Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals

Minsu Cho, Suha Kwak, Cordelia Schmid, Jean Ponce

Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals

Minsu Cho[1,*] Suha Kwak[1,*] Cordelia Schmid[1,†] Jean Ponce[2,*]

[1]Inria [2]École Normale Supérieure / PSL Research University

Abstract

This paper addresses unsupervised discovery and localization of dominant objects from a noisy image collection with multiple object classes. The setting of this problem is fully unsupervised, without even image-level annotations or any assumption of a single dominant class. This is far more general than typical colocalization, cosegmentation, or weakly-supervised localization tasks. We tackle the discovery and localization problem using a part-based region matching approach: We use off-the-shelf region proposals to form a set of candidate bounding boxes for objects and object parts. These regions are efficiently matched across images using a probabilistic Hough transform that evaluates the confidence for each candidate correspondence considering both appearance and spatial consistency. Dominant objects are discovered and localized by comparing the scores of candidate regions and selecting those that stand out over other regions containing them. Extensive experimental evaluations on standard benchmarks demonstrate that the proposed approach significantly outperforms the current state of the art in colocalization, and achieves robust object discovery in challenging mixed-class datasets.

1. Introduction

Object localization and detection is highly challenging because of intra-class variations, background clutter, and occlusions present in real-world images. While significant progress has been made in this area over the last decade, as shown by recent benchmark results [11, 16], most state-of-the-art methods still rely on strong supervision in the form of manually-annotated bounding boxes on target instances. Since those detailed annotations are expensive to acquire and also prone to unwanted biases and errors, recent work has explored the problem of weakly-supervised object discovery where instances of an object class are found in a collection of images without any box-level annotations. Typically, weakly-supervised localization [9, 35, 36, 43, 45, 56] requires positive and negative image-level labels for a target object class. On the other hand, cosegmentation [25, 29, 40] and colocalization [12, 27, 51] assume less supervision and only require the image collection to contain a single dominant object class, allowing noisy images to some degree.

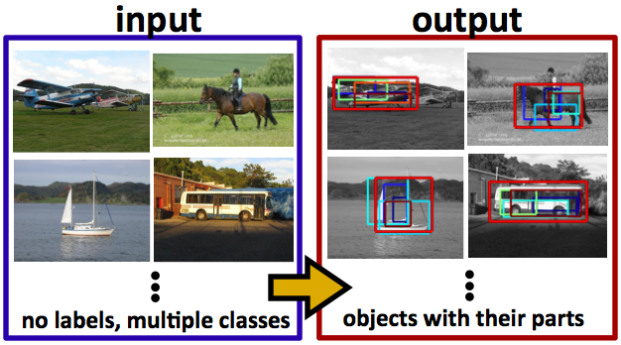


Figure 1. Unsupervised object discovery in the wild. We tackle object localization in an unsupervised scenario without any type of annotations, where a given image collection may contain multiple dominant object classes and even outlier images. The proposed method discovers object instances (red bounding boxes) with their distinctive parts (smaller boxes). (Best viewed in color.)

This paper addresses unsupervised object localization in a far more general scenario where a given image collection contain *multiple dominant object classes* and even *noisy images* without any target objects. As illustrated in Fig. 1, the setting of this problem is fully unsupervised, without any image-level annotations, an assumption of a single dominant class, or even a known number of object classes. In spite of this generality, the proposed method markedly outperforms the state of the arts in colocalization [27, 51] on standard benchmarks [16, 40], and closely competes with current weakly-supervised localization [9, 43, 56].

We advocate a part-based matching approach to unsupervised object discovery using bottom-up region proposals. Multi-scale region proposals have been widely used before to restrict the search space for object bounding boxes in ob-

*WILLOW project-team, Département d'Informatique de l'Ecole Normale Supérieure, ENS/Inria/CNRS UMR 8548.

†LEAR project-team, Inria Grenoble Rhône-Alpes, LJK, CNRS, Univ. Grenoble Alpes, France.

ject recognition [9, 20, 53] and localization [9, 27, 51, 54]. We go further and propose here to use these regions to form a set of candidate regions not only for objects, but also for object parts. We use a probabilistic Hough transorm [2] to match those candidate regions across images, and assign them confidence scores reflecting both appearance and spatial consistency. This can be seen as an unsupervised and efficient variant of both deformable part models [18, 19] and graph matching methods [5, 14]. Objects are discovered and localized by selecting the most salient regions that contain corresponding parts. To this end, we introduce a score that measures how much a region stands out over other regions containing it. The proposed algorithm alternates between part-based region matching and foreground localization, improving both over iterations.

The main contributions of this paper can be summarized as follows: (1) A part-based region matching approach to unsupervised object discovery is introduced. (2) An efficient and robust matching algorithm based on a probabilistic Hough transform is proposed. (3) A standout score for robust foreground localization is introduced. (4) Object discovery and localization in a fully unsupervised setup is explored on challenging benchmark datasets [16, 40].

2. Related work

Unsupervised object discovery has long been attempted in computer vision. Sivic *et al*. [48] and Russell *et al*. [42] apply statistical topic discovery models. Grauman and Darrel [21] use partial correspondence and clustering of local features. Kim and Torralba [28] employ a link analysis technique. Faktor and Irani [17] propose clustering by composition. Unsupervised object discovery, however, has proven extremely difficult "in the wild"; all of these previous approaches have been successfully demonstrated in a restricted setting with a few distinctive object classes, but their localization results turn out to be far behind weakly-supervised results on challenging benchmarks [12, 28, 51].

