Co-localization with Category-Consistent Features and Geodesic Distance Propagation

Hieu Le *[1], Chen-Ping Yu[3], Gregory Zelinsky[1,2], and Dimitris Samaras[1]

[1]Department of Computer Science, Stony Brook University
[2]Department of Psychology, Stony Brook University
[3]Department of Psychology, Harvard University

Abstract

Co-localization is the problem of localizing objects of the same class using only the set of images that contain them. This is a challenging task because the object detector must be built without negative examples that can lead to more informative supervision signals. The main idea of our method is to cluster the feature space of a generically pre-trained CNN, to find a set of CNN features that are consistently and highly activated for an object category, which we call category-consistent CNN features. Then, we propagate their combined activation map using superpixel geodesic distances for co-localization. In our first set of experiments, we show that the proposed method achieves state-of-the-art performance on three related benchmarks: PASCAL 2007, PASCAL-2012, and the Object Discovery dataset. We also show that our method is able to detect and localize truly unseen categories, on six held-out ImageNet categories with accuracy that is significantly higher than previous state-of-the-art. Our intuitive approach achieves this success without any region proposals or object detectors, and can be based on a CNN that was pre-trained purely on image classification tasks without further fine-tuning.

1. Introduction

In recent years, deep learning methods have dominated state-of-the-art results in object detection and localization tasks [21, 25, 31, 29, 6, 34, 4, 26, 36]. However, one major drawback for deep learning methods is that they require a large amount of labeled training data, and data labels are expensive as they depend on extensive human efforts [32, 33]. Therefore, there is great value and importance if deep learning methods can learn to detect and localize objects accu-

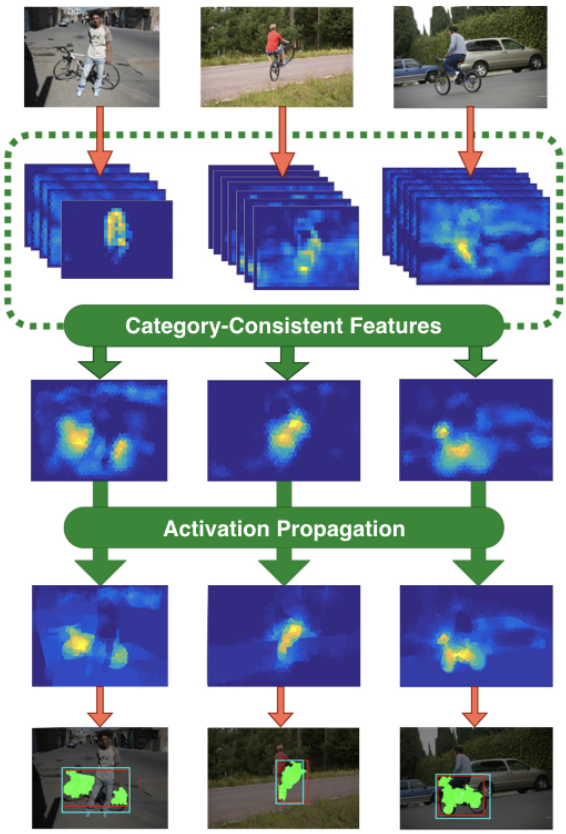

Figure 1. **Object co-localization with CCFs and geodesic distance propagation**. From a set of images containing a common object, first we find the CCFs - the group of features that consistently have high responses to the object images of the same class. The CCFs then are used to form an activation map for each image, followed by geodesic distance propagation to highlight the exact regions of the objects.

*hle@cs.stonybrook.edu

rately with unlabeled data. Object co-localization is one such problem that is close to the goal of learning with unlabeled data. In the problem of object co-localization, one must learn to detect and localize the common object from a set of same-class images, without any other image information. Successful methods for the co-localization problem should be able to localize an object through image search using a single class-name keyword, without the need of any pixel-level labels (e.g. bounding box and segmentation) or negative examples. Such ability can be useful for automatically generating large-scale datasets.

Recent co-localization methods typically utilize existing region proposal methods for generating a number of candidate regions for objects and object parts [15, 32, 5, 20]. From the bag of candidate proposals, the set of object-related proposals are determined based on different criteria, such as those having high mutual similarity scores [20, 32] or those minimizing the entropy of the score distribution of a classifier [5]. Using object proposals significantly increases processing time and hinders the scalability of the algorithms as they depend on the quality of the object proposals. In fact, region and object proposals are part of a research problem of its own, and have drawbacks such as lack of repeatability, reduced detection performance with a large number of proposals, and difficulties in balancing precision and recall [11].

Our method, however, does not require any object proposals to perform co-localization, but only utilizes features from a CNN that has been pre-trained on classification tasks. The main idea in our work is that objects of the same class share common features or parts. Moreover, these commonalities are central to both category representation and detection and localization of the object. By finding those object categorical features, their joint locations can act as a single-shot object detector. This idea is also grounded in human visual learning, where it is suggested that people detect common features from examples of the category, as part of the object-learning process [37]. We do this by obtaining the CNN features of the provided set of positive images, in order to select the features that are highly and consistently activated, which we denote as Category-Consistent CNN Features (CCFs). We then use these CCFs to discover the rough object locations, and demonstrate an effective way to propagate the feature activations into a stable object for precise co-localization, using the output of the boundary detection algorithm [24]. Figure 1 illustrates the pipeline of our proposed framework.

