BiCoS: A Bi-level Co-Segmentation Method for Image Classification

Yuning Chai
Electrical Engineering Dept.
ETH Zurich
chaiy@ee.ethz.ch

Victor Lempitsky
Dept. of Engineering Science
University of Oxford
vilem@robots.ox.ac.uk

Andrew Zisserman
Dept. of Engineering Science
University of Oxford
az@robots.ox.ac.uk

Abstract

The objective of this paper is the unsupervised segmentation of image training sets into foreground and background in order to improve image classification performance. To this end we introduce a new scalable, alternation-based algorithm for co-segmentation, BiCoS*, which is simpler than many of its predecessors, and yet has superior performance on standard benchmark image datasets.*

We argue that the reason for this success is that the co-segmentation task is represented at the appropriate levels – pixels and color distributions for individual images, and super-pixels with learnable features at the level of sharing across the image set – together with powerful and efficient inference algorithms (GrabCut and SVM) for each level.

We assess both the segmentation and classification performance of the algorithm and compare to previous results on Oxford Flowers 17 & 102, Caltech-UCSD Birds-200, the Weizmann Horses, Caltech-4 benchmark datasets.

1. Introduction

Co-segmentation of image collections has recently become a topic of active research [16, 19, 25, 31, 37, 38]. Co-segmentation methods consider sets of images where the appearance of foreground and/or background share some similarities, and try to leverage these similarities to obtain accurate foreground-background segmentations either totally-unsupervised, or with a small amount of interactive supervision [5]. Most of the proposed co-segmentation methods (with the exception of [19]) assume close similarity of the foreground color histograms essentially requiring foreground objects to be the same through the image set.

Our goal in this paper is to develop a method for unsupervised foreground-background co-segmentation of image sets that is scalable to large datasets of thousands of images. Our research is motivated by the desire to use unsupervised co-segmentation for the task of background removal within the training image sets for image classification systems. There is abundant evidence that accurate foreground-background segmentation can benefit visual recognition [24]. This is particularly true when image classification deals with *subordinate* visual categories, e.g. flower species [27] or bird species [40]. In this scenario, background visual features tend to be similar across categories and typically act as a distraction to statistical learning rather than as useful context. Removing background at training time therefore gives a significant boost to the classification performance.

To this end we introduce a new co-segmentation algorithm which is scalable and operates at two levels of representation: at the bottom level, it treats each image separately and uses the well-known GrabCut algorithm [30] applied to the RGB values of individual pixels, whereas at the top level a discriminative classification is performed on high-dimensional descriptors of superpixels. The top layer operates on all images jointly and propagates information about foreground and background appearances across the dataset. Fig. 1 illustrates the algorithm schematically. This simple algorithm, which we term *BiCoS* for Bi-level Co-Segmentation, scales linearly with the number of images. Unlike most co-segmentation algorithms it does not assume the similarity of either global geometric shape or foreground color distributions throughout the image set.

Motivated by our ultimate application, we also propose a *multi-task* modification of BiCoS for co-segmenting *multiple* image sets each corresponding to a different class. This (BiCoS-MT) algorithm solves the co-segmentation tasks jointly, ensuring the similarity of the background appearances across the classes.

We then evaluate both BiCoS and BiCoS-MT in the context of building an image classification system. The evaluation proves convincingly that applying unsupervised co-segmentation to the training sets can benefit the classification accuracies. As an outcome, we report better than state-of-the-art classification accuracy on the popular Oxford-17 [27] and Oxford-102 [29] flowers datasets. For these datasets, we compare BiCoS and BiCoS-MT with GrabCut [30], the state-of-the-art co-segmentation system [19] as well as with the system of [28] that is designed specifically for flower segmentation and is trained in a supervised

