Unsupervised Multi-Class Joint Image Segmentation

Fan Wang
Stanford University
fanw@stanford.edu

Qixing Huang
Stanford University
huangqx@stanford.edu

Maks Ovsjanikov
LIX, Ecole Polytechnique
maks@lix.polytechnique.fr

Leonidas J. Guibas
Stanford University
guibas@cs.stanford.edu

Abstract

Joint segmentation of image sets is a challenging problem, especially when there are multiple objects with variable appearance shared among the images in the collection and the set of objects present in each particular image is itself varying and unknown. In this paper, we present a novel method to jointly segment a set of images containing objects from multiple classes. We first establish consistent functional maps across the input images, and introduce a formulation that explicitly models partial similarity across images instead of global consistency. Given the optimized maps between pairs of images, multiple groups of consistent segmentation functions are found such that they align with segmentation cues in the images, agree with the functional maps, and are mutually exclusive. The proposed fully unsupervised approach exhibits a significant improvement over the state-of-the-art methods, as shown on the co-segmentation data sets MSRC, Flickr, and PASCAL.

1. Introduction

Image segmentation is a fundamental problem in computer vision. Traditional methods have focused on single images and typically utilize segmentation clues, such as color changes or the presence of sharp edges, to divide a given image into locally coherent pieces. However, such techniques do not always obtain satisfactory results [4] since different parts of the same object may exhibit heterogeneous appearance.

Recently, there has been growing interest in unsupervised image co-segmentation, where the segments are forced to be consistent across a collection of similar images, e.g. [14, 6, 19, 15]. This is a common setting, as many natural image collections contain similar or related objects. For example, spatial and temporal coherence in user photo albums leads to shared entities in the images, photo collections of a particular theme (e.g., "grazing animals") invariably contain shared content, etc. In this multi-image setting, the key idea is to establish relations across images, and obtain consistent segmentations that agree with the segmentation clues provided by all the images together. This formulation turns out to perform much better than single image segmentation methods [17]. However, existing techniques are generally restricted to the setting where the input images must all contain exactly the same set of objects or, in other words, when all input images are similar with to other in terms of object content.

In this paper, we consider the problem of co-segmenting a heterogenous image collection, where each input image may contain an arbitrary subset of the objects of interest. Such image collections are easy to obtain (e.g., from internet image collections). We show that the advantage of co-segmentation still applies in this challenging heterogenous setting, and that a careful formulation yields significant improvements over segmenting each image in isolation.

Co-segmenting a heterogenous collection poses fundamental challenges both in how to establish reliable relations across the images and in how to identify objects that only appear in subsets of the input collection. We propose to address these two issues using the functional maps machinery, which was recently introduced to the vision community by Wang et al. [20]. Unlike traditional image matching techniques which establish correspondences between image pixels/superpixels, functional maps establish maps between functions defined over the images. Since image segmentation can be considered as computing binary segment indicator functions on pixels/superpixels, the functional map framework is particularly suitable for the purpose of image co-segmentation as it provides a handy platform for simultaneously expressing image segmentation and image matching desiderata.

The proposed image co-segmentation framework consists of two stages. The first stage establishes consistent functional maps across the input images. In this stage, building upon the framework of [13] and [20], we introduce

a novel formulation that explicitly models partial similarity across images. Given the optimized consistent functional maps between the images, the second stage optimizes multiple groups of consistent segmentation functions across the image collection. Our novel two-stage approach exhibits a significant improvement over existing techniques on several challenging datasets.

1.1. Related Works

The problem of joint segmentation has attracted a lot of attention recently, starting with the early work by Rother et al. [14], who used color histogram matching to find common objects in a pair of images. Later on, other kinds of features were also utilized to exploit the relationship between image foregrounds, such as SIFT [11], saliency [1], and Gabor features [5]. To address the co-segmentation of multiple images, Joulin et al. formulated the co-segmentation task as a discriminative clustering problem by clustering the image pixels into foreground and background [6]. Vicente and colleagues [19] proposed to extract objects from a group of images by using an object recognition scheme to generate a pool of object-like segmentations, and then selecting the most likely segmentations using a learned pairwise consistency term. In contrast, Chang et al. [1] established an MRF optimization model, by introducing a co-saliency prior as a hint about possible consistent foreground locations. The proposed model was then optimized using graph cut techniques. Rubio et al. proposed a method based on first establishing correspondences between regions in the images, and then estimating the appearance distributions of both the foreground and the background for better joint segmentation [15].