In more detail, our approach begins with a CNN that has been pre-trained for image classification on ImageNet. Then, the images of the target category are passed through the network. From the set of the CNN kernels, we group them based on their maximum activations across the whole set of images. In fact, since these images all contain the categorical object, all CCFs tend to have high and similar activations, which encourage them to be grouped into the same cluster. Thus, we simply identify the group of kernels with the highest average activation score as CCFs and compute a single normalized activation probability map that associates to this CCF set. Then, we employ an over-segmentation into superpixels of the input image to propagate the values in the activation probability map to the entire image, weighted by the similarities between superpixels. Such similarities are computed via a boundary detection algorithm and are presented as geodesic distances between superpixels. Finally, the precise object location can be obtained by placing a tight bounding box around the thresholded object-likelihood map.

We test our method on three popular datasets for the co-localization problem. To show that our method is able to generalize to truly unseen categories, we test our method on six held-out Imagenet categories that were unseen, and not part of the pre-trained CNN's training categories. Our experimental results show that our method achieves state-of-the-art co-localization scores even on these "unseen" categories.

We make three contributions in this work:

1. We propose a novel CCF extraction method that can automatically select a set of representative features for a category using only positive images. To the best of our knowledge, our method is the *first* successful framework for extracting representative features inside a CNN framework.

2. We show that the set of CCFs highlights the rough initial object regions, and can act as a single-shot detector. This result is further refined with an effective feature propagation method using superpixel geodesic distances, that results in a distinctive object region using superpixel on the original image.

3. Our method achieves state-of-the-art performance for object co-localization on the VOC 2007 and 2012 datasets [8], the Object Discovery dataset [27], and the six held-out ImageNet subset categories, demonstrating that we are able to accurately localize objects obviating the need for region proposals.

2. Related work

Co-localization is related to work on weakly supervised object localization (WSOL) [31, 29, 6, 34, 4, 26, 36, 38] since both share the same objective: to localize objects from an image. However, since WSOL allows the use of negative examples, designing the objective function to discover the information of the object-of-interest is less challenging: WSOL-based methods achieve higher performance on the same datasets as compared to co-localization methods, due

to the allowed supervised training. For instance, Wang et al. [34] uses image labels to evaluate the discrimination of discovered categories in order to localize the objects. Ren et al. [26] adopts a discriminative multiple instance learning scheme to compensate for the lack of object-level annotations to localize the objects based on the most discriminative instances. Two recent works using CNN that share the objective with our work are GAP [38] and GMP [22], which both try to obtain the activation map from a CNN representation. However, both GAP & GMP learn to find *discriminative* features for object locations using additional supervised learning, while the features discovered through our CCF framework are *representative* in nature (representative may or may not be discriminative, and are similar in concept to generative models). Our method obviates the need for negative examples, and does not require additional training as in GAP and GMP. Because of the supervision that is required by those methods, it is not trivial for WSOL approaches to be directly applied to the co-localization scenarios.

One challenge of co-localization is to define the criteria for discovering the objects without any negative examples. To fill the gap, state-of-the-art co-localization methods such as [20, 32, 5, 15] employ object proposals as part of their object discovery and co-localization pipelines. Tang et al. [32] use the measure of objectness [2] to generate multiple bounding boxes for each image, followed by an objective function to simultaneously optimize the image-level labels and box-level labels. Such settings allow the use of a discriminative cost function [13]. This is also used in the work of co-localization on video frames [15]. Cho et al. [5]'s method also starts from object proposals, sharing the same spirit with deformable part models [9] where objects are discovered and localized by matching common object parts. Most recently, Li et al. [20] study the confidence score distribution of a supervised object detector over the set of object proposals to define an objective function, that learns a common object detector with similar confidence score distribution. All the aforementioned methods heavily depend on the quality of object proposals.

Our work approaches the problem from a different perspective. Instead of trying to fill in the gap of the negative data and annotations that are unavailable, we find the common features shared by the objects from the positive images. Then, we use the joint locations of those features as our single-shot object detector. Our subsequent step refines the detected object features into a stable object by propagating their activations together. We describe the details of our 2-step approach in the following sections.

3. Extracting Category-Consistent CNN Features

Our proposed method consists of two main steps. The first step is to find the CCFs of a category, and obtain their combined feature map that contains aggregated CCF activations over the rough object region. Then, the CCF activations are propagated into a stable object using superpixel geodesic distances on the original images.

Our CCFs extraction step is indeed a feature selection method in which we select a set of representative features for a category from the variable pool. The variable pool can be the set of CNN kernels in any layer, whether a convolutional layer or a fully connected layer. Given a set of $\boldsymbol{n}$ object images from the same class and a pre-trained CNN, we first compute the CNN representations for all these n images. Assuming that there are $\boldsymbol{m}$ kernels from the layer we want to extract the CCFs, we compute $\boldsymbol{m}$ activation vectors to represent the activation behavior of these kernels, i.e., to which degree each kernel is *activated* for each image. Specifically, for each kernel $\mathbf{i}$, its activation vector $\boldsymbol{A_i}$ is defined as: $\boldsymbol{A_i} = [a_{i,1}, a_{i,2}, \ldots, a_{i,n}]^T$ where $a_{i,j} = \max(F(i,j))$, and $F(i,j)$ is the CNN representation of kernel i for image j.

