

Toward Unsupervised, Multi-Object Discovery in Large-Scale Image Collections

Huy V. Vo^{1,2,3}, Patrick Pérez³, and Jean Ponce^{1,2}

¹ INRIA, Paris, France

² Département d’informatique de l’ENS, ENS, CNRS, PSL University, Paris, France

³ Valeo.ai

Abstract. This paper addresses the problem of discovering the objects present in a collection of images without any supervision. We build on the optimization approach of Vo *et al.* [34] with several key novelties: (1) We propose a novel saliency-based region proposal algorithm that achieves significantly higher overlap with ground-truth objects than other competitive methods. This procedure leverages off-the-shelf CNN features trained on classification tasks without any bounding box information, but is otherwise unsupervised. (2) We exploit the inherent hierarchical structure of proposals as an effective regularizer for the approach to object discovery of [34], boosting its performance to significantly improve over the state of the art on several standard benchmarks. (3) We adopt a two-stage strategy to select promising proposals using small random sets of images before using the whole image collection to discover the objects it depicts, allowing us to tackle, for the first time (to the best of our knowledge), the discovery of multiple objects in each one of the pictures making up datasets with up to 20,000 images, an over five-fold increase compared to existing methods, and a first step toward true large-scale unsupervised image interpretation.

Keywords: Object discovery, large-scale, optimization, region proposals, unsupervised learning.

1 Introduction

Object discovery, that is finding the location of salient objects in images without using any source of supervision, is a fundamental scientific problem in computer vision. It is also potentially an important practical one, since any effective solution would serve as a reliable free source of supervision for other tasks such as object categorization, object detection and the like. While many of these tasks can be tackled using massive amounts of annotated data, the manual annotation process is complex and expensive at large scales. Combining the discovery results with a limited amount of annotated data in a semi-supervised setting is a promising alternative to current data-hungry supervised approaches [35].

Vo *et al.* [34] posit that image collections possess an implicit graph structure. The pictures themselves are the nodes, and an edge links two images when they

share similar visual content. They propose the object and structure discovery framework (OSD) to localize objects and find the graph structure simultaneously by solving an optimization problem. Though demonstrating promising results, [34] has several shortcomings, e.g., the use of supervised region proposals, the limitation in addressing large image collections (See Section 2). Our work is built on OSD, aims to alleviate its limitations and improves it to effectively discover multiple objects in large image collections. Our contributions are:

- We propose a simple but effective method for generating region proposals directly from CNN features (themselves trained beforehand on some auxiliary task [29] *without* bounding boxes) in an unsupervised way (Section 3.1). Our algorithm gives on average half the number of region proposals per image compared to selective search [33], edgeboxes [40] or randomized Prim [23], yet significantly outperforms these off-the-shelf region proposals in object discovery (Table 3).
- Leveraging the intrinsic structure of region proposals generated by our method allows us to add an additional constraint into the OSD formulation that acts as a regularizer on its behavior (Section 3.2). This new formulation (rOSD) significantly outperforms the original algorithm and allows us to effectively perform multi-object discovery, a setting never studied before (to the best of our knowledge) in the literature.
- We propose a two-stage algorithm to make rOSD applicable to large image collections (Section 3.3). In the first stage, rOSD is used to choose a small set of good region proposals for each image. In the second stage, these proposals and the full image collection are fed to rOSD to find the objects and the image graph structure.
- We demonstrate that our approach yields significant improvements over the state of the art in object discovery (Tables 4 and 5). We also run our two-stage algorithm on a new and much larger dataset with 20,000 images and show that it significantly outperforms plain OSD in this setting (Table 7).

The only supervisory signal used in our setting are the image labels used to train CNN features in an auxiliary classification task (see [21,35] for similar approaches in the related colocalization domain). We use CNN features trained on ImageNet classification [29], *without* any bounding box information. Our region proposal and object discovery algorithms are otherwise fully unsupervised.

2 Related Work

Region proposals have been used in object detection/discovery to serve as object priors and reduce the search space. In most cases, they are found either by a bottom-up approach in which low-level cues are aggregated to rank a large set of boxes obtained with sliding window approaches [1,33,40] and return the top windows as proposals, or by training a model to classify them (as in randomized Prim [23], see also [26]), with *bounding box supervision*. Edgeboxes [40] and selective search [33] are popular off-the-shelf algorithms that are used to generate region proposals in object detection [13,14], weakly supervised object detection [7,31] or image colocalization [21]. Note, however, that the features

used to generate proposals in these algorithms and those representing them in the downstream tasks are generally different in nature: Typically, region proposals are generated from low-level features such as color and texture [33] or edge density [40], but CNN features are used to represent them in downstream tasks. However, the Region Proposal Network in Faster-RCNN [26] shows that proposals generated directly from the features used in the object detection task itself give a great boost in performance. In the object discovery setting, we therefore propose a novel approach for generating region proposals in an unsupervised way from CNN features trained on an auxiliary classification task without bounding box information. Features from CNNs trained on large-scale image classification have also been used to localize object in the weakly supervised setting. Zhou *et al.* [39] and Selvaraju *et al.* [28] fine-tune a pre-trained CNN to classify images and construct class activation maps, as weighted sums of convolutional feature maps or their gradient with respect to the classification loss, for localizing objects in these images. Tang *et al.* [32] generate region proposals to perform weakly supervised object detection on a set of labelled images by training a proposal network using the images' labels as supervision. Contrary to these works, we generate region proposals using only pre-trained CNN features without fine-tuning the feature extractor. Moreover, our region proposals come with a nice intrinsic structure which can be exploited to improve object discovery performance.

