Discussion Leader Cover Letter

Rex Stayer-Suprick

Paper: Pohland, S., Tomlin, C. (2024). Understanding the Dependence of Perception Model Competency on Regions in an Image. In: Longo, L., Lapuschkin, S., Seifert, C. (eds) Explainable Artificial Intelligence. xAI 2024. Communications in Computer and Information Science, vol 2154. Springer, Cham. https://doi.org/10.1007/978-3-031-63797-1$_8$.

The paper studies explainability for DNN-based perception model competency. In clearer terms it seeks to understand *why* a model lacks confidence in its predictions, not merely *that* it does, bridging uncertainty quantification, OOD detection, and explainable AI. The central question asks which methods best estimate *"the dependence of the competency score on regions in the input image to understand the extent to which different regions contribute to low model competency."* The paper is framed as a comparative evaluation across three failure scenarios: unfamiliar objects, unseen classes, and unexplored environments.

1. Problem

Existing uncertainty and OOD methods tell a system *when* a model is incompetent but not *why*, limiting autonomous systems from responding appropriately. In safety-critical applications (robotics, autonomous vehicles, remote sensing) a system that can spatially localize the source of incompetency can make far more informed decisions about whether to proceed, defer to a human, or seek additional data.

2. Claims

Reconstruction Loss, an autoencoder trained to reconstruct image regions from the classifier's feature vector, where high reconstruction error signals unfamiliarity, is the most reliable method overall for explaining areas of low model competency. **Competency Gradients**, which computes partial derivatives of the competency score averaged over image segments, provides complementary localization and combining both yields the highest TPR for unexplored environment detection. Both run in ≤ 0.1 seconds, making them viable for real-time pipelines unlike the Cropping, Masking, and Perturbation baselines.

3. Evidence

Dataset	Train Distribution	Test Novelty
Lunar (simulated)	Uninhabited moon	Astronauts & structures
German speed signs	Signs $\geq$ 30 km/h	20 km/h sign (unseen class)
Outdoor park	Forest & grass	Pavilion area

Evaluation compares five proposed methods against each other using qualitative saliency map comparisons and pixel-level binary classification metrics against manual ground-truth annotations; an appendix additionally benchmarks against ten CAM baselines.

4. Statistical Analysis

Accuracy, TPR, TNR, PPV, and NPV are averaged across test images descriptively. No inferential statistics or significance tests are reported. Reconstruction achieves 96% overall accuracy on the lunar dataset (TPR 89%, TNR 99%) versus 60–84% for all other methods. Performance degrades on the speed limit and park datasets, attributed to visual similarity between novel and familiar features and to artifacts from the Felzenszwalb segmentation algorithm.