Our goal in this step is to identify a subset of representative kernels for a category from the set of global candidate kernels, that contain common features from the positive images of the same class. Since there is at least one object of the category on every image, the activation vectors of the representative kernels should have relatively high values over all vector elements. Furthermore, the representative kernels should have similar activation behavior, given that their values associate to the same object instances. Thus, we aim to find a set of kernels that have similar activation behavior and high average activation value. To find the CCFs, we compute the pair-wise L_p similarities between all pairs of kernels' activation vectors, and cluster them using k-means. The kernels from the cluster with the highest mean activation correspond to the CCFs.

The CCF kernels can then be used to generate the rough object location in an image in a single-shot: given an image from the target category, the feature maps that associates to the CCFs are combined to form a single activation map. Conceptually, the kernels that we seek correspond to object parts, or some object-associated features. Thus the densely activated area of the activation map indicates the rough location of the target object. The final activation map is normalized into a probability map whose values are in the $[0, 1]$ range. In Figure 2, we show the identified CCFs for the bus category, where the activated regions describe bus-related features and all fall within the spatial extent of the objects.

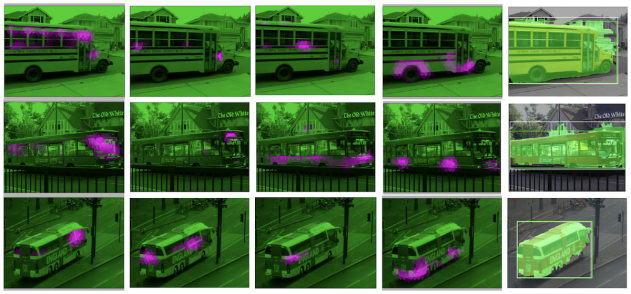

Figure 2. **Examples of our CCFs for *bus* category.** Each row is a different example image, and each column is the activation (in violet) of a single CNN feature in the set of our CCFs. The last column shows the final co-localization results.

4. Stable Object Completion via Propagating CCF Activations

The activation probability map from the CCFs automatically points out only the rough location of the object. It does not ensure a reliable object localization because:

1. The higher layer of a CNN does not guarantee a kernel's receptive-field size to cover the area of an entire object.

2. While the feature maps contain spatial information, they have imprecise object locations due to previous max-pooling layers.

3. The CNN was trained discriminatively. Hence, only the discriminative features of each object may be localized rather than the the whole object.

The geodesic distance has been shown to be effective for many segmentation problems [17, 3, 35, 19]. Given the locations of some representative parts of the object, we utilize the superpixels of the input image and their geodesic distances computed using the image boundary map. The boundary map and the superpixels of the input image can then be used to effectively propagate the initial activation probability values into a complete object region. The geodesic distance between two superpixels is the shortest path between two superpixels where the edge weight is the likelihood of an object boundary between two adjacent superpixels, computed from the boundary probability map. In essence, the geodesic distance compactly encodes the similarity relationship between the two superpixels' image contents. Therefore, the smaller the geodesic distance between the two superpixels, the more likely that they belong to the same object. Based on this characteristic, we propose a simple and effective method to highlight the object region from the activation probability map, that is both low-resolution and contains non-smooth feature activations.

Next, we proceed to localize the object by combining superpixels that belong to the same object using their geodesic distances, which we call geodesic distance propagation. The main idea is that if two superpixels belong to the same object, they should have 1. a small geodesic distance, and 2. similar activation values. These concepts have been similarly adopted by various works in terms of interactive image segmentation and matting [3, 10], and are especially useful in producing well differentiated object regions in this case. Specifically, we seek to obtain a resulting global activation map that has clear and highly boosted target object regions, by propagating activations based on superpixel geodesic distances.

Given an input image, we oversegment it into superpixels using [23]. We take the combined activation map that was obtained in the CCF identification step, and assign an energy value to each superpixel by averaging its corresponding pixel values found in the activation map. For k superpixels, we denote the resulting flattened $k \times 1$ superpixel activation vector as $\mathbf{E}$. Vector $\mathbf{E}$ can be considered as the initial likelihood of each superpixel being within the object. Then, We compute a $k \times k$ propagation matrix $\mathbf{W}$, such that $W_{i,j}$ is the normalized amount of propagation between superpixel i and j, with their geodesic distance denoted as $d_{i,j}$, and a parameter μ for controlling the amount of activation diffusion:

$$W_{i,j} = \frac{\exp(-d_{i,j} \times \mu^{-1})}{\sum_{k=1}^{N} \exp(-d_{i,k} \times \mu^{-1})}. \quad (1)$$

The propagation matrix W can then be applied to the activation vector E directly:

$$E' = \mathbf{WE}, \quad (2)$$

where E' is the propagated activation vector of the image, containing the globally boosted activations of the superpixels based on their pair-wise geodesic distances to all other superpixels. This allows us to fill in each superpixel on the image with their respective values from E', and normalize the propagated superpixel map by dividing every pixel by the max value of the map. The result is an object-likelihood map, on which we apply a global threshold to obtain the region as our final object co-localization result. Finally, a tight bounding box is placed around the maximum coverage of the thresholded regions within an image.