Early work on object discovery [12,15,20,27,30] focused on a restricted setting where images are from only a few distinctive object classes. Cho *et al.* [6] propose an approach for object and structure discovery by combining a part-based matching technique and an iterative *match-then-localize* algorithm, using off-the-shelf region proposals as primitives for matching. Vo *et al.* [34] reformulate [6] in an optimization framework and obtain significantly better performance. Image colocalization can be seen as a narrow setting of object discovery where all images in the collection contain objects from the same class. Observing that supervised object detectors often assign high scores to only a small number of region proposals, Li *et al.* [21] propose to mimic this behavior by training a classifier to minimize the entropy of the scores it gives to region proposals. Wei *et al.* [35] localize objects by clustering pixels with high activations in feature maps from CNNs pre-trained in ImageNet. All of the above works, however, focus on discovering only the main object in the images and target small-to-medium-scale datasets. Our approach is based on a modified version of the OSD formulation of Vo *et al.* [34] and pre-trained CNN features for object discovery, offers an effective and efficient solution to discover multiple objects in images in large-scale datasets. The recent work of Hsu *et al.* [18] for instance co-segmentation can also be adapted for localizing multiple objects in images. However, it requires input images to contain an object of a single dominant class while images may instead contain several objects from different categories in our setting.

Object and structure discovery (OSD) [34]. Since our work is built on [34], we give a short recap of this work in this section. Given a collection of n images, possibly containing objects from different categories, each equipped with p region proposals (which can be obtained using selective search [33], edgeboxes [40],

randomized Prim [23], etc.) and a set of potential neighbors, the unsupervised object and structure discovery problem (OSD) is formalized in [34] as follows: Let us define the variable e as an element of $\{0, 1\}^{n \times n}$ with a zero diagonal, such that $e_{ij} = 1$ when images i and j are linked by a (directional) edge, and $e_{ij} = 0$ otherwise, and the variable x as an element of $\{0, 1\}^{n \times p}$, with $x_i^k = 1$ when region proposal number k corresponds to visual content shared with neighbors of image i in the graph. This leads to the following optimization problem:

$$\max_{x, e} S(x, e) = \sum_{i=1}^n \sum_{j \in N(i)} e_{ij} x_i^T S_{ij} x_j, \text{ s.t. } \sum_{k=1}^p x_i^k \leq \nu \text{ and } \sum_{j \neq i} e_{ij} \leq \tau \quad \forall i, \quad (1)$$

where $N(i)$ is the set of potential neighbors of image i , S_{ij} is a $p \times p$ matrix whose entry S_{ij}^{kl} measures the similarity between regions k and l of images i and j , and ν and τ are predefined constants corresponding respectively to the maximum number of objects present in an image and to the maximum number of neighbors an image may have. This is however a hard combinatorial optimization problem. As shown in [34], an approximate solution can be found by (a) a dual gradient ascent algorithm for a continuous relaxation of Eq. (1) with exact updates obtained by maximizing a supermodular cubic pseudo-Boolean function [4, 24], (b) a simple greedy scheme, or (c) a combination thereof. Since solving the continuous relaxation of Eq. (1) is computationally expensive and may be less effective for large datasets [34], we only consider the version (b) of OSD in our analysis.

OSD has some limitations: (1) Although the algorithm itself is fully unsupervised, it gives by far its best results with region proposals from randomized Prim [23], a region proposal algorithm trained with bounding box supervision. (2) Vo *et al.* use whitened HOG (WHO) [16] to represent region proposals in their implementation although CNN features work better on the similar image colocalization problem [21, 35]. In our experiments, naively switching to CNN features does not give consistent improvement on common benchmarks. OSD with CNN features gives a CorLoc of 82.9, 71.5 and 42.8 compared to 87.1, 71.2 and 39.5 given by OSD with WHO, respectively on OD, VOC_6x2 and VOC_all data sets respectively. (3) Finally, due to its high memory cost, the algorithm cannot be applied to large datasets without compromising its final performance. In the next section, we describe our approach to addressing these limitations, as well as extending OSD to solve multi-object discovery.

3 Proposed Approach

3.1 Region Proposals from CNN Features

We address the limitation of using off-the-shelf region proposals of [34] with insights gained from the remarkably effective method for image colocalization proposed by Wei *et al.* [35]: CNN features pre-trained for an auxiliary task, such as ImageNet classification, give a strong, *category-independent* signal for unsupervised tasks. In retrospect, this insight is not particularly surprising, and it

is implicit in several successful approaches to image retrieval [38] or co-saliency detection [2,3,19,36]. Wei *et al.* [35] use it to great effect in the image colocalization task. Feeding an image to a pre-trained convolutional neural network yields a set of feature maps represented as a 3D tensor (e.g., a convolutional layer of VGG16 [29] or ResNet [17]). Wei *et al.* [35] observe that the “image” obtained by simply adding the feature maps gives hints to the locations of the objects it contains, and identify objects by clustering pixels with high activation. Similar but different from them, we observe that local maxima in the above “images” correspond to salient parts of objects in the original image and propose to exploit this observation for generating region proposals directly from CNN features. As we do not make use of any annotated bounding boxes, our region proposal itself is indeed unsupervised. Our method consists of the following steps. First, we feed the image to a pre-trained convolutional neural network to obtain a 3D tensor of size $(H \times W \times D)$, noted F . Adding elements of the tensor along its depth dimension yields a $(H \times W)$ 2D saliency map, noted as s_g (*global* saliency map), showing salient locations in the image with each location in s_g being represented by the corresponding D -dimensional feature vector from F . Next, we find robust local maxima in the previous saliency map using *persistence*, a measure used in topological data analysis [5,8,9,25,41] to find critical points of a function (see Section 4.2 for details). We find regions around each local maximum y using a *local* saliency map s_y of the same size as the global one. The value at any location in s_y is the dot product between normalized feature vectors at that location and the local maximum. By construction, the local saliency map highlights locations that are likely to belong to the same object as the corresponding local maximum. Finally, for each local saliency map, we discard all locations with scores below some threshold and the bounding box around the connected component containing the corresponding local maximum is returned as a region proposal. By varying the threshold, we can obtain tens of region proposals per local saliency map. An example illustrating the whole process is shown in Fig. 1.

