

ZerO-shot Aerial Object Detection

Ilan Motiei, Niv Ron, Ron Gaist, Amit Nadam, Meir Akabayov

The Problem

Recognizing
anything,
anywhere

Training is
expensive

Demand changes
frequently

Goal

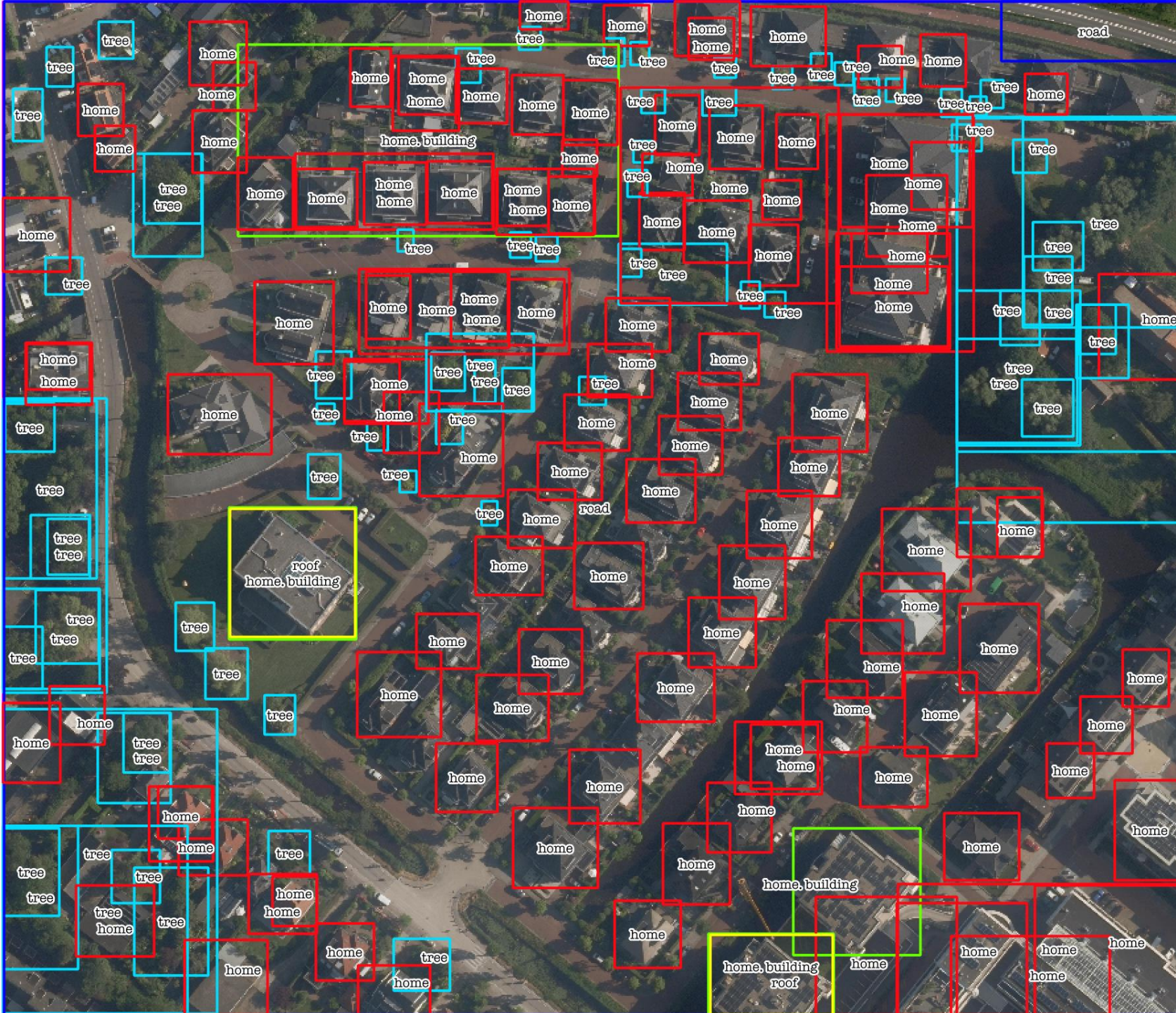
- Developing a (semi)autonomic ability to detect and segment any type of object out of aerial imagery by applying open-source tools
- Analyzing and evaluating key quality parameters of the method
- Scale-oriented pipeline



Pipeline

Grounding Dino – Recognize anything

- Zero-shot detection
- Input – ANY natural language prompt (e.g. Roof. Road. Car. Tree. Water. Airplane. Truck. Grass)
- Output – bounding box for each detected object
- No task-specific training – capable of ANY vision and text domain



Detection

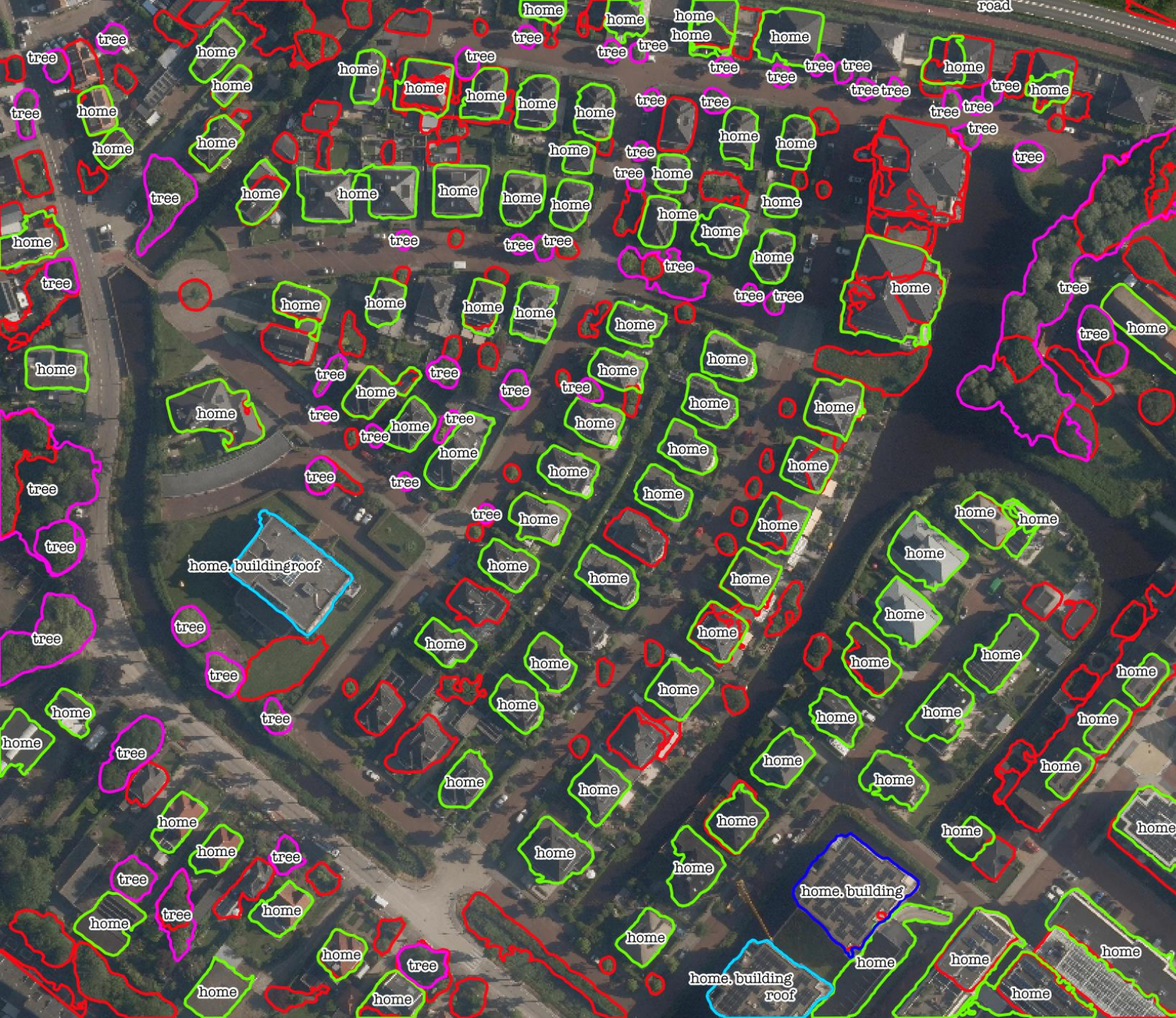


מקרא:	
Home	
Building	
Road	
Roof	

SAM – Segment Anything

- Zero-shot segmentation
- Foundation model for segmentation
- Developed by Meta to generalize across images types and tasks
- Trained on over 1 billion masks across 11 million images (~100 masks per image)





