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

Minsu Cho, Suha Kwak, Cordelia Schmid, Jean Ponce

▶ **To cite this version:**

Minsu Cho, Suha Kwak, Cordelia Schmid, Jean Ponce. Unsupervised Object Discovery and Localization in the Wild: Part-based Matching with Bottom-up Region Proposals. 2015. hal-01110036v1

HAL Id: hal-01110036

https://inria.hal.science/hal-01110036v1

Preprint submitted on 27 Jan 2015 (v1), last revised 4 May 2015 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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[4,*]

[1]Inria [4]École Normale Supérieure / PSL Research University

Abstract

This paper addresses unsupervised discovery and localization of dominant objects from a noisy image collection of 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 significantly more general than typical colocalization, cosegmentation, or weakly-supervised localization tasks. We tackle the discovery and localization problem using a part-based 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 in each candidate region considering both appearance similarity 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 [9, 12], most state-of-the-art methods still rely on strong supervision in the form of manually-annotated bounding boxes on target instances. Recent work has begun to explore 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. Weakly-supervised localization [8, 28, 29, 36, 37, 45] requires positive and negative image-level labels for a target object class. On the other hand, cosegmentation [19, 20, 23, 33] and colocalization [10, 21, 40] assume less supervision and only require the image collection to contain a single dominant object class, allowing noisy images to some degree.

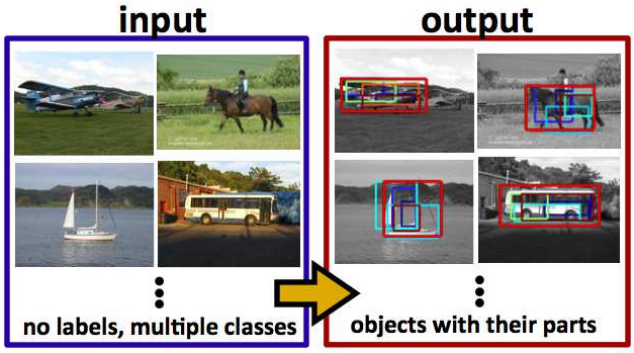


Figure 1. Unsupervised object discovery in the wild. We tackle object localization in an unsupervised scenario without any types of annotations, where a given image collection may contain multiple dominant object classes and even outlier images. The proposed method discovers object instances (red bounding 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 unsupervised, without any image-level annotations, an assumption of a single dominant class, or even a given number of object classes. In spite of this generality, the proposed method markedly outperforms the state of the arts in colocalization [21, 40] on standard benchmarks [12, 33], and closely competes with current weakly-supervised localization [8, 36, 45].

We advocate a part-based matching approach to unsupervised object discovery using bottom-up region proposals. Unlike previous proposal-based approaches [8, 21, 40, 43], we use region proposals [27] to form a set of candidate re-

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

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

gions not only for objects, but also for object parts.

The fact that multi-scale region proposals often include meaningful portions of the objects (*e.g.*, the object bounding box) in images has been used before to restrict the search space in object recognition tasks [8, 15, 42]. We go further and propose here to use these regions as part and object candidates for part-based matching. We use a probabilistic Hough transorm [2] to match those candidate regions across images, and assign them confidence scores reflecting both appearance similarity and spatial consistency. This can be seen as an unsupervised and efficient variant of both deformable part models [13, 14] and graph matching methods [4, 11]. Objects are discovered and localized by selecting the most salient regions that contain corresponding parts. To this end, we use a score that measures how the confidence in a region stands out over confidences of other boxes containing it. The proposed algorithm alternates between part-based region matching and foreground localization, improving both.

The main contributions of this paper can be summarized as follows: (1) A part-based region matching approach for unsupervised object discovery is introduced. (2) An efficient and robust matching algorithm based on a probabilistic Hough transform is proposed. (3) An unsupervised setup with mixed classes is explored on challenging benchmark datasets. (4) An extensive experimental evaluation on standard benchmarks demonstrates that the proposed approach gives significantly better localization performance than the state of the art in colocalization, and achieves robust object discovery in challenging mixed-class datasets.

2. Related work

Unsupervised object discovery is most closely related to cosegmentation and colocalization, and also to weakly-supervised localization. Cosegmentation is the problem of segmenting common foreground regions out of images. It has been first introduced by Rother *et al*. [31] 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 [3, 5, 16, 44], and extended to handle more general cases.