3.2 Regularized OSD

Due to the greedy nature of OSD [34], its block-coordinate ascent iterations are prone to bad local maxima. Vo *et al.* [34] attempt to resolve this problem by using a larger value of ν in the optimization than the actual number of objects they intend to retrieve (which is one in their case) to diversify the set of retained regions in each iteration. The final region in each image is then chosen amongst its retained regions in a post processing step by ranking these using a new score solely based on their similarity to the retained regions in the image’s neighbors. Increasing ν in fact gives limited help in diversifying the set of retained regions. Since there is redundancy in object proposals with many highly overlapping regions, the ν retained regions are often nearly identical (see supplementary document for a visual illustration). This phenomenon also prevents OSD from retrieving multiple objects in images. One can use the ranking in OSD’s post processing step with non-maximum suppression to return more than one region

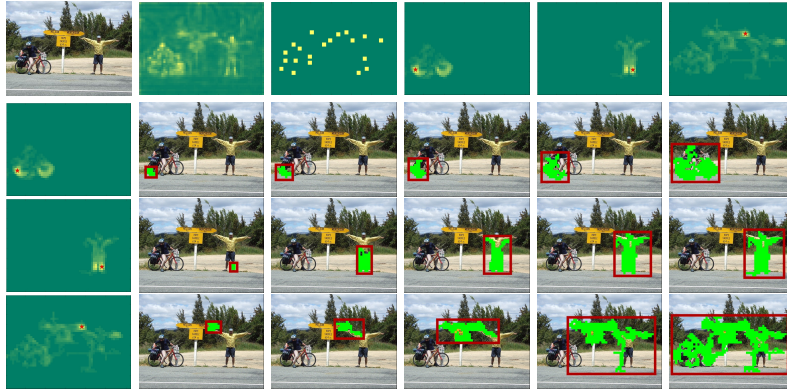


Fig. 1: Illustration of the unsupervised region proposal generation process. The top row shows the original image, the global saliency map s_g , local maxima of s_g and three local saliency maps s_y from three local maxima (marked by red stars). The next three rows illustrate the proposal generation process on the local saliency maps: From left to right, we show in green the connected component formed by pixels with saliency above decreasing thresholds and, in red, the corresponding region proposals.

from ν retained regions but since ν regions are often highly overlapping, this fails to localize multiple objects.

By construction, proposals produced by our approach also contain many highly overlapping regions, especially those generated from the same local maximum in the saliency map. However, they come with a nice intrinsic structure: Proposals in an image can be partitioned into groups labelled by the local maximum from which they are generated. Naturally, it makes sense to impose that at most one region in a group is retained in OSD since they are supposed to correspond to the same object. This additional constraint also conveniently helps to diversify the set of proposals returned by the block-coordinate ascent procedure by avoiding to retain highly overlapping regions. Concretely, let G_{ig} be the set of region proposals in image i generated from the g -th local maximum in its global saliency map s_g , with $1 \leq g \leq L_i$ where L_i is the number of local maxima in s_g , we propose to add the constraints $\sum_{k \in G_{ig}} x_i^k \leq 1 \ \forall i, g$ to Eq. (1). We coin the new formulation regularized OSD (rOSD). Similar to OSD, a solution to rOSD can be obtained by a greedy block-coordinate ascent algorithm whose iterations are illustrated in the supplementary document. We will demonstrate the effectiveness of rOSD compared to OSD and the state of the art in Section 4.

3.3 Large-Scale Object Discovery

The optimization algorithm of Vo *et al.* [34] requires loading all score matrices S_{ij} into the memory (they can also be computed on-the-fly but at an unacceptable computational cost). The corresponding memory cost is $M = (\sum_{i=1}^n |N(i)|) \times K$, decided by two main factors: The number of image pairs considered $\sum_{i=1}^n |N(i)|$ and the number of positive entries K in matrices S_{ij} . To reduce the cost on larger datasets, Vo *et al.* [34] pre-filter the neighborhood of each image ($|N(i)| \leq 100$

for classes with more than 1000 images) and limit K to 1000. This value of K is approximately the average number of proposals in each image, and it is intentionally chosen to make sure that S_{ij} is not too sparse in the sense that approximately every proposal in image i should have a positive match with some proposal in image j . Further reducing the number of positive entries in score matrices is likely to hurt the performance (Table 7) while a number of 100 potential neighbors is already small and can not be significantly lowered. Effectively scaling up OSD⁴ therefore requires lowering considerably the number of proposals it uses. To this end, we propose two different interpretations of the image graph and exploit both to scale up OSD.

Two different interpretations of the image graph. The *image graph* $G = (x, e)$ obtained by solving Eq. (1) can be interpreted as capturing the “true” structure of the input image collection. In this case, ν is typically small (say, 1 to 5) and the discovered “objects” correspond to maximal cliques of G , with instances given by active regions ($x_i^k = 1$) associated with nodes in the clique. But it can also be interpreted as a *proxy* for that structure. In this case, we typically take ν larger (say, 50). The active regions found for each node x_i of G are interpreted as the most promising regions in the corresponding image and the active edges e_{ij} link it to other images supporting that choice. We dub this variant *proxy* OSD.

For small image collections, it makes sense to run OSD only. For large ones, we propose instead to split the data into random groups with fixed size, run proxy OSD on each group to select the most promising region proposals in the corresponding images, then run OSD using these proposals. Using this two-stage algorithm, we reduce significantly the number of image pairs in each run of the first stage, thus permitting the use of denser score matrices in these runs. In the second stage, since only a very small number of region proposals are considered in each image, we need to keep only a few positive entries in each score matrix and are able to run OSD on the entire image collection. Our approach for large-scale object discovery is summarized the supplementary material.