Image co-segmentation with multiple objects has only been explored in the last few years. To handle multiple object classes, Kim et al. [8] model the segmentation task as temperature maximization on anisotropic heat diffusion. The submodular property of the formulation guarantees a constant factor approximation to the optimal solution. Joulin et al. propose an effective energy-based objective that combines a spectral-clustering term with a discriminative one, allowing the objective to be optimized using an efficient expectation-minimization algorithm [7]. Both works can handle multiple object classes; however, they still assume that all objects appear in each image, which is not realistic in many applications. To segment images containing an unknown subset of objects, Kim et al. proposed to alternate between foreground modeling and region assignment steps [9]. The foreground modeling step learns the appearance models of the foregrounds and the background, and the region assignment step is formulated as welfare maximization in a combinatorial auction. Finally, Li et al. generate unknown object-like proposals by ensemble clustering and solve the cosegmentation problem by a multi-label energy minimization [10].

Unlike these methods, our technique provides a principled framework for co-segmenting a heterogenous image collection. We do not pose any constraints on the association between objects and images. Moreover, the segmentations of all classes are optimized simultaneously, obtaining significant improvement over state-of-the-art techniques.

1.2. Notations

Throughout this paper, we use the following convention for linear algebra notations. We use bold face capital characters (e.g., $\boldsymbol{A}, \boldsymbol{B}, \cdots$) to denote matrices, and use bold face lowercase characters (e.g., $\boldsymbol{d}, \boldsymbol{s}, \cdots$) to denote vectors. With $\|\cdot\|_F$ we denote the matrix Frobienius norm, i.e., $\|\boldsymbol{A}\|_F = (\sum_{i,j} a_{ij}^2)^{\frac{1}{2}}$. In contrast, we use $\|\cdot\|_1$ to denote the column-wise 1-norm, i.e., for matrix $\boldsymbol{A} = (\boldsymbol{a}_1, \cdots, \boldsymbol{a}_m)$, $\|\boldsymbol{A}\|_1 = \sum_{i=1}^m \|\boldsymbol{a}_i\|_1$.

2. Problem Statement and Overview

The input to our algorithm is a collection of N related images $\mathcal{I} = \{I_1, \cdots, I_N\}$. The images are related in the sense that each image contains one or multiple objects from an unknown set of classes. Nothing is known about these classes, except their total number M. The output consists of (i) the classification result: a collection of M image sets $\mathcal{C}_k \subset \{1, \cdots, N\}, 1 \leq k \leq M$, collecting the images that contain one object (or more) of each class, and (ii) the corresponding segments $s_{ik}, \forall i \in \mathcal{C}_k$.We represent each s_{ik} as a binary indicator function on image I_i, indicating the location of object(s) of class k in image i, and call these the *segmentation functions*.

2.1. Functional Map Representation

Following the work of Wang et al. [20] we assume that each image $I_i = (\mathcal{P}_i, \mathcal{E}_i)$ is represented by the dual graph of its super-pixel decomposition. We use the normalized cut algorithm [17] to compute the decomposition and set $m = 200$ to be the number of superpixels in all the experiments.

Functional Space and Segmentation Functions. The key concept of the functional map framework is to equip each image I_i with a linear functional space $\overline{\mathcal{F}}_i$. Here we consider $\overline{\mathcal{F}}_i$ to be the space of functions, which are piecewise constant on each super-pixel. Thus, for an image with m super-pixels $\overline{\mathcal{F}}_i \cong \mathbb{R}^m$. Moreover, following [20] we approximate $\overline{\mathcal{F}}_i$ by only considering the subspace $\mathcal{F}_i$ spanned by the first $K = 30$ eigenvectors of the normalized cut Laplacian matrix $\boldsymbol{L}_i$. We use these eigenvectors as the standard basis, and encode each $f \in \mathcal{F}_i$ as a vector of coefficients $\boldsymbol{f} \in \mathbb{R}^K$. Note that in the remainder of this paper, we will project any function in the original space $\overline{f} \in \overline{\mathcal{F}}_i$ into this reduced space $f \in \mathcal{F}_i$.