Given the difficulty of fully unsupervised discovery, recent work has more focused on weakly-supervised approaches from different angles. Cosegmentation is the problem of segmenting common foreground regions out of a set of images. It has been first introduced by Rother *et al*. [38] who fuse Markov random fields with color histogram matching to segment objects common to two images. Since then, this approach has been improved in numerous ways [4, 6, 23, 55], and extended to handle more general cases [7, 25, 40, 54]. Given the same type of input as cosegmentation, colocalization seeks to localize objects with bounding boxes instead of pixel-wise segmentations. Tang *et al*. [51] use the discriminative clustering framework of [25] to localize common objects in a set of noisy images, and Joulin *et al*. [27] extend it to colocalization of video frames. Weakly-supervised localization [9, 12, 35, 36, 46, 49] shares the same type of output as colocalization, but assumes a more supervised scenario with image-level labels that indicate whether a target object class appears in the image or not. These labels enable to learn more discriminative localization methods, *e.g*., by mining negative images [9]. Recent work on discriminative patch discovery [15, 44, 50] learns mid-level visual representations in a weakly-supervised mode, and use them for object recognition [15, 44] and discovery [13, 50].

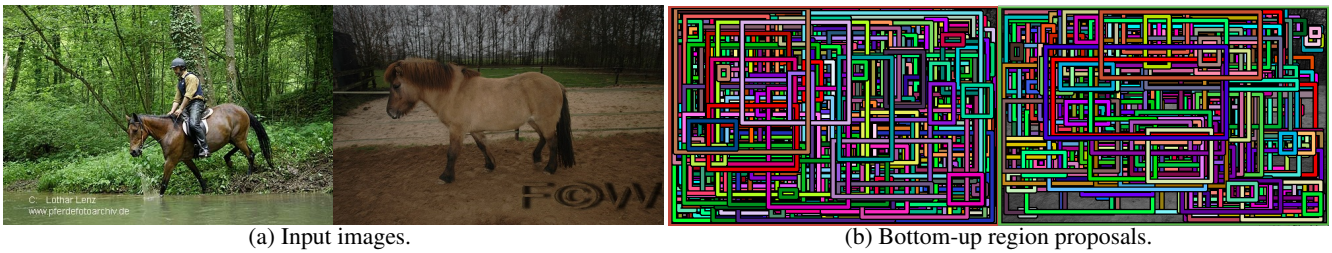
Region proposals have been used in many of the methods discussed so far, but most of them [12, 27, 28, 42, 51, 54] use relatively a small number of the best proposals (typically, less than 100 for each image) to form whole object hypotheses, often together with generic objectness measures [1]. In contrast, we use a large number of region proposals (typically, between 1000 and 4000) as primitive elements for matching without any objectness priors. While many other approaches [7, 40, 41] also use correspondences between image pairs to discover object regions, they do not use an efficient part-based matching approach such as ours. Many of them [7, 21, 40] are driven by correspondence techniques, *e.g*., the SIFT flow [32], based on generic local regions. In the sense that semi-local or mid-level parts are crucial for representing generic objects [18, 30], we believe segment-level regions are more adequate for object matching and discovery. The work of Rubio *et al*. [41] introduces such a segment-level matching term in their cosegmentation formulation. Unlike ours, however, it requires a reasonable initialization by an objectness measure [1], and does not scale well with a large number of segments and images.

3. Proposed approach

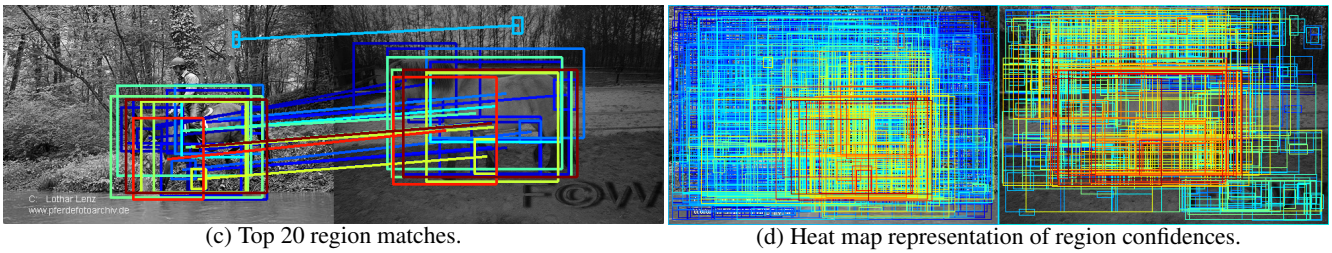
For unsupervised object discovery, we combine an efficient part-based matching technique with a foreground localization scheme. In this section we first introduce the two main components of our approach, and then describe the overall algorithm for unsupervised object discovery.

3.1. Part-based region matching

For part-based matching in an unsupervised setting, we use off-the-shelf region proposals [34] as candidate regions for objects and object parts: Diverse multi-scale proposals include meaningful parts of objects as well as objects themselves. Let us assume that two sets of region proposals R and R' have been extracted from two images $\mathcal{I}$ and $\mathcal{I}'$, respectively. Let $r = (f, l) \in R$ denote a region with feature f observed at location l. We use 8×8 HOG descriptors [10, 22] for f to describe the region patches, and center position and scale for l to specify the location. A match between r' and r is represented by (r, r'). For the sake of brevity, we use short notations $\mathcal{D}$ for two region sets, and m for a match: $\mathcal{D} = (R, R')$, $m = (r, r')$ in $R \times R'$. Then, our probabilistic model of a match confidence for m is rep-



(a) Input images. (b) Bottom-up region proposals.



(c) Top 20 region matches. (d) Heat map representation of region confidences.