Segmentation



+



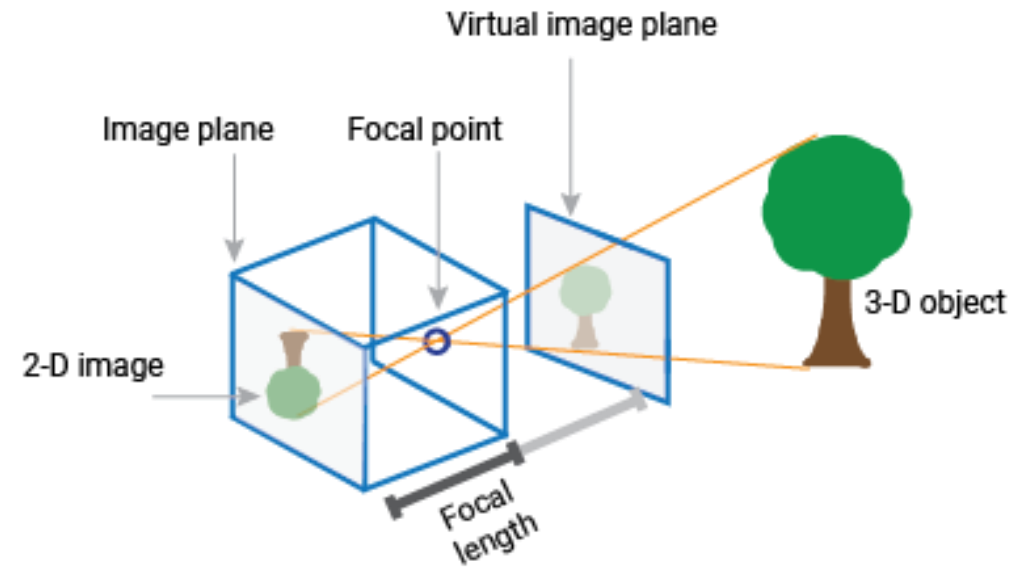
מקרא:

Home	
Building	
Road	
Roof	
Tree	
NULL	

Image to Ground

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix}$$

2D Image Coordinates Intrinsic properties (Optical Centre, scaling) Extrinsic properties (Camera Rotation and translation) 3D World Coordinates





Data Collection

- Google satellite imagery – 15cm/pixel RGB anchored imagery
- POTK service – the netherlands geo database (8cm/pixel RGB imagery). Anchored
- OSM - Reliable, open data source. Used for assessing accuracy