Joulin *et al*. [19] propose a discriminative clustering framework that can handle multiple images, and Cho *et al*. [6] use a match-growing approach to cosegmentation even for a single image. Kim and Xing [23] introduce an efficient sub-modular optimization approach for even larger datasets. Vicente *et al*. [43] combine cosegmentation with a notion of objectness by learning generic pairwise similarity between foreground segments. Rubinstein *et al*. [33] propose a robust cosegmentation method to discover common objects from noisy datasets collected by Internet search. Given the same type of input as cosegmentation, colocalization seeks to localize objects with bounding boxes instead of pixel-wise segmentations. Kim and Torralba [22] use a link analysis technique to discover regions of interest in a bounding-box representation. Tang *et al*. [40] extend the work of [19] to label object proposals among noisy images. Joulin *et al*. [21] introduce an efficient optimization approach and apply it to colocalization of video frames. Weakly-supervised localization [8, 10, 28, 29, 38] 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. Region proposals have been used in some of the methods discussed so far [10, 21, 22, 35, 40, 43], but relatively a few number of the best proposals (typically, less than 100 for each image) are typically used 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.

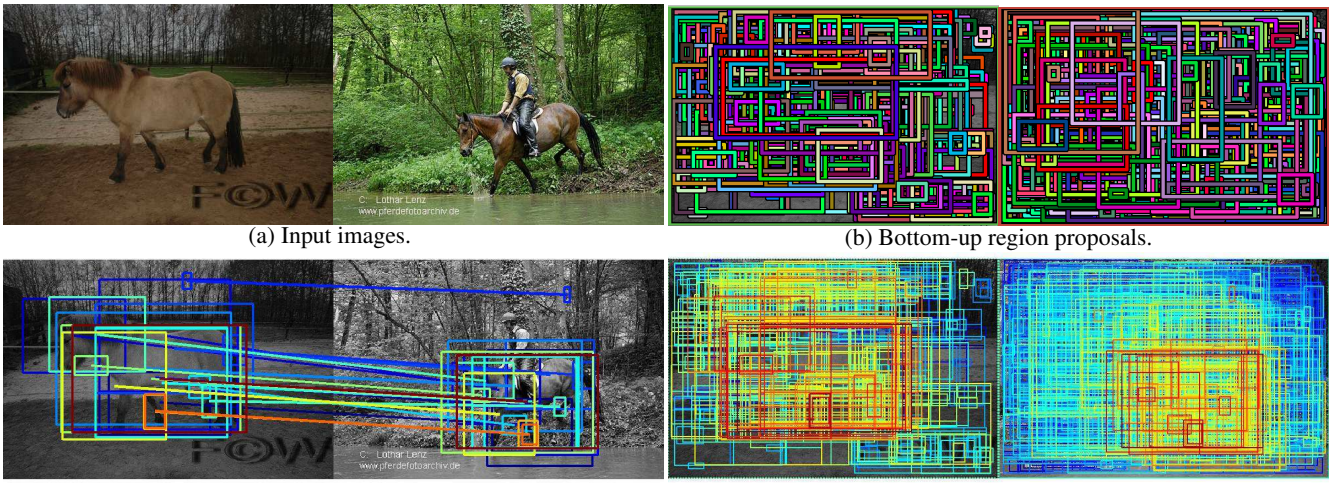
While many previous approaches [6, 33, 34] use correspondences between image pairs to find common foreground (dominant) object regions, they do not use an efficient part-based matching approach such as ours. Many of them are driven by correspondence techniques based on generic local regions [6, 33], *e.g.*, the SIFT flow [26]. In the sense that semi-local or mid-level parts are crucial for representing generic objects [13, 24], segment-level regions are more adequate for object discovery. The work of Rubio *et al*. [34] is close to this direction, and introduces a graph matching term in their cosegmentation formulation to enforce consistent region correspondences. Unlike ours, however, it requires a reasonable initialization by a generic objectness measure [1], and does not scale well with a large number of segments and images.

3. Proposed approach

For robust and unsupervised object discovery, we combine an efficient part-based matching technique with a part-aware foreground localization scheme that considers part regions contained in the object region. 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 [27] as candidate regions for objects and object parts. While actively studied in recent years, region proposals have been mostly used as candidates for an entire object region in detection or segmentation. Our insight is that diverse multi-scale proposals include meaningful parts of objects as well as objects themselves [27]. These regions not only reduce the search space but also provide primitive elements for part-based matching.



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

(c) Input images and 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 [27], the proposed matching algorithm efficiently evaluates candidate matches between two sets of regions (1044×2205 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.)