Functional Maps. A functional map between images I_i and I_j is a linear map $X_{ij} : \mathcal{F}_i \to \mathcal{F}_j$. In the remainder of this paper, we will use bold face $\boldsymbol{X}_{ij}$ to denote the

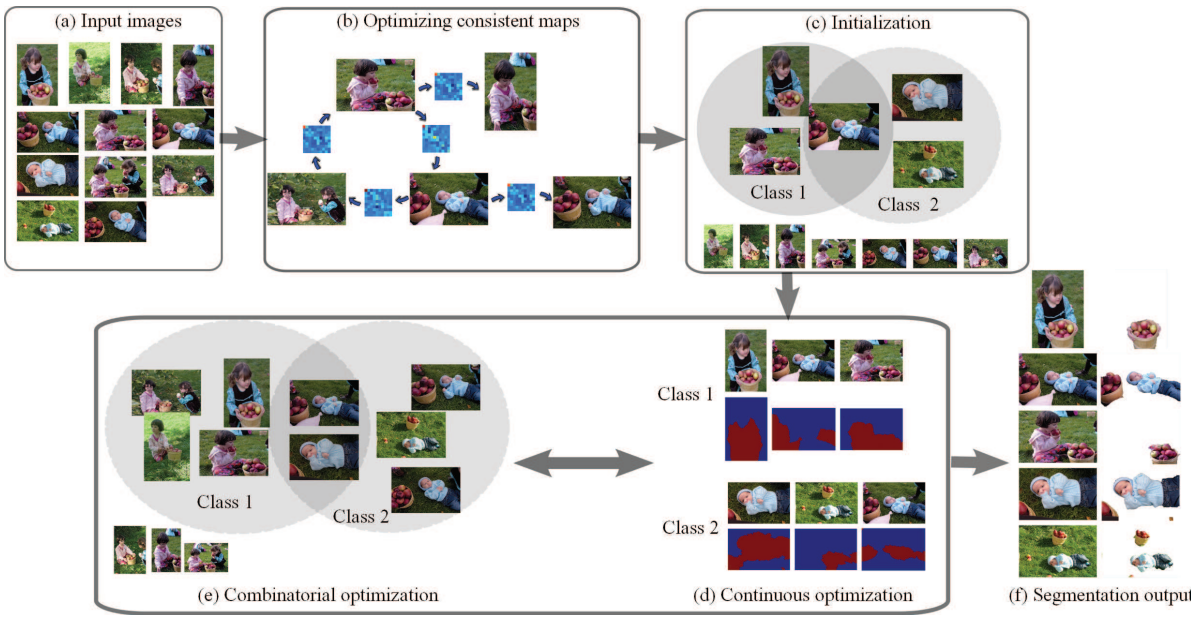


Figure 1. The pipeline of the proposed image co-segmentation framework. Our method begins by computing consistent functional maps between similar input images. Given the optimized functional maps, it extracts for each class an initial seed set of images and the corresponding segmentation functions. It then alternates between jointly optimizing the segmentation functions and using the optimized segmentation functions to refine and extend the image set associated with each class.

matrix representation of X_{ij} in the standard basis associated with $\mathcal{F}_i$ and $\mathcal{F}_j$. As demonstrated in the single-class co-segmentation work [20], high-quality functional maps can be computed using linear constraints on X_{ij} that enforce descriptor preservation across pairs of images, and global consistency constraints on the entire collection of maps reflecting the presence of a shared object.

In the case of multiple object classes, it is not reasonable to expect globally consistent functional maps. Therefore, we adapt the formulation in [20] to the setting where only subsets of maps are consistent, which is significantly more challenging both conceptually and algorithmically. We also show how the segmentation functions can be optimized for and diffused to only appropriate subsets of images.

2.2. Approach overview

The proposed joint image co-segmentation technique consists of two major stages (Fig. 1), as summarized below.

Consistent partial functional maps. We extend the consistent functional map framework of Wang et al. [20] to handle the case where there exist only partial similarities between images in terms of shared objects. We introduce a formulation that utilizes both continuous and discrete latent variables to model partial similarities and show how to optimize the induced objective function via two-level alternating optimizations.