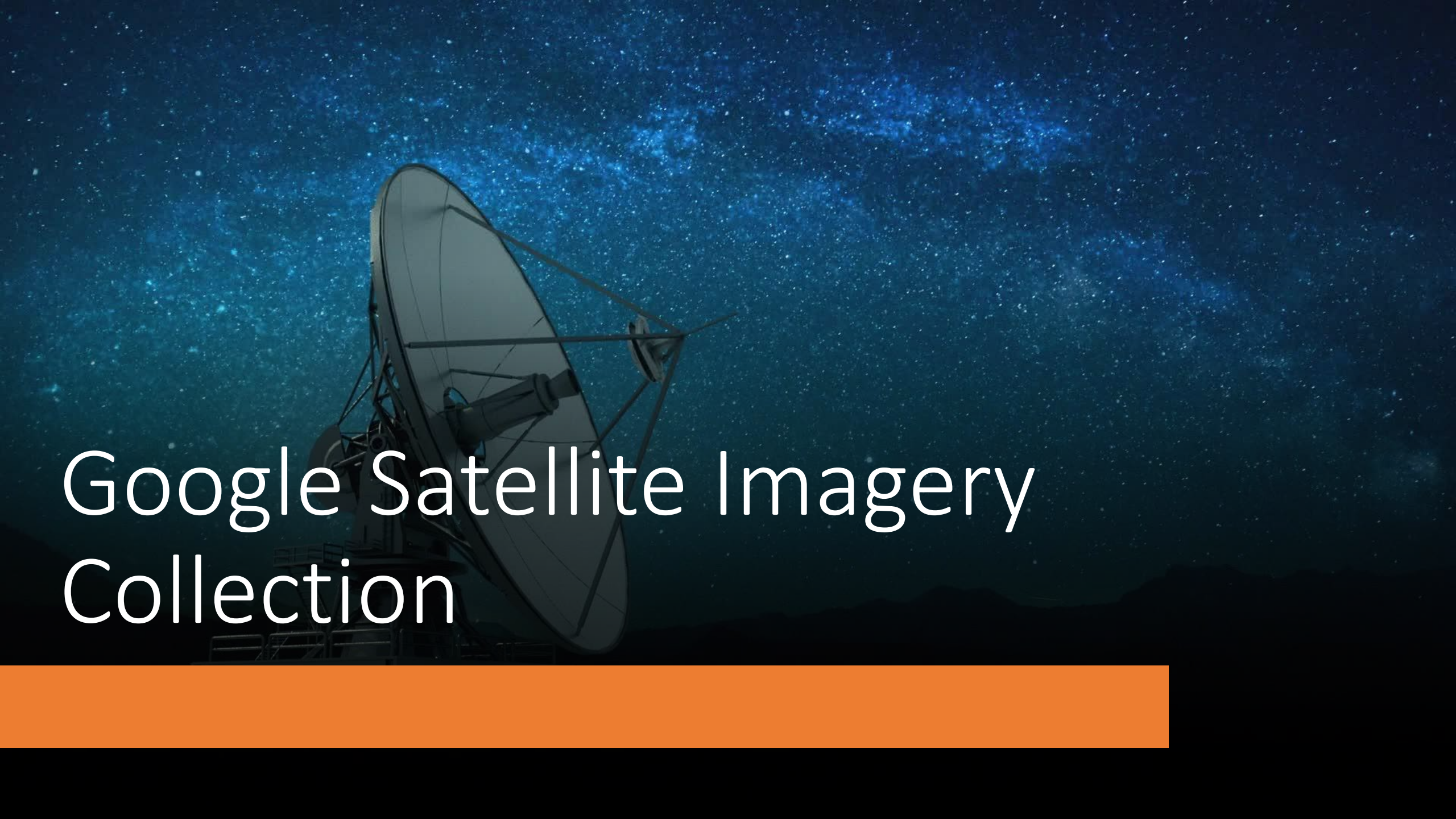
Testing Aerial Imagery



Roofs Detection

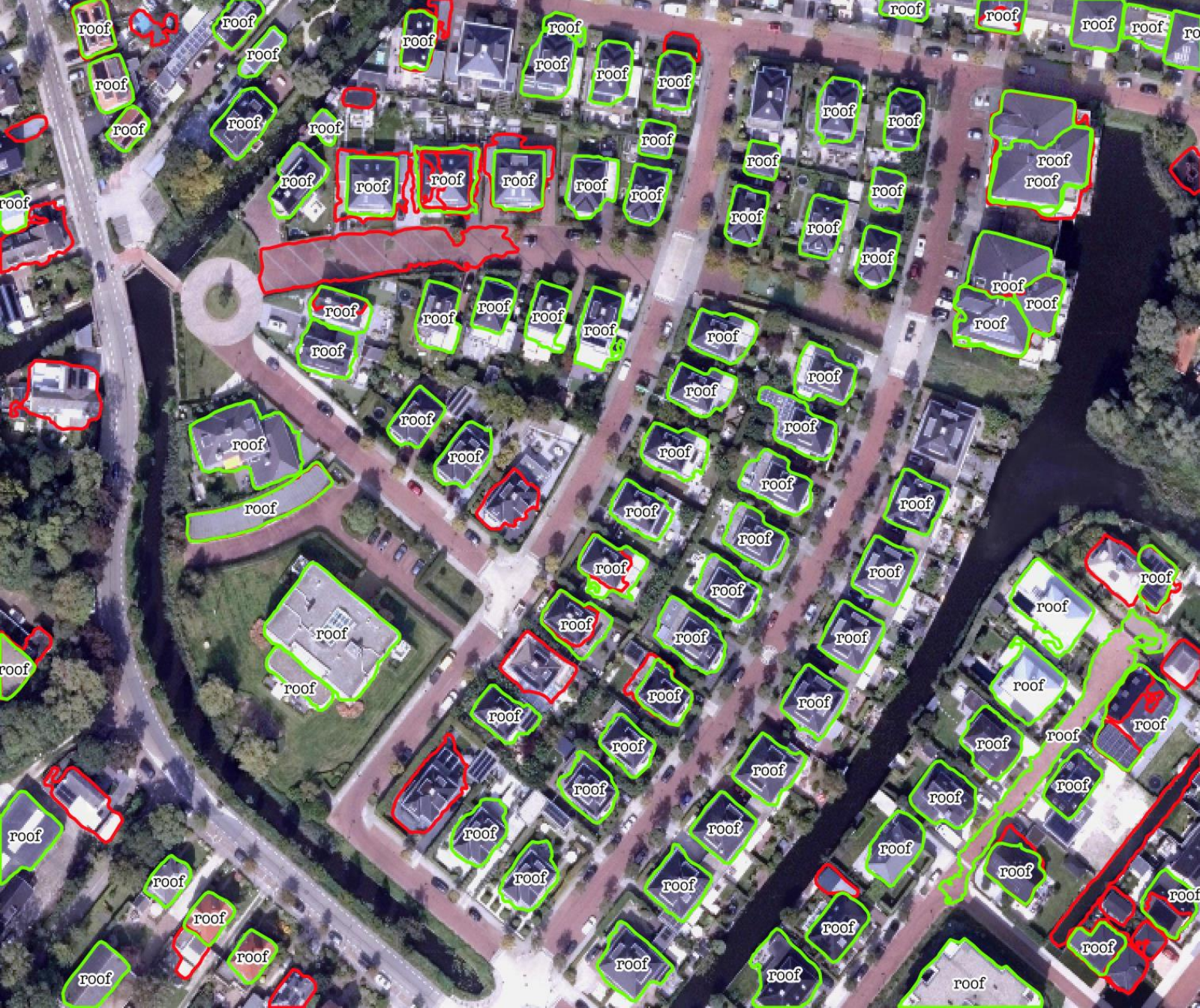


The Netherlands



Google Satellite Imagery Collection





Ideal size

Area ~ 50 dunam (250x250m)

GSD ~ 73cm/pixel

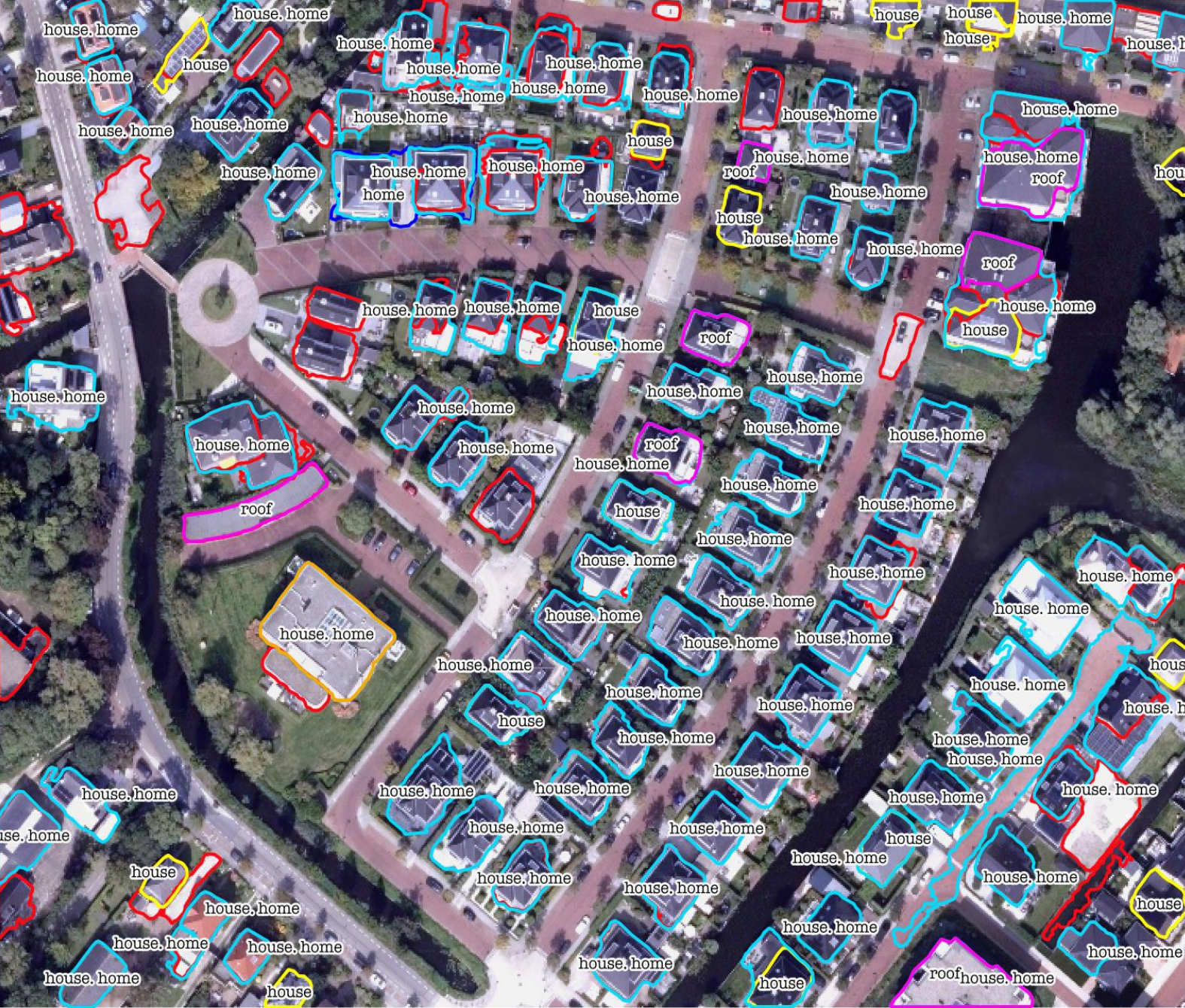
Prompt := "Roof."