Our propagation step can be implemented in $O(n \log(n) + n \times e)$ using Johnson's algorithm [12] to compute all pair shortest paths (where n is the number of superpixels, and e is the number of edges between superpixels on an image). Thus, our method can potentially be applied on real-time applications. The effect of parameter μ, which controls the amount of the activation diffusion, is discussed in more detail in the supplementary material.

5. Experiments

We evaluate our proposed 2-step framework with different parameter settings to illustrate different characteristics of our method. We also evaluate our method on multiple benchmarks, with intermediate and final results to show the localization effects of our proposed method. In all of our experiments, we used two different CNN models: AlexNet[18] and VGG-19[30], both pre-trained on ImageNet [28]. We use the *fc6* of AlexNet and the last convolutional layer of a VGG-19 network [30] as variable pool. We used $k = 5$ for k-means clustering in the CCF identification step. The final global threshold for obtaining the object region from the object-likelihood map was set at 0.25 for all images. It is worth mentioning that the recent state-of-the-art co-localization method of Li et al. [20] also uses the *fc6* feature of AlexNet and employs the output of *EdgeBox*[39] in their experiments, which is a direct application of [24]. Our experiments using AlexNet *fc6* features therefore are directly comparable to theirs. There is, unfortunately, no publicly available implementation of their method using VGG19. We elaborate on some technical implementation details in our supplementary material, including the method to obtain the normalized activation map from *fc6* kernels of AlexNet.

5.1. Evaluation metric and datasets

We use the conventional CorLoc metric [7] to evaluate our co-localization results. The metric measures the percentage of images that contain correctly localized results. An image is considered correctly localized if there is at least one ground truth bounding box of the object-of-interest having more than a 50% Intersection-over-Union (IoU) score with the predicted bounding box. To benchmark our method performance, we evaluate our method on three commonly used datasets for the problem of co-localization. These are the VOC 2007 and 2012 [8], and the Object Discovery dataset [27]. We also follow [20] to test our method on the six held-out ImageNet subsets. For experiments on the VOC datasets, we followed previous work [5, 15, 20] that used all images on the *trainval* set excluding the images that only contain the object instances annotated as *difficult* or *truncated*. For experiments on the Object Discovery dataset, we used the 100-image subset following [27] in order to make an appropriate comparison with related methods. The ground truth bounding box for each image in the Object Discovery dataset is defined as the smallest bounding box covering all the segmentation ground truth of the object.

5.2. Comparison to state-of-the-art co-localization methods

We first evaluate our method on the 100-image subset of Object Discovery dataset which contains objects of three classes, namely *Airplane*, *Car*, and *Horse*. There are 18, 11, and 7 noisy images in each class, respectively. Table 3 reports the co-localization performance of our approach in comparison with the state-of-the-art methods on image co-localization [13, 27, 14, 15, 5, 20]. In this small scale setting, our method outperformed other methods in both in individual object classes and overall.

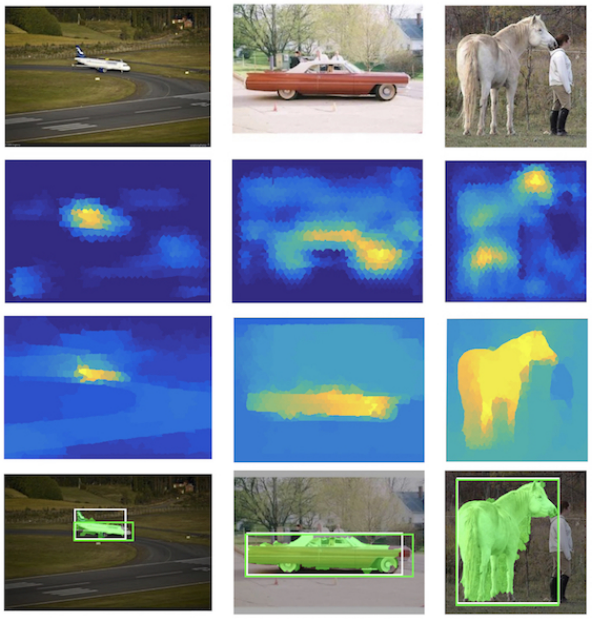

Figure 3. **Object co-localization results on the Object Discovery dataset.** From the top row to the bottom row: input image, combined activation map from the identified CCFs, propagated object-likelihood map, and resulting bounding boxes. We also depict in green the pixels with the object-region that is predicted by our method. Our predicted bounding boxes are colored as white while the ground truth bounding boxes are green.

