UNSUPERVISED WHOLE OBJECT DISCOVERY BY CONTEXTUAL GROUPING WITH REPULSION

Anonymous authors
Paper under double-blind review

ABSTRACT

It is challenging to discover and segment *whole* objects from *unlabeled* images, as features unsupervisedly learned on images tend to focus on distinctive appearances (e.g., the *face* rather than the *torso*), and grouping by *feature similarity* could reveal only these representative parts, not the whole objects (e.g., the *entire human body*). Our key insight is that, *an object of distinctive parts* pops out as a whole, due not only to *how similar they are to each other*, but also to *how different they are from their contexts* within an image or across related images. The latter could be crucial for binding different parts into a coherent whole without preconception of objects. We formulate our idea for unsupervised object segmentation in a spectral graph partitioning framework, where nodes are patches and edges are grouping cues between patches, measured by feature similarity for attraction, and by feature dissimilarity for repulsion. We seek the graph cuts that maximize within-group attraction and figure-ground repulsion while minimizing figure/ground attraction and within-group repulsion. Our simple method consistently outperforms the state-of-the-art on unsupervised object discovery, figure/ground saliency detection, and unsupervised video object segmentation benchmarks. In particular, it excels at discovering whole objects instead of salient parts.

1 INTRODUCTION

We consider the unsupervised learning task of discovering and segmenting *whole* objects from a set of unlabeled images. Any computational model that achieves this goal is not only useful in practical applications, where segmentation annotations are tedious and costly to obtain, but also illuminating in understanding how infants make sense of their visual world from initial undivided sensations.

Existing works (Yang et al., 2019b; 2021b; Liu et al., 2021) accomplish this task by learning from unlabeled videos. AMD (Liu et al., 2021) assumes that a video contains different views of the same

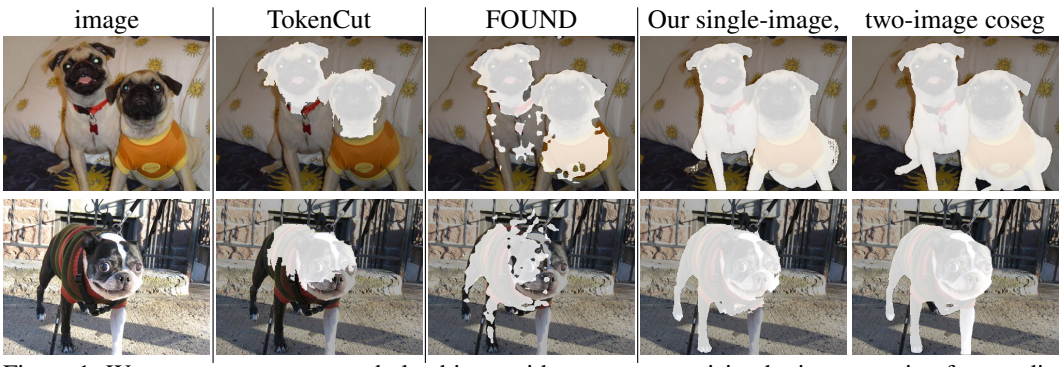

Figure 1: We propose to segment whole objects without any supervision by incorporating feature dissimilarity (repulsion) as cues. Existing methods TokenCut (Wang et al., 2023) and FOUND (Siméoni et al., 2023), which rely solely on feature similarity, often segment partial objects, like the *dog's face*, while missing other components like *legs* or *bodies*. In contrast, we capture the nexus of feature similarities and dissimilarities within and across images in a joint weighted graph. This enables the segmentation of entire objects from their backgrounds.

scene related by moving components, and the right region segmentation and region flow can be learned concurrently by image reconstruction over time. However, the performance of AMD is constraint by its piece-wise constant motion assumption within the same segment. Recent works (Wang et al., 2023; Melas-Kyriazi et al., 2022) show that objectness can be discovered from unlabeled images in attention maps of self-supervised ViT such as DINO (Caron et al., 2021). Nevertheless, features learned in such a self-supervised manner (Wu et al., 2018; Chen et al., 2020a; He et al., 2020; Misra & Maaten, 2020) tend to focus on distinctive appearances. If we cluster patches by feature similarity via e.g., TokenCut (Wang et al., 2023), we can only discover parts of characteristic appearances such as *faces*, but miss parts of plain appearances such as *torso* and *legs* of a whole object (Fig. 1).