Confusion Matrix

	14 (FP)
43 (FN)	122 (TP)

Aggregating Metrics

Precision	89.7%
Recall	73.9%
F1	81.1%



Another prompt

Area ~ 50 dunam (250x250m)

GSD ~ 73cm/pixel

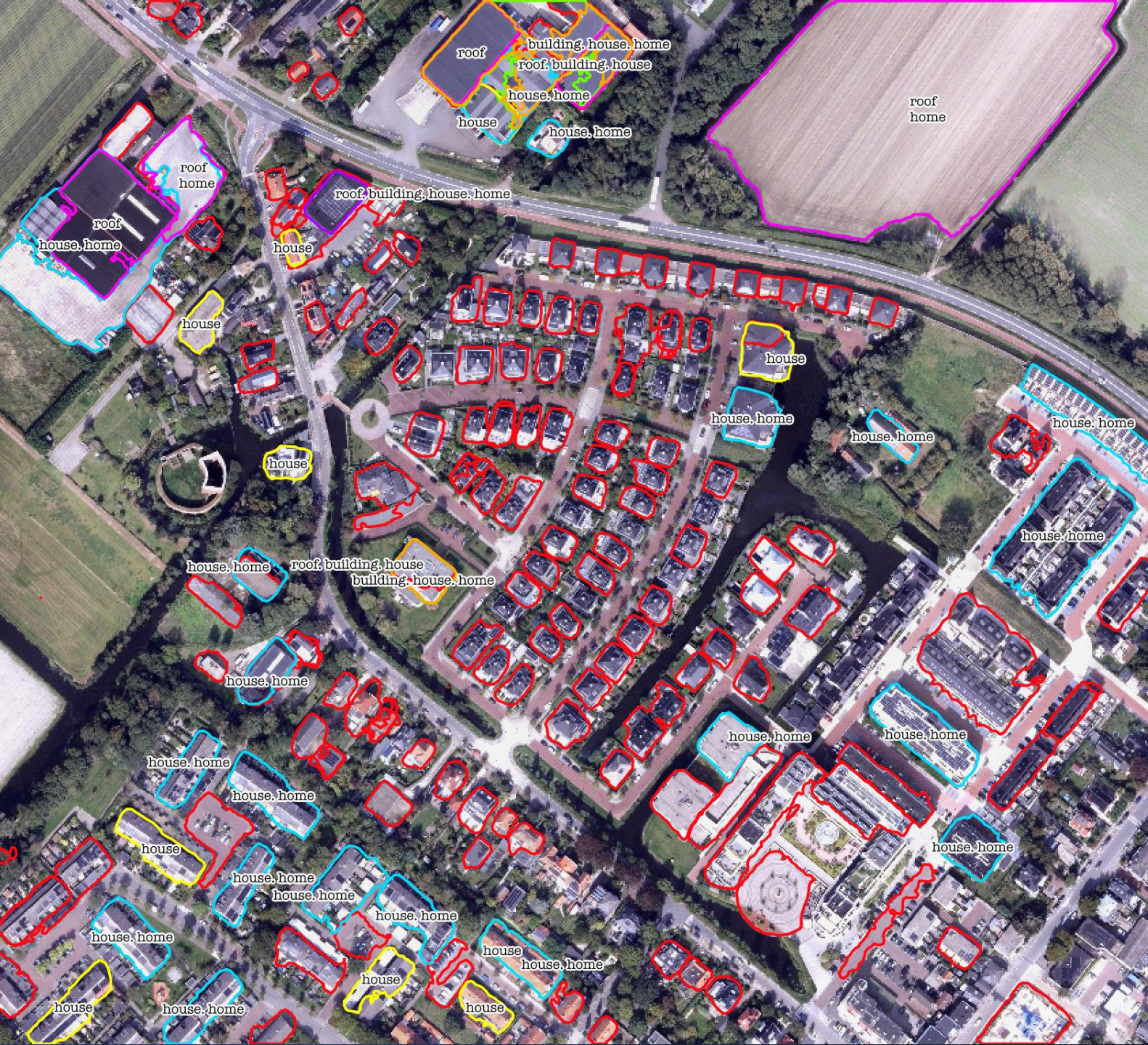
Prompt := “Roof. Building. House. Home”

Confusion Matrix

	22 (FP)
35 (FN)	163 (TP)

Aggregating Metrics

Precision	88.1%
Recall	82.3%
F1	85.117%



Zooming out

Area ~ 300 dunam (650x550m)

GSD ~ 1.6m/pixel

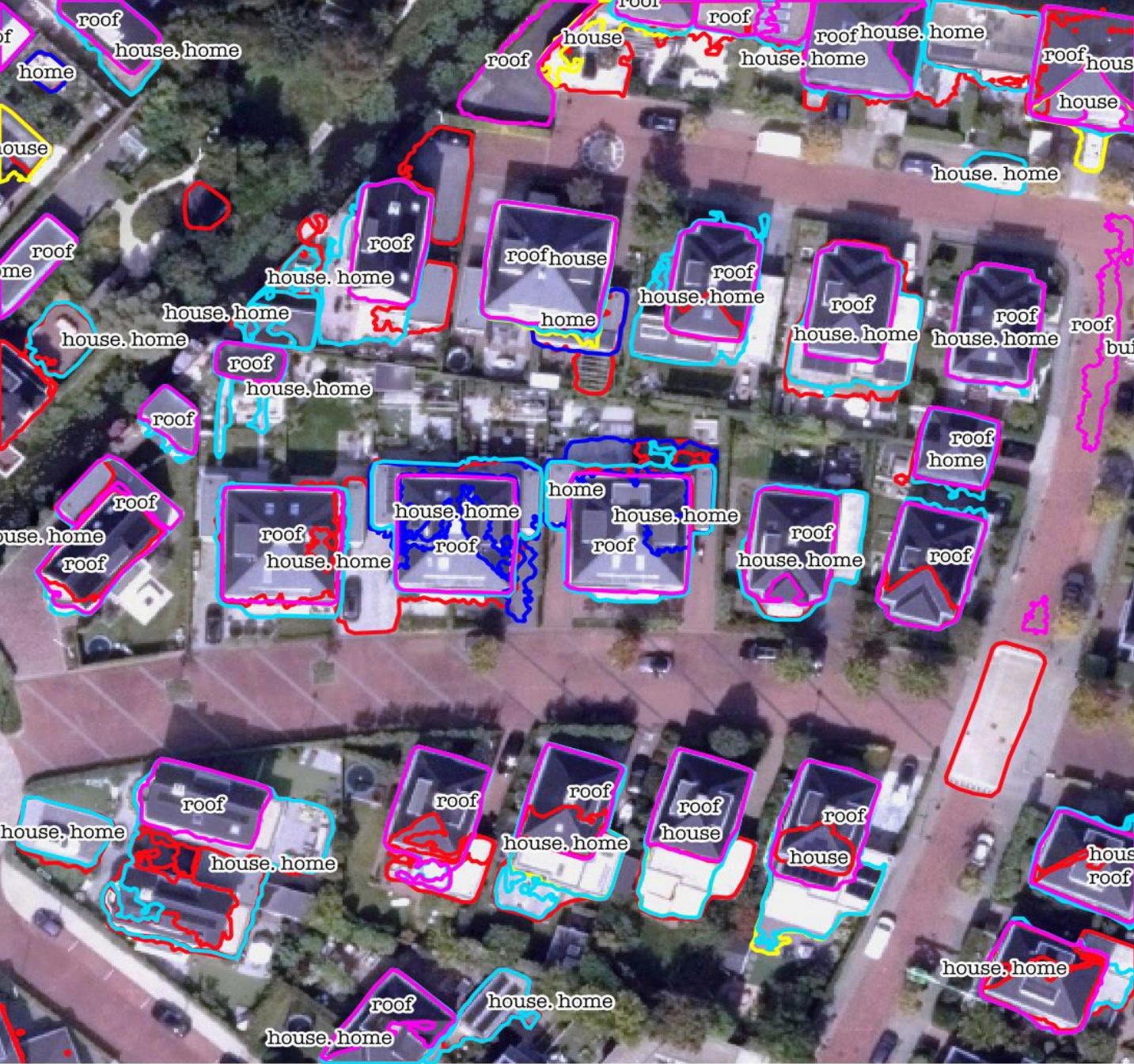
Prompt := “Roof. House. Building. Home.”