Three examples of our co-localization approach using VGG19 on the Object Discovery dataset are illustrated in figure 3. The second row shows the combined activation map from the set of identified CCFs, that acted as our single-shot object detector. It is apparent that the combined activation maps already provided object estimates that were quite accurate to the location of the actual object in the images, with different parts of each object getting high values such as the tail of the airplane, the wheel of the car, or head and tail of the horse. All these values were propagated based on the superpixel geodesic distances, resulting in the images shown in the third row. The propagated object-likelihood maps on the third row show that the sporadic activations on the background have been smoothed out evenly, and that the non-smooth object parts and their associated activations have been boosted and completed into complete and stable objects with significantly higher activa-

Method	aero	bicy	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
Joulin et al. [15]	32.80	17.30	20.90	18.20	4.50	26.90	32.70	41.00	5.80	29.10	**34.50**	31.60	26.10	40.40	17.90	11.80	25.00	27.50	35.60	12.10	24.60
Cho et al. [5]	50.30	42.80	30.00	18.50	4.00	62.30	**64.50**	42.50	**8.60**	**49.00**	12.20	44.00	64.10	57.20	15.30	9.40	30.90	**34.00**	61.60	**31.50**	36.60
Li et al. [20] - AlexNet	**73.10**	45.00	43.40	27.70	6.80	53.30	58.30	45.00	6.20	48.00	14.30	47.30	**69.40**	66.80	**24.30**	**12.80**	**51.50**	25.50	65.20	16.80	40.00
Ours - AlexNet	69.60	51.67	43.80	**30.05**	5.11	55.74	60.00	58.50	6.20	**49.00**	16.30	**51.26**	58.74	67.38	22.60	11.60	47.06	27.40	58.93	16.20	40.36
Ours - VGG19	71.90	**61.67**	**48.20**	27.66	**11.90**	**63.90**	59.30	**71.50**	5.70	37.00	12.20	44.80	66.50	**71.70**	18.48	11.11	36.76	29.25	**66.96**	22.84	**41.97**

Table 1. **CorLoc scores of our approach and state-of-the-art co-localization methods on Pascal VOC 2007 dataset.**

Method	aero	bicy	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mean
Cho et al. [5]	57.00	41.20	36.00	26.90	5.00	81.10	54.60	50.90	**18.20**	54.00	31.20	44.90	61.80	48.00	13.00	11.70	51.40	**45.30**	64.60	**39.20**	41.80
Li et al. [20] - AlexNet	65.70	57.80	47.90	28.90	6.00	74.90	48.40	48.40	14.60	54.40	23.90	50.20	**69.90**	68.40	**24.00**	**14.20**	**52.70**	30.90	**72.40**	21.60	43.80
Ours - AlexNet	72.58	67.40	48.73	31.22	8.22	74.05	52.42	64.75	16.04	52.09	26.09	**58.64**	67.40	71.97	23.99	13.07	37.55	34.25	64.20	15.64	45.02
Ours - VGG19	**75.94**	**76.20**	**55.70**	**37.56**	**22.24**	**83.97**	**55.41**	**67.63**	14.08	**54.88**	**31.88**	53.65	69.28	**76.26**	14.04	10.23	42.86	33.70	69.96	18.94	**48.22**

Table 2. **CorLoc scores of our approach and state-of-the-art co-localization methods on Pascal VOC 2012 dataset.**

Methods	Airplane	Car	Horse	Mean
Kim et al. [16]	21.95	0.00	16.13	12.69
Joulin et al. [13]	32.93	66.29	54.84	51.35
Joulin et al. [14]	57.32	64.04	52.69	58.02
Rubinstein et al. [27]	74.39	87.64	63.44	75.16
Joulin et al. [15]	71.95	93.26	64.52	76.58
Cho et al. [5]	82.93	**94.38**	75.27	84.19
Ours - AlexNet	81.71	**94.38**	77.42	84.50
Ours - VGG19	**84.15**	**94.38**	**79.57**	**86.03**

Table 3. **Experiment on the Object Discovery Dataset.** Highest performance scores are labeled in bold.

tion magnitudes. This shows that our two-step framework was able to generate informative single-shot object detections using the CCFs, and the subsequent stable object region via activation propagation.

The PASCAL VOC 2007 and 2012 datasets both contain realistic images of 20 object classes with significantly larger numbers of images per class. These datasets are more challenging than the Object Discovery dataset due to the diversity of viewpoints, and the complexity of the objects. Table 1 reports our performance on the VOC 2007 dataset. Our approach using VGG19 outperforms the state-of-the-art method [20] by 2% on average. Our method works particularly well on *bicycle* and *cat* with 16% and 26% higher Corloc score respectively. With AlexNet, our method still shows a mild improvement from the state-of-the-art method.

Results on VOC 2012 dataset are shown in table 2, showing a more pronounced improvement over previous methods. Notice that VOC 2012 has twice the number of images than the VOC 2007 dataset. Our method using AlexNet and VGG19 achieves significantly better results than state-of-the-art methods with 1.2% and 4.42% increase on average.

