Image Segmentation-based Unsupervised Multiple Objects Discovery

Sandra Kara Hejer Ammar Florian Chabot Quoc-Cuong Pham

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France
{firstname.lastname}@cea.fr

Abstract

Unsupervised object discovery aims to localize objects in images, while removing the dependence on annotations required by most deep learning-based methods. To address this problem, we propose a fully unsupervised, bottom-up approach, for multiple objects discovery. The proposed approach is a two-stage framework. First, instances of object parts are segmented by using the intra-image similarity between self-supervised local features. The second step merges and filters the object parts to form complete object instances. The latter is performed by two CNN models that capture semantic information on objects from the entire dataset. We demonstrate that the pseudo-labels generated by our method provide a better precision-recall trade-off than existing single and multiple objects discovery methods. In particular, we provide state-of-the-art results for both unsupervised class-agnostic object detection and unsupervised image segmentation.

1. Introduction

Deep learning methods have shown tremendous advances in resolving several computer vision tasks such as object detection and image segmentation. However, massive amounts of carefully labeled images are necessary to train reliable deep learning models that can reach high performances. Due to the high cost of such manual annotations, several approaches were proposed to use only limited amounts of annotated data, such as semi-supervised learning, weakly supervised learning or few-shot learning. In this work, we address the problem of localizing objects in images without any supervision, called Unsupervised Object Discovery (UOD).

UOD can be useful for other vision tasks related to object localization. Pseudo-labels generated without supervision have been shown to provide reliable object priors for image instance retrieval in [34]. For object detection, they can be used either to initialize an object detector without additional annotation [24], or in a semi-supervised setting, when combined with few labeled data [32]. Providing robust pseudo-labels with limited noise is key to the success of these tasks. However, this remains a major challenge, especially in a completely unsupervised context, where no prior knowledge is provided about the semantics and localization of objects present in an image.

Many approaches solve this problem by leveraging inter-image similarities [28, 29] between pairs of object proposals in different images. Without carefully designed optimization mechanisms, these methods come with a computational cost and complexity that compromise their scalability. Moreover, these methods have shown to be dependent on supervised CNN features for the calculation of similarities.

Recently, vision transformers (ViT) have achieved excellent performances, outperforming CNN architectures in both supervised tasks [8, 7, 1] and self-supervised learning [2, 3]. Particularly, strong objects localization hints emerge from training ViT models using a self-distillation scheme, in DINO [2]. These self-supervised features were so far explored only for solving single object discovery task [24, 31]. TokenCut [31] demonstrated the effectiveness of spectral-clustering applied on self-supervised ViT features, for saliency detection, and significantly improved the state-of-the-art for single object discovery.

In this work, we propose a new approach to address multiple objects discovery without any supervision. We explore self-supervised Vision Transformer (SS-ViT) features to localize and segment multiple object instances in the image. Discovering multiple objects in each image is not straightforward as it requires a clear definition of what an object is. In fact, objects are either defined as the annotated regions in the supervised setting, or as the salient region in each image, in unsupervised single object discovery approaches. To address the localization of multiple objects in a fully unsupervised way, we propose to recognise object regions using a semantic information captured at the dataset level. In other terms, an object is defined as belonging to one of the discovered semantic categories, in the image collection. Concretely, the semantic categories present in the dataset are discovered in an unsupervised way. This information is encoded using classification models. Object discovery is then designed as the activation of object parts in each image using SS-ViT features, and the merging of these object parts using the self-supervised classifiers, to discover complete object instances. The effectiveness of the proposed framework is demonstrated through extensive experiments on the object detection benchmarks PASCAL VOC [9] and MSCOCO [17]. Since by design, our method provides pixel-wise mask proposals, we also show that the same framework solves the unsupervised image segmentation task.

Our contributions can be formulated as follows :

- We propose a fully unsupervised, bottom-up approach, for multiple objects discovery. We first discover object parts using intra-image similarities. Object parts are merged using a dataset-driven information, to form complete object instances. Both stages exploit self-supervised ViT features to produce instance masks. To the best of our knowledge, this is the first work that builds on SS-ViT features to solve the multi-object discovery task.

- We generalize the saliency-based approach in TokenCut [31] for the discovery of local fine semantic concepts (object parts) of multiple objects in an image.

- We propose a novel semantic object proposal method for the self-supervised learning of a region classifier. This visual model encodes the dataset-level semantic information.

- We improve the state-of-the-art in unsupervised multiple objects discovery, unsupervised class-agnostic object detection and unsupervised image segmentation, on challenging object detection benchmarks.

2. Related work

2.1. Unsupervised object discovery/co-localization

We can distinguish, from previous works, two distinct tasks: object discovery and object co-localization. The former consists in localizing objects in an image without any prior knowledge of the image content. This is the *real* object discovery task, which is much more challenging than object co-localization [29]. On the other hand, co-localization aims at localizing common objects between images, that share the same semantic content. Algorithms in this setting are fed with perfect image clusters derived from the ground-truth. It is therefore a weakly supervised version of object discovery.

