Co-localization in Real-World Images

Kevin Tang[1] Armand Joulin[1] Li-Jia Li[2] Li Fei-Fei[1]
[1]Computer Science Department, Stanford University [2]Yahoo! Research
{kdtang,ajoulin,feifeili}@cs.stanford.edu lijiali@yahoo-inc.com

Abstract

In this paper, we tackle the problem of co-localization in real-world images. Co-localization is the problem of simultaneously localizing (with bounding boxes) objects of the same class across a set of distinct images. Although similar problems such as co-segmentation and weakly supervised localization have been previously studied, we focus on being able to perform co-localization in real-world settings, which are typically characterized by large amounts of intra-class variation, inter-class diversity, and annotation noise. To address these issues, we present a joint image-box formulation for solving the co-localization problem, and show how it can be relaxed to a convex quadratic program which can be efficiently solved. We perform an extensive evaluation of our method compared to previous state-of-the-art approaches on the challenging PASCAL VOC 2007 and Object Discovery datasets. In addition, we also present a large-scale study of co-localization on ImageNet, involving ground-truth annotations for 3,624 classes and approximately 1 million images.

1. Introduction

Object detection and localization has long been a cornerstone problem in computer vision. Given the variability of objects and clutter in images, this is a highly challenging problem. Most state-of-the-art methods require extensive guidance in training, using large numbers of images with human-annotated bounding boxes [11, 28]. Recent works have begun to explore weakly-supervised frameworks [9, 14, 21, 22, 27, 29], where labels are only given at the image level. Inspired by these works, we focus on the problem of unsupervised object detection through co-localization, which further relaxes the need for annotations by only requiring a set of images that each contain *some* common object we would like to localize.

We tackle co-localization in real-world settings where the objects display a large degree of variability, and worse, the labels at the image level can be noisy (see Figure 1). Although recent works have tried to explicitly deal with annotation noise [24, 30, 31], most previous works related to co-localization have assumed clean labels, which is not a realistic assumption in many real-world settings where we have to analyze large numbers of Internet images or discover objects with roaming robots. Our aim is therefore to overcome the challenges posed by noisy images and object variability.

Figure 1. The co-localization problem in real-world images. In this instance, the goal is to localize the airplane within each image. Because these images were collected from the Internet, some images do not actually contain an airplane.

We propose a formulation for co-localization that combines an image model and a box model into a joint optimization problem. Our image model addresses the problem of annotation noise by identifying incorrectly annotated images in the set, while our box model addresses the problem of object variability by localizing the common object in each image using rich correspondence information. The joint image-box formulation allows the image model to benefit from localized box information, and the box model to benefit by avoiding incorrectly annotated images.

To illustrate the effectiveness of our method, we present results on three challenging, real-world datasets that are representative of the difficulties of intra-class variation, inter-class diversity, and annotation noise present in real-world images. We outperform previous state-of-the-art approaches on standard datasets, and also show how the joint

image-box model is better at detecting incorrectly annotated images. Finally, we present a large-scale study of co-localization on ImageNet [8], involving ground-truth annotations for 3,624 classes and 939,542 images. The largest previous study of co-segmentation on ImageNet consisted of ground-truth annotations for 446 classes and 4,460 images [18].

2. Related Work

Co-localization shares the same type of input as co-segmentation [15–18,24,33], where we must find a common object within a set of images. However, instead of segmentations, we seek to localize objects with bounding boxes. Considering boxes allows us to greatly decrease the number of variables in our problem, as we label boxes instead of pixels. It also allows us to extract rich features from within the boxes to compare across images, which has shown to be very helpful for detection [32].

Co-localization shares the same type of output as weakly supervised localization [9, 21, 22, 27], where we draw bounding boxes around objects without any strong supervision. The key difference is that in co-localization we have a more relaxed scenario, where we do not know what the object contained in our set of images is, and are not given negative images for which we know do not contain our object. Most similar is [9], which generates candidate bounding boxes and tries to select the correct box within each image using a conditional random field. Object co-detection [3] also shares similarities, but is given additional bounding box and correspondence annotations.