4 Experiments

4.1 Datasets and Metrics

Similar to previous works on object discovery [6,34] and image colocalization [21,35], we evaluate object discovery performance with our proposals on four datasets: Object Discovery (OD), VOC_6x2, VOC_all and VOC12. OD is a small dataset with three classes *airplane*, *car* and *horse*, and 100 images per class, among which 18, 11 and 7 images are outliers (images not including an object of the corresponding class) respectively. VOC_all is a subset of the PASCAL VOC 2007 dataset [11] obtained by eliminating all images containing only *difficult* or *truncated* objects as well as *difficult* or *truncated* objects in retained images. It has 3550 images and 6661 objects. VOC_6x2 is a subset of VOC_all which

⁴ Since the analysis in this section applies to both OSD and rOSD, we refer to both as OSD for ease of notation.

contains images of 6 classes *aeroplane*, *bicycle*, *boat*, *bus*, *horse* and *motorbike* divided into 2 views *left* and *right*. In total, VOC_6x2 contains 463 images of 12 classes. VOC12 is a subset of the PASCAL VOC 2012 dataset [10] and obtained in the same way as VOC_all. It contains 7838 images and figures 13957 objects. For large-scale experiments, we randomly choose 20000 images from the training set of COCO [22] and eliminate those containing only *crowd* bounding boxes as well as bounding boxes marked as *crowd* in retained images. The resulting dataset, which we call COCO_20k, has 19817 images and 143951 objects.

As single-object discovery and colocalization performance measure, we use *correct localization* (CorLoc) defined as the percentage of images correctly localized. In our context, this means the intersection over union (*IoU*) between one of the ground-truth regions and one of the predicted regions in the image is greater than 0.5. Since CorLoc does not take into account multiple detections per image, for multi-object discovery, we use instead *detection rate* at the *IoU* threshold of 0.5 as measure of performance. Given some threshold ζ , detection rate at $IoU = \zeta$ is the percentage of ground-truth bounding boxes that have an *IoU* with one of the retained proposals greater than ζ . We run the experiments in both the colocalization setting, where the algorithm is run separately on each class of the dataset, and the average CorLoc/detection rate over all classes is computed as the overall performance measure on the dataset, and the true *discovery* setting where the whole dataset is considered as a single class.

4.2 Implementation Details

Features. We test our methods with the pre-trained CNN features from VGG16 and VGG19 [29]. For generating region proposals, we apply the algorithm described in Section 3.1 separately to the layers right before the last two max pooling layers of the networks (*relu4_3* and *relu5_3* in VGG16, *relu4_4* and *relu5_4* in VGG19), then fuse proposals generated from the two layers as our final set of proposals. Note that using CNN features at multiple layers is important as different layers capture different visual patterns in images [37]. One could also use more layers from VGG16 (e.g., layers *relu3_3*, *relu4_2* or *relu5_2*) but we only use two for the sake of efficiency. In experiments with OSD, we extract features for the region proposals by applying the RoI pooling operator introduced in Fast-RCNN [13] to layer *relu5_3* of VGG16.

Region Proposal Generation Process. For finding robust local maxima of the global saliency maps s_g , we rank its locations using persistence [5,8,9,25,41]. Concretely, we consider s_g as a 2D image and each location in it as a pixel. We associate with each pixel a cluster (the 4-neighborhood connected component of pixels that contains it), together with a “birth” (its own saliency) and “death” time (the highest value for which one of the pixels in its cluster also belongs to the cluster of a pixel with higher saliency, or, if no such location exists, the lowest saliency value in the map). The persistence of a pixel is defined as the difference between its birth and death times. A sorted list of pixels in decreasing persistence order is computed, and the local maxima are chosen as the top pixels in the list. For additional robustness, we also apply non maximum suppression on the list

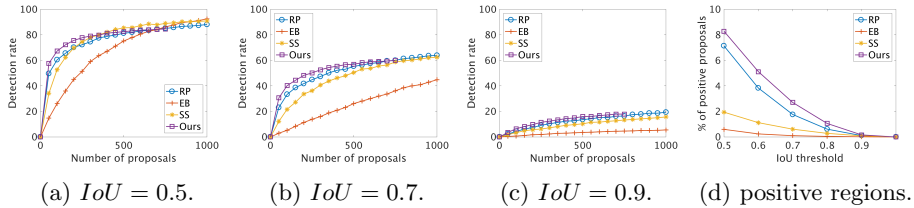


Fig. 2: Quality of proposals by different methods. (a-c): Detection rate by number of proposals at different IoU thresholds of randomized Prim (RP) [23], edgeboxes (EB) [40], selective search (SS) [33] and ours; (d): Percentage of positive proposals for the four methods.

over a 3×3 neighborhood. Since the saliency map created from CNN feature maps can be very noisy, we eliminate locations with score in s_g below $\alpha \max s_g$ before computing the persistence to obtain only good local maxima. We also eliminate locations with score smaller than the average score in s_g and whose score in s_g is smaller than β times the average score in s_g . We choose the value of the pair (α, β) in $\{0.3, 0.5\} \times \{0.5, 1\}$ by conducting small-scale object discovery on VOC_6x2. We find that $(\alpha, \beta) = (0.3, 0.5)$ yields the best performance and gives local saliency maps that are not fragmented while eliminating well irrelevant locations across settings and datasets. We take up to 20 local maxima (after non-maximum suppression) and use 50 linearly spaced thresholds between the lowest and the highest scores in each local saliency map to generate proposals.