DDT [32] addresses the co-localization task and is the first work that demonstrated the reusability of supervised CNN features for object co-localization. In DDT, objects are selected from regions of high correlation within a given cluster (semantic category).

Other methods address both tasks, and many of them leverage inter-image similarities between off-the-shelf region proposals. Cho *et al.* [4] formulated the problem as a structure and objects discovery, by iterating part-matching and object localization. Similarly, OSD [28] simultaneously localizes objects and discovers structures of the image collection. It formalizes the task as an optimization problem. Although OSD brought a large improvement, it has shown to be highly dependent on supervised proposals provided by [18]. The method also suffers from overlapping region proposals, which prevents it from proposing multiple objects per image. These limitations were addressed in rOSD [29] by providing unsupervised proposals corresponding to regions of high activations around local maxima, within CNN feature maps. rOSD also constrains the number of proposals per local maximum, and performs non-maximum suppression (NMS) [21] post-processing, to address the problem of overlapping proposals. Note that these methods, while unsupervised, are built on supervised CNN features, from the ImageNet [6] classification task. LOD [30] formalized the task as a ranking problem, and focused on ensuring the scalability of the proposed approach. It also demonstrated the utility of self-supervised CNN features for single and multiple objects discovery.

Other methods [24, 31] tackled the single object discovery problem, and showed the potential of self-supervised features, especially from ViT models, for saliency detection. LOST [24] proposed a seed expansion heuristic based on inter-patch correlation. TokenCut [31] investigated the use of spectral-clustering on self-supervised ViT features, which are projected into a new space that allows for a more accurate binary clustering [23].

In previous multi-object discovery methods, relying on inter-image similarities in the computation of an objectness score could favor the discovery of the most frequent objects. In our approach, even though we also use a dataset-driven information, we overcome this issue by training visual classification models to encode the information of semantic classes. This results in a better separation between object and non-object regions, and a better detection of under-represented classes.

2.2. Unsupervised image segmentation

Image segmentation is the task of grouping all pixels of an image into meaningful regions, where pixels sharing the same characteristics are assigned to the same region [16]. Due to the very high cost of such a dense annotation, weakly supervised and fully unsupervised methods were explored. In the weakly supervised setting, [33] takes as input the image-level labels of the class categories present in the image, and utilizes a vision-language embedding model to create a rough segmentation map for each class.

Other approaches do not use any kind of supervision. On one hand, we find classic methods such as k-means [11], that focuses on pixels clustering based on color and texture features, and assigns each pixel to the cluster with the nearest mean. Moreover, graph-based segmentation (GS) [10] generates image segments, while ensuring that these segments are not being too coarse or too detailed, based on regions comparison. More recently, methods based on unsupervised learning for image segmentation have been introduced. For example, IIC [13] learns to maximize the mutual information between an image and its augmentations on a patch-level cluster. Kim *et al.* [14, 16] trains a CNN by iterating features clustering and network parameters tuning. The method is based on three criteria to maximize the features similarity between spatially continuous pixels and pixels assigned to the same cluster, while imposing a large number of clusters. The authors proposed two solutions for label assignment without any supervision (i) by superpixel extraction using simple linear iterative clustering in [14] and (ii) by the use of a spacial continuity loss in [16] to address the limitation of fixed segment boundaries.

These methods discover multiple objects by proposing dense object candidates. Several other methods address the semantic segmentation task in an unsupervised way, without proposing dense object discovery. We do not consider these approaches in our study as we solve a different task.

2.3. Self-supervised vision transformers

The self-supervised setting aims at learning useful representations with no real label. It was first used for pretraining CNN models [12, 19, 27], and showed a strong generalization ability to downstream tasks. More recently, the self-attention based encoding of images using transformers for vision [7] was proved to be effective for a large spectrum of supervised vision tasks such as classification [7], semantic segmentation [25] and dense prediction tasks [20]. ViT has also become a reference architectural choice of neural nets for visual representation learning. In addition to the classic masked auto-encoding paradigm inspired from NLP, MoCo-v3 [3] among others, demonstrated a strong potential of training ViT with a contrastive approach.

Recently, a self-distillation scheme was used in DINO [2] to train ViT with no labels. The choices made during training result in effective semantic separation and local-global alignment of the learned features. In particular, the resulting attention maps strongly activate the object regions, which provide clues to the localization of objects in the image. SS-ViT features have been explored recently to perform saliency detection and single object discovery tasks [24, 31].

To the best of our knowledge, our method is the first to exploit self-supervised ViT features in a fully unsupervised multiple object discovery pipeline. The method outputs object instance masks resolving also the unsupervised image segmentation task. Such results can be used as pseudo-labels to initialize the training of a class-agnostic object detector, without any supervision.

3. Method

3.1. Overview

Recently, SS-ViT features showed to generalize well to saliency-based tasks [24, 31]. In this work, we aim to demonstrate the potential of using those features for multiple objects discovery, without any supervision. We adopt a bottom-up approach, illustrated in figure 1, starting with an intra-image analysis, for the discovery of object parts. At the dataset level, two CNN models are trained in a self-supervised manner, using carefully selected, and semantic object proposals. These models are used to merge and filter object parts, to form complete object instances.

