DINO Soars: DINOv3 for Open-Vocabulary Semantic Segmentation of Remote Sensing Imagery

Supplementary Material

Table 1. The prompt class name of the evaluation datasets.

Dataset	Class Name
OpenEarthMap	barren, grass, pavement, road, tree, water, cropland, building
LoveDA	building, road, water, barren, tree, farm
Potsdam, Vaihingen	road, building, low vegetation, tree, car

1. Prompt Details

Every dataset contains a named set of classes. Instead of using these names directly, we have found it beneficial to make some logical modifications to the existing names, which we list here in Tab. 1 for reproducibility. Our heuristic in making substitutions is to target classes with complex or abstract semantic names (e.g. "agriculture") and replace them with something simple and concrete (e.g. "farm"). There may be more gains to be realized with sophisticated prompt tuning, but this is beyond the scope of our work. Note that the semantic names used for Potsdam and Vaihingen are left unmodified. We believed that the model would struggle to apply the "low vegetation" class, so we attempted some simple substitutions (grass, greenery, etc.) only to find that using low vegetation gave the best performance. For LoveDA, we realized small performance gains by substituting "forest" and "agriculture" with "tree" and "farm", respectively. For OpenEarthMap, we substituted "bareland", "rangeland", "developed space", and "agriculture land" with "barren", "grass", "pavement", and "cropland", respectively.

Besides class names, the other variable in forming prompts is the choice of surrounding sentence. Ensembling is a common approach, in which a group of "wrapper" prompts is submitted to the text encoder for each semantic class, and the mean of all embeddings is used downstream. We use the set of wrappers defined by [**?**] in their code implementation, and repeat them here in Tab. 2 for easy reference.

We show aggregated cost maps for Potsdam, Vaihingen, LoveDA, and OEM in Fig. 1, Fig. 2 ,Fig. 3, Fig. 4, respectively, for both vanilla DINOv3 and CAFe-DINO. Aggregated cost maps are effectively a per-class probability score, so they are a good empirical heuristic of model uncertainty and identifying problem classes. For example, the "car" cost map for Potsdam is much sharper than the one for Vaihingen for both models, and therefore the predictions are weaker on the Vaihingen image (we attribute this particular phenomenon to the non-RGB spectra of the Vaihingen dataset). A core limitation of CAFe-DINO (and RS-training-free methods in general) can be gleaned from the cost maps of OEM, particularly the near-identical Grass and Cropland maps. We conjecture that the difference between texture-characterized classes such as grass and crops in a satellite image is too fine-grained for a model trained on natural imagery to distinguish, though more shape-characterized greenery such as trees are more easily identified.

CAFe-DINO cost maps are generally sharper and higher-contrast than DINOv3 cost maps as a result of cost aggregation. Some classes are segmentable by DINOv3 out-of-the-box (such as cars in Potsdam), while others produce almost completely unstructured cost maps (such as Low Veg. in Potsdam). These unstructured cost maps benefit the most from cost aggregation.

Note that the cost map of a class is aggregated relative to other classes in the given semantic set, so changing the input set of classes will change the cost map of a given class.

Table 2. Prompt wrappers for CAFe-DINO. "{}" is a placeholder for a desired semantic class.

a photo of {}
an image of {}
a photograph of {}
a picture of {}
a photo of a {}
an image of a {}
a photo of the {}
an image of the {}
a close-up photo of {}
a cropped image featuring {}

2. COCO-Stuff Subset

We use a subset of the COCO-Stuff dataset containing only the following classes:

Classes: bicycle, car, motorcycle, airplane, bus, train, truck, boat, bridge, building, bush, dirt, fence, grass, gravel, ground, hill, house, leaves, metal, mountain, mud, pavement, plant, platform, playing field, railing, railroad, river, road, rock, roof, sand, sea, skyscraper, snow, stone, structural, tree, water, wood.

Our heuristic in forming this dataset was to remove semantic classes that could never exist in satellite imagery at current resolution capabilities, such as household objects and food items. Using a reduced semantic set brings the benefit of smaller cost volumes, resulting in a substantial decrease in GPU memory and FLOPs for both training and inference. Additionally, we find that training convergence with the reduced set is very short (45,000 iterations).

3. Additional Cost Maps

On the following pages, we provide additional cost maps for samples from the Potsdam (Fig. 1), Vaihingen (Fig. 2), LoveDA (Fig. 3), and OEM (Fig. 4) datasets.