Our VGG-19 and AlexNet model was pre-trained on ImageNet's 1000 classes. While VOC 2007 and 2012 are different datasets from ImageNet, there are significant overlaps between the object categories in VOC and ImageNet. For example, the "motorbike" class of VOC datasets is equivalent to the "moped" class of ILSVRC 2012 dataset. The six subsets of the ImageNet dataset, chosen by Li et al. [20], are held-out categories from the 1000-label classification task, which means they do not overlap with the 1000 classes used to train VGG-19. The images and the corresponding bounding box annotations were downloaded from ImageNet website [1]. We show that our method is generalizable to truly novel object categories with the six held-out ImageNet subset classes. Table 4 reports the co-localization of our method compares to [20] on the six ImageNet subset. The results show that our method significantly outperforms the competing methods. This result demonstrates that our method can robustly detect and localize truly unseen categories using previously learned CNN features. Our approach outperforms the state-of-the-art method by 1.3% using a pre-trained AlexNet, and 12.8% using a pre-trained VGG19 CNN model. This experiment shows that our method can perform co-localization task accurately on truly unseen classes, implying the strong generalization ability of the CNN features. For example, our results that used the pre-trained VGG19 show noticeable improvements over the state-of-the-art method with a 19% CorLoc score increase in the *chipmunk* class and 50% CorLoc score increase in the *racoon* class.

Based on our experimental results on all three datasets, our method using AlexNet performs comparably to the state-of-the-art method, and shows mild improvements on the VOC2007 and six held-out ImageNet subsets and an average of 1.2% CorLoc score increase on the VOC2012 dataset. It is worth mentioning that the performance of the previous state-of-the-art method [20] significantly relies on the object proposal algorithm, the EdgeBox method [39]. Unfortunately there is no available implementation of [20] with different base CNN models other than AlexNet, to measure how it scales with the better CNN models. Our method scales up well, and it appears to show bigger performance gains as the models and datasets get more complex.

ImageNet	chipmunk	rhino	stoat	racoon	rake	wheelchair	mean
Cho et al. [5]	26.60	81.80	44.20	30.10	8.30	35.30	37.72
Li et al. [20] - AlexNet	44.00	81.80	**67.30**	41.80	14.50	39.30	48.12
Ours - AlexNet	44.94	86.36	56.73	66.02	10.34	32.37	49.46
Ours - VGG19	**63.29**	**89.77**	56.73	**91.26**	**20.69**	**43.93**	**60.95**

Table 4. CorLoc scores [7] (%) of our approach and state-of-the-art co-localization methods on the 6 held-out subsets of ImageNet collected by Li et al. [20].

5.3. Category-consistent CNN features selection analysis

In this section, we provide an analysis to justify our CCF selection method, where we conduct additional experiments with the same configurations but using different subsets of CNN features for the initial object detection step. We use VGG19 for the analysis in this section. After clustering the last-layer CNN kernels based on their image-level activations, these clusters were sorted in a decreasing order by the clusters' average activations. We then obtained the rough object locations using individual clusters and thresholded on those maps directly to obtain the object locations. Their respective average CorLoc scores on the VOC 2007 and 2012 dataset are reported in table 5. The table shows that co-localization performance correlates strongly with the level of average activations of the clusters, suggesting that the most representative features were indeed members of the top cluster. We visualize some examples in figure 4, with an image from the *bus* and *motorbike* category, respectively. For each image, the first row is the results of our method when using the first cluster (ranked by their average activation) and the second row shows the results of our method when using the third cluster. It is clear that the combined activation maps from the third cluster failed to detect and estimate the object locations, and ultimately lead to incorrect object localization results. This indicates that the selection of the top cluster is essential, and the CCFs could not be chosen arbitrarily.

We furthermore validate our feature selection method by evaluating the performance of our approach when the CCFs are not selected from the first cluster, but from the top k clusters instead, where k varies from 1 to 5. As shown in table 6, the results indicate that a larger number of kernels does not provide enough object specificity as the performance decreases when more clusters are added.

Dataset	1st	2nd	3rd	4th	5th
VOC07	41.97	39.55	32.15	25.44	21.40
VOC12	48.22	41.47	38.71	32.43	24.80

Table 5. **Co-localization performance of our method on the VOC 2007 and 2012 dataset.** Each column indicates which top cluster was taken as the CCF cluster (out of 5 total clusters), and the corresponding average CorLoc score (%) by using the selected cluster of features for co-localization.

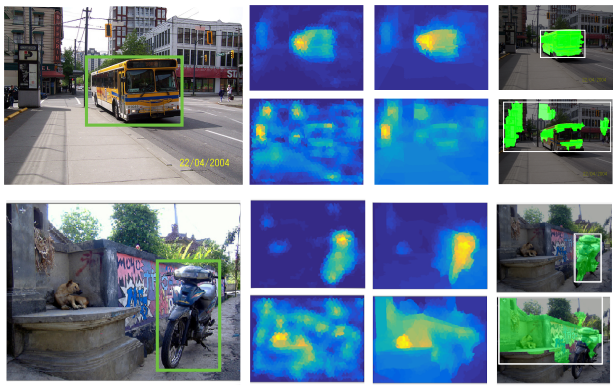

Figure 4. **Two examples illustrating the effect of our feature selection method.** For each image, the first row is the results of our method when using the first cluster and the second row is the results of our method when using the third cluster. The green bounding box is the ground truth and the predicted bounding boxes are colored as white, the predicted object regions are masked as green.

Dataset	1	1-2	1-3	1-4	1-5
VOC07	41.97	41.10	39.30	32.25	31.20
VOC12	48.22	46.68	42.25	40.15	38.45

Table 6. **Co-localization performance of our methods on the VOC 2007 and 2012 dataset.** Each column indicates how many top clusters were taken as the CCF clusters (out of 5 total clusters), and the corresponding average corLoc score (%) by using the selected clusters of features for co-localization. For example, the last column indicates that all available features were used for co-localization.