The intra-image analysis can be seen as a generalization of TokenCut [31] to the multiple object discovery task. Similar to TokenCut, we perform spectral clustering using SS-ViT features, to decompose the image into eigen vectors with useful information. Different from TokenCut: (i) Since we focus on the localization of multiple objects, we look for more localization clues than just saliency. Thus we use multiple eigen vectors, as the feature space to apply local clustering, instead of only using the second eigen vector. (ii) The number of local clusters is no longer known as we try to solve the multi-object discovery task (2 clusters in saliency detection task). To manage this, we propose an algorithm for choosing an *optimal* number of clusters, without any knowledge about the number of objects, or semantic concepts, in each image. The algorithm is detailed in section 3.2 and aims at discovering multiple object parts, while limiting over-segmentation.

The goal of the dataset-level analysis is to build two classifiers that capture the dominant semantic classes in the image collection. One classifier is used for merging object parts, resulting from the local segmentation, and associates a confidence score to each discovered object. The second classifier separates foreground/background classes and is used to filter remaining noise after the merging phase. We perform image clustering to get pseudo-labels for training both models. Since images may contain several semantic concepts, instead of using the whole images, we apply clustering on selected object proposals from Selective Search [26]. For proposals selection, we build an objectness score detailed in section 3.3. The retained top proposals are grouped into clusters, which are used for training the classifiers.

Finally, the classifiers are used in cascade to merge and denoise the discovered object parts. Both stages use self-supervised ViT features trained using DINO [2]. We show in section 3.3 how these features are particularly relevant to our approach because of some properties like semantic separation, local-global alignment, and object regions activation.

3.2. Discovery of intra-image semantic concepts

In this step, we extend TokenCut [31] to discover multiple objects in each image, instead of solving saliency detection. TokenCut constructs a weighted graph where nodes are ViT embeddings of image patches and edges correspond to the cosine similarity between tokens. Single object discovery is then formalized as a normalized graph-cut (Ncut) problem, which is solved using spectral clustering: features are projected into a new space via eigen decomposition. In this space, the second smallest eigen vector provides a solution to the Ncut problem for binary clustering, as demonstrated by Shi and Malik [23]. Likewise, we create a similarity graph based on SS-ViT features. The image is then decomposed into eigen vectors with useful information. We consider N eigen vectors ($N \geq 2$) for local clustering, since we aim to capture multiple objects in the image. The choice of N is studied in section 4.6. The selected N eigen vectors represent the feature space where local clustering of image pixels is performed: each pixel is represented with a new feature vector f_i' of size N, where i varies between 1 and the total number of pixels per image n_p.

Since in a fully unsupervised setting the number of semantic concepts in each image is unknown, we determine an optimal number of clusters K using an iterative process as detailed in algorithm 1. We apply k-means clustering to the image pixels, in the new feature space of eigen vectors $\mathcal{F} = \{f_i'; 1 \leq i \leq n_p\}$. This partitions the image into K groups, which we denote C_K. We consider the background cluster as the one occupying the biggest area in the image. The background id is denoted as b_id. All the remaining clusters represent the *objects area*. K is incremented, starting from $K = 2$, until no significant *object area* is newly activated. The goal is to activate multiple object regions in the image, while limiting over-segmentation. Examples of the results of this step are provided in figure 3, first column. In particular, we see in the last row that, in some cases, the algorithm directly outputs an optimal segmentation of the image. This shows its effectiveness compared to a simple over-segmentation, where a predefined number of clusters is used, without adapting to the content of each image.

3.3. Dataset-level semantic object proposals

As stated above, we use Selective Search [26] (SeSe) region proposals as object priors to discover the semantic classes in the dataset, through proposals clustering. These proposals provide a fairly high recall. However, their ranking is rather naive: given an over-segmentation of the image, the regions merged first, based on color and texture similarities, are ranked first. This makes even the top proposals subject to a lot of noise. We thus propose a new ranking of SeSe proposals, to select the most relevant ones. To do this, we build an objectness score, based on assumptions about object-like regions.

Algorithm 1 Iterative clustering for intra-image discovery of semantic concepts

```
Initialize:
    $K \leftarrow 2$
    $C_K \leftarrow Kmeans(\mathcal{F}, K)$
    $b\_id \leftarrow \arg\max_k \{area(C_K(k)), 1 \leq k \leq K\}$
    $obj\_area \leftarrow \sum_{k=1, k \neq b\_id}^{K} area(C_K(k))$
    $add\_semantic\_concepts \leftarrow True$
while $add\_semantic\_concepts$ do
    $K \leftarrow K + 1$
    $C_K \leftarrow Kmeans(\mathcal{F}, K)$
    $new\_obj\_area \leftarrow \sum_{k=1, k \neq b\_id}^{K} area(C_K(k))$
    if $\frac{new\_obj\_area}{obj\_area} > thresh$ then
        $obj\_area \leftarrow new\_obj\_area$
    else $add\_semantic\_concepts \leftarrow False$
return $K$
```

