Quantifying the Limits of Segmentation Foundation Models: Modeling Challenges in Segmenting Tree-Like and Low-Contrast Objects

Yixin Zhang*, Nick Konz*, Kevin Kramer, Maciej A. Mazurowski
*Equal contribution. (Correspondence to: yz696@duke.edu)

TL;DR:

Prior evaluations of segmentation foundation models (e.g., SAM) are largely observational and case-based, offering limited insight into *what features drive performance*.
We identify two key factors all failure cases can be attributed to, and propose quantitative metrics to measure them, showing a strong correlation between these factors and segmentation outcomes.

Task and Challenges

Background: Segmentation Foundation Models (SFMs) demonstrate strong zero-shot performance, raising the prospect of using them for generic run-time perception or to assist the annotation of objects whose labels are scarce or costly.

Challenges: However, without understanding when and why these models fails (i.e., the failure modes), we cannot deploy them with trust in highly automated or safety-critical settings.

Objective: *Identify and quantify* the failure modes of the existing SFMs and potential mitigation methods.

Metric for Tree-Likeness

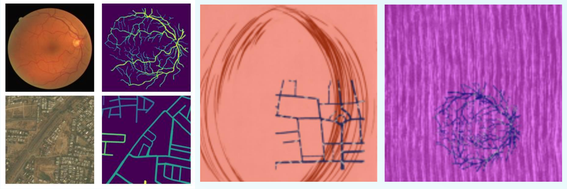
One feature that SFMs tend to fail on is Tree-likeness. For this, we proposed two metrics, namely Contour Pixel Rate (CPR) and Difference of Gini Impurity Deviation (DoGD), such that

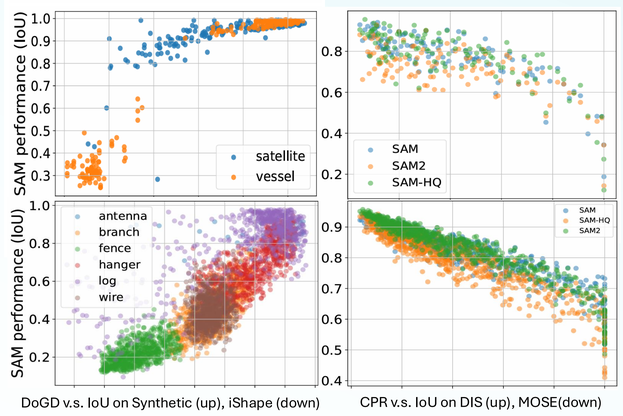
$$CPR = \frac{|C|}{|F|}, \qquad DoGD = \sigma_a^{Gini}(m) - \sigma_b^{Gini}(m)$$

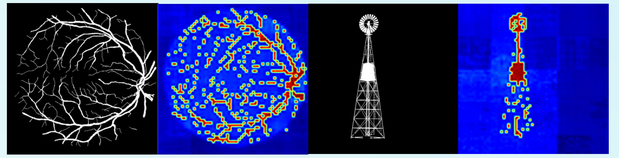
Here, CPR is defined as the ratio of the number of pixels on the object contour to the number of pixels in the object, and DoGD measures the homogeneity of the object's spatial occupancy across different scales.

Failure Mode on Tree-Likeness

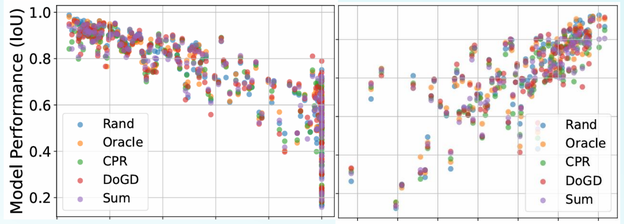
On both synthetic and real datasets, statistically significant correlation between the two metrics and the performance of different SFMs can be observed. High tree-likeness in image consistently yield lower model IoU and broken attention.







The broken mask-guided attention roll-out indicates that existing SFMs cannot fully capture free-form structures.



This property of existing SFMs cannot be easily mitigated through fine-tuning on a small dataset. Fine-tuning on DIS, a dataset containing finer structures, yields similar results when samples are ranked by (1) descending IoU, (2) descending tree-likeness metric, or (3) random selection. However, none removed the model's sensitivity to tree-likeness. In contrast, a baseline model trained from scratch on the same dataset shows no statistically significant correlation. We believe this property is deeply rooted in the training data and the encoder itself, causing such features to be irreversibly suppressed.

Metric for Textural Separability

Another characteristic on which SFMs tend to fail is low textural separability. For a SFM to generalize, it must encode the target object in a representation sufficiently distinct from its surroundings. We define the textural separability metric as:

$$Acc_{clf}(h_{obj}, h_{boundary})$$

Here, h_{obj} denote the feature activation map for pixels inside the objects, while $h_{boundary}$ denotes the feature activation maps of pixels adjacent to, but outside, the object boundary. The function $Acc_{clf}(.)$ represent the train accuracy of a deliberately chosen weak classifier.

Failure Mode on Textural Separability

Across DIS, iShape, Plittersdorf, MOSE, and NST-VOC, a consistent positive correlation between IoU and the proposed textural separability metric is observed across multiple object classes.