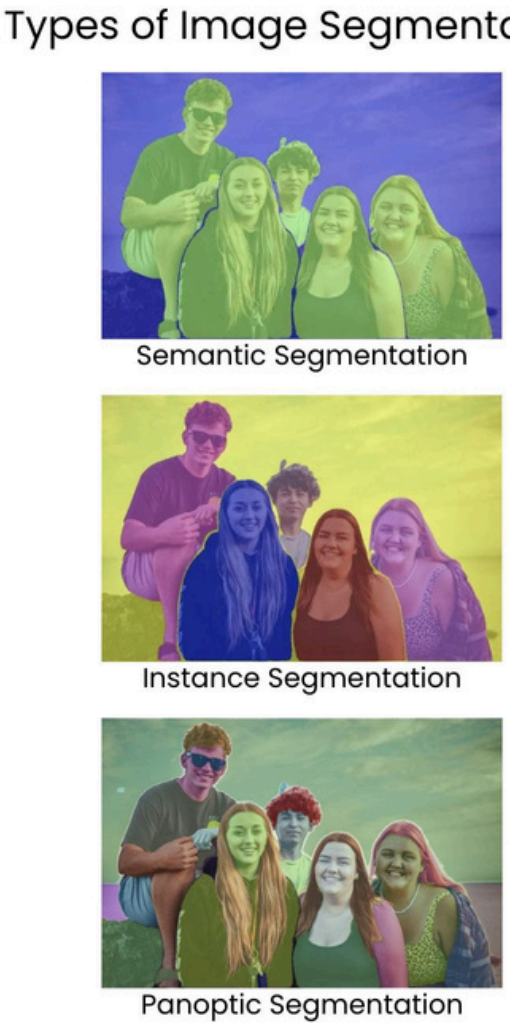
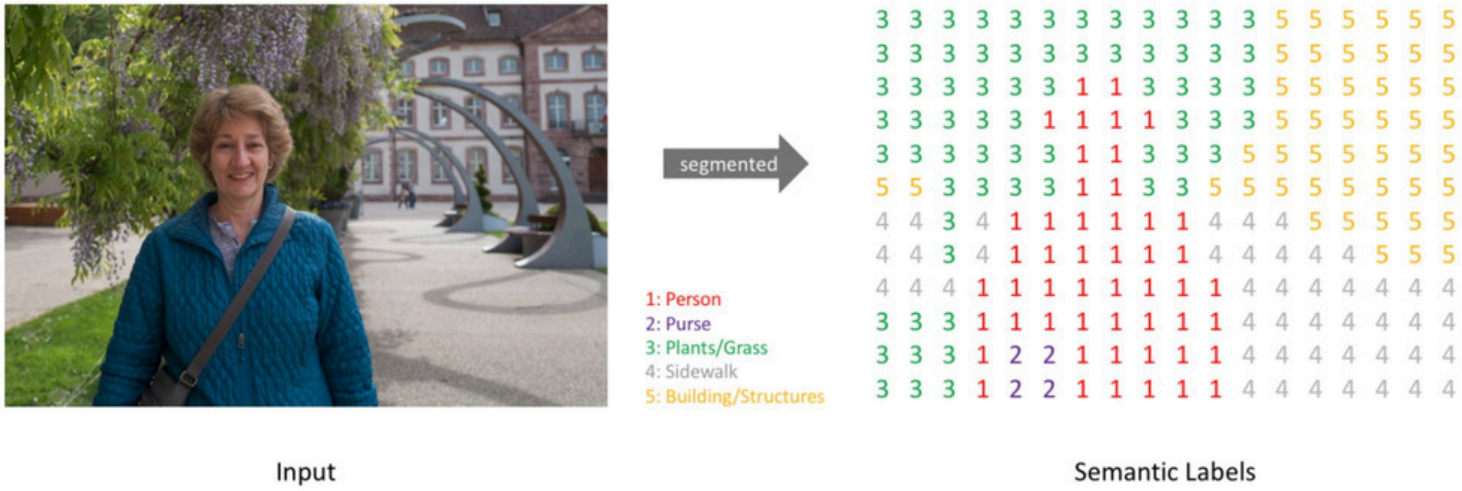