Object Discovery Experiments. For single-object colocalization and discovery, following [34], we use $\nu = 5, \tau = 10$ and apply the OSD’s post processing to obtain the final localization result. For multi-object setting, we use $\nu = 50, \tau = 10$ and apply the post processing with non-maximum suppression at $IoU = 0.7$ to retain at most 5 regions in the final result. On large classes/datasets, we pre-filter the set of neighbors that are considered in the optimization for each image, using the cosine similarity between features from the fully connected layer *fc6* of the pre-trained network, following [2]. The number of potential neighbors of each image is fixed to 50 in all experiments where the pre-filtering is necessary.

4.3 Region Proposal Evaluation

Following other works on region proposals [23,33,40], we evaluate the quality of our proposals on PASCAL VOC 2007 using the *detection rate* at various IoU thresholds. But since we intend to later use our proposals for object discovery, unlike other works, we evaluate directly our proposals on VOC_all instead of the test set of VOC 2007 to reveal the link between the quality of proposals and the object discovery performance. Figure 2(a-c) shows the performance of different proposals on VOC_all. It can be seen that our method performs better than others at a very high overlap threshold (0.9) regardless of the number

of proposals allowed. At medium threshold (0.7), our proposals are on par (or better for fewer than 500 proposals) with those from selective search [33] and randomized Prim [23] and much better than those from edgeboxes [40]. At a small threshold (0.5), our method is still on par with randomized Prim and edgeboxes, but does not fare as well as selective search. It should be noted that randomized Prim is supervised whereas the others are unsupervised.

In OSD, localizing an object in an image means singling out a *positive* proposal, that is, a proposal having an *IoU* greater than some threshold with object bounding boxes. It is therefore easier to localize the object if the percentage of positive region proposals is larger. As shown by Fig. 2(d), our method performs very well according to this criterion: Over 8% of our proposals are positive at an *IoU* threshold of 0.5, and over 3% are still positive for an *IoU* of 0.7. Also, randomized Prim and our method are by far better than selective search and edgeboxes, which explains the superior object discovery performance of the former over the latter (*cf.* [34] and Table 3). Note that region proposals with a high percentage of positive ones could also be used in other tasks, i.e., weakly supervised object detection, but this is left for future work.

4.4 Object Discovery Performance

Single-Object Colocalization and Discovery. An important component of OSD is the similarity model used to compute score matrices S_{ij} , which, in [34], is the Probabilistic Hough Matching (PHM) algorithm [6]. Vo *et al.* [34] introduce two scores, *confidence* score and *standout* score, but use only the latter for it gives better performance. Since our new proposals come with different statistics, we test both scores in our experiments. Table 1 compares colocalization performance on OD, VOC_6x2 and VOC_all of OSD using the confidence and standout scores as well as our proposals. It can be seen that on VOC_6x2 and VOC_all, the confidence score does better than the standout score, while on OD, the latter does better. This is in fact not particularly surprising since images in OD generally contain bigger objects (relative to image size) than those in the other datasets. In fact, although the standout score is used on all datasets in [6] and [34], the authors adjust the parameter γ (see [6]) used in computing their standout score to favor larger regions when running their models on OD. In all of our experiments from now on, we use the standout score on OD and the confidence score on other datasets (VOC_6x2, VOC_all, VOC12 and COCO_20k).

Our proposal generation process introduces a few hyper-parameters. Apart from α and β , two other important hyper-parameters are the number of local maxima u and the number of thresholds v which together control the number of proposals p per image returned by the process. We study their influence on the colocalization performance by conducting experiments on VOC_6x2 and report the results in Table 2. It shows that the colocalization performance does not depend much on the values of these parameters. Using ($u = 50, v = 100$) actually gives the best performance but with twice as many proposals as ($u = 20, v = 50$). For efficiency, we use $u = 20$ and $v = 50$ in all of our experiments.

Table 1: Colocalization performance with our proposals in different configurations of OSD

Config.	Confidence	Standout
OD	83.7 ± 0.4	89.0 ± 0.6
VOC_6x2	73.6 ± 0.6	64.1 ± 0.3
VOC_all	44.7 ± 0.3	41.4 ± 0.1

Table 2: Colocalization performance for different values of hyper-parameters

(u, v)	(20,50)	(20,100)	(50,50)	(50,100)
CorLoc	73.6 ± 0.8	73.4 ± 0.7	73.3 ± 1.1	74.2 ± 0.8
p	760	882	1294	1507

Table 3: Single-object colocalization and discovery performance of OSD with different types of proposals. We use VGG16 features to represent regions in these experiments

Region proposals	Colocalization			Discovery		
	OD	VOC_6x2	VOC_all	OD	VOC_6x2	VOC_all
Edgeboxes [40]	81.6 ± 0.3	54.2 ± 0.3	29.7 ± 0.1	81.4 ± 0.3	55.2 ± 0.3	32.6 ± 0.1
Selective search [33]	82.2 ± 0.2	54.5 ± 0.3	30.9 ± 0.1	81.3 ± 0.3	57.8 ± 0.2	33.0 ± 0.1
Randomized Prim [23]	82.9 ± 0.3	71.5 ± 0.3	42.8 ± 0.1	82.5 ± 0.1	70.6 ± 0.4	44.5 ± 0.1
Ours (OSD)	89.0 ± 0.6	73.6 ± 0.6	44.7 ± 0.3	87.8 ± 0.4	69.2 ± 0.5	48.7 ± 0.3
Ours (rOSD)	89.0 ± 0.5	<u>73.3 ± 0.5</u>	45.8 ± 0.3	<u>87.6 ± 0.3</u>	71.1 ± 0.8	49.2 ± 0.2

We report in Table 3 the performance of OSD and rOSD on OD, VOC_6x2 and VOC_all with different types of proposals. It can be seen that our proposals give the best results on all datasets among all types of proposals with significant margins: 6.1%, 2.1% and 3.0% in colocalization and 5.3%, 0.5% and 4.7% in discovery, respectively. It is also noticeable that our proposals not only fare much better than the unsupervised ones (selective search and edgeboxes) but outperform those generated by randomized Prim, an algorithm trained with bounding box annotation.