Note that the objectness score is computed within each image, independently from all other images in the dataset. Concretely, we use two main measures in this computation: intersection over union (IoU) and cosine similarity between object proposals in the same image. Given M proposals (p_1, p_2, ..., p_M) in an image, we define u_{ij} as the overlap rate and s_{ij} as the similarity between p_i and p_j. For the latter, we use the cosine similarity between the CLS tokens from the last layer of a ViT trained using DINO. Let f_i and f_j be the feature vectors (CLS token) that result from passing p_i and p_j respectively to the SS-ViT. The cosine similarity s_{ij} is defined as:

$$s_{ij} = \frac{f_i . f_j}{||f_i|| ||f_j||} \tag{1}$$

The new objectness score for object proposals re-ranking is the weighted sum of three normalized terms:

$$\text{score}(\text{p}_i) = \frac{\alpha}{2}(Sim_L(p_i) + Dissim_G(p_i)) + (1 - \alpha)H(p_i) \tag{2}$$

Each term of this score is based on a different assumption:

Object-like regions have high local similarity. We define local similarity for a given proposal p_i as its average similarity to its neighbouring proposals, i.e. proposals having an IoU with p_i above a threshold t. We notice that these proposals correspond usually to parts of the same objects. We

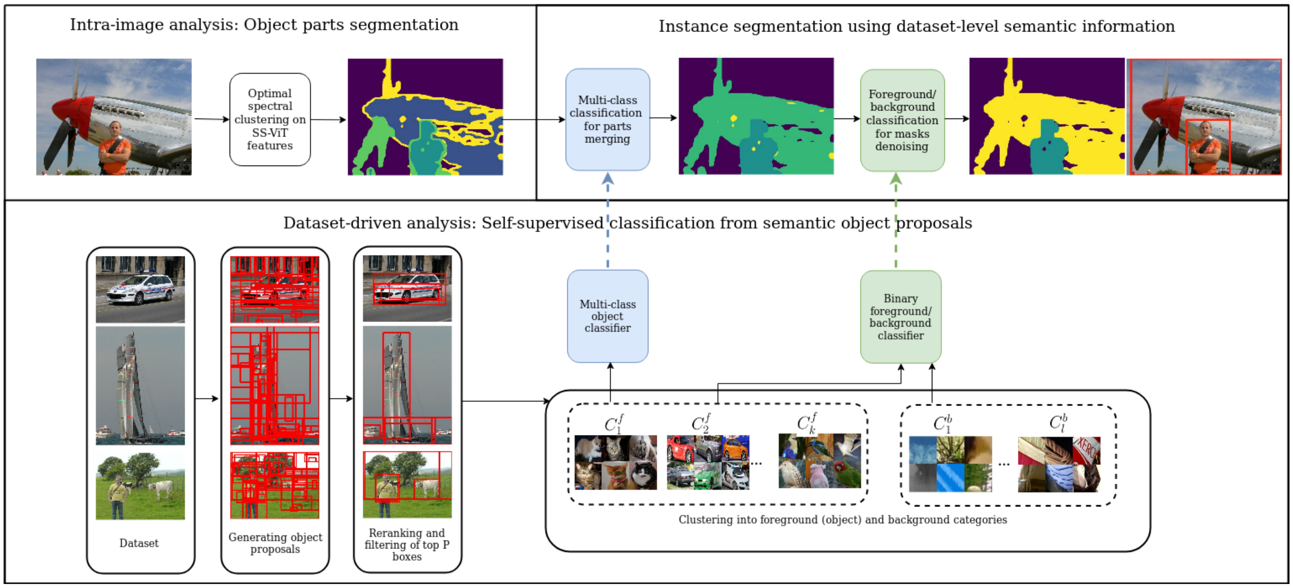
Intra-image analysis: Object parts segmentation
Optimal spectral clustering on SS-ViT features
Instance segmentation using dataset-level semantic information
Multi-class classification for parts merging
Foreground/ background classification for masks denoising
Dataset-driven analysis: Self-supervised classification from semantic object proposals
Multi-class object classifier
Binary foreground/ background classifier
Dataset
Generating object proposals
Reranking and filtering of top P boxes
C_1^f
C_2^f
C_k^f
C_1^b
C_l^b
Clustering into foreground (object) and background categories

Figure 1. **Pipeline of the method.** Top left: Intra-image analysis for the discovery of local semantic concepts. Bottom: Dataset-level analysis for the selection of semantic object proposals to train self-supervised classifiers. Top right: Using the data-driven classifiers on each image for parts merging and denoising.