Segmentation function optimization. Given the optimized functional maps between pairs of images, we proceed to extract consistent segmentation functions (multiple per image). This stage alternates between a combinatorial phase, which determines the existence of each object in each image, and a continuous phase to estimate their locations, by jointly optimizing the segmentation functions. Specifically, the combinatorial phase begins with initializing a few seed images for each class, and then gradually augments the images contained in each class during successive iterations. The objective function in the continuous phase considers the saliency of each segmentation function, the mutual exclusiveness of different segmentation functions on the same image, as well as the consistency between segmentation functions of the same class and the optimized functional maps. We show how to effectively optimize the induced segmentation functions via alternating optimization.

3. Consistent Functional Maps Among a Heterogenous Image Collection

In the first stage of our pipeline, we estimate the functional maps X_{ij} between certain pairs of images in our collection, connecting them into a network. Since the number of input images can be large, computing functional maps between all pairs of images is computationally expensive. Furthermore, for dissimilar images, estimated functional maps can be noisy and may pollute the network. Therefore, we connect each image with its k-nearest neighbors (we use $k = 30$) in terms of the GIST descriptor [12] to form a similarity graph $\mathcal{G}$, and only compute functional maps along the edges of $\mathcal{G}$. In the following, we first describe the formulation, and then show how to solve the optimization

problem.

3.1. Formulation

The objective function consists of a pair-wise term, which forces functional maps to align image clues, and a consistency term, which ensures the consistency of functional maps among the network.

Pair-wise objectives. The pair-wise objective is similar to the one in [20], where we force each functional map to agrees with image descriptors and to transfer functions of similar frequencies:

$$f_{\text{pair}} = \|\boldsymbol{X}_{ij}\boldsymbol{D}_i - \boldsymbol{D}_j\|_1 + \lambda\|\boldsymbol{X}_{ij}\Lambda_i - \Lambda_j\boldsymbol{X}_{ij}\|_F^2, \quad (1)$$

where $\boldsymbol{D}_i \in \mathbb{R}^{K\times 367}$ stacks the image descriptors [20] (e.g., color, BoW) from image I_i; Λ_i is the diagonal matrix of eigenvalues of the normalized Laplacian on the super-pixel graphs of images I_i; in our experiments $\lambda = 100$.

Consistency term. To enforce the consistency of functional maps, we introduce $L = 100$ latent functions $\overline{f}_1, \cdots, \overline{f}_L$ that are shared by the input images, and formulate the consistency term so that pair-wise functional maps link corresponding latent functions on each image. In the presence of partial similarity, the technical challenge is to model the fact that each latent function may only appear in a subset of images. To address this issue, we introduce for each image I_i a discrete latent variable $\boldsymbol{z}_i = \{z_{il} \in \{0,1\}, 1 \le l \le L\}$ and a continuous variable $\boldsymbol{Y}_i = (\boldsymbol{y}_{i1}, \cdots, \boldsymbol{y}_{iL})$. The discrete variables encode the association between the latent functions and input images, i.e., $z_{il} = 1$ if and only if $\overline{f}_l$ appears on I_i. The continuous variables encode the latent functions on each image, i.e., $\boldsymbol{y}_{il}$ is the corresponding function of f_l on image I_i. Note that if $z_{il} = 0$, then $\boldsymbol{y}_{il}$ simply corresponds to the zero function $\mathbf{0}$. It is clear that that these two latent variables satisfy the following constraint:

$$\boldsymbol{Y}_i\text{Diag}(\boldsymbol{z}_i) = \boldsymbol{Y}_i. \quad (2)$$

To model the independence among latent functions, we introduce a big matrix $\boldsymbol{Y}$ that stacks the $\boldsymbol{Y}_i$ in a column and require that

$$\boldsymbol{Y}^T\boldsymbol{Y} = I_L. \quad (3)$$

In other words, the vectors that stack each set of corresponding latent functions are orthogonal with each other.