Let us assume that two sets of candidate regions R and R' have been extracted from two images $\mathcal{I}$ and $\mathcal{I}'$, respectively. Let $r = (d, l) \in R$ denote a region with descriptor d observed at location l. We use 8×8 HOG descriptors to describe the local patches. Then, our probabilistic model of a match confidence from r to r' is represented by $p(r \mapsto r'|R, R')$. Assuming a common object appears in $\mathcal{I}$ and $\mathcal{I}'$, let the offset x denote its pose displacement from $\mathcal{I}$ to $\mathcal{I}'$, related to properties such as position, scale, and aspect ratio. $p(x|R, R')$ becomes the probability of the common object being located with *offset* x from $\mathcal{I}$ to $\mathcal{I}'$. Now, the match confidence is decomposed in a Bayesian manner:

$$\begin{aligned} p(r \mapsto r'|R, R') &= \sum_x p(r \mapsto r', x|R, R') \\ &= \sum_x p(r \mapsto r'|x, R, R')p(x|R, R') \\ &= p(d \mapsto d') \sum_x p(l \mapsto l'|x)p(x|R, R'), \end{aligned} \tag{1}$$

where we suppose that the probability of descriptor matching is independent of that of location matching and an object location offset. Appearance likelihood $p(d \mapsto d')$ is simply computed as the similarity between descriptors d and d'. Geometry likelihood $p(l \mapsto l'|x)$ is estimated by comparing $l' - l$ to the given offset x. In this work, we construct three-dimensional offset bins for translation and scale of a box, and use a Gaussian distribution centered on the offset x for $p(l \mapsto l'|x)$.

Now, the main issue is how to estimate $p(x|R, R')$ without any supervised information about common objects and their locations.

Inspired by the generalized Hough transform [2] and its extensions [25, 46], we propose the Hough space score $h(x)$, that is the sum of individual probabilities $p(r \mapsto r', x|R, R')$ over all possible region matches. The voting is done with an initial assumption of a uniform prior over x, according to the principle of insufficient reason [18]:

$$\begin{aligned} h(x) &= \sum_{\forall i,a} p(r_i \mapsto r'_a|x, R, R') \\ &= \sum_{\forall i,a} p(d_i \mapsto d'_a)p(l_i \mapsto l'_a|x), \end{aligned} \tag{2}$$

which predicts a likelihood of common objects at offset location x. Assuming $p(x|R, R') \propto h(x)$, the match confidence of Eq.(1) can be updated so that the *Hough match confidence* is defined as

$$\mathcal{M}(i, a) = p(d_i \mapsto d'_a) \sum_x p(l_i \mapsto l'_a|x)h(x). \tag{3}$$

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, and the top-down process evaluates each match confidence based on those scores. 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, in our setting with region proposals, foreground part and object regions cast votes for each other and make the regions ob-

tain high confidences all together. This is an efficient part-based matching procedure with computational complexity of $\mathcal{O}(nn')$, where n and n' is 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. In this work, however, we focus not on the final matches but the use of the match confidences for unsupervised object discovery.

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

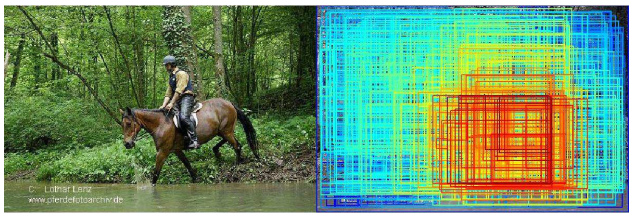
$$f(a) = \max_i \mathcal{M}(i, a), \tag{4}$$

which indicates a foreground likelihood for the region. High region confidences guarantee that corresponding regions have at least single good matches in consideration of both similarity 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. Part-aware foreground localization

Foreground objects do not directly emerge from part-based region matching: a region with the highest confidence is usually only a part of a common object while good object 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 [32] and design [17], 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 focus on detecting a *contrasting boundary* around the foreground, standing out of the background. To measure such a perceptual 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 [30].

First, we generalize the notion of the region confidence to the case of multiple images. The region confidence of Eq.(4) is a function of the region index a with its best correspondence as a latent variable i. Given multiple images, it can be naturally extended with more latent valuables, meaning correspondences from multiple other images to the region. Let us define *neighbor images* $\mathcal{N}_q$ of image I_q as a set of other images where an object in I_q appears. Generalizing Eq.(4), the region confidence for image I_q is defined as



(a) Region confidences with respect to its neighbor images.