Image segmentation

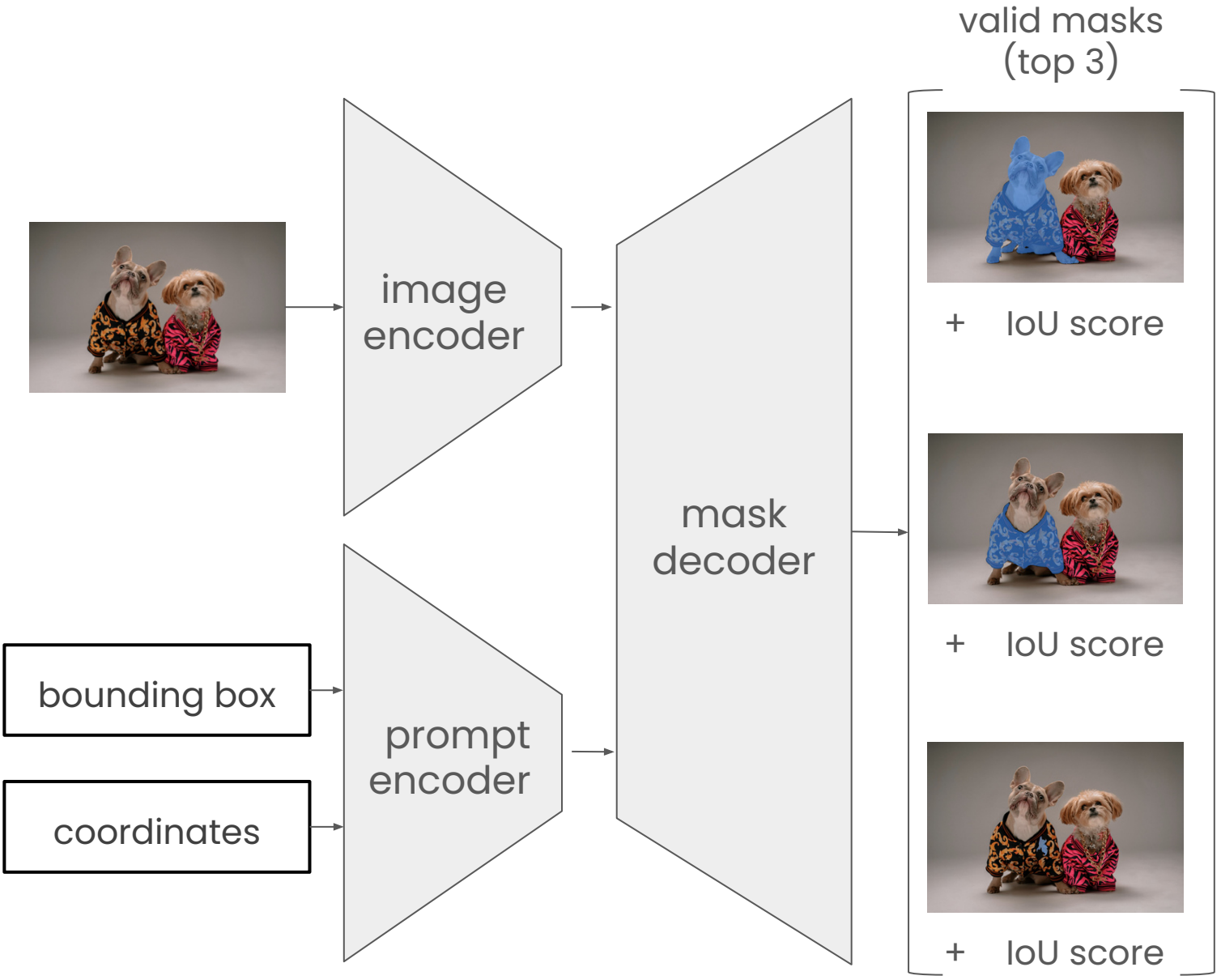


Input

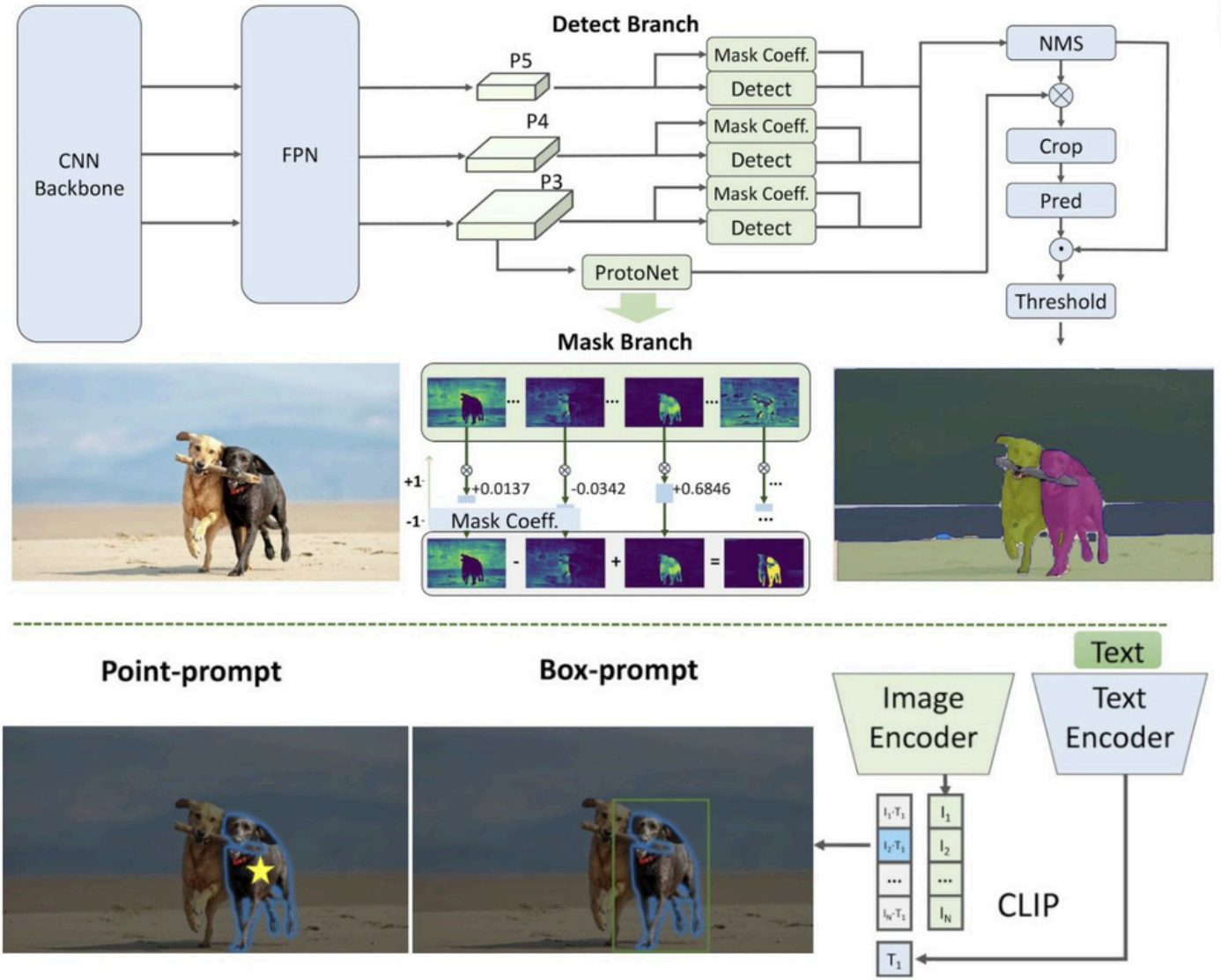
Semantic Labels

Source: Jeremy Jordan
"An overview of semantic image segmentation"
https://www.jeremyjordan.me/semantic-segmentation/