We aim to discover *whole* instead of partial objects without any preconception of objects. Our key insight is that, *an object of distinctive parts* pops out as a whole, due not only to *how similar they are to each other*, but also to *how different they are from their contexts* within an image or across related images. The latter could be crucial for binding different parts into a coherent whole, in a bottom-up data-driven manner. For example, while the *faces* of two different dogs look similar, their *torsos* and *legs* are only mildly similar to the *faces*. However, all these parts are more dissimilar to their surrounding backgrounds. It's this common *repulsion* against the contexts they are embedded in, in addition to *attraction* of varying strengths within the objects, that helps bind object parts of heterogeneous appearances into coherent wholes.

We formulate our approach to unsupervised whole object segmentation within a spectral graph partitioning framework. In this setup, image patches are represented as nodes, and edges are the grouping relationships between these patches. These relationships are quantified using two measures: attraction based on feature similarity, and repulsion based on feature dissimilarity. The goal is to find the graph cuts that simultaneously maximize within-group attraction and between-group repulsion, while minimizing figure-to-background attraction and within-group repulsion.

We investigate this idea not only within a single image but also across related images in a co-segmentation setting, where contextual grouping with repulsion between similar images brings additional power for discovering whole objects together (Fig. 1). These images should be semantically similar yet visually distinct; *if they are identical, co-segmentation lacks new information, and if they are semantically unrelated, co-segmentation becomes ineffective*.

We present a method for unsupervised object segmentation by contextual grouping with repulsion, named CGR . With whole objectness revealed by attraction and repulsion, we further refine the self-supervised features with an attached segmentation head over the whole object masks. Our method consistently outperforms the state-of-the-art on unsupervised object discovery, unsuperivsed saliency detection, and unsupervised video object segmentation benchmarks.

Contributions. **1)** We propose to leverage contextual relationship from both within-image and cross-image to group distinctive parts into coherent whole objects without any annotations. **2)** We optimize the grouping with a new framework using attraction and repulsion cues of self-supervised ViT features. **3)** We further enhance the self-supervised ViT features by re-training the backbone with an attached segmentation head over whole object masks, thereby achieving state-of-the-art performance on multiple unsupervised segmentation benchmarks.

2 Related Work

Unsupervised Object Discovery. Unsupervised object discovery aims at localizing and segmenting objects from a collection of unlabeled images. Most of current works utilize self-supervised features from visual transformers (Caron et al., 2021; Chen et al., 2020b; Caron et al., 2020). SelfMask (Shin et al., 2022) applies spectral clustering on multiple self-supervised features to extract object masks. LOST (Siméoni et al., 2021) utilizes feature similarities to localize an object seed and expands the seed to all similar patches. FreeSOLO (Wang et al., 2022) presents a FreeMask predictions from feature similarities which are ranked and filtered by a maskness score. Deep Spectral Methods (Melas-Kyriazi et al., 2022) and TokenCut (Wang et al., 2023) build a weighted graph using feature similarities (attraction) and conduct graph cut to separate objects from backgrounds. FOUND (Siméoni et al., 2023) first searches a background seed to localize objects and HEAP (Zhang et al., 2024) applies contrastive learning to learn clustered feature embeddings. PEEKABOO (Zunair & Hamza, 2024) presents to hide part of images and localize the objects with remaining image information. However, all these methods are limited in discovering whole objects as self-supervised features only capture