Confusion Matrix

	40 (FP)
310 (FN)	141 (TP)

Aggregating Metrics

Precision	77.9%
Recall	31.3%
F1	44.6%



Zooming in

Area ~ 10 dunam (100x100m)

GSD ~ 34cm/pixel

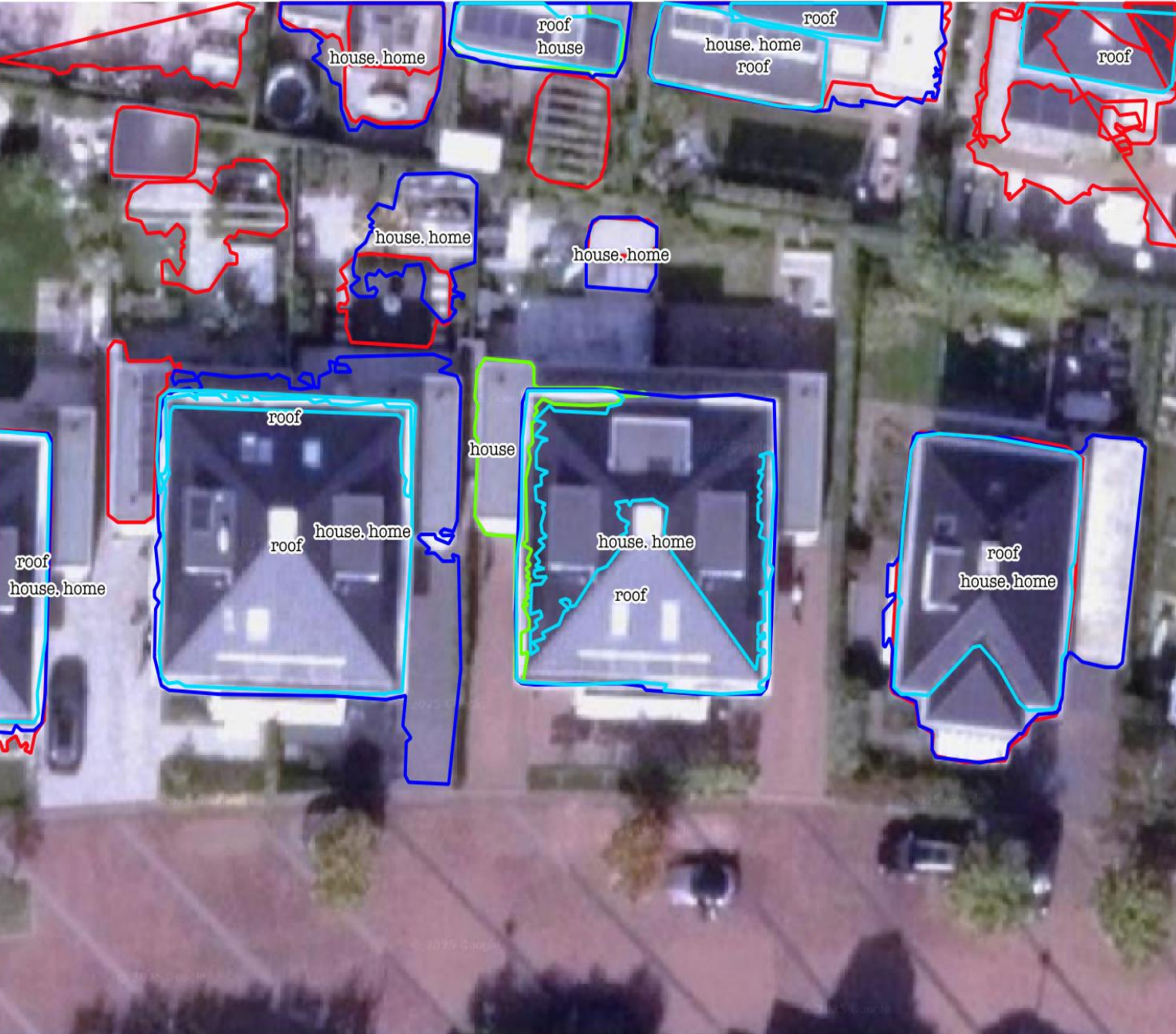
Prompt := "Roof.House.Building.Home."

Confusion Matrix

	26 (FP)
13 (FN)	89 (TP)

Aggregating Metrics

Precision	77.4%
Recall	87.25%
F1	82.0%



Diving in

Area ~ 3 dunam (60x50m)

GSD ~ 15cm/pixel (Rock bottom)

Prompt := "Roof.House.Building.Home."

Confusion Matrix

	2 (FP)
17 (FN)	18 (TP)

Aggregating Metrics

Precision	51.4%
Recall	90%
F1	65.45%

A close-up photograph of a blue ribbon with a gold laurel wreath emblem. The ribbon is draped and folded, creating deep shadows and highlights. The laurel wreath is embroidered in gold thread on the right side of the ribbon. The background is dark and out of focus.

The Netherlands Official Imagery Source



Area ~ 105 dunam (300x350m)

GSD ~ 54cm/pixel

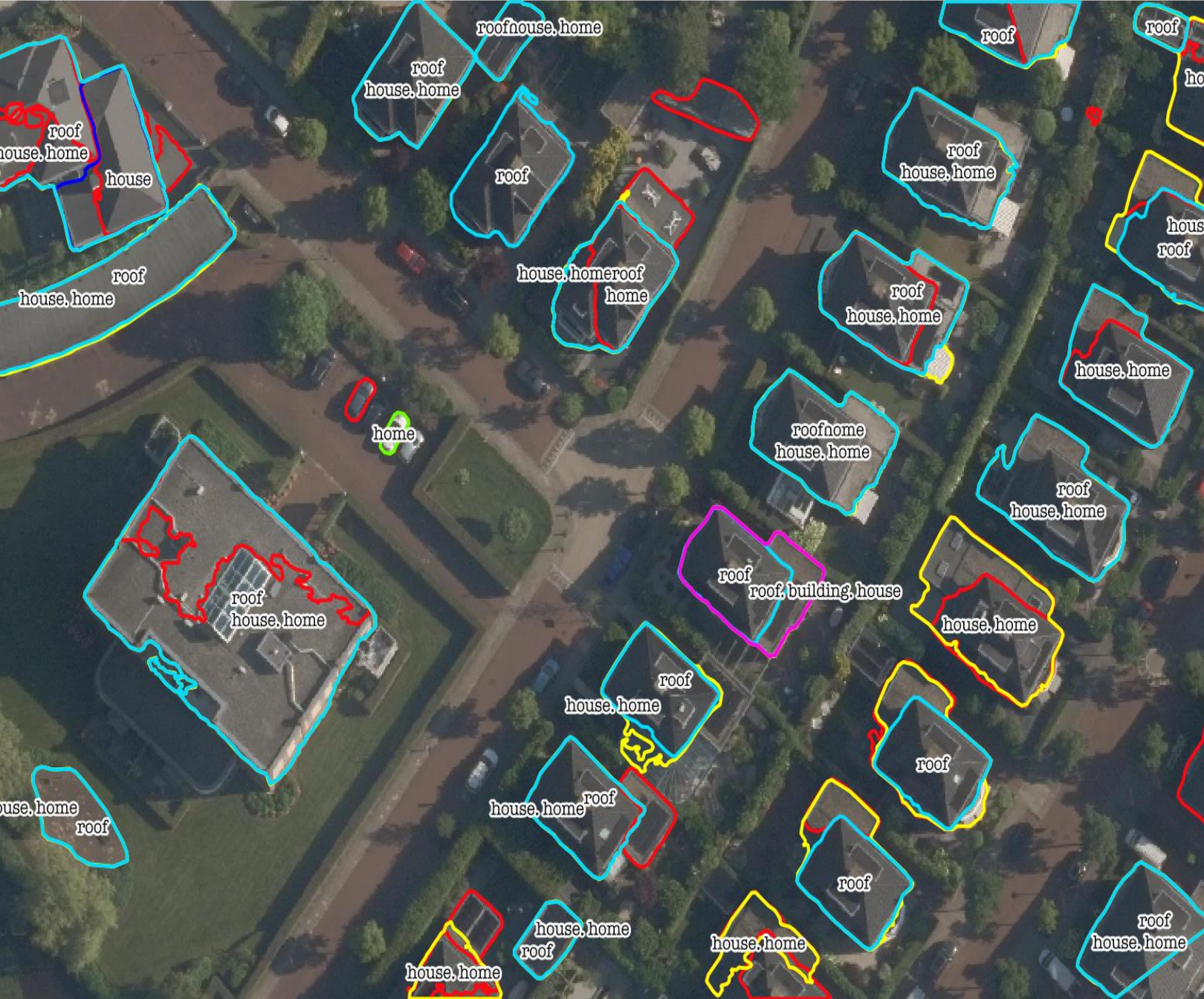
Prompt := "Roof.House.Building.Home."