We compare OSD and rOSD using our region proposals to the state of the art in Table 4 (colocalization) and Table 5 (discovery). In their experiments, Wei *et al.* [35] only use features from VGG19. We have conducted experiments with features from both VGG16 and VGG19 but only present experiment results with VGG19 features in comparisons with [35] due to the space limit. A more comprehensive comparison with features from VGG16 is included in the supplementary material. It can be seen that our use of CNN features (for both creating proposals and representing them in OSD) consistently improves the performance compared to the original OSD [34]. It is also noticeable that rOSD performs significantly better than OSD on the two large datasets (VOC_all and VOC12) while on the two smaller ones (OD and VOC_6x2), their performances are comparable. It is due to the fact that images in OD and VOC_6x2 mostly contain only one well-positioned object thus bad local maxima are not a big problem in the optimization while images in VOC_all and VOC12 contain much more complex scenes and the optimization works better with more regularization. In overall, we obtain the best results on the two smaller datasets, fare better than [21] but are behind [35] on VOC_all and VOC12 in the colocalization setting. It should be noticed that while methods for image colocalization [21,35] suppose that images in the collection come from the same category and explicitly exploit this

Table 4: Single-object colocalization performance of our approach compared to the state of the art. Note that Wei *et al.* [35] outperform our method on VOC_all and VOC12 in this case, but the situation is clearly reversed in the much more difficult discovery setting, as demonstrated in Table 5

Method	Features	OD	VOC_6x2	VOC_all	VOC12
Cho <i>et al.</i> [6]	WHO	84.2	67.6	37.6	-
Vo <i>et al.</i> [34]	WHO	87.1 \pm 0.5	71.2 \pm 0.6	39.5 \pm 0.1	-
Li <i>et al.</i> [21]	VGG19	-	-	41.9	45.6
Wei <i>et al.</i> [35]	VGG19	87.9	67.7	48.7	51.1
Ours (OSD)	VGG19	90.3 \pm 0.3	75.3 \pm 0.7	45.6 \pm 0.3	47.8 \pm 0.2
Ours (rOSD)	VGG19	90.2 \pm 0.3	76.1 \pm 0.7	46.7 \pm 0.2	49.2 \pm 0.1

Table 5: Single-object discovery performance on the datasets with our proposals compared to the state of the art

Method	Features	OD	VOC_6x2	VOC_all	VOC12
Cho <i>et al.</i> [6]	WHO	82.2	55.9	37.6	-
Vo <i>et al.</i> [34]	WHO	82.3 \pm 0.3	62.5 \pm 0.6	40.7 \pm 0.2	-
Wei <i>et al.</i> [35]	VGG19	75.0	54.0	43.4	46.3
Ours (OSD)	VGG19	89.1 \pm 0.4	71.9 \pm 0.7	47.9 \pm 0.3	49.2 \pm 0.2
Ours (rOSD)	VGG19	89.2 \pm 0.4	72.5 \pm 0.5	49.3 \pm 0.2	51.2 \pm 0.2

assumption, rOSD is intended to deal with the much more difficult and general object discovery task. Indeed, in discovery setting, rOSD outperforms [35] by a large margin, 5.9% and 4.9% respectively on VOC_all and VOC12.

Multi-Object Colocalization and Discovery. We demonstrate the effectiveness of rOSD in multi-object colocalization and discovery on VOC_all and VOC12 datasets, which contain images with multiple objects. We compare the performance of OSD and rOSD to Wei *et al.* [35] in Table 6. Although [35] tackles only the single-object colocalization problem, we modify their method to have a reasonable baseline for the multi-object colocalization and discovery problem. Concretely, we take the bounding boxes around the 5 largest connected components of positive locations in the image’s *indicator matrix* [35] as the localization results. It can be seen that our method obtains the best performance with significant margins to the closest competitor across all datasets and settings. It is also noticeable that rOSD, again, significantly outperforms OSD in this task. An illustration of multi-object discovery is shown in Fig. 3. For a fair comparison, we use high values of ν (50) and *IoU* (0.7) in the multi-object experiments to make sure that both OSD and rOSD return approximately 5 regions per image. Images may of course contain fewer than 5 objects. In such cases, OSD and rOSD usually return overlapping boxes around the actual objects. We can often eliminate these overlapping boxes and obtain better qualitative results by using smaller ν and *IoU* threshold values. It can be seen in Fig. 3 that with $\nu = 25$ and *IoU* = 0.3, rOSD is able to return bounding boxes around objects without many overlapping regions. Note however that the quantitative results may worsen due to the reduced number of regions returned and the fact that many images contain objects that highly overlap, e.g., the last two columns of Fig. 3.

Table 6: Multi-object colocalization and discovery performance of rOSD compared to competitors on VOC_all and VOC12 datasets

Method	Features	Colocalization		Discovery	
		VOC_all	VOC12	VOC_all	VOC12
Vo <i>et al.</i> [34]	WHO	40.7 \pm 0.1	-	30.7 \pm 0.1	-
Wei <i>et al.</i> [35]	VGG19	43.3	45.5	28.1	30.3
Ours (OSD)	VGG19	46.8 \pm 0.1	47.9 \pm 0.0	34.8 \pm 0.0	36.8 \pm 0.0
Ours (rOSD)	VGG19	49.4 \pm 0.1	51.5 \pm 0.1	37.6 \pm 0.1	40.4 \pm 0.1

In such cases, a small *IoU* threshold prevents discovering all of these objects. See supplementary document for more visualizations and details.