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

Figure 3. Part-aware foreground localization. (a) Given multiple neighbor images with common objects, region confidences can be computed according to Eq.(5). More positive 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 regions). In the same way, the standout score for the white box takes into account blue, red, and green boxes altogether. Three boxes on the right are ones with the top three standout scores from the region confidence in (a). The red one has the top score. (Best viewed in color.)

$$\begin{aligned} \mathcal{F}_q(a) &= \max_{\{i_p | p \in \mathcal{N}_q\}} \sum_{p \in \mathcal{N}_q} \mathcal{M}_{pq}(i_p, a) \\ &= \sum_{p \in \mathcal{N}_q} \max_{i_p} \mathcal{M}_{pq}(i_p, a), \end{aligned} \tag{5}$$

where $\mathcal{M}_{pq}$ represents the match confidence from I_p to I_q, and i_p denotes the index of a region in I_p. It reduces to the aggregated confidence from the neighbor images. More neighbor images could make better confidences.

Given regions R with these region confidences, we localize an object foreground by selecting the region that most "stands out". The idea is illustrated in Fig. 3b, which derives from the fact that a tight object region (red box) has less background clutter than any other larger region containing it (green boxes), while a part region (white box) has no less background than larger regions within a tight object region (blue boxes). Imagine a region gradually shrinking from a whole image region, to a tight object region, to a part region. Significant increase in maximum region confidence is most likely to occur at the point of taking the tight object region. Based on this insight, we define the *standout score* as to measure how more salient the region is than the most salient region containing it:

$$\begin{aligned} &\mathcal{S}(a) = \mathcal{F}(a) - \max_{b \in L(a)} \mathcal{F}(b), \\ &s.t. \;\; L(a) = \{b \mid r_a \subsetneq r_b, r_b \in R\}, \end{aligned} \tag{6}$$

where $r_a \subsetneq r_b$ means region r_a is contained in region r_b. In practice, we decide $r_a \subsetneq r_b$ by two simple criteria: (1) The box area of r_a is less than 50% of the box area of r_b. (2) 80% of the box area of r_a 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 part-aware foreground localization in a coordinate descent-style algorithm. Given a collection of images $\mathcal{I}$, 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_q for each image I_q, and performs the following three steps at each iteration.

Neighbor image retrieval For each image I_q, k nearest neighbor images $\{I_p | p \in N_q\}$ are retrieved based on the similarity between O_q and O_p. At the first iteration, as the potential object regions become entire image regions, nearest-neighbor matching with the GIST descriptor [41] 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. We use 10 neighbor images for each image ($k = 10$). In our experiments, the use of more neighbor images does not always improve the performance while increasing computation time.

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

Part-aware foreground localization For each image I_q, standout score $\mathcal{S}_q$ is computed so that the set of potential object regions O_q is updated to that of regions with top standout scores. This re-localization advances both neighbor image retrieval and part-based 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 object localization is done by selecting the most standing-out region at the end. Basically, the algorithm is designed based on the idea that better object localization makes better retrieval and matching, and vice versa. As each image is independently processed at each iteration, the proposed 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. For more details of the algorithm, see our supplementary material.

4. Experimental evaluation

The degree of supervision used in visual learning tasks varies from strong (supervised localization [13, 15]) to weak (weakly-supervised localization [8, 38]), very weak (colocalization [21, 40] and cosegmentation [33]), 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 experiments has never been attempted before on challenging real-world datasets. We conduct experiments on two realistic benchmarks, the Object Discovery [33] and the PASCAL VOC 2007 [12], and compare the results with those of the current state of the arts in cosegmentation [23, 19, 20, 33], colocalization [7, 10, 35, 21, 40], and weakly-supervised localization [8, 10, 28, 29, 38, 45].

