Drawbacks of prior work
- Can only localize the most salient object in an image
- Need training to localize multiple instances per image

Motivation
- Self attention of features from last layer of Self-supervised transformers demonstrates interesting properties.

Main Idea
- Leverage similarity maps of features from large self-supervised transformers to localize objects without supervision.
- Fractal analysis to identify similarity maps that belong to foreground objects.

Results
- State-of-the-art localization and discovery performance.
- Effective proposal generator for unsupervised pre-training of object detectors.

Approach Overview
- Entropy-based Box Analysis
  - Features $F$ from an image are extracted using DINO and the outer product $A = FF^T$ is computed.
  - The Entropy-based Box Analysis (EBA) module identifies similarity maps of foreground tokens from $A$.
  - Highly redundant neighboring tokens are grouped using spatial clustering to obtain pools.
  - Similarity maps of tokens from each pool are processed to obtain a single bounding box.
  - MOST can automatically identify multiple objects in an image and can do so without any training.

Entropy-based Box Analysis
- We observe that similarity maps of tokens on foreground objects are “less” random spatially.
- EBA uses box counting to segregate similarity maps of tokens on foreground patches from those of background.
- On each similarity map, we perform a raster scan with increasing box sizes.
- Each map is then average pooled and flattened to compute entropy since we are interested in randomness of maps.
- A similarity map belongs to a token on foreground patch if its entropy $\leq a + b \log(n^2)$ where $a = 1$ and $b = 0.5$.

Effect of Clustering
- EBA module identifies redundant tokens.
- Without clustering localization is very noisy.
- Redundancy can be reduced by spatial clustering.
- Each spatial cluster, aka pool, focuses on a different object/instance.

Approach Overview

Effect of Clustering

Qualitative results
- MOST can also be used to perform unsupervised saliency detection by using the mask obtained from the largest pool.

Acknowledgements
This project was partially supported by DARPA Xanadu (D18AP00007) and DARPA SAIL-ON (W911NF20-20-009), and Amazon Research Award to Abhinav Shrivastava, Rama Chellappa was supported by an ONR MURI (N00014-20-1-2787).