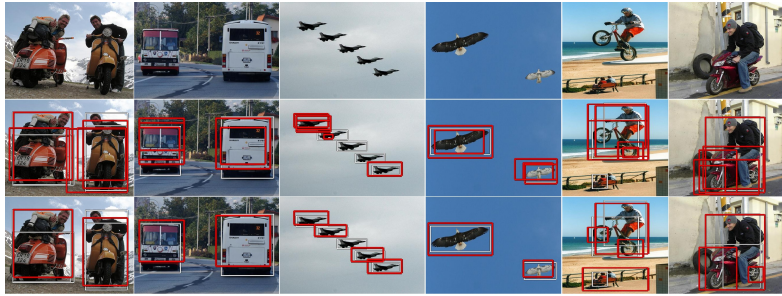


Fig. 3: Qualitative multi-object discovery results obtained with rOSD. White boxes are ground truth objects and red ones are our predictions. Original images are in the first row. Results with $\nu = 50$ and $IoU = 0.7$ are in the second row. Results with $\nu = 25$ and $IoU = 0.3$ are in the third row.

Large-Scale Object Discovery. We apply our large-scale algorithm in the discovery setting on VOC_all, VOC12 and COCO_20k which are randomly partitioned respectively into 5, 10 and 20 parts of roughly equal sizes. In the first stage of all experiments, we prefilter the initial neighborhood of images and keep only 50 potential neighbors. We choose $\nu = 50$ and keep K_1 (which are 250, 500 and 1000 respectively on VOC_all, VOC12 and COCO_20k) positive entries in each score matrix. In the second stage, we run rOSD (OSD) on the entire datasets with $\nu = 5$, limit the number of potential neighbors to 50 and use score matrices with only 50 positive entries. We choose K_1 such that each run in the first stage and the OSD run in the second stage have the same memory cost, hence the values of K chosen above. As baselines, we have applied rOSD (OSD) directly to the datasets, keeping 50 positive entries (baseline 1) and 1000 positive entries (baseline 2) in score matrices. Table 7 shows the object discovery performance on VOC_all, VOC12 and COCO_20k for our large-scale algorithm compared to the baselines. It can be seen that our large-scale two-stage rOSD algorithm yields significant performance gains over the baseline 1, obtains an improvement of 6.6%, 9.3% and 4.0% in single-object discovery and

Table 7: Performance of our large-scale algorithm compared to the baselines. Our method and baseline 1 have the same memory cost, which is much smaller than the cost of baseline 2. Also, due to memory limits, we cannot run baseline 2 on COCO_20k

Method	Single-object			Multi-object		
	VOC_all	VOC12	COCO_20k	VOC_all	VOC12	COCO_20k
Baseline 1 (OSD)	41.1 \pm 0.3	40.5 \pm 0.2	43.6 \pm 0.2	31.4 \pm 0.1	32.4 \pm 0.0	10.5 \pm 0.0
Baseline 1 (rOSD)	42.8 \pm 0.3	42.6 \pm 0.2	44.5 \pm 0.1	35.4 \pm 0.2	37.2 \pm 0.1	11.6 \pm 0.0
Baseline 2 (OSD)	47.9 \pm 0.3	49.2 \pm 0.2	-	34.8 \pm 0.0	36.8 \pm 0.0	-
Baseline 2 (rOSD)	49.3 \pm 0.2	51.2 \pm 0.2	-	37.6 \pm 0.1	40.4 \pm 0.1	-
Large-scale OSD	45.5 \pm 0.3	46.3 \pm 0.2	46.9 \pm 0.1	34.6 \pm 0.0	36.9 \pm 0.0	11.1 \pm 0.0
Large-scale rOSD	49.4 \pm 0.1	51.9 \pm 0.1	48.5 \pm 0.1	38.3 \pm 0.0	41.2 \pm 0.1	12.0 \pm 0.0

2.9%, 4.0% and 0.4% in multi-object discovery, respectively on VOC_all, VOC12 and COCO_20k. Interestingly, large-scale rOSD also outperforms the baseline 2, which has a much higher memory cost, on VOC_all and VOC12.

Execution time. Similar to [34], our method requires computing the similarity scores for a large number of image pairs which makes it computationally costly. It takes in total 478 parallelizable CPU hours, 300 unparallelizable CPU seconds and 1 GPU hour to run single-object discovery on VOC_all with 3550 images. This is more costly compared to only 812 GPU seconds needed by DDT+ [35] but is less costly than [34] using CNN features. The latter requires 546 parallelizable CPU hours, 250 unparallelizable CPU seconds and 4 GPU hours. Note that the unparallelizable computational cost, which comes from the main OSD algorithm, grows very fast (at least linearly in theory, it takes 2.3 hours on COCO_20k in practice) with the data set’s size and is the time bottleneck in large scale.

5 Conclusion

We have presented an unsupervised algorithm for generating region proposals from CNN features trained on an auxiliary and unrelated task. Our proposals come with an intrinsic structure which can be leveraged as an additional regularization in the OSD framework of Vo *et al.* [34]. The combination of our proposals and regularized OSD gives comparable results to the current state of the art in image colocalization, sets a new state-of-the-art single-object discovery and has proven effective in the multi-object discovery. We have also successfully extended OSD to the large-scale case and show that our method yields significantly better performance than plain OSD. Future work will be dedicated to investigating other applications of our region proposals.

Acknowledgments. This work was supported in part by the Inria/NYU collaboration, the Louis Vuitton/ENS chair on artificial intelligence and the French government under management of Agence Nationale de la Recherche as part of the “Investissements d’avenir” program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute). Huy V. Vo was supported in part by a Valeo/Prairie CIFRE PhD Fellowship.