Using these latent functions, we model the consistency of pair-wise functional maps $\boldsymbol{X}_{ij}$ as:

$$\boldsymbol{X}_{ij}\boldsymbol{Y}_i = \boldsymbol{Y}_j\text{Diag}(\boldsymbol{z}_i), \quad (i,j) \in \mathcal{E}. \quad (4)$$

Intuitively, each $\boldsymbol{X}_{ij}$ links shared functions between Y_i and Y_j and maps the remaining functions in $\boldsymbol{Y}_i$ to zero.

The consistency term is formulated to preserve (3) and (4) in the least square sense:

$$\begin{aligned} f_{\text{cons}} =& \mu \sum_{(i,j)\in\mathcal{E}} \|\mathbf{X}_{ij}\mathbf{Y}_i - \mathbf{Y}_j\text{Diag}(\boldsymbol{z}_i)\|^2 \\ &+ \gamma \sum_{i=1}^{N} \|\mathbf{Y}_i - \mathbf{Y}_i\text{Diag}(\boldsymbol{z}_i)\|^2, \end{aligned} \quad (5)$$

where $\mu = 100$ and $\gamma = 10$ for all experiments. To ease the optimization, z_{il} are relaxed so that $0 \le z_{il} \le 1$.

Formulation. Combining f_{pair} and f_{cons}, we write down the following optimization problem for optimizing consistent functional maps

$$\{\boldsymbol{X}_{ij}^\star\} = \arg\min_{\boldsymbol{X}_{ij}} \quad f_{\text{cons}} + \sum_{(i,j)\in\mathcal{E}} f_{\text{pair}} \quad (6)$$

3.2. Optimization

Equation 6 is not convex, however, the special structure in the objective function allows us to effectively optimize it via alternating optimization. In other words, we alternate between optimizing each type of parameter so that in each iteration we solve a much easier sub-optimization problem.

Initializing the variables. We begin by optimizing the functional maps between pairs of images by dropping the consistency term. This amounts to estimating a standard pair-wise functional map, which is convex and can be solved by CVX:

$$\boldsymbol{X}_{ij}^\star = \arg\min_{\boldsymbol{X}_{ij}} \|\boldsymbol{X}_{ij}\boldsymbol{D}_i - \boldsymbol{D}_j\|_1 + \lambda\|\boldsymbol{X}_{ij}\Lambda_i - \Lambda_j\boldsymbol{X}_{ij}\|_F^2. \quad (7)$$

The initial value of $\boldsymbol{z}_i = \mathbf{1}^T$. After that, we fix $\boldsymbol{X}_{ij}$ and $\boldsymbol{z}_i$ to optimize $\boldsymbol{Y}_i$. As described in [20], this amounts to compute the top eigenvectors of a sparse matrix.

Optimizing Latent functions $\mathbf{Y}_i, 1 \le i \le N$. We first fix the indicator vectors $\boldsymbol{z}_i$ and functional maps $\boldsymbol{X}_{ij}$ and optimize the latent functions $\boldsymbol{Y}_i, 1 \le i \le N$. In this case, the objective function is quadratic in $\boldsymbol{Y}_i$, and thus the technical challenge is to enforce the orthornormality constraint $\boldsymbol{Y}^T\boldsymbol{Y} = I_L$. To address this issue, we employ a standard optimization-on-manifold strategy. Specifically, given the current value of $\boldsymbol{Y}$, we seek a displacement of $d\boldsymbol{Y}$ to minimize the objective function. $d\boldsymbol{Y}$ is forced to lie within the tangent plane at $\boldsymbol{Y}$, i.e., it satisfies $\boldsymbol{Y}^T(d\boldsymbol{Y}) = 0$. Since the objective function is quadratic in the variables, this leads to solving a linear system. After obtaining the optimal value of $\boldsymbol{Y} \leftarrow \boldsymbol{Y} + d\boldsymbol{Y}$, we project $\boldsymbol{Y}$ back onto the manifold $\boldsymbol{Y}^T\boldsymbol{Y} = I_L$. This is done by computing SVD of $\boldsymbol{Y} = \boldsymbol{U}\Sigma\boldsymbol{V}^T$ and set $\boldsymbol{Y} \leftarrow \boldsymbol{U}\boldsymbol{V}^T$.