Although co-localization shares similarities with both co-segmentation and weakly supervised localization, an important and new difficulty we address in this paper is the problem of noisy annotations, which has recently been considered [24, 30, 31]. Most similar is [24], where the authors utilize dense correspondences to ignore incorrect images. We combine an image model that detects incorrectly annotated images with a box model that localizes the common object, which sets us apart from previous work. The objective functions in our models are inspired by works from outlier detection [13], image segmentation [26], and discriminative clustering [2, 15, 34]. Previous works have considered combining object detection with image classification [11, 28], but only in supervised scenarios.

3. Our Approach

Given a set of n images $\mathcal{I} = \{I_1, I_2, \ldots, I_n\}$, our goal is to localize the common object in each image. In addition, we also consider the fact that due to noise in the process of collecting this set, some images may not contain the common object. We denote these as *noisy* images, as opposed to *clean* images, which contain the common object. Our goal is to simultaneously identify the noisy images and localize the common object in the clean images.

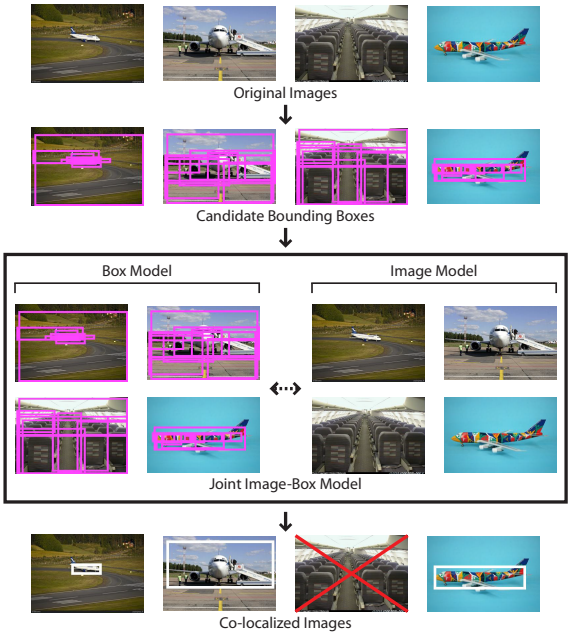

Figure 2. Given a set of images, we start by generating a set of candidate boxes independently for each image. Then, our joint image-box model is able to simultaneously identify the noisy images and select the box from each clean image that contains the common object, resulting in a set of co-localized images. Previous work considers only the box model [9].

An overview of our approach is given in Figure 2. We start by generating a set of candidate boxes for each image that could potentially contain an object. Then, we formulate an image model for selecting the clean images, and a box model for selecting the box in each image that contains an instance of the common object. We denote the boxes that contain an instance of the common object as *positive* boxes, and the ones that don't as *negative* boxes.

Combining the two models into a joint formulation, we allow the image model to prevent the box model from being adversely affected by boxes in noisy images, and allow the box model to help the image model determine noisy images based on localized information in the images. Similar approaches have been considered [9], but only using a box model and only in the context of clean images.

3.1. Generating candidate boxes

We use the measure of objectness [1], but any method that is able to generate a set of candidate regions can be used [5, 32]. The objectness measure works by combining multiple image cues such as multi-scale saliency, color contrast, edge density, and superpixel straddling to generate a set of candidate regions as well as scores associated with each region that denote the probability a generic object is

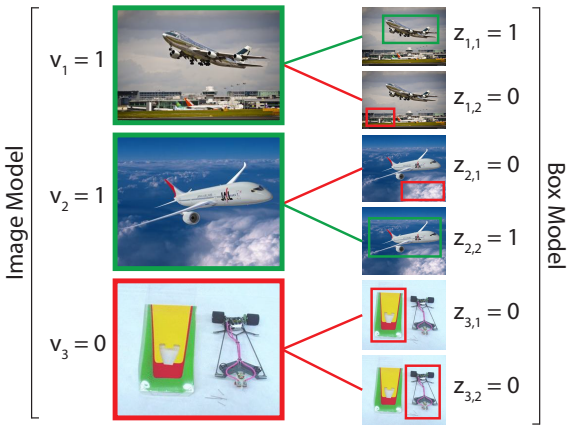

Figure 3. The variables v in the image model relate to the variables z in the box model through constraints that ensure noisy images (red) do not select any boxes, while clean images (green) select a single box as the postive box.

present in the region. Examples of candidate boxes generated by objectness can be seen in Figure 2.