References

1. Alexe, B., Deselaers, T., Ferrari, V.: Measuring the objectness of image windows. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **34**, 2189–2202 (2012) 2
2. Babenko, A., Slesarev, A., Chigorin, A., Lempitsky, V.: Neural codes for image retrieval. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2014) 5, 9
3. Babenko, A., Lempitsky, V.: Aggregating deep convolutional features for image retrieval. In: *Proceedings of the International Conference on Computer Vision (ICCV)* (2015) 5
4. Bach, F.: Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning* **6**(2-3), 145–373 (2013) 4
5. Chazal, F., Guibas, L.J., Oudot, S.Y., Skraba, P.: Persistence-based clustering in riemannian manifolds. *Journal of the ACM* **60**(6), 41:1–41:38 (2013) 5, 8
6. Cho, M., Kwak, S., Schmid, C., Ponce, J.: Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2015) 3, 7, 10, 12
7. Cinbis, R., Verbeek, J., Schmid, C.: Weakly supervised object localization with multi-fold multiple instance learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **39**, 189–203 (2017) 2
8. Edelsbrunner, H., Harer, J.: *Computational Topology: An Introduction*. AMS Press (2009) 5, 8
9. Edelsbrunner, H., Letscher, D., Zomorodian, A.: Topological persistence and simplification. *Discrete and Computational Geometry* (2002) 5, 8
10. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html> 8
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results 7
12. Faktor, A., Irani, M.: Clustering by composition—unsupervised discovery of image categories. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2012) 3
13. Girshick, R.: Fast R-CNN. In: *Proceedings of the International Conference on Computer Vision (ICCV)* (2015) 2, 8
14. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2014) 2
15. Grauman, K., Darrell, T.: Unsupervised learning of categories from sets of partially matching image features. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2006) 3
16. Hariharan, B., Malik, J., Ramanan, D.: Discriminative decorrelation for clustering and classification. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2012) 4
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) 5

18. Hsu, K.J., Lin, Y.Y., Chuang, Y.Y.: Deepco³: Deep instance co-segmentation by co-peak search and co-saliency detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 3
19. Hsu, K.J., Tsai, C.C., Lin, Y.Y., Qian, X., Chuang, Y.Y.: Unsupervised cnn-based co-saliency detection with graphical optimization. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 5
20. Kim, G., Torralba, A.: Unsupervised detection of regions of interest using iterative link analysis. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) (2009) 3
21. Li, Y., Liu, L., Shen, C., Hengel, A.: Image co-localization by mimicking a good detector’s confidence score distribution. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016) 2, 3, 4, 7, 11, 12
22. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common Objects in Context. In: Proceedings of the European Conference on Computer Vision (ECCV) (2014) 8
23. Manen, S., Guillaumin, M., Van Gool, L.: Prime Object Proposals with Randomized Prim’s Algorithm. In: Proceedings of the International Conference on Computer Vision (ICCV) (2013) 2, 4, 9, 10, 11
24. Nedić, A., Ozdaglar, A.: Approximate primal solutions and rate analysis for dual subgradient methods. *SIAM Journal on Optimization* **19**(4), 1757–1780 (2009) 4
25. Oudot, S.: Persistence Theory: From Quiver Representations to Data Analysis. *AMS Surveys and Monographs* (2015) 5, 8
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: Proceedings of the Conference on Neural Information Processing Systems (NeurIPS) (2015) 2, 3
27. Russell, B., Freeman, W., Efros, A., Sivic, J., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2006) 3
28. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the International Conference on Computer Vision (ICCV) (2017) 3
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015) 2, 5, 8
30. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: Proceedings of the International Conference on Computer Vision (ICCV) (2005) 3
31. Tang, P., Wang, X., Bai, S., Shen, W., Bai, X., Liu, W., Yuille, A.: Pcl: Proposal cluster learning for weakly supervised object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* **42**(1), 176–191 (2020) 2
32. Tang, P., Wang, X., Wang, A., Yan, Y., Liu, W., Huang, J., Yuille, A.: Weakly supervised region proposal network and object detection. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 3
33. Uijlings, J.R.R., van de Sande, K.E.A., Gevers, T., Smeulders, A.W.M.: Selective search for object recognition. *International Journal of Computer Vision* **104**(2), 154–171 (2013) 2, 3, 9, 10, 11
34. Vo, H.V., Han, K., Cho, M., Pérez, P., Bach, F., LeCun, Y., Ponce, J.: Unsupervised image matching and object discovery as optimization. In: Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 2, 3, 4, 5, 6, 7, 9, 10, 11, 12, 13, 14

35. Wei, X.S., Zhang, C.L., Wu, J., Shen, C., Zhou, Z.H.: Unsupervised object discovery and co-localization by deep descriptor transforming. *Pattern Recognition (PR)* **88** (2019) 1, 2, 3, 4, 5, 7, 11, 12, 13, 14
36. Wei, X.S., Luo, J.H., Wu, J., Zhou, Z.H.: Selective convolutional descriptor aggregation for fine-grained image retrieval. *IEEE Transactions on Image Processing* **26**(6), 2868–2881 (2017) 5
37. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2014) 8
38. Zhang, D., Meng, D., Li, C., Jiang, L., Zhao, Q., Han, J.: A self-paced multiple-instance learning framework for co-saliency detection. In: *Proceedings of the International Conference on Computer Vision (ICCV)* (2015) 5
39. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)* (2016) 3
40. Zitnick, C.L., Dollár, P.: Edge boxes: Locating object proposals from edges. In: *Proceedings of the European Conference on Computer Vision (ECCV)* (2014) 2, 3, 9, 10, 11
41. Zomorodian, A., Carlsson, G.: Computing persistent homology. *Discrete and Computational Geometry* (2005) 5, 8