also recall that we are using SS-ViT features learned using DINO, with a global local alignment objective. This means that an object is close to its parts in the DINO features space. From this we deduce that a high similarity between p_i and its neighbours increases its chance of containing an object. Thus, we make all the neighbours of p_i vote positively for it, in the following local similarity term:

$$Sim_L(p_i) = \sum_{j=1}^{M} s_{ij}, j \neq i, u_{ij} \geq t \tag{3}$$

Object-like regions have high global dissimilarity. We now consider the global similarity, i.e. the average similarity between p_i and all other proposals that *do not* overlap with p_i. Given the foreground/background imbalance in real-world images, most object proposals in an image occupy the background, and have similar visual content (e.g. 'sky'). Objects, on the contrary, occupy regions that are distinct in the image. If a proposal p_i contains an object, then it has low overall similarity, as all object proposals in the background vote negatively for it. Thus, p_i will be highly dissimilar, hence the following global dissimilarity term:

$$Dissim_G(p_i) = \sum_{j=1}^{M} 1 - s_{ij}, j \neq i, u_{ij} < t \tag{4}$$

Object-like regions have high entropy. The Shannon entropy of a discrete random variable is defined as:

$$H(p) = -\sum_{x} P_x log(P_x) \tag{5}$$

This measure is used to quantify the randomness of a variable [5]. In image processing, P_x refers to the distribution of gray levels x in image p (or colors intensities in RGB images). The previous formula associates higher entropy to images with more details and colors variation. Inversely, homogeneous regions are characterized by a low entropy. We thus associate low-entropy proposals to background, by adding an entropy term in the final objectness score.

It can be seen from figure 2 that the proposed ranking improves the detection rate for a fixed number of proposals, compared to two modes of SeSe. This is especially true when a small number of proposals are selected. We also compare qualitatively the retained top-1 proposals with the two rankings.

We recall that the aim of this new ranking is to reduce the amount of noise in the top proposals, which will be retained for clustering, as explained in section 3.4. We choose to use SeSe proposals for its popularity. However, the proposed ranking should be valid with other proposals, provided that they have a similar distribution, i.e. bounding boxes that occupy the whole image, and thus verify the background dominance condition, discussed above.

3.4. Dataset-driven self-supervised region classification

After re-ranking the SeSe object proposals, the goal is to transform the top P object priors in each image into pseudo-labels to train multi-class classifiers. These will be used to merge and refine the discovered local semantic concepts. We use a value of P large enough to make sure that

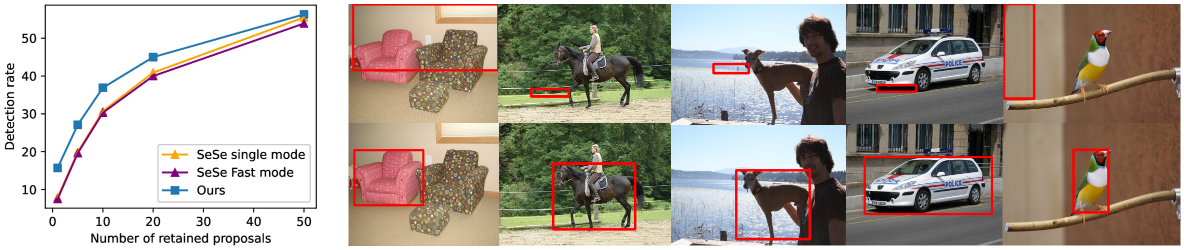

Figure 2. **Results of ranking object proposals.** Left: Comparison of the detection rate given the number of retained top proposals from two modes of SeSe, with our ranking. Right: Examples of the top1 proposal using SeSe score (top) and our score (bottom).

we do not only select objects but also object parts. This is important since the classifier must learn to assign the same semantic class to an object and its parts, for an accurate merging. We use k-means clustering [11] on the SS-ViT features of all the selected object proposals. The optimal number of clusters is chosen by finding the best silhouette score [22], which minimizes the mean intra-cluster distance and maximizes the mean nearest-cluster distance. With the semantic information contained in SS-ViT features, similar semantic concepts are grouped together. Moreover, each cluster contains proposals of both objects and their parts. This is especially due to the multi-crop augmentation technique used in DINO. The obtained clusters capture the global semantic information of the dataset. Note that the number of the clusters is not necessarily equal to the number of classes annotated in the dataset. However, we can still localize instances of undiscovered categories, such as ‘bottle’ and ‘plant’.

Since some of the selected proposals may still belong to background (Bg) regions, some of the discovered pseudo-classes are Bg clusters, that we aim to identify. According to our ranking score detailed in 3.3, the proposals having the lowest scores are the ones representing most probably Bg regions. Each of these proposals is passed to a SS-ViT to extract its features. The average vector of these features is considered as a *pattern* of Bg regions. The clusters whose center has a distance below a threshold t_{bg} with the Bg *pattern*, are considered Bg clusters.

After Foreground (Fg) and Bg groups identification, we associate to the clusters two types of labels (i) Each Fg cluster is assigned an *id* representing one discovered semantic class. (ii) All clusters have a binary label indicating whether it belongs to Fg or Bg. These image clusters are used to train two CNN-based classifiers, with the cluster *id* as a classification target. The first is a multi-class classifier trained using Fg clusters to assign objects and object parts to a specific class. The second classifier is trained using all the discovered clusters, and learns to distinguish between objects and Bg regions.