Using the objectness measure, for each image $I_j \in \mathcal{I}$, we generate a set of m candidate boxes $\mathcal{B}_j = \{b_{j,1}, b_{j,2}, \ldots, b_{j,m}\}$, ordered by their objectness score.

3.2. Model setup

Given a set of images $\mathcal{I}$ and a set of boxes $\mathcal{B}_j$ for each image $I_j \in \mathcal{I}$, our goal is to jointly determine the noisy images and select the positive box from each clean image. To simplify notation, we define the set of all boxes as $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2 \ldots \cup \mathcal{B}_n$ and $n_b = nm$ the total number of boxes.

Feature representation. For each box $b_k \in \mathcal{B}$, we compute a feature representation of the box as $x_k^{box} \in \mathbb{R}^d$, and stack the feature vectors to form a feature matrix $X_{box} \in \mathbb{R}^{n_b \times d}$. Similarly for each image $I_j \in \mathcal{I}$, we compute a feature representation of the image as $x_j^{im} \in \mathbb{R}^d$, and stack the feature vectors to form a feature matrix $X_{im} \in \mathbb{R}^{n \times d}$. We densely extract SIFT features [20] every 4 pixels and vector quantize each descriptor into a 1,000 word codebook. For each box, we pool the SIFT features within the box using 1×1 and 3×3 SPM pooling regions [19], and for each image, we use the same pooling regions over the entire image to generate a $d = 10,000$ dimensional feature descriptor for each box and each image.

Optimization variables. We associate with each image $I_j \in \mathcal{I}$ a binary label variable v_j, which is equal to 1 if I_j is a clean image and 0 otherwise. Similarly, we associate with each box $b_{j,k} \in \mathcal{B}_j$ a binary label variable $z_{j,k}$, which is equal to 1 if $b_{j,k}$ is a positive box and 0 otherwise. We denote by v, the n dimensional vector $v = (v_1, \ldots, v_n)^T$ and by z the n_b dimensional vector obtained by stacking the $z_{j,k}$. Making the assumption that in each clean image there is only one positive box, and in each noisy image there are no positive boxes, we define a constraint that relates the two sets of variables:

$$\forall I_j \in \mathcal{I}, \sum_{k=1}^{m} z_{j,k} = v_j. \tag{1}$$

This constraint is also illustrated in Figure 3, where we show the relationship between image and box variables.

3.3. Model formulation

We begin by introducing and motivating the terms in our objective function that enable us to jointly identify noisy images and select the positive box from each clean image.

Box prior. We introduce a prior for each box that represents our belief that the box is positive. We compute an off-the-shelf saliency map for each image [6, 23], and for each box we compute the average saliency within the box, weighted by the size of the box, and stack these values into the n_b dimensional vector m_{box} to obtain a linear term that penalizes less salient boxes:

$$f_{Pbox}(z) = -z^T \log(m_{box}). \tag{2}$$

Although objectness also provides scores for each box, we found that the saliency measure used in objectness is dated and does not work as well.

Image prior. We introduce a prior for each image that represents our belief that the image is a clean image. For each image, we compute the χ^2 distance, defined further below, from the image feature to the average image feature in the set, and stack these values into the n dimensional vector m_{im} to obtain a linear term that penalizes outlier images:

$$f_{Pim}(v) = v^T m_{im}. \tag{3}$$

We experimented with several measures for outlier detection [13], but found that this simple distance worked well.

Box similarity. We encourage boxes with similar appearances to have the same label through a similarity matrix based on the box feature described above. Since this feature is a histogram, we compute a $n_b \times n_b$ similarity matrix S based on the χ^2-distance:

$$S_{ij} = \exp\left(-\gamma \sum_{k=1}^{d} \frac{(x_{ik}^{box} - x_{jk}^{box})^2}{x_{ik}^{box} + x_{jk}^{box}}\right), \tag{4}$$

where $\gamma = (10d)^{-\frac{1}{2}}$. We set the similarity of boxes from the same image to be 0. We then compute the normalized Laplacian matrix $L_{box} = I - D^{-\frac{1}{2}} S D^{-\frac{1}{2}}$, where D is the diagonal matrix composed of the row sums of S, resulting