Figure 2. Part-based region matching using bottom-up region proposals. (a-b) Given two images and their multi-scale region proposals [34], the proposed matching algorithm efficiently evaluates candidate matches between two sets of regions (2205×1044 regions in this example) and produce match confidences for them. (c) Based on the match confidence, the 20 best matches are shown by greedy mapping with a one-to-one constraint. The confidence is color-coded in each match (red: high, blue: low). (d) The region confidences of Eq.(4) are visualized in the heat map representation. Common object foregrounds tend to have higher confidences than others. (Best viewed in color.)

resented by $p(m|\mathcal{D})$. Assuming a common object appears in $\mathcal{I}$ and $\mathcal{I}'$, let the offset x denote its pose displacement between $\mathcal{I}$ and $\mathcal{I}'$, related to properties such as position and scale change. $p(x|\mathcal{D})$ becomes the probability of the common object being located with *offset* x. Now, the match confidence is decomposed in a Bayesian manner:

$$\begin{aligned} p(m|\mathcal{D}) &= \sum_x p(m,x|\mathcal{D}) = \sum_x p(m|x,\mathcal{D})p(x|\mathcal{D}) \\ &= p(m_\mathrm{a})\sum_x p(m_\mathrm{g}|x)p(x|\mathcal{D}), \end{aligned} \tag{1}$$

where we suppose that appearance matching m_a is independent of geometry matching m_g and an object location offset x. Appearance likelihood $p(m_a)$ is simply computed as the similarity between f and f'. Geometry likelihood $p(m_\mathrm{g}|x)$ is estimated by comparing displacement $l'-l$ to the given offset x. For $p(m_\mathrm{g}|x)$, we construct three-dimensional offset bins for translation and scale change, and use a Gaussian distribution centered on offset x.

The main issue is how to estimate geometry prior $p(x|\mathcal{D})$ without any information about objects and their locations. Inspired by the generalized Hough transform [2] and its extensions [31, 57], we propose the Hough space score $h(x|\mathcal{D})$, that is the sum of individual probabilities $p(m,x|\mathcal{D})$ over all possible region matches $m \in R\times R'$. The voting is done with an initial assumption of a uniform prior over x:

$$\begin{aligned} h(x|\mathcal{D}) &= \sum_m p(m|x,\mathcal{D}) \\ &= \sum_m p(m_\mathrm{a})p(m_\mathrm{g}|x), \end{aligned} \tag{2}$$

which predicts a pseudo likelihood of common objects at offset x. Assuming $p(x|\mathcal{D}) \propto h(x|\mathcal{D})$[1], we rewrite Eq.(1) to define the *Hough match confidence* as

$$c(m|\mathcal{D}) = p(m_\mathrm{a})\sum_x p(m_\mathrm{g}|x)h(x|\mathcal{D}). \tag{3}$$

Interestingly, this formulation can be seen as a combination of bottom-up and top-down processes: The bottom-up process aggregates individual votes into the Hough space scores (Eq.(2)), and the top-down process evaluates each match confidence based on those scores (Eq.(3)). We call this algorithm *Probabilistic Hough Matching* (PHM). Leveraging the Hough space score as a spatial prior, it provides robust match confidences for candidate matches. In particular, given multi-scale region proposals, different region matches on the same object cast votes for each other, and make all the region matches on the object obtain high confidences. This is an efficient part-based matching procedure with complexity of $\mathcal{O}(nn')$, where n and n' are the number of regions in R and R', respectively. As shown in Fig. 2c, reliable matches can be obtained when a proper mapping constraint (*e.g.*, one-to-one, one-to-many, etc.) is enforced on the confidence as a post-processing.[2]

We define the *region confidence* as a max-pooled match confidence for r in R with respect to R':

$$\phi(r) = \max_{r'} c\big((r,r')|(R,R')\big), \tag{4}$$

[1] Although Eq.(1) is a valid probabilistic model, using $h(x|\mathcal{D})$ as defined by Eq.(2) is a heuristic way of estimating $p(x|\mathcal{D})$ in terms of "pseudo likelihood". Like all its "probabilistic Hough transform" predecessors we know of [3, 31, 33], however, it lacks a proper probabilistic interpretation.

[2] Here we focus on the use of match confidence for object discovery rather than the final individual matches. For more details on PHM, see our webpage: http://www.di.ens.fr/willow/research/PHM/.

which derives from the best matches from R' to R under one-to-many mapping constraints. High region confidences guarantee that corresponding regions have at least single good matches in consideration of both appearance and spatial consistency. As shown in Fig. 2d, the region confidence provides a useful measure for common regions between images, thus functioning as a driving force in object discovery.

3.2. Foreground localization

Foreground objects do not directly emerge from part-based region matching: A region with the highest confidence is often a salient part of a common object while good localization is supposed to tightly bound the entire object region. We need a principled and unsupervised way to tackle the intrinsic ambiguity in separating the foreground objects from the background, which is one of the main challenges in unsupervised object discovery. In Gestalt principles of visual perception [39] and design [24], regions that "stand out" are more likely to be seen as a foreground. A high contrast lies between the foreground and background, and a lower contrast between foreground parts or background parts. Inspired by these figure/ground principles, we evaluate a foreground score of a region by its perceptual *contrast* standing out of its potential backgrounds. To measure the contrast, we leverage on the region confidence from part-based matching, which is well supported by the work of Peterson and Gibson, demonstrating the role of object recognition or matching in the figure/ground process [37].

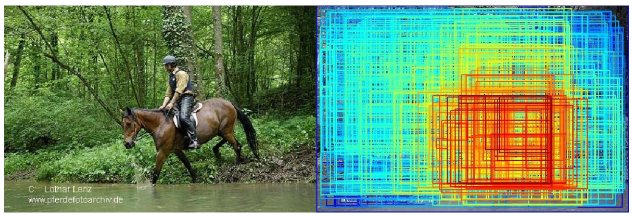
First, we generalize the notion of the region confidence to exploit multiple images. Let us assume $\mathcal{I}$ as a target image, and $\mathcal{I}'$ as a source image. The region confidence of Eq.(4) is a function of region r in target R with its best correspondence r' in source R' as a latent variable. Given multiple source images, it can be naturally extended with more latent variables, meaning the best correspondences from the source images to r. Let us define *neighbor images* N of target image $\mathcal{I}$ as an index set of source images where an object in $\mathcal{I}$ may appear. Generalizing Eq.(4), the region confidence can be rewritten as

$$\begin{aligned}\psi(r) &= \max_{\{r'_i\}_{i\in N}} \sum_{i\in N} c\big((r,r'_i)|(R,R'_i)\big) \\ &= \sum_{i\in N} \max_{r'\in R'_i} c\big((r,r')|(R,R'_i)\big), \end{aligned} \tag{5}$$

which reduces to the aggregated confidence from the neighbor images. More images may give better confidences.