3.5. Instance segmentation using dataset-level information

In this final step, the obtained classifiers are used to merge and refine the object parts identified in the intra-image analysis. The multi-class classifier is first used on each segmented region: Image crops enclosing each object part segment are passed to the CNN-based classifier. Nearby regions assigned to the same category are merged to form complete object instances. The image crop around each merged region segment is then passed to the Fg/Bg classifier to eliminate segments classified as Bg. This binary classification is performed second to avoid incorrect classification of small object parts as Bg, if used before merging. The multi-class classifier also assigns to each object a confidence score, which is necessary for the evaluation metrics (AP@50, odAP). We provide in figure 3 illustrations for each step of the proposed framework.

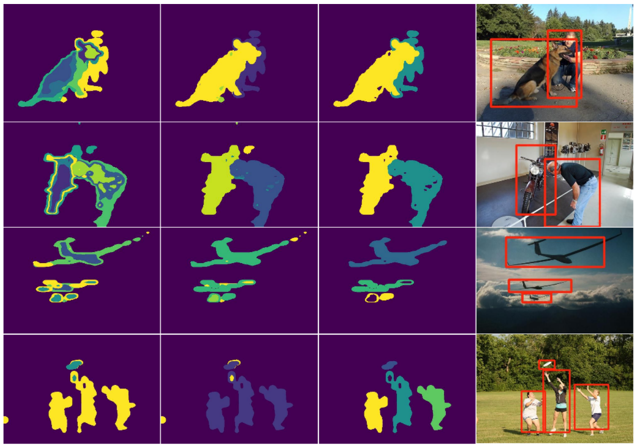

Figure 3. **Example of results.** By column: results of the discovery of local concepts, segmentation result after parts merging, final instance mask segmentation, final bounding boxes.

4. Experiments

4.1. Implementation details

Following previous works [29, 30, 24] we conduct our experiments on three detection and localization benchmarks: VOC2007 trainval, VOC2012 trainval [9] and COCO20k which is composed of 19817 images randomly chosen from COCO2014 trainval dataset [24]. We specify in the following the implementation details and hyper-parameters for each addressed task.

Unsupervised multiple objects discovery. In the intra-image analysis, local clustering is applied on SS-ViT features learned using the DINO training scheme. Based on the conclusions from previous works [31, 24], we use the variant ViT-S with a patch size of 16. Eigen decomposition is performed using the $keys$ features of the last layer. To find the optimal number of local clusters, we set as convergence criterion a fraction of newly activated area of 2%, ($thresh = 1.02$ in algorithm 1). The number of eigen vectors used for clustering is studied in section 4.6, and showed to be invariant to the dataset: 3 eigen vectors for PASCAL VOC and COCO20k. For object proposals re-ranking, the three terms are found to have an equivalent impact on the final re-ranking, with $\alpha = 0.7$, and $t = 0.1$. Proposals from Selective Search single mode are used in this work. For the dataset-level analysis, $P = 20$ top proposals in each image are selected to train the classifiers. A distance threshold $t_{bg} = 0.8$ from the Bg is used to separate Fg and Bg clusters. We use ResNet50 as the backbone of the two classifiers, initialized with DINO pre-training.

Unsupervised class-agnostic object detection. We follow the same configuration described in [24] for training a class-agnostic Faster-RCNN, with our pseudo-labels. We also use the same batch-size and the number of training iterations, for an objective comparison with previous works.

Unsupervised image segmentation. Following [16], this experiment is conducted on VOC2012 validation set, consisting of 1446 images. Masks resulting from multi-object discovery task are evaluated using mIOU, see section 4.2.

4.2. Metrics and evaluation settings

Different metrics are used to evaluate different tasks:

Unsupervised multiple objects discovery. Most of multiple objects discovery methods are based on ranking of object proposals. This makes them able to produce a large number of object candidates. The question then arises as to how many proposals to keep for computing recall, precision, or even the classical AP50 metric, since all of these would be affected by the number of retained top proposals. [30] addressed this issue and proposed an new version of AP, adapted to the object discovery task, called odAP. odAP is presented as the area under the precision-recall curve, where each precision-recall point is computed for a number of retained proposals, starting from 1, to the maximum number of objects in any image in the dataset. Even though by design, our approach outputs a reduced number of proposals, we use odAP to compare with previous works. We report the odAP50 where a detection is considered correct if its overlap rate with a ground truth bounding box is above 50%. And odAP@$[50:95]$, which is the average odAP for 10 values of IoU, varying from 50% to 95%.

Class-agnostic unsupervised object detection. A classical class agnostic Average Precision (AP@50) is calculated.

Unsupervised image segmentation. Following [16], we use the mean intersection over union (mIoU) to evaluate unsupervised image segmentation. mIOU is calculated as the average IOU between each ground truth mask (along with the background) and the detected mask that has the largest IOU with it, without considering any class label.