in a quadratic term that encourages the selection of similar boxes:

$$f_{Sbox}(z) = z^T L_{box} z. \quad (5)$$

This choice is motivated by the work of Shi and Malik [26], who have shown that considering the second smallest eigenvector of a normalized Laplacian matrix leads to clustering z along the graph defined by the similarity matrix, leading to Normalized Cuts when used for image segmentation. Furthermore, Belkin and Niyogi [4] have shown that minimizing Equation 5 under linear constraints results in an equivalent problem. The similarity term can be interpreted as a a generative term that seeks to select boxes that cluster well together.

Image similarity. We also encourage images with similar appearances to have the same label through a similarity matrix based on the image feature described above. Replacing the box features with image features in Equation 4, we compute a $n \times n$ similarity matrix and subsequently the normalized Laplacian matrix L_{im} to obtain a quadratic term that encourages the selection of similar images:

$$f_{Sim}(v) = v^T L_{im} v. \quad (6)$$

Box discriminability. Discriminative learning techniques such as the support vector machine and ridge regression have been widely used within the computer vision community to obtain state-of-the-art performance on many supervised problems. We can take advantage of these methods even in our unsupervised scenario, where we do not know the labels of our boxes [2, 34]. Following [15], we consider the ridge regression objective function for our boxes:

$$\min_{\substack{w \in \mathbb{R}^d, \\ c \in \mathbb{R}}} \frac{1}{n_b} \sum_{j=1}^{n} \sum_{k=1}^{m} ||z_{j,k} - w x_{j,k}^{box} - c||_2^2 + \frac{\kappa}{d} ||w||_2^2, \quad (7)$$

where w is the d dimensional weight vector of the classifier, and c is the bias. The choice of ridge regression over other discriminative cost functions is motivated by the fact that the ridge regression problem has a closed form solution for the weights w and bias c, leading to a quadratic function in the box labels [2]:

$$f_{Dbox}(z) = z^T A_{box} z, \quad (8)$$

where $A_{box} = \frac{1}{n_b}(\Pi_{n_b}(I_{n_b} - X_{box}(X_{box}^T \Pi_{n_b} X_{box} + n_b \kappa I)^{-1} X_{box}^T)\Pi_{n_b})$ and $\Pi_{n_b} = I_{n_b} - \frac{1}{n_b} 1_{n_b} 1_{n_b}^T$ is the centering projection matrix. We know also that A_{box} is a positive semi-definite matrix [12]. This quadratic term allows us to utilize a discriminative objective function to penalize the selection of boxes whose features are not easily linearly separable from the other boxes.

Image discriminability. Similar to the box discriminability term, we also employ a discriminative objective to ensure that the features of the clean images should be easily linearly separable from noisy images. Replacing the box features in Equation 7 with image features, we can similarly substitute the solutions for w and c to obtain:

$$f_{Dim}(v) = v^T A_{im} v, \quad (9)$$

where A_{im} is defined in the same way as A_{box}, replacing box features with image features.

Joint formulation. Combining the terms presented above, we obtain the following optimization problem:

$$\begin{aligned} \underset{z,v}{\text{minimize}} \quad & z^T(L_{box} + \mu A_{box})z - z^T \lambda \log(m_{box}) \\ & + \alpha(v^T(L_{im} + \mu A_{im})v + v^T \lambda m_{im}) \\ \text{subject to} \quad & v \in \{0,1\}, z \in \{0,1\} \\ & \forall I_j \in \mathcal{I}, \sum_{k=1}^{m} z_{j,k} = v_j \\ & K_0 \leq \sum_{i=1}^{n} v_i, \end{aligned} \quad (10)$$

where the constraints in the formulation ensure that only a single box is selected in clean images, and none in noisy images. Using the constant K_0, we can avoid trivial solutions and incorporate an estimate of noise by allowing noisy images to not contain boxes. This prevents the boxes in the noisy images from adversely affecting the box similarity and discriminability terms.

The parameter μ controls the tradeoff between the quadratic terms, the parameter λ controls the tradeoff between the linear and quadratic terms, and the parameter α controls the tradeoff between the image and box models. Since the matrices L_{box}, A_{box}, L_{im}, and A_{im} are each positive semi-definite, the objective function is convex.