5.4. Geodesic distance propagation analysis

The geodesic distance acts as a refinement step in our pipeline to remove the background as well as to boost the activation within the object region. To evaluate the effect of geodesic distance propagation, we simply compared the performance of our method on VOC 2007 and 2012 datasets with and without geodesic distance propagation, and report the results in table 7. The results show that geodesic distance propagation significantly improved the

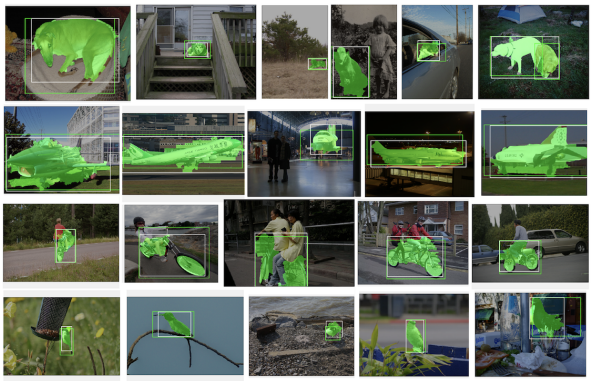

Figure 5. **Some example results of object co-localization with our CCFs and geodesic distance propagation.** The results show the bounding boxes that we generate match the ground truth bounding boxes very well, even when objects are not located centrally in the image. In addition, our co-localized object pixel-level regions (pixels colored in green) show well delineated shape in most cases.

co-localization accuracy by more than 6% in absolute CorLoc scores for both datasets, which means that it is an important step after the initial CCF object detection step.

Dataset	Without GDP	With GDP
VOC07	35.41	41.97
VOC12	41.20	48.22

Table 7. **Co-localization performance of our methods on VOC 2007 and 2012 dataset with and without using geodesic distance propagation.** Using geodesic distance propagation increased the average CorLoc score (%) of our approach by 6.56% for the VOC 2007 dataset, and 7.02% for the VOC 2012 dataset.

5.5. Qualitative results

We show some examples of our co-localization results in figure 5. The results show that the bounding boxes generated by our proposed framework accurately match the ground truth bounding boxes. It is apparent that our results generate well-covered object regions, that has the potential to well delineate the objects in majority of the cases. The figures also show that the objects were able to be accurately co-localized with various sizes and locations.

Figure 6 illustrates three failure scenarios of our approach. While these three examples did not cover the ground truth bounding box sufficiently, but they were not far off. Some analysis suggests that these failures were due to some CCFs that are shared by multiple categories, and that the object boundaries may not have been strong enough (i.e. bottle and boat).

Figure 6. **Some failed examples of our approach.** While these examples did not have significant coverage with the ground truth boudning box in terms of CorLoc, the mistakes were still accurate, in the sense that the areas were detected but not the spatial extent.

6. Conclusion

In this work, we proposed a 2-step approach for the problem of co-localization, that uses only positive images, without any region or object proposals. Our method is motivated by human vision; people implicitly detect the common features of category examples to learn the representation of the class. We show that the identified category-consistent features can also act as an effective first-pass object detector. This idea is implemented by finding the group of CNN features that are highly and consistently activated by a given positive set of images. The result of this first step generates a rough but reliable object location, and acts as a single-shot object detector. Then, we aggregate the activations of the identified CCFs, and propagate their activations so that the activations over the true object region are boosted, while the activations over the background region are smoothed out. This effective activation refinement step allowed us to obtain accurately co-localized objects in terms of the standard CorLoc score with bounding boxes. We achieved new state-of-the-art performance on the three commonly used benchmarks. In the future, we plan to extend our method to generate object co-segmentations.

Acknowledgement. This work was partially supported by the Vietnam Education Foundation, the Stony Brook University SensorCAT, a gift from Adobe, the Partner University Fund 4DVision project, and New York State ITSC.

References

[1] Imagenet. `http://image-net.org/`. Accessed: 2016-11-15.

[2] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2189–2202, November 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.28. URL `http://dx.doi.org/10.1109/TPAMI.2012.28`.

[3] X. Bai and G. Sapiro. Geodesic matting: a framework for fast interactive image and video segmentation and matting. *IJCV*, 2008.

[4] H. Bilen, M. Pedersoli, and T. Tuytelaars. Weakly supervised object detection with convex clustering. In *CVPR*, 2015.

[5] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: part-based matching with bottom-up region proposals. In *CVPR*, 2015.

[6] Ramazan Gokberk Cinbis, Jakob Verbeek, and Cordelia Schmid. Multi-fold MIL Training for Weakly Supervised Object Localization. In *CVPR 2014 - IEEE Conference on Computer Vision & Pattern Recognition*, Columbus, United States, June 2014. IEEE. URL https://hal.inria.fr/hal-00975746.

[7] Thomas Deselaers, Bogdan Alexe, and Vittorio Ferrari. Weakly supervised localization and learning with generic knowledge. *International Journal of Computer Vision*, 100 (3):275–293, 2012. ISSN 1573-1405. doi: 10.1007/s11263-012-0538-3. URL http://dx.doi.org/10.1007/s11263-012-0538-3.