Given regions R with their region confidences, we evaluate a perceptual contrast for region $r \in R$ by computing the increment of its confidence from its potential backgrounds. To this end, we define the *standout score* as

$$\begin{aligned} s(r) &= \psi(r) - \max_{r_\text{b}\in B(r)} \psi(r_\text{b}), \\ s.t.\ \ B(r) &= \{r_\text{b} \mid r \subsetneq r_\text{b}, r_\text{b} \in R\}, \end{aligned} \tag{6}$$



(a) Region confidences with respect to multiple source images.

(b) Measuring the standout score from the region confidences.

Figure 3. Foreground localization. (a) Given multiple source images with common objects, region confidences can be computed according to Eq.(5). More source images may give better region confidences. (b) Given regions (boxes) on the left, the standout score of Eq.(6) for the red box corresponds to the difference between its confidence and the maximum confidence of boxes containing the red box (green boxes). In the same way, the standout score for the white box takes into account blue, red, and green boxes altogether as its potential backgrounds. Three boxes on the right are ones with the top three standout scores from the region confidence. The red one has the top score. (Best viewed in color.)

where $r \subsetneq r_\text{b}$ means region r is contained in region r_b. The idea is illustrated in Fig. 3b. Imagine a region gradually shrinking from a whole image region, to a tight object region, to a part region. Significant increase in region confidence is most likely to occur at the point of taking the tight object region. In practice, we decide the inclusive relation $r \subsetneq r_\text{b}$ by two simple criteria: (1) The box area of r is less than 50% of the box area of r_b. (2) 80% of the box area of r overlaps with the box area of r_b.

The standout score reflects the principle that we perceive a lower contrast between parts of the foreground than that between the background and the foreground. As shown in the example of Fig. 3b, we can localize potential object regions by selecting regions with top standout scores.

3.3. Object discovery algorithm

For unsupervised object discovery, we combine part-based region matching and foreground localization in a coordinate descent-style algorithm. Given a collection of images $\mathcal{C}$, our algorithm alternates between matching image pairs and re-localizing potential object regions. Instead of matching all possible pairs over the images, we retrieve k neighbors for each image and perform part-based matching only from those neighbor images. To make the algorithm robust to localization failure in precedent iterations, we maintain five potential object regions for each image. Both the neighbor images and the potential object regions

are updated over iterations.

The algorithm starts with an entire image region as an initial set of potential object regions O_i for each image $\mathcal{I}_i$, and performs the following three steps at each iteration.

Neighbor image retrieval. For each image $\mathcal{I}_i$, k nearest neighbor images $\{\mathcal{I}_j \mid i \in N_i\}$ are retrieved based on the similarity between O_i and O_j. We use 10 neighbor images ($k = 10$).[3] At the first iteration, as the potential object regions become entire image regions, nearest-neighbor matching with the GIST descriptor [52] is used. From the second iteration, we perform PHM with re-localized object regions. For efficiency, we only use the top 20 region proposals according to region confidences, which are contained in the potential object regions. The similarity for retrieval is computed as the sum of those region confidences.

Part-based region matching. Part-based matching by PHM is performed on $\mathcal{I}_i$ from its neighbor images $\{\mathcal{I}_j \mid j \in N_i\}$. To exploit current localization in a robust way, an *asymmetric matching strategy* is adopted: We use all regions proposals in $\mathcal{I}_i$, whereas for the neighbor image $\mathcal{I}_j$ we take regions only contained in potential object regions O_j. This matching strategy does not restrict potential object regions in target $\mathcal{I}_i$ while effectively utilizing localized object regions at the precedent step.

Foreground localization. For each image $\mathcal{I}_i$, the standout scores are computed so that the set of potential object regions O_i is updated to that of regions with top standout scores. This re-localization improves both neighbor image retrieval and region matching at the subsequent iteration.

These steps are repeated for a few iterations until near-convergence. As will be shown in our experiments, 5 iterations are sufficient as no significant change occurs in more iterations. Final localization is done by selecting the most standing-out region at the end. The algorithm is designed based on the idea that better localization makes better retrieval and matching, and vice versa. As each image is independently processed at each iteration, the algorithm is easily parallelizable in computation. Object discovery on 500 images takes less than an hour with a 10-core desktop computer, using our current parallel MATLAB implementation.

4. Experimental evaluation

The degree of supervision used in visual learning tasks varies from strong (supervised localization [18, 20]) to weak (weakly-supervised localization [9, 46]), very weak (colocalization [27, 51] and cosegmentation [40]), and null (fully-unsupervised discovery). To evaluate our approach for unsupervised object discovery, we conduct two types of experiments: *separate-class* and *mixed-class* experiments. Our separate-class experiments test performance of our approach in a very weakly supervised mode. Our mixed-class experiments test object discovery "in the wild" (in a fully-unsupervised mode), by mixing all images of all classes in a dataset, and evaluating performance on the whole dataset. To the best of our knowledge, this type of localization experiments has never been fully attempted before on challenging real-world datasets. We conduct experiments on two realistic benchmarks, the Object Discovery [40] and the PASCAL VOC 2007 [16], and compare the results with those of the current state of the arts in cosegmentation [29, 25, 26, 40], colocalization [8, 12, 42, 27, 51], and weakly-supervised localization [9, 12, 35, 36, 46, 56].