Segment Anything Model



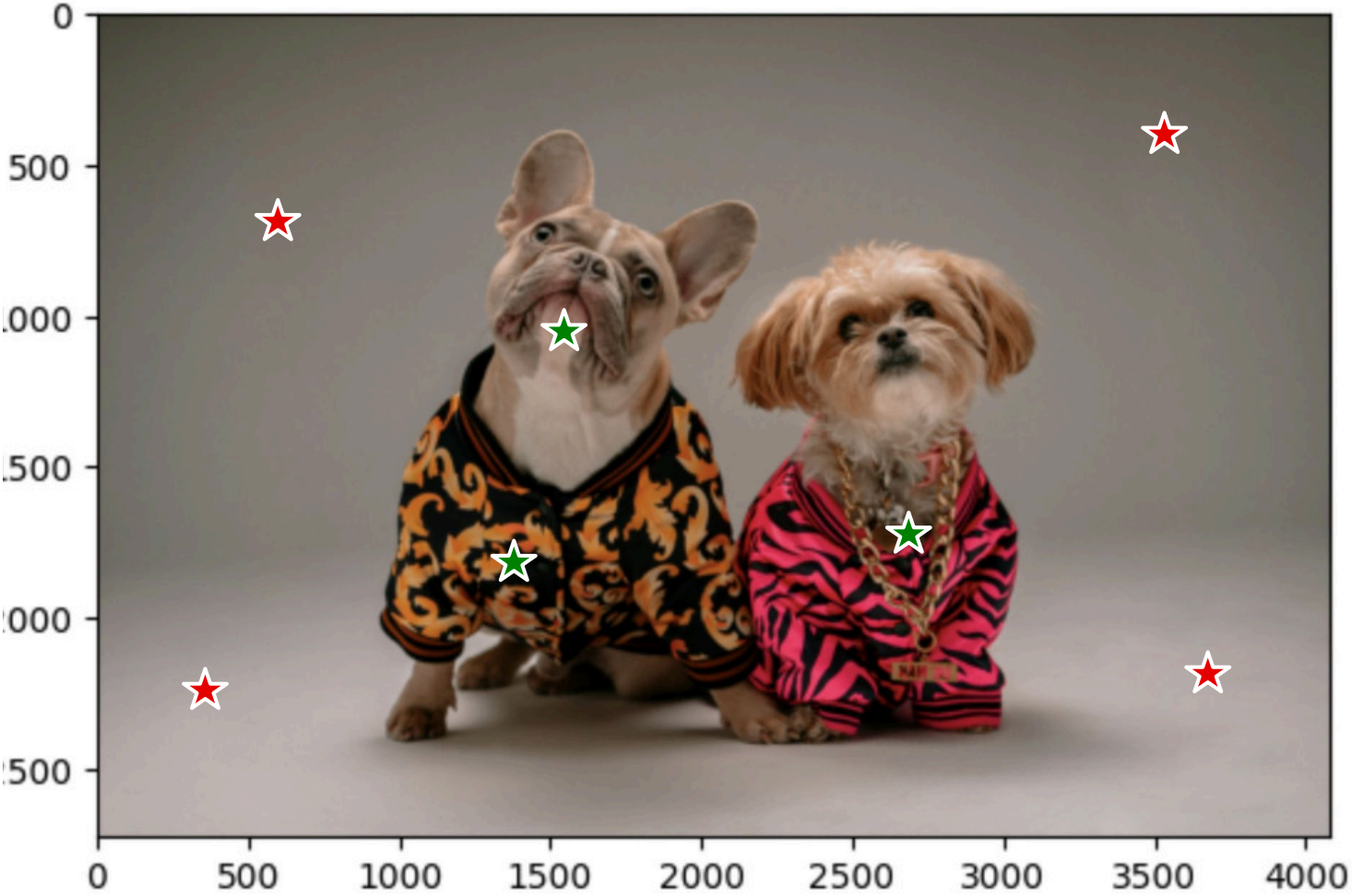
FastSAM



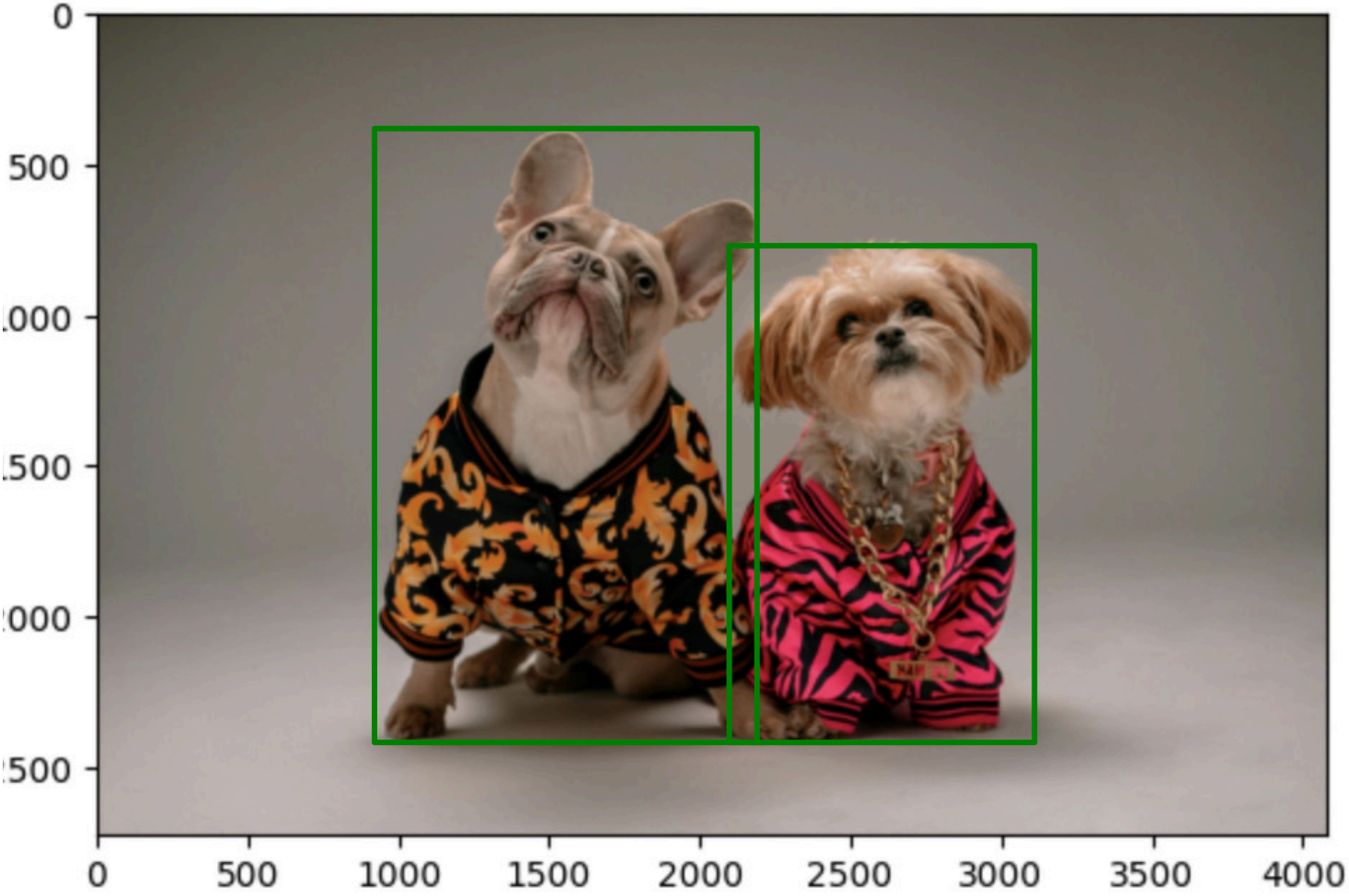
Source: "Fast Segment Anything"
Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, Jinqiao Wang