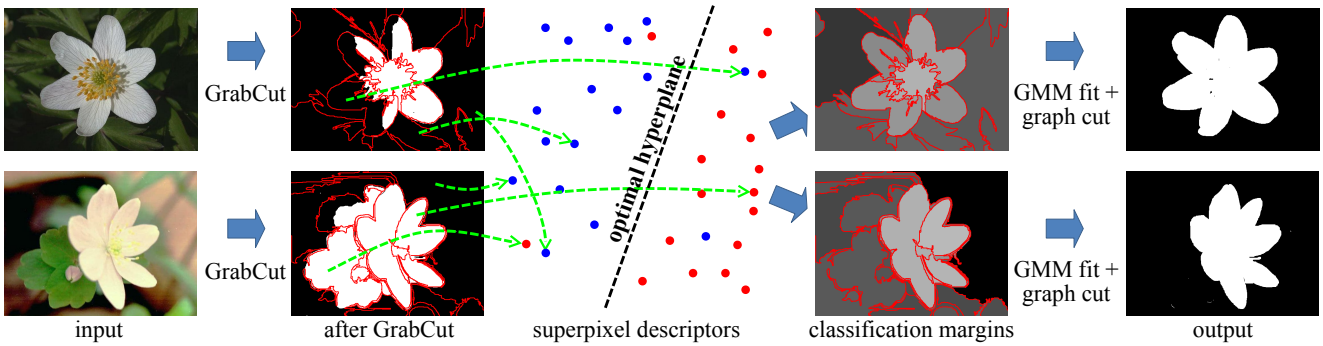

Figure 1. The diagram of our method (BiCoS) shown for 2 (out of 15) training images from one class of the Oxford Flowers-17 dataset. BiCoS starts with GrabCut at the pixel-level applied to each image independently. Each superpixel (superimposed) is then assigned to foreground or background and mapped to a descriptor space, where the optimal separating hyperplane is found. The pixels in each image are then assigned values according to the margins of the respective superpixels, and such soft segmentation map is then used to fit color Gaussian mixtures and apply pixel-level graph cut to each image.

way. Despite their simplicity, the proposed algorithms came out on top both in terms of the *segmentation accuracy* and in terms of the *classification accuracy* obtained by the visual classifier trained on the segmented training set. We then extend the experimental evaluation to the recently introduced Caltech-UCSD bird species dataset [40], which is very difficult for both segmentation and recognition.

Closely related to co-segmentation algorithms, is the group of approaches that deal with an unlabeled image set of an unknown class and aim to simultaneously segment the images and build a loose geometric model of that class [3, 4, 22, 34, 41]. These algorithms are typically evaluated and achieve excellent results on the viewpoint-constrained datasets (Weizmann Horses [7], Caltech-4 [21]), where their geometric modeling is very appropriate. In contrast, BiCoS algorithms do not attempt to build any geometric models and are therefore applicable to image sets of highly non-rigid classes under extreme variations of viewpoints (such as user photographs of flowers taken in unconstrained setting). Nevertheless, we evaluate BiCoS on a number of viewpoint-constrained datasets, finding out that, despite the lack of geometric modeling, the performance of BiCoS is rather competitive with many of the recent geometry-based methods.

2. Related work

In the next section, we review the GrabCut method of Rother *et al*. [30] used at the bottom level of BiCoS. The discriminative learning in the space of superpixel descriptors used at the top level has been employed within *semantic segmentation* methods such as [13, 15, 17], which are trained in a supervised way.

Among the unsupervised co-segmentation algorithms, discriminative learning on superpixels is used by Joulin *et al*. [19]. The optimization framework of [19] simultaneously enforces spatial smoothness within each image as well as finding the foreground-background boundary in the superpixel space. Unlike [19], BiCoS decouples spatial smoothness enforcement and classification of superpixels, so that these two steps are performed consecutively rather than simultaneously. We demonstrate experimentally that, despite the sub-optimality that such alternation-based approach might bring, BiCoS consistently attains higher segmentation accuracies, while being applicable to much larger image sets than [19].

Alexe *et al*. [3] is another very recent work closely related to ours, as their system also uses superpixels to propagate information across multiple images. Such propagation is however achieved through binary-label Conditional Random Field (CRF) with unary potentials derived in a generative fashion as opposed to discriminative learning (SVM) used within BiCoS. Despite the use of explicit geometric modeling within [3], the experimental comparison revealed that BiCoS is able to achieve similar or higher segmentation accuracy for several viewpoint-constrained datasets, where their loose geometric model is appropriate.

Another co-segmentation work that adopts an alternation strategy similar to the bi-level architecture of BiCoS is Batra *et al*. [5]. Their appearance models are however based purely on color (and hence are too limited for many scenarios). Their focus is also on interactive user supervision, rather than the fully unsupervised scenario used by most other co-segmentation works, as well as ours.

3. Segmentation methods

We start with a review of GrabCut [30] that works at single image level. We then introduce the BiCoS method that is based on the combination of GrabCut and discriminative learning in the superpixel descriptor space. After that, we demonstrate how the proposed method may be modified to work with a dataset of multiple image sets with shared background patterns (a scenario typical for image classification as discussed above).