Convex relaxation. In Equation 10, we obtain a standard boolean constrained quadratic program. The only sources of non-convexity in this problem are the boolean constraints on v and z. We relax the boolean constraints to continuous, linear constraints, allowing v and z to take any value between 0 and 1. This becomes a convex optimization problem and can be solved efficiently using standard methods.

Given the solution to the quadratic program, we reconstruct the solution to the original boolean constrained problem by thresholding the values of v to obtain the noisy images, and simply taking the box from each clean image with the highest value of z.

4. Results

We perform experiments on three challenging datasets, the PASCAL VOC 2007 dataset [10], the Object Dis-

Method	aeroplane		bicycle		boat		bus		horse		motorbike		
	left	right	left	right	left	right	left	right	left	right	left	right	Average
Our Method (prior)	13.95	20.51	10.42	8.00	2.27	6.98	9.52	13.04	12.50	13.04	17.95	23.53	12.64
Our Method (prior+similarity)	39.53	35.90	**25.00**	**24.00**	0.00	2.33	23.81	34.78	37.50	43.48	48.72	58.82	31.16
Our Method (full)	**41.86**	**51.28**	**25.00**	**24.00**	**11.36**	**11.63**	**38.10**	**56.52**	**43.75**	**52.17**	**51.28**	**64.71**	**39.31**

Table 1. CorLoc results for various combinations of terms in our box model on PASCAL07-6x2.

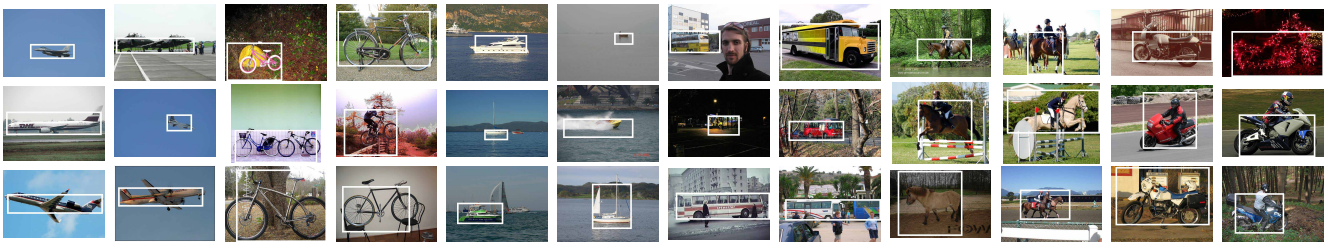

Figure 4. Example co-localization results on PASCAL07-6x2. Each column contains images from the same class/viewpoint combination.

Method	Average CorLoc
Russell *et al.* [25]	22
Chum and Zisserman [7]	33
Deselaers *et al.* [9]	37
Our Method	**39**

Table 2. CorLoc results compared to previous methods on PASCAL07-6x2.

Method	Airplane	Car	Horse	Average CorLoc
Kim *et al.* [17]	21.95	0	16.13	12.69
Joulin *et al.* [15]	32.93	66.29	54.84	51.35
Joulin *et al.* [16]	57.32	64.04	52.69	58.02
Rubinstein *et al.* [24]	**74.39**	87.64	63.44	75.16
Our Method	71.95	**93.26**	**64.52**	**76.58**

Table 3. CorLoc results on the 100 image subset of the Object Discovery dataset.

covery dataset [24], and ImageNet [8]. Following previous works in weakly supervised localization [9], we use the CorLoc evaluation metric, 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. All CorLoc results are given in percentages.

4.1. Implementation details and runtime

We set the parameters of our method to be $\mu = 0.6$, $\lambda = 0.001$, and $\alpha = 1$, and tweaked them slightly for each dataset. We set $\kappa = 0.01$ in the ridge regression objective. Because there are no noisy images for PASCAL and ImageNet, we fix the value of $K_0 = n$ for these datasets. For the Object Discovery dataset, we set $K_0 = 0.8n$. We use 10 objectness boxes for ImageNet, and 20 objectness boxes for the other datasets.