[8] M. Everingham, S. Eslami, L. Gool, C. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *IJCV*, 2015.

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[10] L. Grady. Random walks for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006.

[11] J. Hosang, R. Benenson, P. Dollar, and B. Schiele. What makes for effective detection proposals. *arXiv*, 2015.

[12] Donald B. Johnson. Efficient algorithms for shortest paths in sparse networks. *J. ACM*, 24(1):1–13, January 1977. ISSN 0004-5411. doi: 10.1145/321992.321993. URL http://doi.acm.org/10.1145/321992.321993.

[13] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[14] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[15] A. Joulin, K. Tang, and F.-F. Li. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV*, 2014.

[16] Gunhee Kim, E. P. Xing, Li Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *2011 International Conference on Computer Vision*, pages 169–176, Nov 2011. doi: 10.1109/ICCV.2011.6126239.

[17] Philipp Krähenbühl and Vladlen Koltun. *Geodesic Object Proposals*, pages 725–739. Springer International Publishing, Cham, 2014. ISBN 978-3-319-10602-1. doi: 10.1007/978-3-319-10602-1_47. URL http://dx.doi.org/10.1007/978-3-319-10602-1_47.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[19] Hieu Le, Vu Nguyen, Chen-Ping Yu, and Dimitris Samaras. Geodesic distance histogram feature for video segmentation. In *Computer Vision – ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part I*, pages 275–290, Cham, 2017. Springer International Publishing. ISBN 978-3-319-54181-5. doi: 10.1007/978-3-319-54181-5_18. URL http://dx.doi.org/10.1007/978-3-319-54181-5_18.

[20] Y. Li, L. Liu, C. Shen, and A. van den Hengel. Image co-localization by mimicking a good detector's confidence score distribution. In *ECCV*, 2016.

[21] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. Ssd: Single shot multibox detector. 2016. URL http://arxiv.org/abs/1512.02325. To appear.

[22] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Is object localization for free? weakly-supervised learning with convolutional neural networks. In *CVPR*, 2015.

[23] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740. IEEE Computer Society, 2012. ISBN 978-1-4673-1226-4. URL http://ieeexplore.ieee.org/xpl/mostRecentIssue.jsp?punumber=6235193.

[24] Larry Zitnick Piotr Dollar. Structured forests for fast edge detection. In *ICCV*. International Conference on Computer Vision, December 2013. URL https://www.microsoft.com/en-us/research/publication/structured-forests-for-fast-edge-detection/.

[25] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[26] Weiqiang Ren, Kaiqi Huang, Dacheng Tao, and Tieniu Tan. Weakly supervised large scale object localization with multiple instance learning and bag splitting. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(2):405–416, February 2016. ISSN 0162-8828. doi: 10.1109/TPAMI.2015.2456908. URL http://dx.doi.org/10.1109/TPAMI.2015.2456908.

[27] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[29] Zhiyuan Shi, Timothy M. Hospedales, and Tao Xiang. Bayesian joint topic modelling for weakly supervised object localisation. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2013.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[31] Parthipan Siva and Tao Xiang. Weakly supervised object detector learning with model drift detection. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 343–350, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-1101-5. doi: 10.1109/ICCV.2011.6126261. URL http://dx.doi.org/10.1109/ICCV.2011.6126261.

[32] K. Tang, A. Joulin, L.-J. Li, and F.-F. Li. Co-localization in real-world images. In *CVPR*, 2014.

[33] Tinne Tuytelaars, Christoph H. Lampert, Matthew B. Blaschko, and Wray Buntine. Unsupervised object discovery: A comparison. *International Journal of Computer Vision*, 88(2):284–302, 2010. ISSN 1573-1405. doi: 10.1007/s11263-009-0271-8. URL http://dx.doi.org/10.1007/s11263-009-0271-8.

[34] Chong Wang, Weiqiang Ren, Kaiqi Huang, and Tieniu Tan. Weakly supervised object localization with latent category learning. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, volume 8694 of *Lecture Notes in Computer Science*, pages 431–445. Springer, 2014. ISBN 978-3-319-10598-7. doi: 10.1007/978-3-319-10599-4_28. URL http://dx.doi.org/10.1007/978-3-319-10599-4_28.

[35] Wenguan Wang, Jianbing Shen, and F. Porikli. Saliency-aware geodesic video object segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3395–3402, June 2015. doi: 10.1109/CVPR.2015.7298961.

[36] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. Relaxed multiple-instance svm with application to object discovery. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1224–1232, Washington, DC, USA, 2015. IEEE Computer Society. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.145. URL http://dx.doi.org/10.1109/ICCV.2015.145.

[37] C.-P. Yu, J. Maxfield, and G. Zelinsky. Searching for category-consistent features: a computational approach to understanding visual category representation. *Psychological Science*, 2016.

[38] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[39] Larry Zitnick and Piotr Dollar. Edge boxes: Locating object proposals from edges. In *ECCV*. European Conference on Computer Vision, September 2014. URL https://www.microsoft.com/en-us/research/publication/edge-boxes-locating-object-proposals-from-edges/.