3.1. Image-level segmentation: GrabCut [30]

The GrabCut algorithm [30] is a popular tool for the segmentation of single images. It combines two components: a binary-label random field defined on image pixels, and a generative foreground/background classifier that takes pixel RGB values as features. GrabCut proceeds by iterations: starting from some initialization, it alternates between **(1)** The (re)estimation of foreground and background probability densities via Gaussian mixtures given foreground/background pixel labels [6], and **(2)** The graph cut inference of foreground/background labels in the random field with unary terms defined according to the evidence from the Gaussian mixtures and the pairwise terms defined according to local image gradients cues [8].

The assumption behind GrabCut and in particular behind the step (1) in its alternation scheme, is that the RGB distributions of foreground and background are shared across the entire image (so-called *global color modeling*). As demonstrated in [30] and subsequent evaluations, GrabCut is capable of producing accurate segmentation even when initialized in a very crude way (e.g. a rectangular mask overlapping the true foreground).

3.2. Class-level co-segmentation

Given a set of images of the same class, co-segmentation algorithms attempt to improve the segmentation accuracy by assuming the similarity of the visual appearance of foreground and/or background across the images in the dataset. A number of approaches [16, 25, 31, 37] work with pixel RGB values to model and propagate the distributions of visual appearance. However, in realistic scenarios for the vast majority of classes, RGB distributions of foreground and background pixels across the entire set of images overlap too much for such approaches to be useful. In other words, the color-based representation which works very well at the level of individual images is often unsuitable at the dataset level.

Rather than share a simple descriptor (RGB) at the level of individual pixels across the dataset, we share a richer descriptor at the level of super-pixels. This richer descriptor (a high dimensional feature vector) is suitable for discriminative learning at the dataset level. Thus, each image is partitioned into a set of superpixels (via the graph-based method [12], preferred for its time efficiency) and described by a real-valued descriptor (with the dimensionality $D = 1076$) stacked from five fairly standard sub-descriptors, representing the superpixels' color distribution, SIFT [23] distribution, size, location within the image and shape (more details below).

The super-pixels are then classified into foreground and background using standard linear support vector machine (SVM) training. Assume that a set $\{\mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x}_N\}$ of N descriptors corresponding to all superpixels are given. Let also $\{y_1, y_2 \dots y_N\}$ be the binary labels where $y_i = +1$ ($y_i = -1$) corresponds to superpixels with the majority of pixels assigned to foreground (background). Then, the separating hyperplane $\mathbf{w}$ in the descriptor space may be obtained by solving the standard SVM (convex quadratic) optimization program:

$$\frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{N} \ell(y_i \cdot \mathbf{w}^T\mathbf{x}_i) \longrightarrow \min_{\mathbf{w}}, \qquad (1)$$

where $\ell(t) = \max(0, 1-t)$ is the hinge loss function, and C is the regularization constant set to 10 in all our experiments.

Our co-segmentation method then integrates this discriminative learning step that propagates the information about appearance distributions across images with GrabCut steps that propagate the information within images. Overall, BiCoS has the following steps (Fig. 1):

1. **Initialization at the image level.** BiCoS starts with GrabCut being applied to each image independently. GrabCut here is intialized with the rectangle in the center (50% of the image size, unless noted otherwise) assigned to the foreground and the rest to background. The output of this step is a pixel level segmentation of each image.

2. **Propagation at the dataset level.** Each superpixel is assigned in each image to either foreground or background depending on the label of the majority of its pixels. Given the set of superpixels and their labels (over all images), the separating hyperplane $\mathbf{w}$ is then learnt using a linear SVM. Each superpixel is then reclassified according to its sign w.r.t. to the hyperplane (so that superpixels on the "wrong" side of the hyperplane effectively switch label). The output of this step is a linear classifier and a classification of all superpixels across the dataset.

3. **Update at the image level.** Within each image independently, the Gaussian mixture distribution are reestimated from the new foreground and background regions using the pixels that belong to superpixels classified as foreground or background in the previous step. Each pixel is weighted according to the distance from the separating hyperplane, i.e. the margin $\mathbf{w}^T\mathbf{x}_i$. Given the new mixtures, a single graph cut [8] is then applied with the potentials defined as in GrabCut. The output of this step is again a pixel level segmentation of each image.

One can iterate the sequence of steps 2 and 3. For efficiency reasons, however, we adopted the single-pass version of BiCoS for our experiments (unless noted otherwise). Overall, scalability to large datasets is an attractive property

of the method. Essentially, the scaling is linear in the number of images: the SVM in step 2 can be trained in linear time in the number of superpixels [33] and all other steps are done independently within each image.