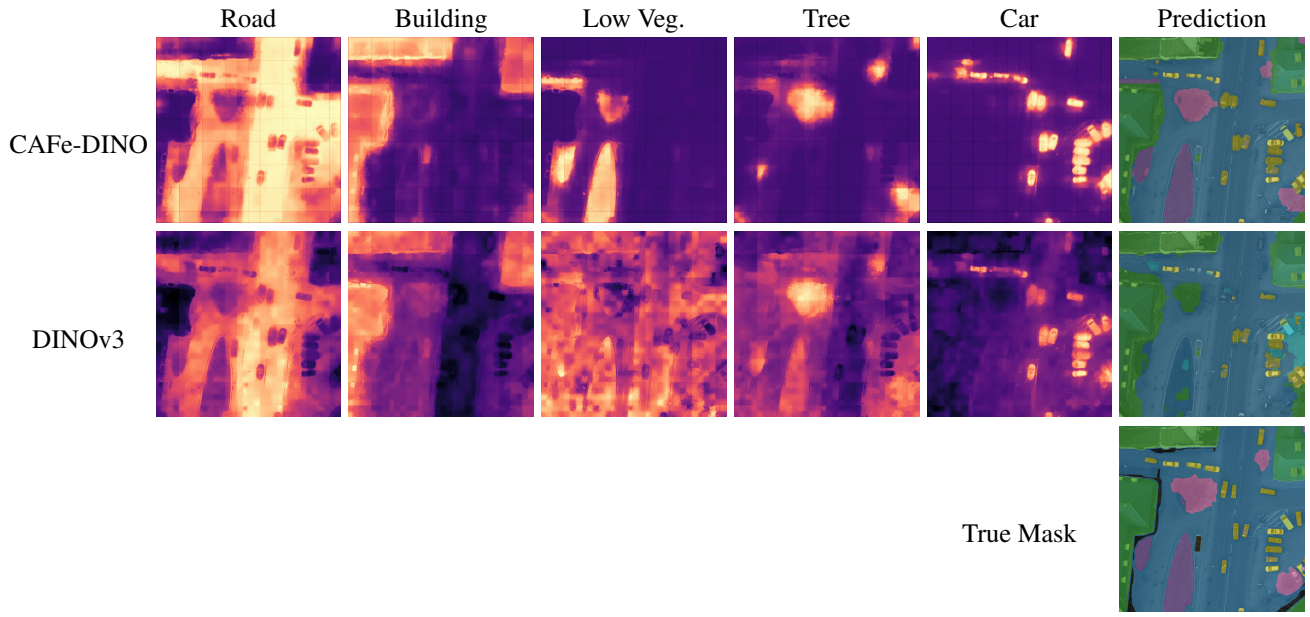

Figure 1. Cost maps for a Potsdam image.

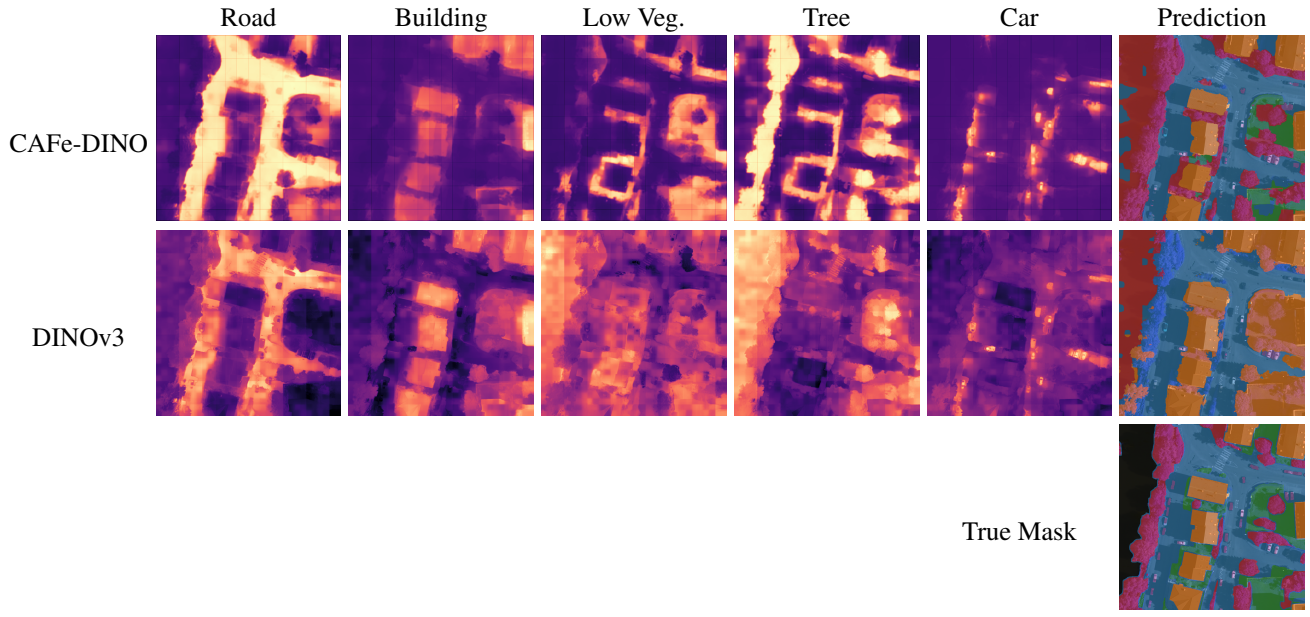

Figure 2. Cost maps for a Vaihingen image.