4.1. Evaluation metrics

The correct localization (CorLoc) metric is an evaluation metric widely used in related work [12, 27, 46, 51], and defined as the percentage of images correctly localized according to the PASCAL criterion: $\frac{area(b_p \cap b_{gt})}{area(b_p \cup b_{gt})} > 0.5$, where b_p is the predicted box and b_{gt} is the ground-truth box. The metric is adequate for a conventional separate-class setup: As a given image collection contains a single target class, only object localization is evaluated per image. In a mixed-class setup, however, we have another dimension involved: As different images may contain different object classes, associative relations across the images need to be evaluated. As such a metric orthogonal to CorLoc, we propose the *correct retrieval* (CorRet) evaluation metric defined as follows. Given the k nearest neighbors identified by retrieval for each image, CorRet is defined as the mean percentage of these neighbors that belong to the same (ground-truth) class as the image itself. This measure depends on k, fixed here to a value of 10. CorRet may also prove useful in other applications that discover the underlying "topology" (nearest-neighbor structure) of image collections. CorRet and CorLoc metrics effectively complement each other in the mixed-class setup: CorRet reveals how correctly an image is associated to other images, while CorLoc measures how correctly an object is localized in the image.

4.2. The Object Discovery dataset

The Object Discovery dataset [40] was collected by the Bing API using queries for airplane, car, and horse, resulting in image sets containing outlier images without the query object. We use the 100 image subsets [40] to enable comparisons to previous state of the art in cosegmentation and colocalization. In each set of 100 images, airplane, car, horse have 18, 11, 7 outlier images, respectively. Following [51], we convert the ground-truth segmentations and cosegmentation results of [29, 25, 26, 40] to localization boxes.

We conduct separate-class experiments as in [12, 51], and a mixed-class experiment on a collection of 300 images from all the three classes. The mixed-class image collection

[3] In our experiments, the use of more neighbor images does not always improve the performance while increasing computation time.

Table 1. CorLoc (%) on separate-class Object Discovery dataset.

Methods	Airplane	Car	Horse	Average
Kim *et al.* [29]	21.95	0.00	16.13	12.69
Joulin *et al.* [25]	32.93	66.29	54.84	51.35
Joulin *et al.* [26]	57.32	64.04	52.69	58.02
Rubinstein *et al.* [40]	74.39	87.64	63.44	75.16
Tang *et al.* [51]	71.95	93.26	64.52	76.58
Ours	**82.93**	**94.38**	**75.27**	**84.19**

Table 2. Performance on mixed-class Object Discovery dataset.

Evaluation metric	Airplane	Car	Horse	Average
CorLoc	81.71	94.38	70.97	82.35
CorRet	73.30	92.00	82.80	82.70

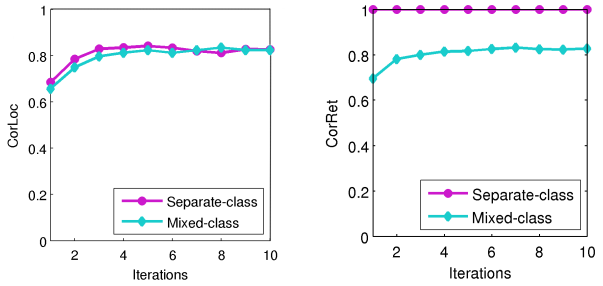


Figure 4. Average CorLoc (left) and CorRet (right) vs. # of iterations on the Object Discovery dataset.

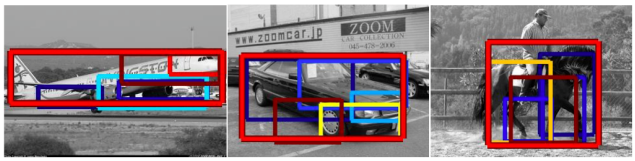


Figure 5. Examples of localization on unlabeled Object Discovery dataset. Small boxes inside the localized object box (red box) represents five most confident part regions. (Best viewed in color.)

contains 3 classes and 36 outlier images. Figure 4 shows the average CorLoc and CorRet over iterations, where we see the proposed algorithm quickly improves both localization (CorLoc) and retrieval (CorRet) in early iterations, and then approaches a steady state. In the separate-class setup, CorRet is always perfect because no other object class exists in the retrieval. As we have found no significant change in both localization and retrieval after 4-5 iterations in all our experiments, we measure all performances of our method in this paper after 5 iterations. The separate-class results are quantified in Table 1, and compared to those of state-of-the-art cosegmentation [29, 25, 26] and colocalization [40, 51] methods. Our method outperforms all of them in this setup. The mixed-class result is in Table 2, and examples of localization are shown in Fig. 5. Remarkably, our localization performance in the mixed-class setup is almost the same as that in the separate-class setup. Localized objects are visualized in red boxes with five most confident regions inside the object, indicating parts most contributing to discovery. Table 2 and Fig. 4 show that our localization is robust to noisy neighbor images retrieved from different classes.

Table 5. CorLoc comparison on PASCAL07-6x2.

Method	Data used	Avg. CorLoc (%)
Chum and Zisserman [8]	P + N	33
Deselaers *et al.* [12]	P + N	50
Siva and Xiang [47]	P + N	49
Tang *et al.* [51]	P	39
Ours	P	**68**
Ours (mixed-class)	unsupervised	54

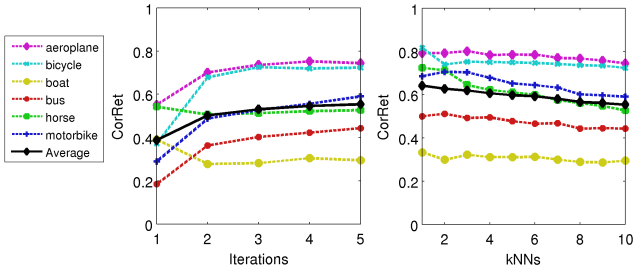


Figure 6. CorRet variation on mixed-class PASCAL07-6x2.

