Objects in Focus: Predicting Object-Based Attention from Spatial Features

Elizabeth Hall, Meta; Zoe Loh, UC Merced

ICCV OCT 19-23, 2025 HONOLULU HAWAII

Background

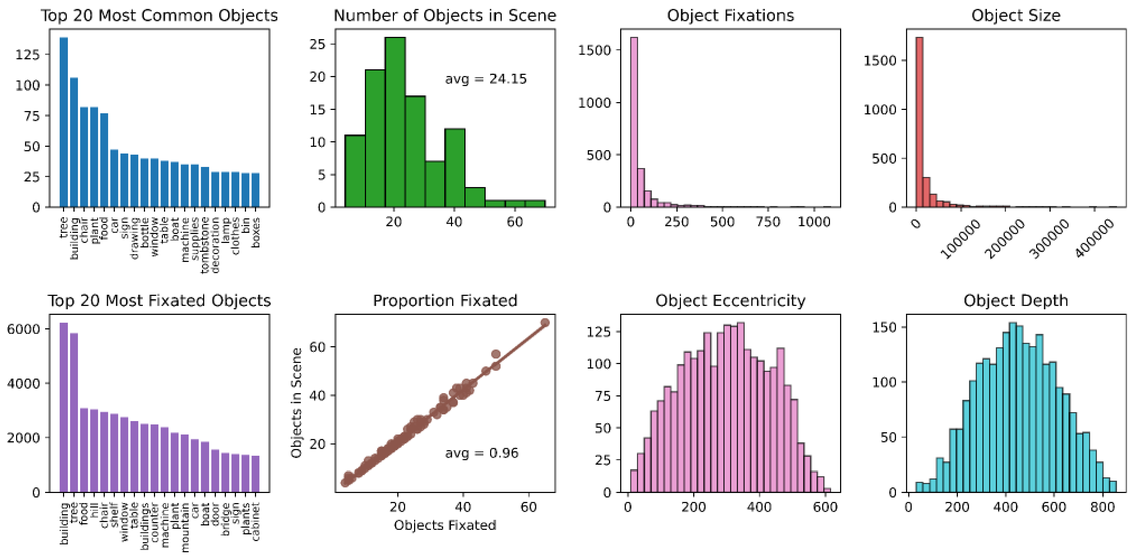
Visual objects are discrete, bounded entities distinguished by features, processed by the brain as units segmenting scenes along physical boundaries

Channel and spatial attention guide models on *what* and *where* to focus, guiding segmentation tasks by prioritizing relevant visual information.

Object-based attention focuses on goal-relevant objects, even when they occupy a small part of the visual field.

Datasets

OiF: 100 real-world scenes (50 indoor, 50 outdoor), fully segmented at 1024×768 px resolution with 360 semantically unique ConceptNet labels; viewed by 100 participants for 12 s per scene.

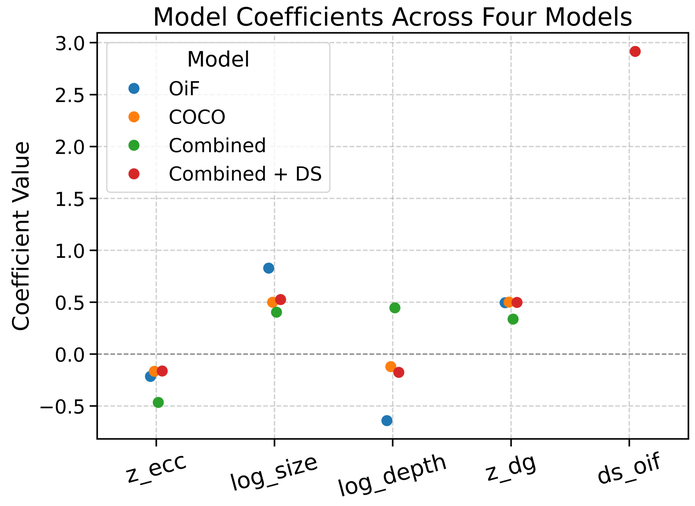


COCO-Freeview: 2243 images, scaled to 1050 px height. Masks for identical IDs were separated; objects <0.05% image area were filtered out; arrays were downscaled to 1/30th, connected component analysis was applied, then masks were upscaled to compute values for each of 168 unique labels.