Example image

Prompting with coordinates



Prompting with bounding boxes



Embeddings

"Ships at a distance have every man's wish on board." → [0.12, -0.31, 0.79, 0.05, ..., -0.41]

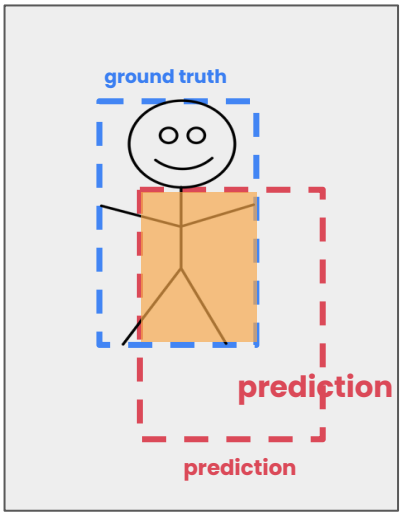
"Too much sanity may be madness — and maddest of all: to see life as it is, and not as it should be!" → [0.92, 0.31, -0.22, -0.39, ..., 0.03]

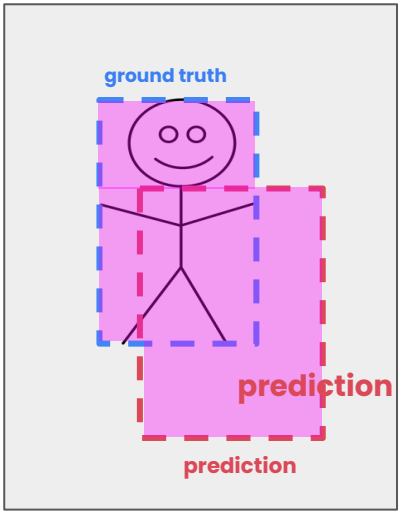
 → [-0.72, -0.05, 0.82, 0.74, ..., 0.06]

 → [0.75, -0.93, -0.27, 0.40, ..., 0.08]