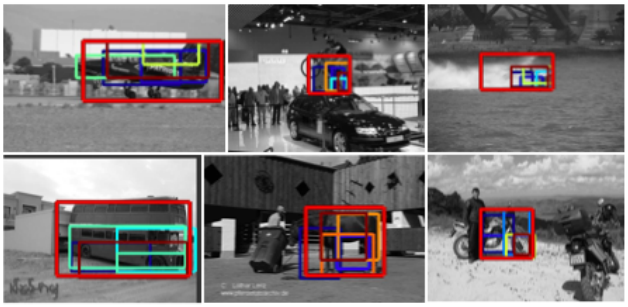


Figure 7. Example results on mixed-class PASCAL07-6x2.

4.3. PASCAL VOC 2007 dataset

The PASCAL VOC 2007 [16] contains realistic images of 20 object classes. Compared to the Object Discovery dataset, it is significantly more challenging due to considerable clutter, occlusion, and diverse viewpoints. To facilitate a scale-level analysis and comparison to other methods, we conduct experiments on two subsets of different sizes: PASCAL07-6x2 and PASCAL07-all. The PASCAL07-6x2 subset [12] consists of all images from 6 classes (aeroplane, bicycle, boat, bus, horse, and motorbike) of train+val dataset from the left and right aspect each. Each of the 12 class/viewpoint combinations contains between 21 and 50 images for a total of 463 images. For a large-scale experiment with all classes following [9, 12, 36], we take all train+val dataset images discarding images that only contain object instances marked as "difficult" or "truncate". Each of the 20 classes contains between 49 and 1023 images for a total of 4548 images. We refer to it as PASCAL07-all.

Experiments on PASCAL07-6x2. In the separate-class setup, we evaluate performance for each class in Table 3, where we also analyze each component of our method by removing it from the full version: 'w/o MOR' eliminates the use of multiple object regions over iterations, thus main-

Table 3. CorLoc performance (%) on separate-class PASCAL07-6x2

Method	aeroplane		bicycle		boat		bus		horse		motorbike		Average
	L	R	L	R	L	R	L	R	L	R	L	R	
Ours (full)	62.79	71.79	77.08	62.00	25.00	32.56	66.67	91.30	83.33	86.96	82.96	70.59	**67.68**
Ours w/o MOR	62.79	74.36	52.08	42.00	15.91	27.91	61.90	91.30	85.42	76.09	48.72	8.82	53.94
Ours w/o PHM	39.53	38.46	54.17	60.00	6.82	9.30	42.86	73.91	68.75	82.61	33.33	2.94	42.72
Ours w/o STO	34.88	0.0	2.08	0.0	0.0	4.65	0.0	8.70	64.58	30.43	2.56	0.0	12.32

Table 4. CorLoc and CorRet performance (%) on mixed-class PASCAL07-6x2.

Metric	aeroplane		bicycle		boat		bus		horse		motorbike		Average
	L	R	L	R	L	R	L	R	L	R	L	R	
CorLoc	62.79	66.67	54.17	56.00	18.18	18.60	42.86	69.57	70.83	71.74	69.23	44.12	53.73
CorRet	61.40	42.56	48.75	56.80	19.09	13.02	13.33	30.87	41.46	41.74	38.72	43.24	37.58
CorRet (class)	74.39		72.35		29.43		44.32		52.66		59.04		55.36

taining only a single potential object region for each image. 'w/o PHM' substitutes PHM with appearance-based matching without any geometric consideration. 'w/o STO' replaces the standout score with the maximum confidence. As expected, we can see that the removal of each component damages performance substantially. In particular, it clearly shows both part-based matching (using PHM) and foreground localization (using the standout score) are crucial for robust object discovery. In Table 5, we quantitatively compare ours to previous results [8, 12, 47, 51] on PASCAL07-6x2. Our method significantly outperforms those with a large margin. Note that our method does not incorporate any form of object priors such as off-the-shelf objectness measures [12, 47, 51], and only use positive images (P) without more training data, *i.e.*, negative images (N) [12, 47]. For the mixed-class experiment, we run our method on a collection of all class/view images in PASCAL07-6x2, and evaluate its CorLoc and CorRet peformance in Table 4. To better understand our retrieval performance per class, we measure CorRet for classes (regardless of views) in the third row, and analyze it by increasing the numbers of iterations and neighbor images in Fig. 6. This shows that our method achieves better localization and retrieval simultaneously, and benefits from each other. In Fig. 7, we show example results of our mixed-class experiment on PASCAL07-6x2. In spite of a small size of objects even partially occluded, our method is able to localize instances from cluttered scenes, and discovers confident object parts as well. From Table 5, we see that even without using the separate-class setup, the method localizes target objects markedly better than recent colocalization methods.

Experiments on PASCAL07-all. Here we tackle a much more challenging and larger-scale discovery task, using all the images from the PASCAL07 dataset. We first report separate-class results, and compare our results to those of the state of the arts in weakly-supervised localization [9, 36, 43, 47, 45, 46, 56] and colocalization [27] in Table 6. Note that weakly-supervised methods use more training data, *i.e.*, negative images (N). Also note that the best performing method [56] uses CNN features pretrained on the ImageNet dataset [11], thus additional supervised data (A). Surprisingly, the performance of our method is very close to the best of weakly-supervised localization [9] not using such additional data.