4.3. Unsupervised multiple objects discovery

We follow previous works and evaluate our method using odAP 4.2. Note that this metric is particularly adapted to the methods that propose a large number of object candidates, based on a ranking of object proposals. Since our approach is built on image segmentation, a limited number of boxes are proposed: 3 per image on average in PASCAL VOC dataset [9]. Hence, our approach is disadvantaged by this metric regarding the recall. Despite that, we show in table 1 the superiority of our method on both odAP@50 and the much more demanding odAP[50-95] metric.

The higher odAP[50-95] demonstrates the accuracy of our returned pseudo-boxes: Since these are generated from instance masks, they better enclose objects, and thus remain valid for a higher IoU threshold condition. Also, our method uses self-supervised features, which makes it fully unsupervised, unlike previous methods, which showed to be dependant on supervised features.

Method	Features	odAP@50			odAP@[50-95]		
		VOC07	VOC12	COCO20k	VOC07	VOC12	COCO20k
Kim et al. [15, 24]	Sup	9.5	11.8	3.93	2.5	3.1	0.96
DDT+ [32, 24]	Sup	8.7	11.1	2.41	3.0	4.1	0.73
rOSD [29, 24]	Sup	13.1	15.4	5.18	4.3	5.3	1.62
LOD [30, 24]	Sup	13.9	16.1	**6.63**	4.5	5.3	1.98
Ours	Self	**15.4**	**17.6**	5.44	**6.8**	**8.1**	**2.11**

Table 1. Multi-object discovery performance in odAP (Average Precision for object discovery)

4.4. Class-agnostic unsupervised object detection

State-of-the-art multiple objects discovery methods (MOD) usually rely on a ranking of object proposals based on inter-image similarities. These methods output a large number of object candidates and the question then arises as to how many bounding boxes to keep for the initialization of an object detector. Inversely, single object discovery methods (SOD) have a very limited recall. We argue that our method provides a better precision/recall trade-off than the previous methods in both settings. To prove this, we train a class-agnostic object detector using our generated pseudo-labels. Results are presented in table 2. We notice a clear improvement with our approach compared to MOD methods on all tested datasets. The gap however gets smaller when comparing with the SOD methods on PASCAL VOC [9] dataset. This can be explained by the presence of a dataset bias in PASCAL VOC: A large proportion of images in this dataset contain one object, which gives a clear advantage to SOD methods. On the more challenging COCO20k dataset, our method exceeds both categories (MOD and SOD). This demonstrates the superiority of our pseudo-labels, even for datasets with complex scenes.

Method	VOC07	VOC12	COCO20K
Selective Search [26]	3.6	4.8	1.8
EdgeBoxes [35]	2.9	4.2	1.6
rOSD + CAD [29]	24.2	29.0	8.4
LOD + CAD [30]	22.7	28.4	8.8
LOST + CAD [24]	**29.0**	33.5	9.9
TokenCut + CAD [31]	26.2	35.0	10.5
Ours + CAD	27.9	**36.2**	**13.8**

Table 2. Class-agnostic unsupervised object detection in AP50%

4.5. Unsupervised image segmentation

We further evaluate the performance of our method on VOC12 [9] validation set for unsupervised image segmentation task (see Table 3). Our method significantly outperforms previous state-of-the-art methods for discovering object masks in a fully unsupervised way. More qualitative results are provided in the supplementary material.

Method	VOC12
k-means clustering [11], k=2	0.3166
k-means clustering [11], k=17	0.2383
Graph-based segmentation (GS) [10], $\tau = 100$	0.2682
Graph-based segmentation (GS) [10], $\tau = 500$	0.3647
IIC [13], k=2	0.2729
IIC [13], k=20	0.2005
Kim *et al.* with superpixels [14]	0.3082
Kim *et al.* with continuity loss [16], $\nu = 5$	0.3520
Ours	**0.4247**

Table 3. Unsupervised image segmentation results in mIOU

4.6. Ablation study

In table 4, we provide an ablation study on different terms of the ranking score presented in 3.3. We evaluate the recall@50 (recall at IoU=50%) for different numbers of the retained top proposals. We compare the results of the overall ranking score, with the ranking obtained when one of the terms is removed from the final score. The best results are achieved by considering all 3 terms, which supports the assumptions made in section 3.3. We also compare our score with the original SeSe ranking of proposals from two settings. Using this new ranking, we ensure that the top proposals are more reliable for the classifiers training.

Method	Recall@50			
Number of boxes	1	4	10	20
SeSe Fast mode [26]	7.5	19.6	30.3	40.0
SeSe Single mode [26]	7.9	19.9	30.7	40.9
Ours: lSim + gDissim	13.9	23.5	34.4	44.1
Ours: lSim + E	12.8	24.9	35.9	44.6
Ours: Overall score	**15.7**	**27.1**	**36.9**	**45.0**

Table 4. Ablation study on the impact of the different terms composing the re-ranking score, evaluated on VOC07 testset

We also provide a study of the number of eigen vectors to be used in the intra-image analysis, in order to activate multiple objects, while limiting the amount of noise. In table 5, we evaluate the AP@50 of the generated pseudo-boxes, to choose the best precision/recal trade-off. We conduct the study on PASCAL VOC and COCO since they present different distributions of objects. Considering this study, the reported results are obtained with 3 eigen vectors in the intra-image analysis, for both datasets.