We have used a discriminative classifier. Due to the high dimensionality of superpixel descriptors and obvious interdependencies in their dimensions, generative modeling and classification of foreground and background distributions would be problematic, although one can use naive-Bayes approximation [3] as well as topic models [9, 32].

Implementation details. We use the popular Felzenswalb's code for superpixel segmentation [12], with sigma=1, k=200, minSize=640. Parameter values were chosen on another, but similar, problem, and are used unchanged for all segmentation experiments. We observe that the system is not too sensitive to the method of superpixel segmentation or its parameter settings.

The superpixel descriptor is composed from the following parts: (1) a 200-dimensional color and a 800-dimensional SIFT histograms that are obtained by vector-quantizing and pooling the densely-sampled *Lab* color values and multi-scaled dense SIFT descriptors respectively. (2) The size is encoded as a tiny 2 bins-histogram with the 1 value placed depending on whether the superpixel is "big" or "small" (the meaning of "big" and "small" is obtained by clustering the sizes of superpixels in the training set into two clusters). (3) The location in the image is represented with a 36-D descriptor obtained by linearly downsampling the image-size binary mask, indicating pixels belonging to the superpixel, to 6×6 size. (4) The shape descriptor is obtained in the same way, except that the mask is cropped with the superpixel bounding box prior to downsampling. (5) Finally, two 0/1 values encoding whether the superpixel contains the center of the image and whether it touches the boundary of the image. The resulting descriptor is quite high-dimensional (1076-D), but is relatively cheap to compute and permits the similarity computation by a simple dot product (hence no need for the kernelization of the algorithms). The vector-quantization is done with locally-constrained linear coding [39].

3.3. Co-segmenting multiple class image sets

The GrabCut algorithm achieves a remarkable accuracy by making foreground and background regions of an image share appearance distributions. Co-segmentation algorithms including BiCoS are capable of improving segmentations by enforcing the sharing of such distributions across the images of the same class. In this subsection, we make one step further and consider the scenario where one is given a dataset that contains multiple sets of images $\mathcal{S}_1, \mathcal{S}_2 \ldots \mathcal{S}_K$ corresponding to K classes, with each set $\mathcal{S}_k$ containing N_k images ($\mathcal{S}_k = \{\mathcal{I}_1, \mathcal{I}_2, \ldots \mathcal{I}_{N_k}\}$), where N_k may be as small as 1 and as large as several hundred.

Clearly, such datasets may be segmented by running co-segmentation algorithm for each set S_k independently. However, one may attempt to improve the segmentation even further by enforcing appearance sharing across classes. As discussed in the introduction, we assume that the ultimate goal in this scenario is the improvement of the recognition accuracy for classifiers trained from the respective training sets after background removal. The consequence of this observation is that a "good" co-segmentation process should have a tendency to assign regions with the appearance patterns that are ubiquitous across multiple image sets to background (as these patterns would not be discriminative but rather confusing for class discrimination).

Our second approach (BiCoS-MT) enforces background sharing by modifying the SVM formulation in BiCoS. For each class-specific dataset S_k, the algorithm finds a separate hyperplane that discriminates between the background and the foreground appearances typical for that class. The hyperplanes for different classes are however not computed independently. Instead, for the class k, the separating hyperplane is found as the sum of two vectors $\mathbf{w}_k + \mathbf{w}_-$, where $\mathbf{w}_k$ is the class-specific part and $\mathbf{w}_-$ is the component shared across all classes. The objective of the joint SVM formulation is then naturally defined as:

$$\frac{\mu}{2}||\mathbf{w}_-||^2 + \sum_{k=1}^{K} \left[\frac{1}{2}||\mathbf{w}_k||^2 + C \sum_{i=1}^{N_k} \ell\left(y_i^k \cdot (\mathbf{w}_k + \mathbf{w}_-)^T \mathbf{x}_i^k\right) \right] \rightarrow \min_{\{\mathbf{w}_1 \ldots \mathbf{w}_K, \mathbf{w}_-\}}. \tag{2}$$

Thus, (2) is defined as a sum of K independent SVM objectives for each class along with the vector component $\mathbf{w}_-$ (and its regularization term) shared across the SVMs. (In (2), $\mathbf{x}_i^k$ and y_i^k denote superpixels within the dataset $\mathcal{S}_k$; N_k denotes the number of superpixels in image set $\mathcal{S}_k$; μ is an additional regularization parameter set to $\frac{1}{2}$ in all our experiments.)