In the mixed-class setting, we face an issue linked to the potential presence of multiple dominant labeled (ground-truth) objects in each image. Basically, both CorLoc and CorRet are defined as a per-image measure, *e.g.*, CorLoc assigns an image true if any true localization is done in the image. For images with multiple class labels in the mixed-class setup, which is the case of PASCAL-all with highly overlapping class labels (*e.g.*, persons appear in almost 1/3 of images), CorLoc needs to be extended in a natural manner. To measure a class-specific average CorLoc in such a multi-label and mixed-class setup, we take all images containing the object class and measure their average CorLoc for the class. The upper bound of this class-specific average CorLoc may be less than 100% because only one localization exists for each image in our setting. To complement this, as shown at the last column of Table 7, we add the 'any'-class average CorLoc, where we assign an image true if any true localization of any class exists in the image. The similar evaluation is also done for CorRec. Both 'any'-class CorLoc and CorRet have an upper bound of 100% even when images have multiple class labels, whereas those in 'Av.' (average) may not. Note that the mixed-class PASCAL07-all dataset has a very imbalanced class distribution: the 20 classes have very different numbers of images, from 49 (sheep) to 1023 (person). Yet, as quantified in Table 7, our method still performs well even in this unsupervised mixed-class setting, and its localization performance is comparable to that in the separate-class setup. To better understand this, we visualize in Fig. 8 a confusion matrix of retrieved neighbor images based on the mixed-class result, where each row corresponds to the average retrieval ratios (%) by each class. Note that the matrix reflects class frequency so that the person class appears dominant. We clearly see that despite relatively low retrieval accuracy, many of retrieved images

Table 6. CorLoc (%) on separate-class PASCAL07-all, compared to the state of the arts in weakly-supervised / co- localization.

Method	Data used	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Pandey & Lazebnik [36]	P + N	50.9	56.7	-	10.6	0	56.6	-	-	2.5	-	14.3	-	50.0	53.5	11.2	5.0	-	34.9	33.0	40.6	-
Siva & Xiang [47]	P + N	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	30.5	19.0	34.0	48.8	65.3	8.2	9.4	16.7	32.3	54.8	5.5	30.4
Siva *et al.* [45]	P + N	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8	27.5	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Shi *et al.* [43]	P + N	67.3	54.4	34.3	17.8	1.3	46.6	60.7	68.9	2.5	32.4	16.2	58.9	51.5	64.6	18.2	3.1	20.9	34.7	63.4	5.9	36.2
Cinbis *et al.* [9]	P + N	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
Wang *et al.* [56]	P + N + A	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Joulin *et al.* [27]	P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.6
Ours	P	50.3	42.8	30.0	18.5	4.0	62.3	64.5	42.5	8.6	49.0	12.2	44.0	64.1	57.2	15.3	9.4	30.9	34.0	61.6	31.5	36.6

Table 7. CorLoc and CorRet performance (%) on mixed-class PASCAL07-all. (See text for 'any').

Evaluation metric	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.	any
CorLoc	40.4	32.8	28.8	22.7	2.8	48.4	58.7	41.0	9.8	32.0	10.2	41.9	51.9	43.3	13.0	10.6	32.4	30.2	52.7	21.8	31.3	37.6
CorRet	51.1	45.3	12.7	12.1	11.4	21.2	61.9	11.6	19.2	9.7	3.9	17.2	29.6	34.0	43.7	10.2	8.1	9.9	23.7	27.3	23.2	36.6

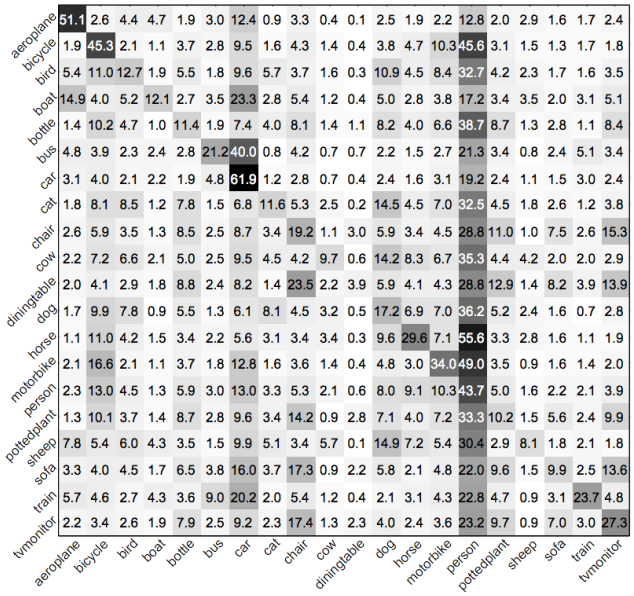


Figure 8. Confusion matrix of retrieval on mixed PASCAL07-all.

Figure 9. Localization in an example and its neighbor images on mixed-class PASCAL07-all. A bus is successfully localized in the image (red dashed box) from its neighbors (10 images) containing even other classes (car, sofa). Boxes in the neighbors show potential objects at the final iteration. (Best viewed in color.)

come from other classes with partial similarity, *e.g.*, bicycle - motorbike, bus - car, etc. Figure 9 shows a typical example of such cases. These results strongly suggest that our part-based approach to object discovery effectively benefits from different but similar classes without any class-specific supervision. Interestingly, the significant difference in retrieval performance (CorRet) from 100% in the separate-class setup influences much less on localization (CorLoc). Further analysis of our experiments also reveals that in the case of an imbalanced distribution of classes, a class with lower frequency is harder to be localized than a class with higher frequency. To see this, consider 'the highest' (person, car, chair, dog, cat) and 'the lowest' (sheep, cow, boat, bus, dinningtable) in class frequency. We have measured how much the average performance changes between the separate-class (clean) and mixed-class (imbalanced) settings. The average CorLoc of 'the highest' only drops by 1.2%, while that of 'the lowest' drops by 9.4%. This clearly indicates that a class with lower class frequency is harder to localize in the mixed-class setting. Retrieval performance of 'the lowest' (CorRet 11.0%) is also much worse than that of 'the highest' (CorRet 30.7%). For more information, see our project webpage: http://www.di.ens.fr/willow/research/objectdiscovery/.