Optimizing indicator vectors $\mathbf{z}_i, 1 \le i \le N$. When the latent functions $\boldsymbol{Y}_i, 1 \le i \le N$ and the functional maps $\boldsymbol{X}_{ij}, (i,j) \in \mathcal{E}$ are fixed, it is easy to see that all indicator variables (i.e., elements of the indicator vectors) are decoupled in the objective function As the objective

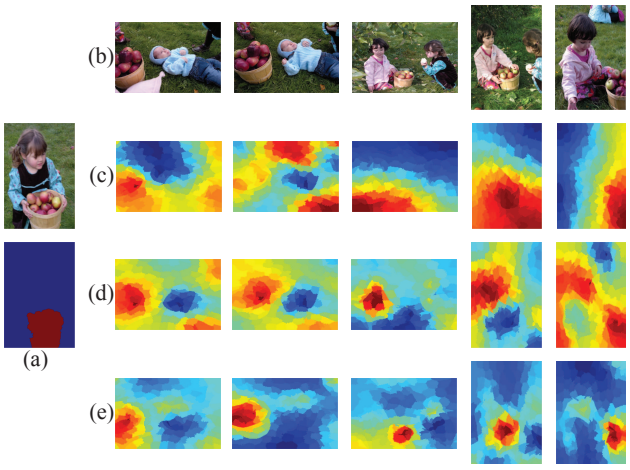


Figure 2. Given an input image in (a), we map its ground truth segmentation function for "apple bucket" to other images in (b). The mapped results are shown in (c) when the maps are optimized independently without any consistency enforced; (d) when the maps are optimized with global consistency as in [20]; (e)when the consistency term Eq. 5 is included. The maps optimized with the proposed consistency term are capable of correctly matching similar parts of other images.

function is quadratic in indicators variables, we can write the optimal value of each indicator variable analytically as

$$z_{il}^{\star} = \arg\min_{0\le z_{il}\le 1} \mu\|\mathbf{y}_{il} - z_{il}\mathbf{y}_{il}\|^2 + \lambda \sum_{j\in\mathcal{N}(i)} \|\mathbf{X}_{ij}\mathbf{y}_{il} - z_{il}\mathbf{y}_{jl}\|^2$$
$$= \max\big(0, \min\big(1, \frac{\mu\|\mathbf{y}_{il}\|^2 + \lambda \sum_{j\in\mathcal{N}(i)} \langle \mathbf{X}_{ij}\mathbf{y}_{il}, \mathbf{y}_{jl}\rangle}{\mu\|\mathbf{y}_{il}\|^2 + \lambda \sum_{j\in\mathcal{N}(i)} \|\mathbf{y}_{jl}\|^2}\big)\big) \tag{8}$$

Optimizing functional maps $\mathbf{X}_{ij}, (i,j) \in \mathcal{E}$. When the latent variables $\boldsymbol{Y}_i, \boldsymbol{z}_i, 1 \le i \le N$ are fixed, we can optimize each pair-wise functional map $\mathbf{X}_{ij}$ independently by solving the following convex optimization problem:

$$\boldsymbol{X}_{ij}^{\star} = \arg\min_{\boldsymbol{X}_{ij}} \|\boldsymbol{X}_{ij}\boldsymbol{D}_i - \boldsymbol{D}_j\|_1 + \lambda\|\boldsymbol{X}_{ij}\Lambda_i - \Lambda_j\boldsymbol{X}_{ij}\|_F^2$$
$$+ \mu\|\boldsymbol{X}_{ij}\boldsymbol{Y}_i - \boldsymbol{Y}_j\mathrm{Diag}(\boldsymbol{z}_i)\|_F^2. \tag{9}$$

Convergence detection. The alternating optimization described above is guaranteed to converge to a local optimal of f. We detect the convergence by checking

$$\max_{(i,j)\in\mathcal{E}} \frac{\|\boldsymbol{X}_{ij} - \boldsymbol{X}_{ij}^{\text{prev}}\|}{\|\boldsymbol{X}_{ij}\|} \le 10^{-3}.$$

Typically, the program converges in 8-10 iterations.