Intersection Over Union

$$IoU = \frac{\text{intersection}}{\text{union}}$$





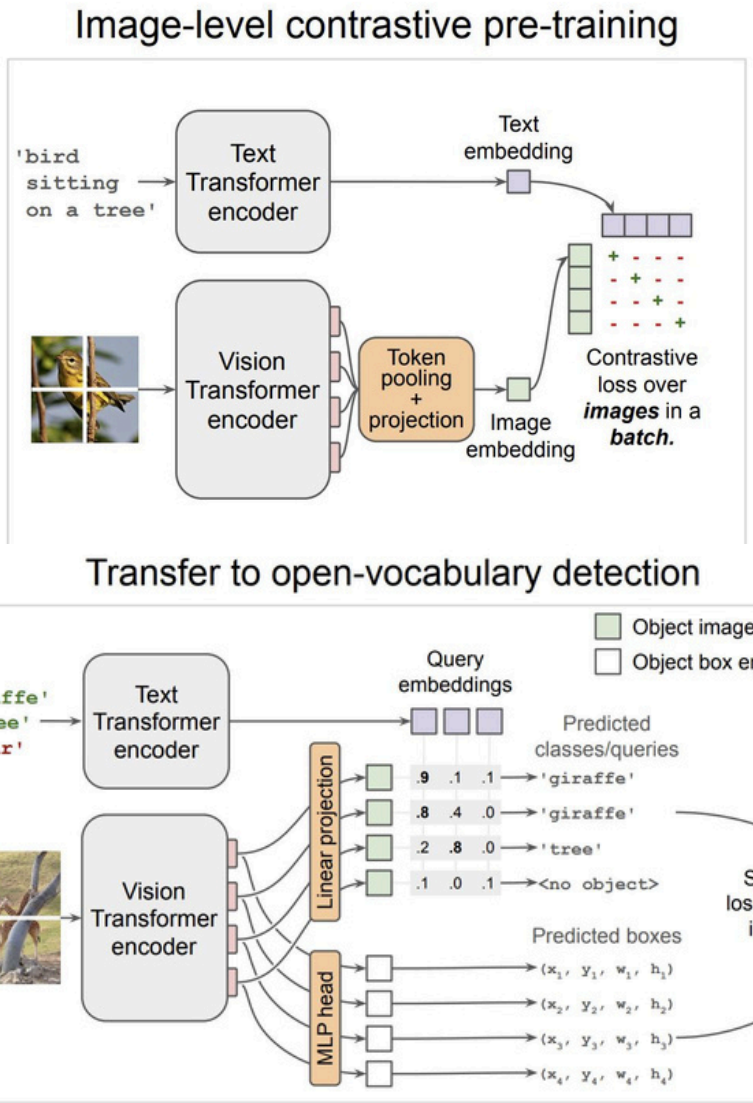
bounding boxes

[[[x1, y1], [x2, y2]]]

[[[xmin, ymin, xmax, ymax]]]

OWL-ViT

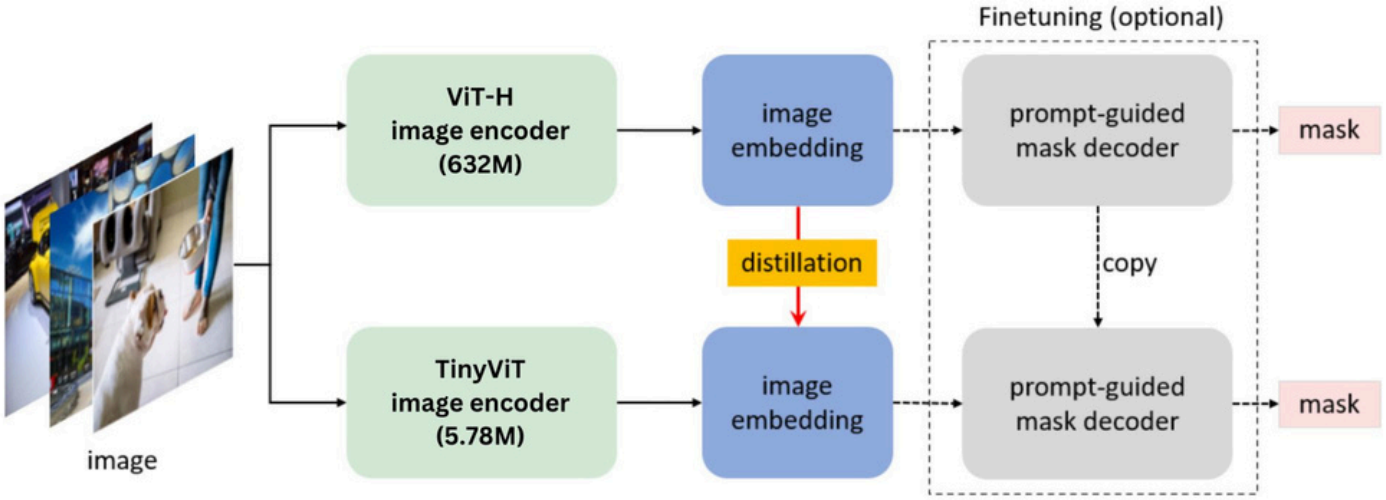
Text prompt → Bounding Boxes



"Simple Open-Vocabulary Object Detection with Vision Transformers"

by Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby

MobileSAM



Model distillation is the process of transferring knowledge from a large model to a smaller one. Model distillation is different from other model compression techniques in that it doesn't actually change the model format, but trains an entirely new (and smaller) model.

Source: "MobileSAMv2: Faster Segment Anything to Everything" Chaoning Zhang, Dongshen Han, Sheng Zheng, Jinwoo Choi, Tae-Ho Kim, Choong Seon Hong