Confusion Matrix

	16 (FP)
73 (FN)	151 (TP)

Aggregating Metrics

Precision	90.4%
Recall	67.4%
F1	77.3%



Diving in

Area ~ 15 dunam (150x100m)

GSD ~ 20cm/pixel

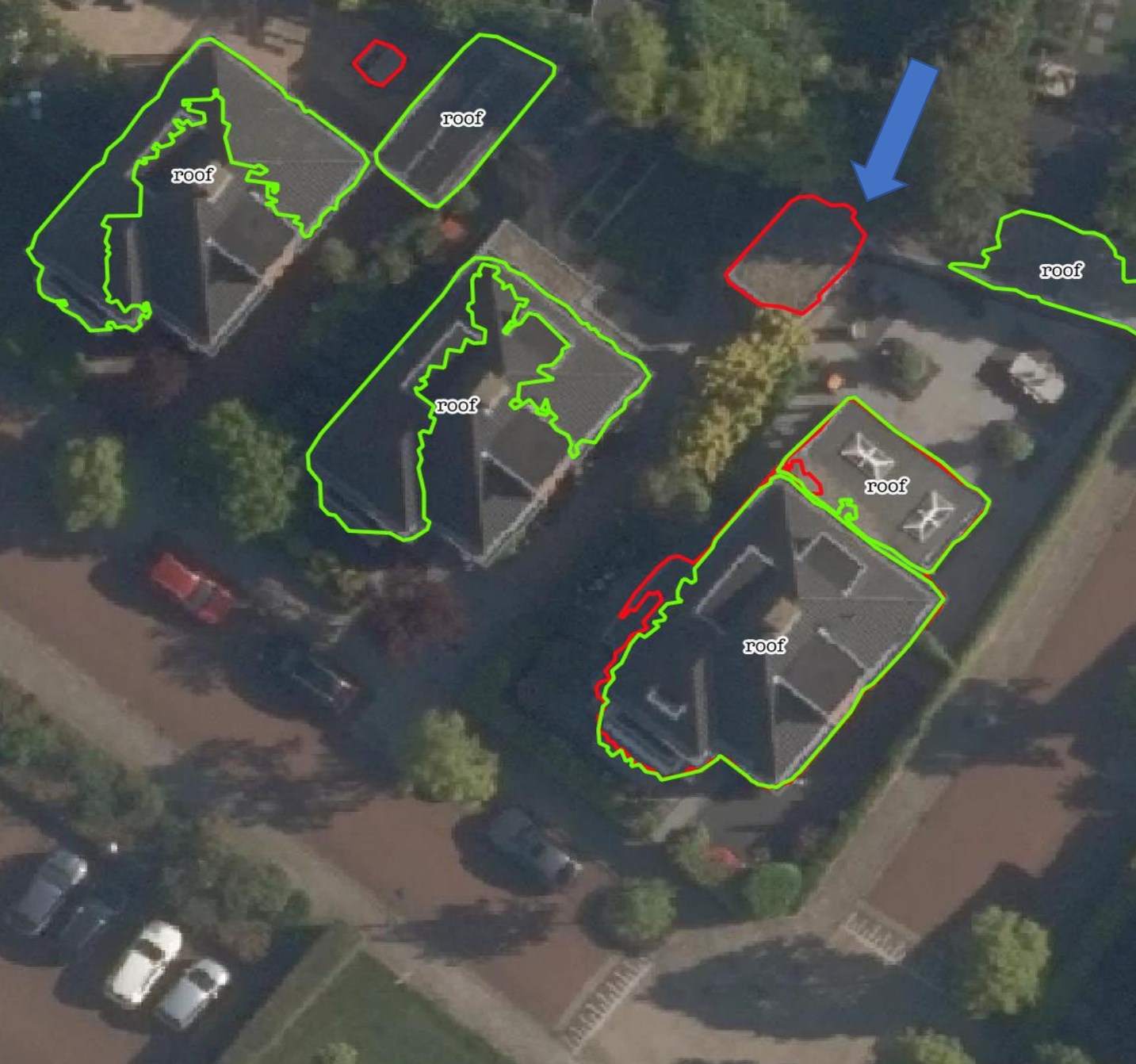
Prompt := "Roof.House.Building.Home."

Confusion Matrix

	11 (FP)
2 (FN)	69 (TP)

Aggregating Metrics

Precision	86.3%
Recall	97.2%
F1	91.4%



Diving in

Area ~ 2.5 dunam (50x50m)

GSD ~ 8cm/pixel (Rock bottom)

