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.

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 [2, 3], 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 without any box-level annotations [1, 4, 6]. 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 (Fig. 1), without even image-level annotations or any assumption of a single dominant class. This is significantly more general than typical colocalization [4], cosegmentation [6], or weakly-supervised localization tasks [1].

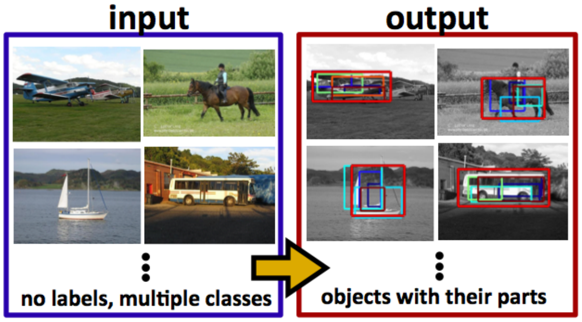

Figure 1: Unsupervised object discovery in the wild. We tackle object localization in an unsupervised scenario without any type of annotations, where a given image collection may contain multiple dominant object classes and even outlier images. The proposed method discovers object instances (red bounding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

Main idea. We tackle the unsupervised discovery and localization problem using a part-based region matching approach: We use off-the-shelf region proposals [5] to form a set of candidate bounding boxes for *objects* and *object parts* (Fig. 2a-b). These regions are efficiently matched across images using a probabilistic Hough transform that evaluates the confidence for each candidate correspondence considering both appearance similarity and spatial consistency (Fig. 2c-d). Dominant objects are discovered and localized by comparing the scores of candidate regions and selecting those that stand out over other regions containing them (Fig. 2d-e).

Algorithm overview. For efficient and robust object discovery, we combine part-based region matching and foreground localization in a coordinate descent-style algorithm. Given a collection of images, 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 region matching only from those neighbor images. The algorithm starts with an entire image region as an initial set of potential object regions for each image, and performs the three steps at each iteration: *neighbor image retrieval*, *part-based region matching*, and *foreground localization*. As each image is independently processed at each iteration, the algorithm is easily parallelizable in computation.

Experimental evaluation. Extensive evaluations on standard benchmarks demonstrate that the proposed approach significantly outperforms the current state of the art in colocalization, and achieves robust object discovery even in a fully unsupervised setting. Table 1 compares our result to those of the state of the arts in weakly-supervised localization [1, 7] and colocalization [4] on the PASCAL VOC 2007 dataset [3]. 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 [7] uses CNN features pretrained on the ImageNet dataset [2], thus additional supervised data (A). Surprisingly, the performance of our method is very close to the best of weakly-supervised localization [1] not using such additional data. Moreover, our method successfully discovers objects even in a fully unsupervised setting (bottom), where we mix all images of all classes into a dataset and evaluate performance on the whole dataset.

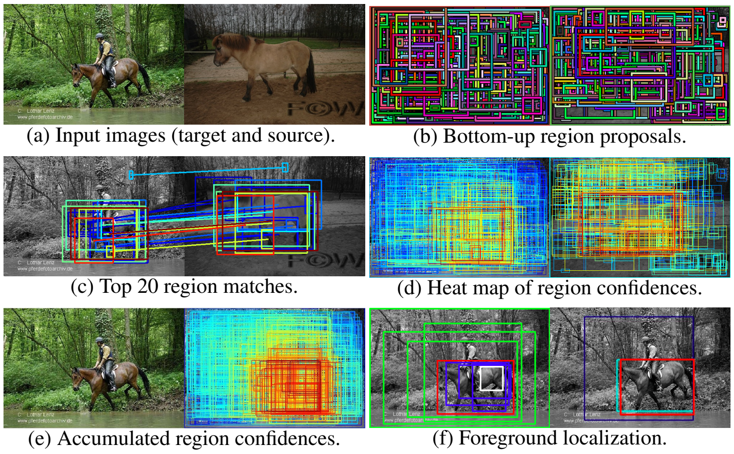
(a) Input images (target and source). (b) Bottom-up region proposals.
(c) Top 20 region matches. (d) Heat map of region confidences.
(e) Accumulated region confidences. (f) Foreground localization.

Figure 2: Region matching and foreground localization. (a-b) Given images and their region proposals [5], the proposed matching method efficiently evaluates candidate matches between two sets of regions and produce match confidences. (c) The top 20 matches are shown based on the match confidence. The confidence is color-coded in each match (red: high, blue: low). (d) The region confidences from matching are visualized in the heat map. Common object foregrounds tend to have higher confidences than others. (e) Using multiple source images with common objects, region confidences are aggregated as more source images may give better region confidences. (f) Given regions (boxes) on the left, the standout score for the red box corresponds to the difference between its confidence and the maximum confidence of boxes containing the red box (green boxes). Three boxes on the right are ones with the top three standout scores. (Best viewed in color.)

Table 1: Object localization on PASCAL VOC 2007.

Method	Data used	Avg. CorLoc (%)
Cinbis et al. [1]	P + N	38.8
Wang et al. [7]	P + N + A	48.5
Joulin et al. [4]	P	24.6
Ours	P	**36.6**
Ours (mixed-class)	unsupervised	**31.3**

http://www.di.ens.fr/willow/research/objectdiscovery/

[1] Ramazan G. Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, 2014.

[2] Jia Deng, Wei Dong, Richard. Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] Mark. Everingham, Luc. Van Gool, Christopher. K. I. Williams, John. Winn, and Andrew. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results.

[4] Armand Joulin, Kevin Tang, and Li Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014.

[5] Santiago Manen, Matthieu Guillaumin, and Luc Van Gool. Prime object proposals with randomized Prim's algorithm. In *ICCV*, 2013.

[6] Michael Rubinstein and Armand Joulin. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[7] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage. * 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.