4.1. Evaluation metrics

The correct localization (CorLoc) metric is an evaluation metric widely used in related work [10, 21, 38, 40], 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 de-

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

Methods	Airplane	Car	Horse	Average
Kim *et al.* [23]	21.95	0.00	16.13	12.69
Joulin *et al.* [19]	32.93	66.29	54.84	51.35
Joulin *et al.* [20]	57.32	64.04	52.69	58.02
Rubinstein *et al.* [33]	74.39	87.64	63.44	75.16
Tang *et al.* [40]	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

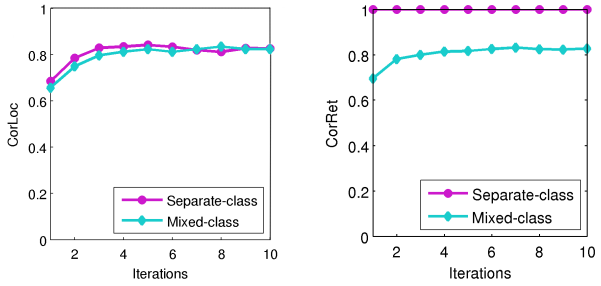


Figure 4. Average CorLoc (left) and CorRet (right) vs. # of iterations on the Object Discovery dataset.

fined as follows. Given the k nearest neighbors identified by our retrieval procedure 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 a 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 [33] 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 [33] 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 [40], we convert the ground-truth segmentations and cosegmentation results of [23, 19, 20, 33] to localization boxes.

We conduct separate-class experiments as in [10, 40], and a mixed-class experiment on a collection of 300 images from all the three classes. The mixed-class image collection contains 3 classes and 36 outlier images. Figure 4 shows the average CorLoc and CorRet over iterations, where we see

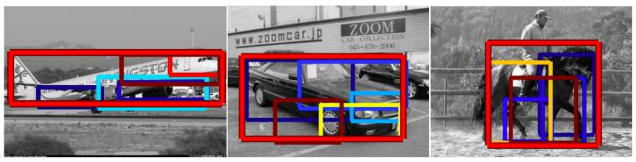


Figure 5. Examples of localization on mixed-class Object Discovery dataset. Small boxes inside the localized object box (red box) represents five most confident part regions. (Best viewed in color.)

Table 5. CorLoc comparison on PASCAL07-6x2.

Method	Average CorLoc (%)
Russell *et al.* [35]	22
Chum and Zisserman [7]	33
Deselaers *et al.* [10]	37
Tang *et al.* [40]	39
Ours	**68**
Ours (mixed-class)	54

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 [23, 19, 20] and colocalization [33, 40] methods. The proposed method outperforms all the other methods in this setup. The mixed-class results are in Table 2, and examples of the localization result 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 object instances are visualized in red boxes with five most confident regions inside the object, indicating parts most contributing to object discovery. Table 2 and Fig. 4 show that our localization is robust to noisy neighbor images retrieved from different classes.

4.3. PASCAL VOC 2007 dataset

The PASCAL VOC 2007 [12] 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 previous methods, we conduct experiments on two subsets of different sizes: PASCAL07-6x2 and PASCAL07-all. The PASCAL07-6x2 subset [10] 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 [8, 10, 29], we take all train+val dataset images discarding images that only

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

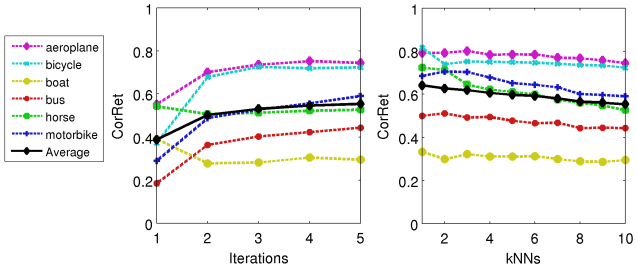


Figure 6. CorRet variation on mixed-class PASCAL07-6x2.

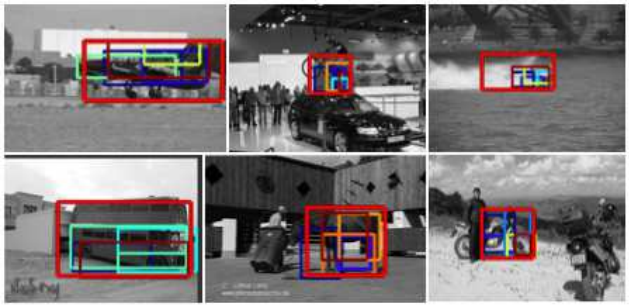


Figure 7. Example results on mixed-class PASCAL07-6x2. (Best viewed in color.)

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 maintaining 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 part-aware localization (using the standout score) are crucial for robust object discovery. In Table 5, we quantitatively compare our method to previous colocalization methods [7, 10, 35, 40] on PASCAL07-6x2. Our method significantly outperforms the state of the art [40] with a large margin for all classes. Note that our method does not incorporate any form of object priors such as off-the-shelf objectness measures [10, 40]. 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 performance 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 relatively small size of objects even partially occluded, our method is able to localize instances from considerably 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.