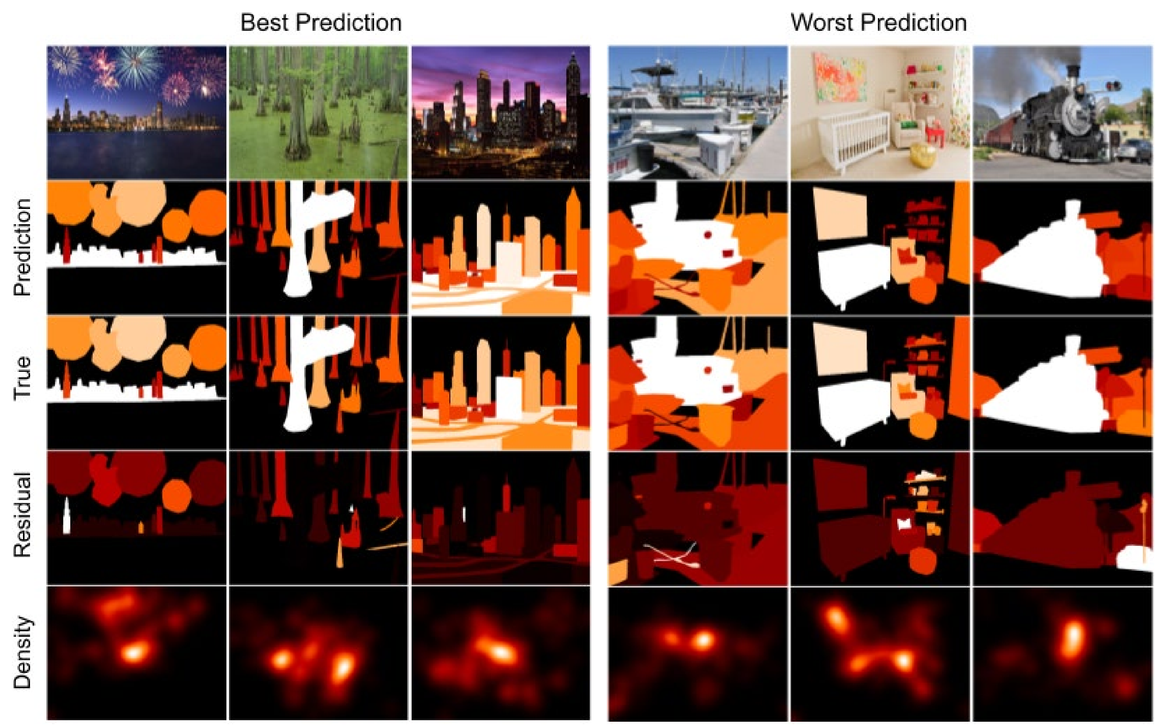
Results

Predicted eye-movement distributions across objects using object size, eccentricity, depth, and salience; size = pixel count of mask, eccentricity = Euclidean distance from object centroid to image center, depth = MONODEPTH2 estimate at centroid, salience = DeepGaze IIE value at centroid.

Skewed variables (fixations, size, depth) were log-transformed, and a multiple linear regression modeled log cumulative fixations as a function of log(depth), log(size), z-scored eccentricity, and z-scored salience.



Best and worst performing images for object-attention prediction in OiF — rows show (1) model-predicted attention, (2) true attention distribution, (3) absolute residuals per object, and (4) Gaussian fixation density; each heatmap independently normalized (black = min, white = max).



Discussion

Penultimate ImageNet-trained CNN layers can model object-level features analogously to human vision, but generalization to unseen fixation data is limited. Improving performance may require object detection–based features, larger fixation pretraining, and viewpoint-aware (static vs. egocentric) modeling.