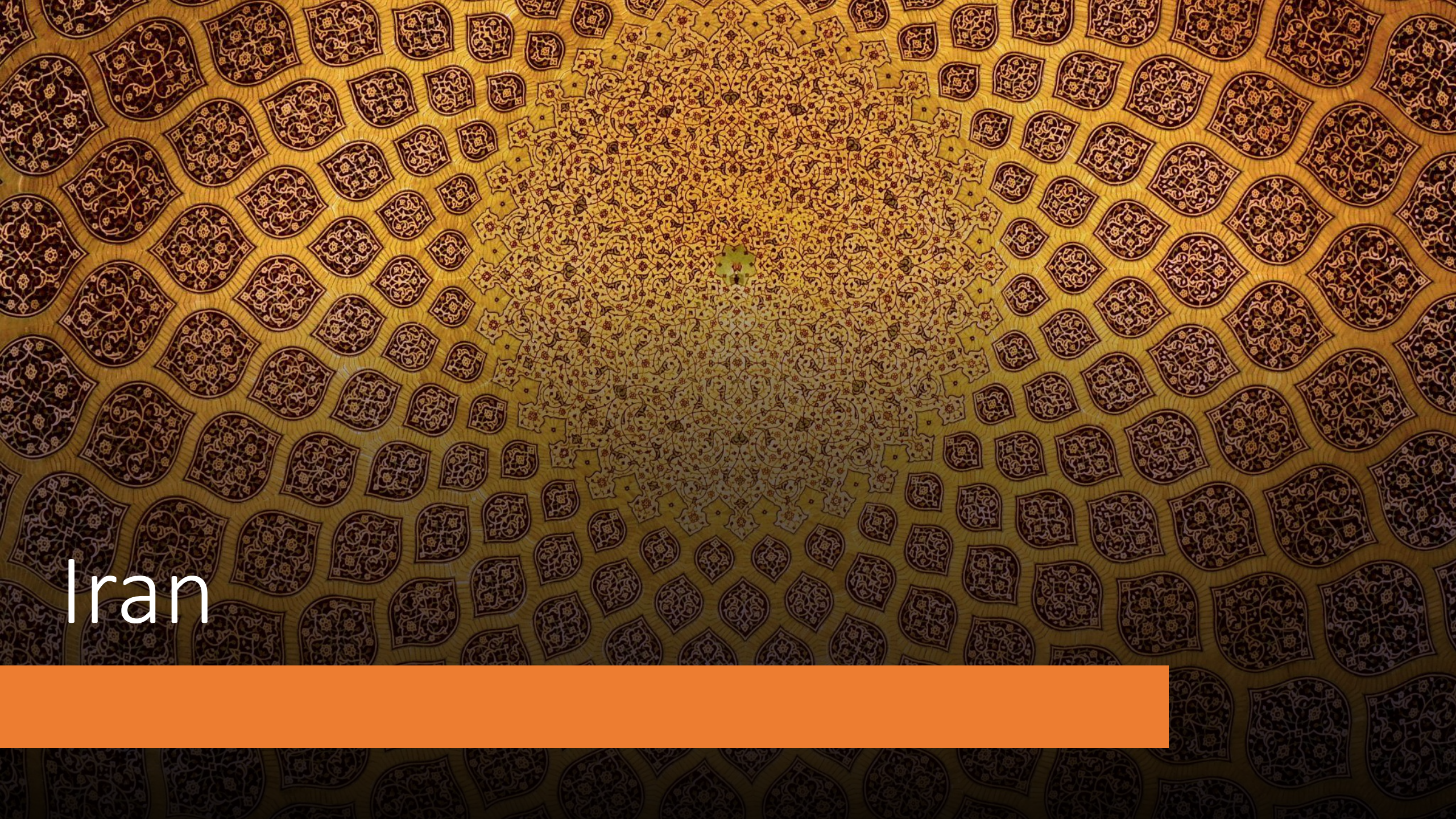
Prompt := "Roof.House.Building.Home."

Confusion Matrix

	1 (FP)
3 (FN)	15 (TP)

Aggregating Metrics

Precision	93.75%
Recall	83.3%
F1	88.23%



Iran



Area ~ 500 dunam (250x250m)

GSD ~ 50cm/pixel

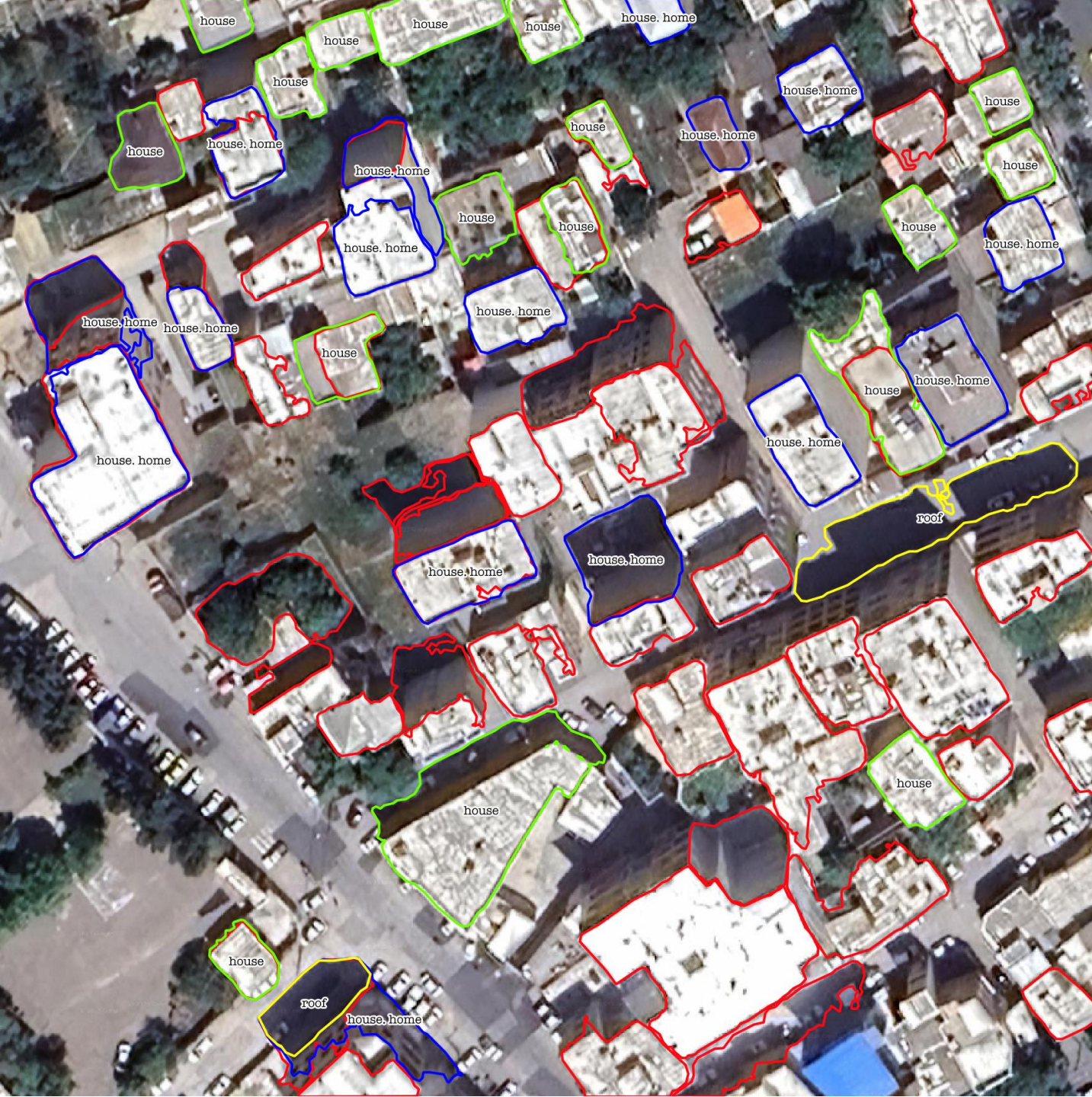
Prompt := "Roof.House.Building.Home."

Confusion Matrix

	? (FP)
? (FN)	? (TP)

Aggregating Metrics

Precision	?
Recall	?
F1	?