After computing candidate object boxes using objectness and densely extracting SIFT features, we are able to co-localize a set of 100 images with 10 boxes per image in less than 1 minute on a single machine using code written in Python and a quadratic program solver written in C++.

4.2. PASCAL VOC 2007

Following the experimental setup defined in [9], we evaluate our method on the PASCAL07-6x2 subset to compare to previous methods for co-localization. This subset consists of all images from 6 classes (aeroplane, bicycle, boat, bus, horse, and motorbike) of the PASCAL VOC 2007 [10] 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.

In Table 1, we analyze each component of our box model by removing various terms in the objective. As expected, we see that results using stripped down versions of our model do not perform as well. In Table 2, we show how our full method outperforms previous methods for co-localization that do not utilize negative images. In addition, our method does not incorporate dataset-specific aspect ratio priors for selecting boxes. In Figure 4, we show example visualizations of our co-localization method for PASCAL07-6x2. In the bus images, our model is able to co-localize instances in the background, even when other objects are more salient. In the bicycle and motorbike images, we see how our model is able to co-localize instances over a variety of natural and man-made background scenes.

4.3. Object Discovery dataset

The Object Discovery dataset [24] was collected by automatically downloading images using the Bing API using queries for airplane, car, and horse, resulting in noisy images that may not contain the query. Introduced as a dataset for co-segmentation, we convert the ground-truth segmentations and results from previous methods to localization boxes by drawing tight bounding boxes around the segmentations. We use the 100 image subset [24] to enable comparisons to previous state-of-the-art co-segmentation meth-

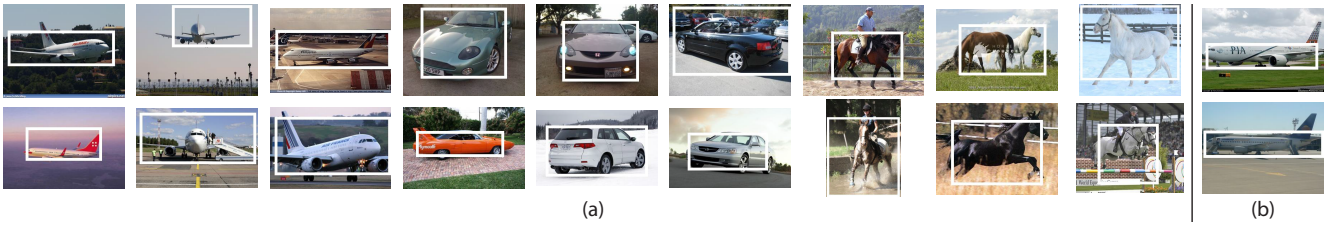

Figure 5. (a) Example co-localization results on the Object Discovery dataset, with every three columns belonging to the same class; (b) Images from the airplane class that were incorrectly localized.

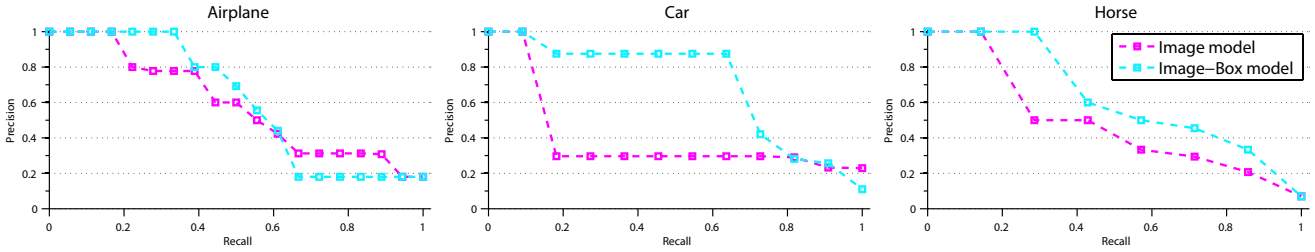

Figure 6. Precision-recall curves illustrating the effectiveness of our image-box model (blue) compared to the image model (pink) at identifying noisy images on the Object Discovery dataset. The joint optimization problem allows the box model to help correct errors made by the image model.

ods. CorLoc results are given in Table 3, and example co-localization results are visualized in Figure 5(a). From the visualizations, we see how our model is able to handle intra-class variation, being able to co-localize instances of each object class from a wide range of viewpoints, locations, and background scenes. This is in part due to our quadratic terms, which consider the relationships between all pairs of images and boxes, whereas previous methods like [24] rely on sparse image connectivity for computational efficiency.