Since we want to encourage the sharing of background and discourage the sharing of the foreground appearance patterns, we impose additional constraints on (2):

$$\mathbf{w}_k^j \geq 0, \quad \forall j = 1 \ldots D \tag{3}$$

$$\mathbf{w}_-^j \leq 0, \quad \forall j = 1 \ldots D \tag{4}$$

An intuition behind the constraints (4) is that they make the separating hyperplanes for different classes share negative coefficients (that are indicative of background superpixels), while positive coefficients can be inferred for each class independently. With respect to the original SVM formulation, it is easy to show that for the case of a single class $K = 1$ and equal regularization on $\mathbf{w}_1$ and $\mathbf{w}_-$ (i.e. when $\mu = 1$), the standard SVM formulation (1) and the new formulation

Figure 2. Bird segmentations: (top) original images, (middle) using BiCoS and (bottom) using BiCoS-MT. In the left 5 images, BiCoS-MT performs better because it takes into account background elements (*e.g.* tree branches) in the background of images from other classes. But due the existence of confusing foreground elements (*e.g.* blue birds) in other classes, sky can be mis-classified using BiCoS-MT.

(2)–(4) are exactly equivalent and produce the same separating hyperplane.

Based on the augmented SVM formulation, we devise the BiCoS-MT algorithm that uses the joint program (2)–(4) to estimate the separating hyperplane $\mathbf{w}_k + \mathbf{w}_-$ for each class k at step 2. The rest of the steps are unchanged from BiCoS. We note that the joint SVM formulation can be naturally interpreted as an instance of multi-task learning [10, 11] (hence the name BiCoS-MT). Within BiCoS-MT, each task corresponds to learning a superpixel classifier for a particular image set $\mathcal{S}_k$. We also note that "softer" version of background sharing may be implemented, where each class possesses both positive component $\mathbf{w}_k^+$ and negative component $\mathbf{w}_k^-$ specific to that class, whereas the extra term $\nu \sum_k ||\mathbf{w}_k^- - \mathbf{w}^-||^2$ penalizes the deviation of the negative components from their joint mean vector $\mathbf{w}^-$ that does not enter in any other terms (c.f. [11]).

The program (2)–(4) may be rather large (many tens of thousands of variables in some of our experiments), and we therefore implemented the modification of a popular stochastic gradient descent-based method (Pegasos [33]) to handle the task. This allows (2)–(4) to scale to a large number of classes and/or a large number of images per class. A detailed pseudocode description of the modified Pegasos solver for (2)–(4) is given at [2].

As illustrated in Fig. 2, sharing the background among all classes may lead to better segmentation, but may also cause mis-segmentation. But for the majority of our experiment setups, BiCoS-MT has an advantage over BiCoS.

4. Experimental results

Performance measures. During the evaluation of co-segmentation systems, we are interested in two measures: the segmentation accuracy and the accuracy of the recognition achieved by a visual classifier trained on the segmented dataset, and we report these measures for three image classification datasets and several methods.

Given a ground truth foreground mask, the segmentation accuracy can be expressed in several ways. One is the ratio between the size of the intersection area between the estimated foreground and ground truth foreground over the size of their union (**Seg. I**). The other is the percentage of pixels classified correctly as either foreground or background (**Seg. II**). We report both numbers for our methods, and whichever is published for the competitors. For the average recognition accuracy (**Rec.**), we take the mean value of the relative numbers of correctly classified images of each category (i.e. the average class accuracy) by a state-of-the-art visual classifier briefly described below.

Evaluation on image recognition. In order to evaluate the effect of co-segmentation on the recognition accuracy we apply the following training pipeline: (i) BiCoS, BiCoS-MT or one of the baseline segmentation algorithms is used to obtain foreground segmentation masks on the training data (note, in all cases the algorithm can make use of the class labels of the training data – BiCoS can be applied to each class independently, and BiCoS-MT is formulated to use this information); (ii) a 1-vs-rest SVM image classifier is learnt for each category (see below) using the foregrounds of that category as positives, and the foregrounds of all other categories as negatives; (iii) a generic segmenter is learnt (which will be applied to the test images) as an SVM for super-pixels using the foreground regions of *all* categories as positives and the background regions of *all* categories as negatives (note, at test time we do not know the class label so cannot use a BiCoS algorithm just for that class). The test pipeline then proceeds as: (i) segmenting into foreground background using the generic segmenter SVM on superpixels followed by step 3 of BiCoS; (ii) classifying the foreground image using each of the 1-vs-rest SVM image classifiers; and (iii) returning the class with the maximum score over all the image classifiers.