Larger-scale experiments on PASCAL07-all In the separate-class setup, we compare our results to those of the state of the arts in weakly-supervised localization [8, 29, 36, 39, 37, 38, 45] and colocalization [21] in Table 6. Note that beside positive images (P) for a target class, weakly-supervised methods use more training data, *i.e.*, negative images (N). Also note that the best performing method [45] uses CNN features pretrained on the ImageNet dataset [9], thus additional training data (A). Surprisingly, the performance of our method is very close to the best of weakly-supervised localization [8] not using such additional data. For the mixed-class experiment on a collection of all images in PASCAL07-all, we handle an evaluation issue as follows. 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

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 [29]	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 [39]	P + A	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.* [37]	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.* [36]	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.* [8]	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.* [45]	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.* [21]	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

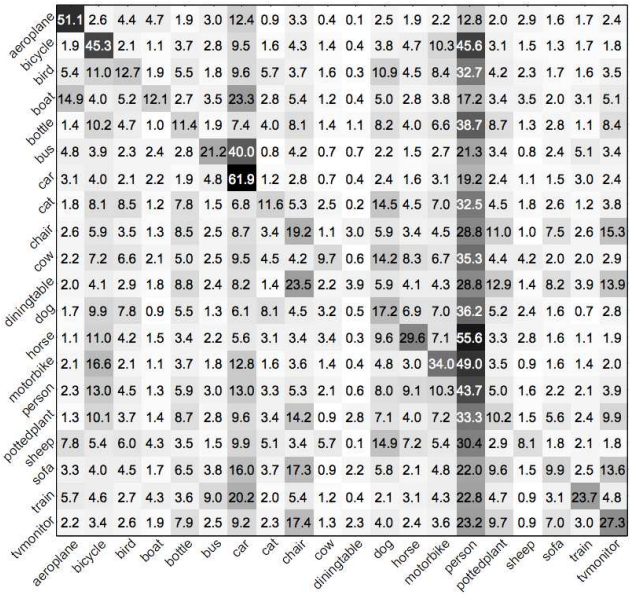


Figure 8. Confusion matrix of retrieval on mixed PASCAL07-all.

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 object regions at the final iteration. (Best viewed in color.)

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. The quantified results in Table 7 show that 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. Interestingly, the significant difference in retrieval performance (CorRet) from 100% in the separate-class setup influences much less on localization (CorLoc). 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 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.

5. Discussion and conclusion

The proposed part-based approach to object discovery markedly outperforms the state of the art in colocalization and closely compete with weakly-supervised localization. In particular, we demonstrate unsupervised object localization in the mixed-class setup, which has never been at-

tempted before on the challenging real-world dataset such as [12]. In future, we will advance this direction and further explore handling multiple object instances per image as well as building visual models for classification and detection. In this paper, we intend to evaluate the unsupervised algorithm per se, and thus abstain 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] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010. 2

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

[5] M. Cho, Y. M. Shin, and K. M. Lee. Co-recognition of image pairs by data-driven monte carlo image exploration. In *ECCV*, pages IV: 144–157, 2008. 2

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

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

[8] R. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL Training for Weakly Supervised Object Localization. In *CVPR*, 2014. 1, 2, 5, 6, 7, 8

[9] 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

[10] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 2012. 1, 2, 5, 6, 7

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

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 1, 5, 6, 9

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

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

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

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

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

[18] E. T. Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003. 3

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

[20] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 1, 5, 6

[21] 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

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

[23] G. Kim and E. Xing. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 1, 2, 5, 6

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

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

[26] C. Liu, J. Yuen, and A. Torralba. Sift flow: dense correspondence across scenes and its applications. *TPAMI*, 2011. 2

[27] S. Manen, M. Guillaumin, and L. V. Gool. Prime object proposals with randomized prim's algorithm. In *ICCV*, 2013. 1, 2, 3

[28] 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

[29] 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

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

[31] 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

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

[33] M. Rubinstein and A. Joulin. Unsupervised Joint Object Discovery and Segmentation in Internet Images. In *CVPR*, 2013. 1, 2, 5, 6

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

[35] 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, 6, 7

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

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

[38] 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

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

[40] K. Tang, A. Joulin, and L.-j. Li. Co-localization in Real-World Images. In *CVPR*, 2014. 1, 2, 5, 6, 7

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

[42] 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

[43] S. Vicente. Object cosegmentation. In *CVPR*, 2011. 1, 2

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

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

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