Scenes are complex stimuli, rich with information that provide the fundamental visual input that allows humans and many artificial systems to reason about the world and act in their environments. Current models in cognitive science do not capture this complexity and typically only account single dimensions (e.g., geometric layout). I will argue that scene understanding and underlying neural mechanisms for scene processing result from the extraction of scene features that have strong, statistically robust associations learned over an organism’s lifetime of experience (e.g., spatial relations – mirror is above the sink, co-occurring objects – oven is found in a kitchen, and diagnostic mid-level visual features – bamboo forest has vertical lines).

Human neuroimaging reveals that associative processing of spatial relations and the prevalence of co-occurring objects can account for the kinds of neural responses we observe during scene understanding. However, some information about associations within scenes may not be carried by objects or their spatial relations. Instead, I posit that mid-level scene attributes support scene understanding. To explicate the nature of such mid-level features, I compared several different artificial vision models, including some that leverage the web-scale analysis of images. These models provide visual vocabularies of potentially critical features that define a scene and may be used in its neural representation. Data from functional MRI revealed that different models had different levels of explanatory power and that, for the best performing models, the model features played a role in accounting for fine-grained neural activity.

Associative processing offers a new framework for explaining the functional and neural mechanisms underlying scene understanding. More generally, the framework I have developed provides roles for both bottom-up and top-down processing in scene understanding in that associations between scene tokens can be used to generate predictions and constrain the space of possible scene contexts that are likely to occur in our environment. Insights from this framework show promise for improving our understanding of biological scene processing and, ultimately, enhancing artificial vision systems’ performance.