Implementation Details. The features for the image recognition part are obtained by concatenating the Bag-Of-Words (BoW) histograms of *Lab* color and SIFT descriptors, which are extracted solely from the foreground area given by the segmentation. We use three different histograms for SIFT feature corresponding to dense sampling at several scales, sampling at interest points, and sampling along the foreground boundary. Similar to the superpixel descriptors, the vector-quantization uses locally-constrained linear coding [39]. The vocabulary sizes are 800 for *Lab* and 8000 for each of the 3 SIFT descriptors. Before concatenating the four descriptors together, we apply the homogeneous kernel map with approximation order 1 and sampling step 0.7 [36] onto it, which results in a 74400 dimensional feature vector, that is used within the linear SVM (with $C = 1$). The implementation uses VL-Feat [35].

4.1. Oxford Flowers 17

The Oxford Flowers 17 dataset contains 17 different flower species with 80 images per category. The dataset provides three different data splits with each having 60 training and 20 test images. And 818 out of the 1360 images have hand-annotated ground truth segmentations. The fact that these ground truth segmentations are unequally distributed among all categories, makes it difficult to compare segmentation and recognition accuracies simultaneously.

We first compare our algorithms to the recently published co-segmentation algorithm by Joulin *et al*. [19]. We use their provided code and their suggested parameter settings for this dataset. The memory demands of this algorithm are too excessive to use the predefined 60/20 train/test split. Instead, a new data split having 15 training and 65 test images per category is used. This split has 15 training images with ground truth segmentations for 16 of the categories (out of the total of 17). Thus, for the new split we can compare both segmentation and recognition accuracies. We note that other published co-segmentation methods either do not scale to even this number of images or require user supervision [5], thus making [19] the natural choice for the comparison.

The numerical comparison between our own methods and method [19] is shown in Tab. 1, while some of the segmented images are shown in Fig. 3. In Tab. 1 and thereafter, we also provide the accuracies for the cases when no pre-segmentation of training and test images is performed, and when the training dataset is pre-segmented with GrabCut.

We also compare the recognition accuracy of our full system that uses BiCoS to co-segment training image with the state-of-art recognition results using the predefined data splits. We show our recognition accuracies in Tab. 2 along with the results provided by Gehler and Nowozin [14] and Nilsback [26].

Methods	Seg. I	Seg. II	Rec.
All foreground	32.8	32.8	62.1
Joulin *et al*. CVPR'10 [19]	75.8	86.6	74.1
GrabCut [30]	89.3	96.3	73.3
BiCoS (this work)	94.1	98.1	79.3
BiCoS-MT (this work)	94.7	98.3	80.5

Table 1. Performance on Oxford Flowers 17 (with the alternative data split – 15 images per class). Our algorithms achieves the highest results in all three measures. For this number of training images, BiCoS-MT also outperforms BiCoS.

Methods	Rec. Accuracy
Gehler and Nowozin ICCV'09 [14]	85.5 ± 3.0
Nilsback' thesis [26]	88.1 ± 1.9
BiCoS (this work)	91.1 ± 1.5
BiCoS-MT (this work)	90.4 ± 2.3

Table 2. Performance on Oxford Flowers 17 (with the original data splits – 60 training images per class). Our algorithms once again lead to the highest recognition accuracies. For 60 images per class, BiCoS-MT no longer perform better than BiCoS.

4.2. Oxford Flowers 102

The Oxford Flowers 102 dataset has 8289 images divided into 102 categories with 40 to 250 images per category. For each category, 10 training and 10 validation images are predefined, while the rest is left for testing. There is no segmentation ground truth provided in the dataset, therefore, we use the recognition accuracy as the only measurement.

Among all recent approaches evaluated on this dataset [18, 20], the best performance is reported in [26]. The authors proposed a model-based foreground segmentation approach which segments images independently. In the following experiments, we can show that this baseline can be surprisingly outperformed in our experiment framework even if the images are segmented independently using GrabCut alone. We also show that the accuracy can be pushed even further with the proposed methods (Tab. 3). We used the results of the segmentation [26] available online to evaluate the combination of their segmentation system and our image classifier. The improvement obtained by BiCoS and BiCoS-MT over [26] in this experiment is all the more important, as the system [26] required around 800 hand-annotated segmentation masks as training data.

4.3. Caltech-UCSD Birds 200

The recently published Caltech-UCSD Birds 200 [40] contains 200 bird categories and 6033 images in total. There are *rough* segmentation masks provided with the dataset, which allows us to compare segmentation accuracies. However, we should note that those values are only approximated because of the rough segmentation ground truth ob-

Figure 3. Flower segmentations: (top) original images, (middle) segmentations by Joulin *et al*. [19] and (bottom) the segmentations of BiCoS-MT. The results of BiCoS are very similar to BiCoS-MT for these particular images.

