We propose SORNet: Spatial Object-centric Representation Network to learn object-centric embeddings that encode spatial relationships.

SORNet generalizes zero-shot to scenes with unseen objects and different number of objects.

SORNet is trained only on classification of logical predicates, but captures continuous spatial relationships.

Training objective (logical)

Testing objective (continuous)