We see that our method outperforms previous methods in all cases except for the airplane class. In Figure 5(b), we see that since our method localizes objects based on boxes instead of segmentations [24], the airplane tail is sometimes excluded from the box, as including the tail would also include large areas of the background. This causes our method to fail in these images due to the non-convex shape of the airplane and the height of the tail.

Detecting noisy images. We also quantitatively measure the ability of our joint image-box model to identify noisy images. Because the solution to the quadratic program gives continuous values for the image variables v, we can interpret the values as a detection score for each image and plot precision-recall curves that measure our ability to correctly detect noisy images, as shown in Figure 6. To make comparisons fair, we compare using the best parameters for the image model alone, and the best parameters for our joint image-box model. By jointly optimizing over both image and box models, we see how the box model can correct errors made by the image model by forcing images that have good box similarity and discriminability to be clean, even if the image model believes them to be noisy.

Method	Average CorLoc
Top objectness box [1]	37.42
Our Method	**53.20**

Table 4. CorLoc results on ImageNet evaluated using ground-truth annotations for 3,624 classes and 939,542 images.

4.4. ImageNet

ImageNet [8] is a large-scale ontology of images organized according to the WordNet hierarchy. Each node of the hierarchy is depicted by hundreds and thousands of images. We perform a large-scale evaluation of our co-localization method on ImageNet by co-localizing all images with ground-truth bounding box annotations, resulting in a total of 3,624 classes and 939,542 images. A similar large-scale segmentation experiment [18] only considered ground-truth annotations in 446 classes and 4,460 images. At this scale, the visual variability of images is unprecedented in comparison to previous datasets, causing methods specifically tuned to certain datasets to work poorly.

Due to the scale of ImageNet and lack of code available for previous methods, we compare our method to the highest scoring objectness box [1], which gives a strong baseline for generic object detection. To ensure fair comparisons, we use the objectness score as the box prior for our model in these experiments, with CorLoc results shown in Table 4 and visualizations for 104 diverse classes in Figure 7.

Box selection. In Figure 8(a), we show the distribution over objectness boxes that our method selects. The boxes are ordered by decreasing objectness score, so objectness simply selects the first box in every image. By considering box similarity and discriminability between images, our

Figure 7. Example co-localization results on ImageNet. Each image belongs to a different class, resulting in a total of 104 classes ranging from lady bug to metronome. White boxes are localizations from our method, green boxes are ground-truth localizations.

method identifies boxes that may not have very high objectness score, but are more likely to be the common object.

Effect of ImageNet node height. We also evaluate the performance of our method on different node heights in ImageNet in Figure 8(b). Here, a height of 1 is a leaf node, and larger values result in more generic object classes. We see that our method seems to perfom better as we go up the ImageNet hierarchy. This could be because generic objects have more images, and thus our method has more examples to leverage in the box similarity and discriminability terms.

CorLoc difference between methods. In Figure 9, we show the CorLoc difference between our method and objectness for all 3,624 classes. From the best CorLoc differences, we find that our method performs much better than objectness on large rooms and objects, which is probably because objectness tries to select individual objects or object parts within these large scenes, whereas our model is able to understand that the individual objects are not similar, and select the scene or object as a whole.

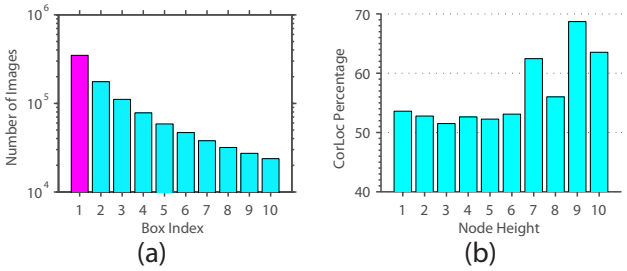

Figure 8. (a) Boxes selected by our method on ImageNet, ordered by descending objectness score; (b) CorLoc performance of our method separated into differing node heights of ImageNet.