Segmentation	Classification	Rec. Accuracy
no segmentation	ours	64.7
Nilsback [26]	Nilsback [26]	76.3
Nilsback [26]	ours	76.8
Kanan and Cottrell [20]	Kanan [20]	72.8
Ito and Cubota[18]	Ito [18]	74.8
GrabCut [30]	ours	77.0
BiCoS	ours	79.4
BiCoS-MT	ours	80.0

Table 3. Recognition performance for different combination of segmentaiton and classification approaches on Oxford Flowers 102. For a number of methods, we vary the (co-)segmentation approach while keeping the classification approach fixed. The proposed methods once again lead to the highest recognition accuracy.

Methods	Seg. I	Seg. II	Rec.
no segmentation	17.0	17.0	6.7
GrabCut [30]	38.8	73.5	13.6
BiCoS	41.1	78.3	15.7
BiCoS-MT	39.9	77.3	16.2
using ground truth seg.	-	-	23.3

Table 4. Performance on Caltech-UCSD Birds 200. BiCoS algorithms outperform GrabCut in all measures. BiCoS-MT leads to slightly higher recognition accuracy. The bottom line provides the performance of our classification approach given ground truth segmentation.

tained with MTurk. The dataset is extremely challenging, and its authors report only 19% recognition accuracy *when using ground truth masks*. Tab. 4 gives the segmentation and classification accuracies (using the suggested 20 training images per class split) of GrabCut, BiCoS, and BiCoS-MT paired with our recognition approach. We hope that they may serve as a baseline for further research on segmentation and co-segmentation.

4.4. Weizmann Horses and Caltech-4

We also evaluate BiCoS on datasets with fixed viewpoint orientation and classes that have well-defined geometric shape. This scenario is quite different from our main application and it is inevitable that methods interleaving segmentation and construction of class-specific geometric models are likely to have an advantage over BiCoS or any other method not having a geometric model. Still, it is interesting to see how well BiCoS can fare against such methods, in particular against Alexe *et al*. [3] to which it is most similar. We therefore evaluated BiCoS on Weizmann Horses [7] and 3 out of 4 Caltech-4 classes [21] (the grayscale 'cars' class was left out). We keep all the parameters fixed from the Flowers and Birds experiments, except that GrabCut is initialized with the central foreground rectangle occupying 25% (rather than 50%) of the image area.

The comparison with state-of-the-art is in Tab. 5. Here we also give results for the iterated BiCoS (repeating step 2 to step 3 for 5 times). In the experiments, BiCoS performed poorly on the motorbike class, where GrabCut often fails due to unnatural framing within many photographs. Apart from that, the performance of BiCoS is surprisingly competitive, in particular outperforming Alexe *et al*. [3] on the remaining 4 categories. The top performing method of [22] outperforms BiCoS on all categories except airplanes where BiCoS (one iteration) is better.

5. Discussion

We presented BiCoS: a co-segmentation method that is arguably very simple, yet is scalable and performs remarkably well in our experimental comparisons. BiCoS outperformed GrabCut that treats images independently and the state-of-the-art co-segmentation [19] in all comparisons. Image classification systems trained on datasets co-segmented with BiCoS achieved state-of-the-art accuracy on the Flowers-17 and Flowers-102 datasets. For the small

Methods	horse		airplane		face		motorbike	
	Seg I	Seg II	Seg I	Seg II	Seg I	Seg II	Seg I	Seg II
All background	0.0	71.0	0.0	83.1	0.0	80.0	0.0	72.9
Joulin *et al*. [19]	-	80.1	-	-	-	-	-	-
Cao&Fei-Fei [9]	-	81.8	-	-	-	-	-	-
Alexe *et al*. [3]	-	86.2	-	89.8	-	89.0	-	90.3
Arora *et al*. [4]	-	-	-	93.1	-	**92.4**	-	83.1
LOCUS [41]	-	93.1	-	-	-	-	-	-
Liu *et al*. [22]	-	**95.9**	63.9	-	-	-	**71.6**	-
GrabCut [30]	60.6	86.3	59.2	90.7	60.1	86.2	41.1	80.6
BiCoS (1 iter.)	65.2	87.6	**64.5**	93.0	66.2	89.3	44.7	82.8
BiCoS (5 iter.)	68.7	90.0	63.8	**93.2**	69.0	91.1	41.8	82.4

Table 5. (Co-)segmentation performance on Weizmann Horses and Caltech-4 datasets reported in the literature. Despite the lack of the geometric model, BiCoS is competitive with the methods that have such models (see text for more discussion).

number of images per class (Tab. 1, Tab. 3) the multi-task version of BiCoS leads to even higher recognition accuracy.