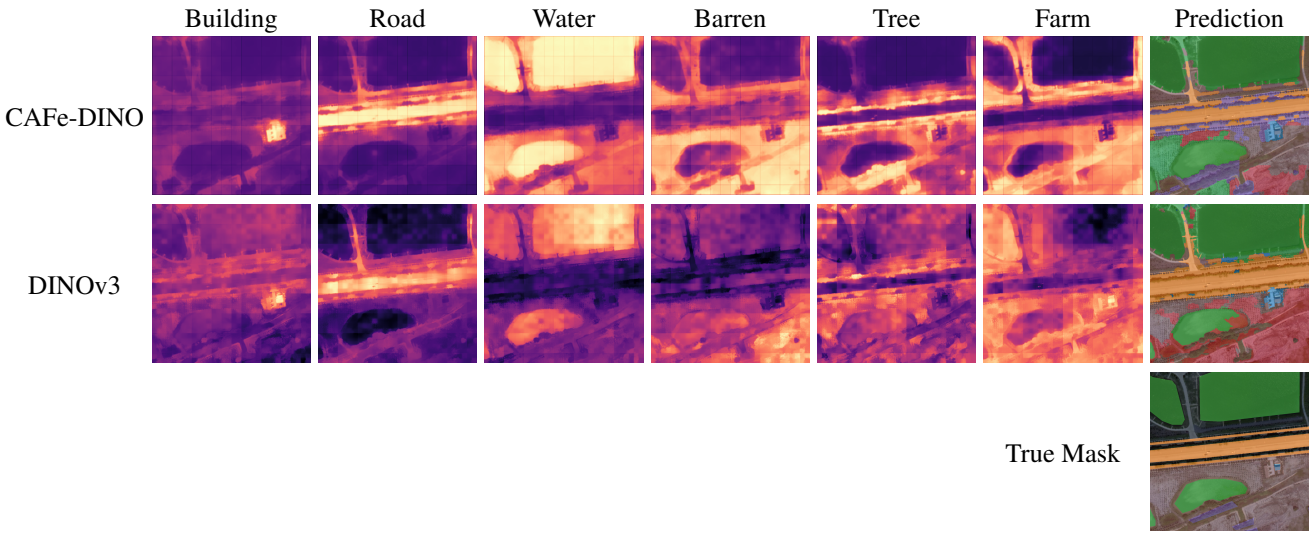

Figure 3. Cost maps for a LoveDA image.

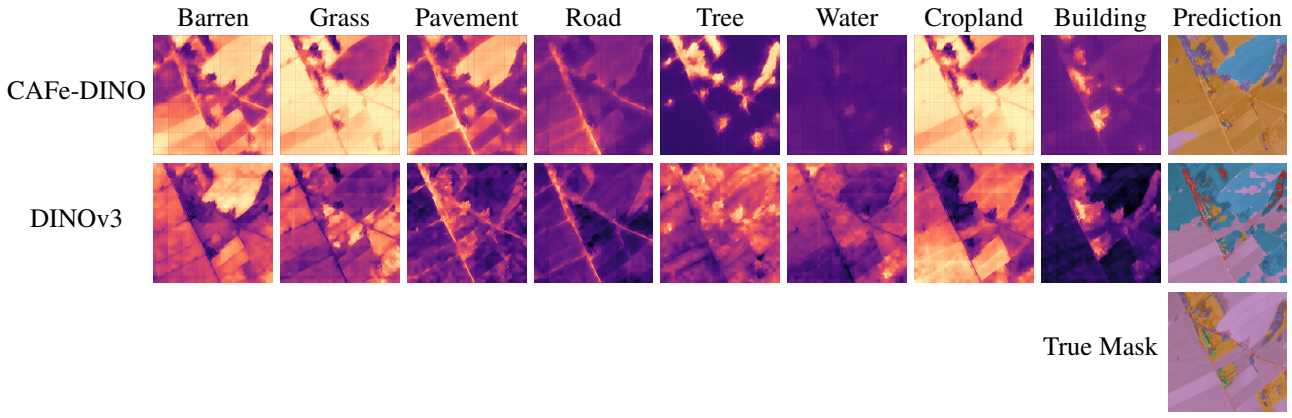

Figure 4. Cost maps for an OEM image.