5. Conclusion

In this paper, we introduce a method for co-localization in real-world images that combines terms for the prior, similarity, and discriminability of both images and boxes into a joint optimization problem. Our formulation is able to account for noisy images with incorrect annotations. We performed an extensive evaluation of our method on standard datasets, and also performed a large-scale evaluation using ground-truth annotations for 3,624 classes from ImageNet.

For future work, we would like to extend our model to the pixel level for tasks such as co-segmentation, and to handle multiple instances of objects.

Acknowledgments. We especially thank V. Ferrari for helpful comments, suggestions, and discussions. We also thank N. Liu for implementation help, A. Alahi, J. Johnson,

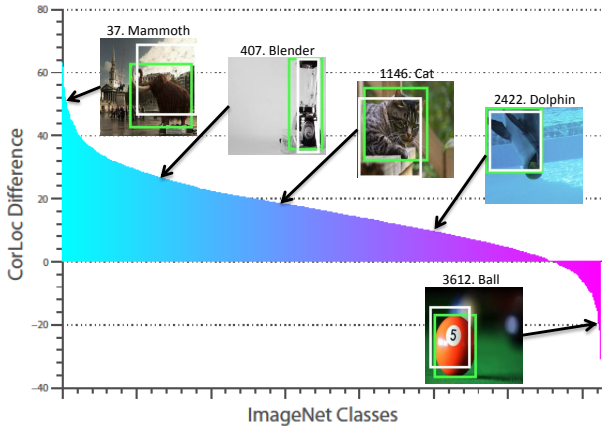

Figure 9. CorLoc difference between our method and objectness on all 3,624 classes from ImageNet that we evaluate on.

O. Russakovsky, V. Ramanathan for paper comments. This research is partially supported by an ONR MURI grant, the DARPA Mind's Eye grant, the Yahoo! FREP, and a NSF GRFP under grant no. DGE-114747 (to K.T.).

References

[1] B. Alexe, T. Deselaers, and V. Ferrari. Measuring the objectness of image windows. *IEEE T-PAMI*, 34(11):2189–2202, 2012. 2, 6

[2] F. Bach and Z. Harchaoui. Diffrac: a discriminative and flexible framework for clustering. In *NIPS*, 2007. 2, 4

[3] S. Y. Bao, Y. Xiang, and S. Savarese. Object co-detection. In *ECCV*, 2012. 2

[4] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003. 4

[5] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010. 2

[6] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011. 3

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

[8] 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. 2, 5, 6

[9] T. Deselaers, B. Alexe, and V. Ferrari. Weakly supervised localization and learning with generic knowledge. *IJCV*, 100(3):275–293, 2012. 1, 2, 5

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. 4, 5

[11] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009. 1, 2

[12] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, 2001. 4

[13] V. J. Hodge and J. Austin. A survey of outlier detection methodologies. *AI Rev.*, 22(2):85–126, 2004. 2, 3

[14] A. Joulin and F. Bach. A convex relaxation for weakly supervised classifiers. In *ICML*, 2012. 1

[15] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *CVPR*, 2010. 2, 4, 5

[16] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012. 2, 5

[17] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 2, 5

[18] D. Küttel, M. Guillaumin, and V. Ferrari. Segmentation propagation in imagenet. In *ECCV*, 2012. 2, 6

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006. 3

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[21] 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

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

[23] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 3

[24] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. *CVPR*, 2013. 1, 2, 5, 6

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

[26] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE T-PAMI*, 22(8):888–905, 2000. 2, 4

[27] P. Siva, C. Russell, T. Xiang, and L. de Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, 2013. 1, 2

[28] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011. 1, 2

[29] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012. 1

[30] K. Tang, R. Sukthankar, J. Yagnik, and L. Fei-Fei. Discriminative segment annotation in weakly labeled video. In *CVPR*, 2013. 1, 2

[31] A. Vahdat and G. Mori. Handling uncertain tags in visual recognition. In *ICCV*, 2013. 1, 2

[32] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011. 2

[33] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 2

[34] L. Xu, J. Neufeld, B. Larson, and D. Schuurmans. Maximum margin clustering. In *NIPS*, 2004. 2, 4