Many co-segmentation methods attempt to segment images and infer the foreground and background appearance models at the same time. BiCoS follows a different, perhaps less principled way, and performs within-image and across-image appearance propagation consecutively rather than jointly. Such a two-step approach, however, allows a proper feature representation to be used at each level (color-based at the level of individual images, discriminative classification of rich superpixel descriptors at the dataset level). The success of BiCoS in our experiments therefore seems to concur with the well-known fact that devising appropriate feature representations has a higher impact on the performance of computer vision systems than the choice of a principled optimization algorithm.

All the segmentations from Sec. 4 are provided at [2], and a classification demo on the Oxford Flower 102 dataset is available at [1].

Acknowledgment. This work was supported by ERC grant VisRec no. 228180. Victor Lempitsky is also supported by Microsoft Research programs in Russia.

References

[1] Flower classification demo. http://www.robots.ox.ac.uk/~vgg/research/flowers_demo.

[2] Segmentation data. http://www.robots.ox.ac.uk/~vgg/data/bicos.

[3] B. Alexe, T. Deselaers, and V. Ferrari. Classcut for unsupervised class segmentation. In *ECCV*, 2010.

[4] H. Arora, N. Loeff, D. A. Forsyth, and N. Ahuja. Unsupervised segmentation of objects using efficient learning. In *CVPR*, 2007.

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

[6] A. Blake, C. Rother, M. Brown, P. Pérez, and P. H. S. Torr. Interactive image segmentation using an adaptive GMMRF model. In *ECCV*, 2004.

[7] E. Borenstein and S. Ullman. Class-specific, top-down segmentation. In *ECCV*, 2002.

[8] Y. Boykov and M.-P. Jolly. Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. In *ICCV*, 2001.

[9] L. L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. In *ICCV*, 2007.

[10] R. Caruana. Multitask learning. *Machine Learning*, 28(1), 1997.

[11] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *KDD*, 2004.

[12] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59(2), 2004.

[13] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[14] P. V. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, pages 221–228, 2009.

[15] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

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

[17] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. In *ICCV*, 2005.

[18] S. Ito and S. Kubota. Object classification using heterogeneous co-occurrence features. In *ECCV*, 2010.

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

[20] C. Kanan and G. W. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *CVPR*, 2010.

[21] F. F. Li, R. Fergus, and P. Perona. One-shot learning of object categories. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(4), 2006.

[22] G. Liu, Z. Lin, X. Tang, and Y. Yu. A hybrid graph model for unsupervised object segmentation. In *ICCV*, pages 1–8, 2007.

[23] D. G. Lowe. Object recognition from local scale-invariant features. In *ICCV*, 1999.

[24] T. Malisiewicz and A. A. Efros. Improving spatial support for objects via multiple segmentations. In *BMVC*, 2007.

[25] L. Mukherjee, V. Singh, and C. R. Dyer. Half-integrality based algorithms for cosegmentation of images. In *CVPR*, 2009.

[26] M.-E. Nilsback. *An Automatic Visual Flora – Segmentation and Classification of Flowers Images*. PhD thesis, University of Oxford, 2009.

[27] M. E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.

[28] M.-E. Nilsback and A. Zisserman. Delving into the whorl of flower segmentation. In *BMVC*, 2007.

[29] M. E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008.

[30] C. Rother, V. Kolmogorov, and A. Blake. "grabcut": interactive foreground extraction using iterated graph cuts. *ACM Trans. Graph.*, 23(3), 2004.

[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.

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

[33] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient SOlver for SVM. In *ICML*, volume 227, 2007.

[34] S. Todorovic and N. Ahuja. Extracting subimages of an unknown category from a set of images. In *CVPR*, 2006.

[35] A. Vedaldi and B. Fulkerson. Vlfeat – an open and portable library of computer vision algorithms. In *ACM int. conf. on Multimedia*, 2010.

[36] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010.

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

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

[39] J. Wang, J. Yang, K. Yu, F. Lv, T. S. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[40] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD Birds 200. Technical Report CNS-TR-2010-001, California Institute of Technology, 2010.

[41] J. M. Winn and N. Jojic. Locus: Learning object classes with unsupervised segmentation. In *ICCV*, 2005.