Area ~ 25 dunam (160x160m)

GSD ~ 34cm/pixel

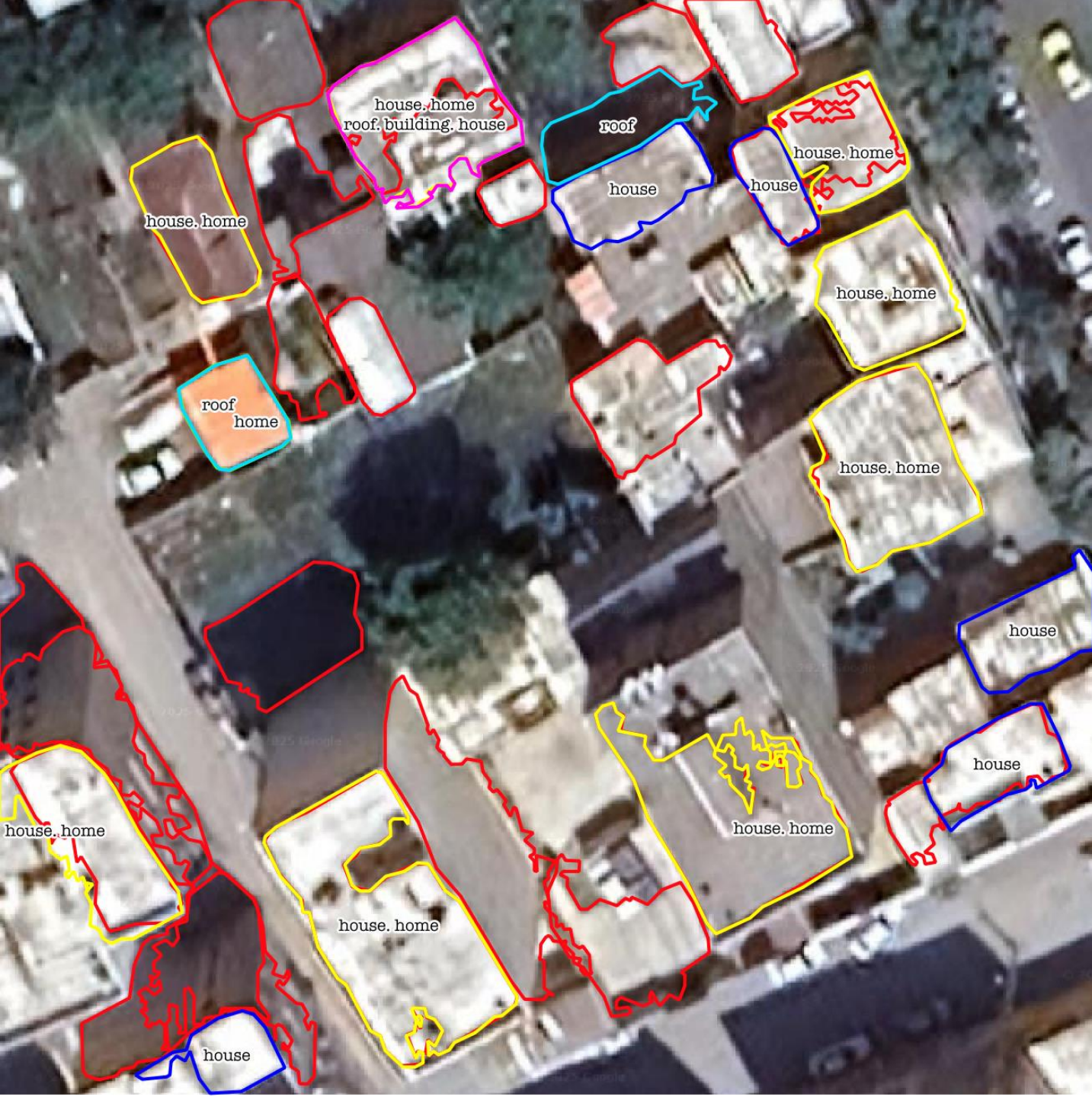
Prompt := “Roof.House.Building.Home.”

Confusion Matrix

	? (FP)
? (FN)	? (TP)

Aggregating Metrics

Precision	?
Recall	?
F1	?



Area ~ 5 dunam (75x75m)

GSD ~ 16cm/pixel (Rock bottom)

Prompt := "Roof.House.Building.Home."

Confusion Matrix

	? (FP)
? (FN)	? (TP)

Aggregating Metrics

Precision	?
Recall	?
F1	?



Did we say
ZERO-SHOT?



Area ~ 40 dunam (200x200m)

GSD ~ 50cm/pixel

Prompt := "Car. Tree. Water. Crosswalk.
Soccer field. Ship. Bridge."



Area ~ 1K dunam (1000x1000m)

GSD ~ 2.7m/pixel

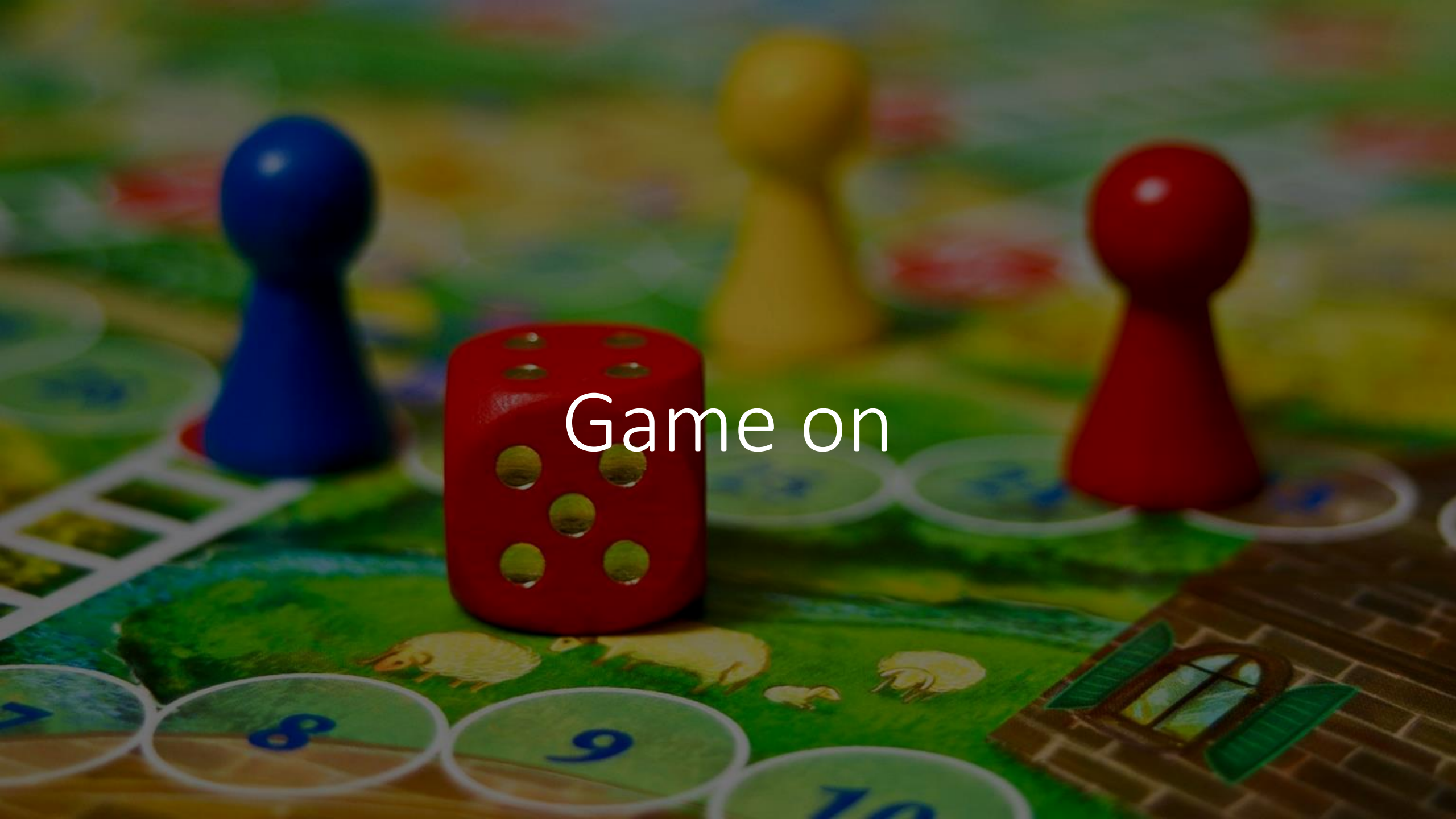
Prompt := "Airplane"

Future Enhancements

Better post-processing for segmentation masks

Faster inference times (GPU required)

Tiling mechanism to preserve resolution



Game on