Number of eigen vectors	VOC07	COCO20k
2	22.1	5.9
3	**22.5**	**6.3**
4	21.3	6.0
5	20.3	5.8

Table 5. AP@50 as a function of the number of eigen vectors used for local analysis

5. Conclusion and future work

We presented a fully unsupervised approach for multiple objects discovery. The aim of this work was to address some of the limitations observed in existing methods. Namely, low recall in saliency detection-oriented methods, and the high amount of noise when several object candidates are proposed. We have shown that formulating the problem as an unsupervised segmentation is particularly suitable for reducing the noise in the generated pseudo-boxes. This provides a better precision-recall trade-off, which leads to a better initialization of an object detector. Still in this direction, we can further investigate the use of these pseudo-labels as an initial seed in a pseudo-labelling approach. Similarly, we can investigate the use of these object candidates with noise handling mechanisms.

6. Acknowledgements

This work benefited from the FactoryIA supercomputer financially supported by the Ile-deFrance Regional Council.

References

[1] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9650–9660, 2021.

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9640–9649, 2021.

[4] Minsu Cho, Suha Kwak, Cordelia Schmid, and Jean Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1201–1210, 2015.

[5] Thomas M Cover and Joy A Thomas. Information theory and statistics. *Elements of information theory*, 1(1):279–335, 1991.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

[8] Alaaeldin El-Nouby, Natalia Neverova, Ivan Laptev, and Hervé Jégou. Training vision transformers for image retrieval. *CoRR*, abs/2102.05644, 2021.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, June 2010.

[10] Pedro Felzenszwalb and Daniel Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59:167–181, 09 2004.

[11] J. A. Hartigan and M. A. Wong. A k-means clustering algorithm. *JSTOR: Applied Statistics*, 28(1):100–108, 1979.

[12] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.

[13] Xu Ji, Joao F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9865–9874, 2019.

[14] Asako Kanezaki. Unsupervised image segmentation by backpropagation. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1543–1547, 2018.

[15] Gunhee Kim and Antonio Torralba. Unsupervised detection of regions of interest using iterative link analysis. *Advances in neural information processing systems*, 22, 2009.

[16] Wonjik Kim, Asako Kanezaki, and Masayuki Tanaka. Unsupervised learning of image segmentation based on differentiable feature clustering. *IEEE Transactions on Image Processing*, 29:8055–8068, 2020.

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[18] Santiago Manen, Matthieu Guillaumin, and Luc Van Gool. Prime object proposals with randomized prim's algorithm. In *2013 IEEE International Conference on Computer Vision*, pages 2536–2543, 2013.

[19] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016.

[20] Jaonary Rabarisoa, Valentin Belissen, Florian Chabot, and Quoc-Cuong Pham. Self-Supervised Pre-training of Vision Transformers for Dense Prediction Tasks. *arXiv e-prints*, page arXiv:2205.15173, May 2022.

[21] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[22] Peter Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20(1):53–65, 1987.

[23] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997.

[24] Oriane Sim'eoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick P'erez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *BMVC*, 2021.

[25] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[26] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[27] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103, 2008.

[28] Huy V Vo, Francis Bach, Minsu Cho, Kai Han, Yann LeCun, Patrick Pérez, and Jean Ponce. Unsupervised image matching and object discovery as optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8287–8296, 2019.

[29] Huy V Vo, Patrick Pérez, and Jean Ponce. Toward unsupervised, multi-object discovery in large-scale image collections. In *European Conference on Computer Vision*, pages 779–795. Springer, 2020.

[30] Huy V. Vo, Elena Sizikova, Cordelia Schmid, Patrick Pérez, and Jean Ponce. Large-scale unsupervised object discovery. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2021.

[31] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022.

[32] Xiu-Shen Wei, Chen-Lin Zhang, Yao Li, Chen-Wei Xie, Jianxin Wu, Chunhua Shen, and Zhi-Hua Zhou. Deep descriptor transforming for image co-localization. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3048–3054, 2017.

[33] Nir Zabari and Yedid Hoshen. Semantic segmentation in-the-wild without seeing any segmentation examples. *CoRR*, abs/2112.03185, 2021.

[34] Z. Zhang, L. Wang, Y. Wang, L. Zhou, J. Zhang, and F. Chen. Dataset-driven unsupervised object discovery for region-based instance image retrieval. *IEEE Transactions on Pattern Analysis Machine Intelligence*, (01):1–1, jan 5555.

[35] C. Lawrence Zitnick and Piotr Dollár. Edge boxes: Locating object proposals from edges. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 391–405, Cham, 2014. Springer International Publishing.