Dense Out-of-Distribution Detection

**Goal:** detect out-of-distribution objects in semantic segmentation data.

**Challenges:**
- Complex in-distribution data, few classes but diverse appearance.
- Data is not object-centric, entities interact in the scene.
- Need for fast and accurate dense predictions.

**Dense OoD Detection with Deep Nearest Neighbors**

1. Select a trained semantic segmentation model.
2. Use the model to collect a set of $N$ local reference features $R \in \mathbb{R}^{N \times C}$ from in-distribution samples (training dataset). Each feature vector encodes an image patch.
3. At inference time, collect features for the test sample: $T \in \mathbb{R}^{H \times W \times C}$.
4. For each test feature, compute the expected $L_2$ distance to its nearest neighbors. The resulting distance map will serve as the anomaly scores.

As baseline we use model uncertainty, computed as LogSumExp(logits).

**Selecting Good Features**

We compare representations from convolutional (ResNet, ConvNeXt) and attention-based (ViT, M2F-EAM) models.

- All networks are trained for supervised semantic segmentation.
- No out-of-distribution data is used.

Key findings:
- kNN distances are overall better than model uncertainty.
- Transformer features greatly outperform CNN ones.
- Attention features (e.g., keys) perform best.
- kNN anomaly scores can be combined with parametric model uncertainty.

**Runtime**

- Overall and added runtime depends on the architecture (feature size and resolution).
- Search time for kNNs with 100k reference features (as for other results) is a small fraction.

**Analysis**

What makes attention features well suited for OoD detection with kNNs?

Hypotheses:
- Self-attention is an implicit RBF kernel machine, similar to SVMs [2, 3].
- The multi-head architecture reduces the effective feature dimensionality.

Limitations of our approach:
- kNNs resolution is typically lower than model uncertainty and depends on the architecture. This harms performance on very small objects.
- Performance is tied to a specific class of models (transformers).
- Additional time and memory requirements, model dependent.

**Qualitative examples**

State-of-the-Art comparison:
- Separate mask prediction gives smoother anomaly scores (row 1).
- It also worsens false positives/negatives (rows 2-4).

**References**


**Acknowledgment:** Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 417962828