5. Discussion and conclusion

We have demonstrated unsupervised object localization in the challenging mixed-class setup, which has never been fully attempted before on a challenging dataset such as [16]. The result shows that the effective use of part-based matching is a crucial factor for object discovery. In the future, we will advance this direction and further explore how to handle multiple object instances per image as well as build visual models for classification and detection. In this paper, our aim has been to evaluate our unsupervised algorithm per se, and have thus abstained from any form of additional supervision such as off-the-shelf saliency/objectness measures, negative data, and pretrained features. The use of such information will further improve our results.

Acknowledgments. This work was supported by the ERC grants Activia, Allegro, and VideoWorld, and the Institut Universitaire de France.

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *TPAMI*, 2012. 2

[2] D. Ballard. Generalizing the Hough transform to detect arbitrary shapes. *Pattern Recognition*, 1981. 2, 3

[3] O. Barinova, V. Lempitsky, and P. Kohli. On detection of multiple object instances using Hough transforms. In *CVPR*. IEEE, 2010. 3

[4] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 2

[5] M. Cho, K. Alahari, and J. Ponce. Learning graphs to match. In *ICCV*, 2013. 2

[6] M. Cho, Y. M. Shin, and K. M. Lee. Co-recognition of image pairs by data-driven Monte Carlo image exploration. In *ECCV*, 2008. 2

[7] M. Cho, Y. M. Shin, and K. M. Lee. Unsupervised detection and segmentation of identical objects. In *CVPR*, 2010. 2

[8] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007. 5, 6, 7

[9] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, 2014. 1, 2, 5, 6, 7, 8

[10] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 2

[11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 1, 7

[12] T. Deselaers, B. Alexe, and V. Ferrari. Localizing objects while learning their appearance. In *ECCV*, 2010. 1, 2, 5, 6, 7

[13] C. Doersch, A. Gupta, and A. A. Efros. Context as supervisory signal: Discovering objects with predictable context. In *ECCV*, 2014. 2

[14] O. Duchenne, A. Joulin, and J. Ponce. A graph-matching kernel for object categorization. In *ICCV*, 2011. 2

[15] I. Endres, K. J. Shih, J. Jiaa, and D. Hoiem. Learning collections of part models for object recognition. In *CVPR*, 2013. 2

[16] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 1, 2, 5, 6, 8

[17] A. Faktor and M. Irani. "Clustering by composition" – unsupervised discovery of image categories. In *ECCV*, 2012. 2

[18] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010. 2, 5

[19] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2003. 2

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014. 2, 5

[21] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006. 2

[22] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *ECCV*, 2012. 2

[23] D. S. Hochbaum and V. Singh. An efficient algorithm for co-segmentation. In *ICCV*, 2009. 2

[24] I. Jackson. Gestalt-a learning theory for graphic design education. *IJADE*, 2008. 4

[25] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 1, 2, 5, 6

[26] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 5, 6

[27] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014. 1, 2, 5, 7, 8

[28] G. Kim and A. Torralba. Unsupervised detection of regions of interest using iterative link analysis. In *NIPS*, 2009. 2

[29] G. Kim and E. Xing. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 1, 5, 6

[30] S. Lazebnik, C. Schmid, and J. Ponce. Semi-local affine parts for object recognition. In *BMVC*, 2004. 2

[31] B. Leibe, A. Leonardis, and B. Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 2008. 3

[32] C. Liu, J. Yuen, and A. Torralba. SIFT Flow: dense correspondence across scenes and its applications. *TPAMI*, 2011. 2

[33] S. Maji and J. Malik. Object detection using a max-margin Hough transform. In *CVPR*, 2009. 3

[34] S. Manen, M. Guillaumin, and L. V. Gool. Prime object proposals with randomized Prim's algorithm. In *ICCV*, 2013. 2, 3

[35] M. H. Nguyen, L. Torresani, F. de la Torre, and C. Rother. Weakly supervised discriminative localization and classification: a joint learning process. In *ICCV*, 2009. 1, 2, 5

[36] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011. 1, 2, 5, 6, 7, 8

[37] M. A. Peterson and B. S. Gibson. Object recognition contributions to figure-ground organization: Operations on outlines and subjective contours. *Perception & Psychophysics*, 1994. 4

[38] C. Rother, T. P. Minka, A. Blake, and V. Kolmogorov. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *CVPR*, 2006. 2

[39] E. Rubin. Figure and ground. *Visual Perception*, 2001. 4

[40] M. Rubinstein and A. Joulin. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013. 1, 2, 5, 6

[41] J. C. Rubio, J. Serrat, A. López, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, 2012. 2

[42] B. Russell, W. Freeman, A. Efros, J. Sivic, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *CVPR*, 2006. 2, 5

[43] Z. Shi, T. M. Hospedales, and T. Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *ICCV*, 2013. 1, 7, 8

[44] S. Singh, A. Gupta, and A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012. 2

[45] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV*, 2012. 1, 7, 8

[46] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013. 2, 5, 7

[47] P. Siva and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011. 6, 7, 8

[48] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. In *ICCV*, 2005. 2

[49] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell. On learning to localize objects with minimal supervision. In *ICML*, 2014. 2

[50] J. Sun and J. Ponce. Learning discriminative part detectors for image classification and cosegmentation. In *ICCV*, 2013. 2

[51] K. Tang, A. Joulin, and L.-j. Li. Co-localization in real-world images. In *CVPR*, 2014. 1, 2, 5, 6, 7

[52] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In *CVPR*, 2008. 5

[53] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders. Selective search for object recognition. *IJCV*, 2013. 2

[54] S. Vicente. Object cosegmentation. In *CVPR*, 2011. 2

[55] S. Vicente, V. Kolmogorov, and C. Rother. Cosegmentation revisited: Models and optimization. In *ECCV*. 2010. 2

[56] C. Wang, W. Ren, K. Huang, and T. Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014. 1, 5, 7, 8

[57] Y. Zhang and T. Chen. Efficient kernels for identifying unbounded-order spatial features. In *CVPR*, 2009. 3