4. Optimizing Consistent Segmentations

In this section, we describe how to compute the association between each image and each class, i.e., $\mathcal{C}_k, 1 \le k \le M$, and the segmentation functions $\boldsymbol{s}_{ik}, i \in \mathcal{C}_k$ of the corresponding objects in each image, by optimizing the following three objectives:

- The segmentation functions should be consistent with the optimized functional maps, i.e.,

$$\boldsymbol{X}_{ij}\boldsymbol{s}_{ik} \approx \boldsymbol{s}_{jk}, \quad (i,j) \in \mathcal{G}, 1 \le k \le M. \tag{10}$$

  Note that $\boldsymbol{s}_{ik} = 0$, if $i \notin \mathcal{C}_k$.

- The segmentation functions should align with sharp edges in each image. As in [20], this is formulated using the normalized cut Laplacian $\boldsymbol{L}$ as minimizing

$$\boldsymbol{s}_{ik}^T\boldsymbol{L}_i\boldsymbol{s}_{ik}, \quad \forall i \in \mathcal{C}_k, 1 \le k \le M. \tag{11}$$

- The segmentation functions for different classes should be mutually exclusive, i.e.,

$$\boldsymbol{s}_{ik}^T\boldsymbol{s}_{ik'} \approx 0, \quad \forall i \in \mathcal{C}_k, 1 \le k \ne k' \le M. \tag{12}$$

Note that the unknowns include both discrete variables, i.e., $\mathcal{C}_k, 1 \le k \le M$ and continuous ones, i.e., the segmentation functions $\boldsymbol{s}_{ik}, i \in \mathcal{C}_k, 1 \le k \le K$. Thus, we deploy an iterative and decoupled optimization strategy. Specifically, we begin by initializing the classes with a small set of highly confident images, and then alternate between optimizing the segmentation functions (Fig.4) and expanding the sets for each object (Fig.3).

4.1. Initialization

To initialize the associations $\mathcal{C}_k$ and segmentation functions $\boldsymbol{s}_{ik}$, we solve a relaxed problem, where we only optimize the mutual exclusiveness of the concatenated segmentation function $\boldsymbol{s}_k = (\boldsymbol{s}_{ik}), i \in \mathcal{C}_k$ of each class, i.e., $\boldsymbol{s}_{ik}\boldsymbol{s}_{ik'}$. In this case, to obtain segmentation functions for each class we minimize the same quadratic form

$$f_{seg} = \frac{1}{|\mathcal{G}|}\sum_{(i,j)\in\mathcal{G}} \|\boldsymbol{X}_{ij}\boldsymbol{s}_{ik} - \boldsymbol{s}_{jk}\|_F^2 + \frac{\gamma}{N}\sum_{i=1}^{N} \boldsymbol{s}_{ik}^T\boldsymbol{L}_i\boldsymbol{s}_{ik}$$
$$= \boldsymbol{s}_k\overline{\boldsymbol{L}}\boldsymbol{s}_k, \tag{13}$$

using the combined Laplacian matrix $\overline{\boldsymbol{L}}$ and setting $\gamma = 10$ in this paper. Thus, a reasonable initialization of the segmentation functions is to set $\boldsymbol{s}_k, 1 \le k \le M$, to be the first M smallest eigenvectors of $\overline{\boldsymbol{L}}$.

Given these initial segmentation functions $\boldsymbol{s}_{ik}, 1 \le i \le N, 1 \le k \le M$, we initialize each set $\mathcal{C}_k$ as

$$\mathcal{C}_k = \{i, \text{ s.t. } \|\boldsymbol{s}_{ik}\| \ge \max_i \|\boldsymbol{s}_i\|/2\}.$$

This tries to select images for which we have high confidence that the corresponding class is present, as if $i \notin \mathcal{C}_k$, then $\boldsymbol{s}_{ik}$ must have a small magnitude.

4.2. Continuous Optimization

Given the fixed associations between the classes and the input images, we optimize the corresponding segmentations $\boldsymbol{s}_{ik}$ via constrained optimization. The objective function consists of three terms. The first term measures the consistency between $\boldsymbol{s}_{jk}$ and $\boldsymbol{X}_{ij}\boldsymbol{s}_{ik}, (i,j) \in \mathcal{E}$, i.e., the induced segmentations from its neighbors:

$$f_{ij}^{\text{cons}} = \|\boldsymbol{X}_{ij}\boldsymbol{s}_{ik} - \boldsymbol{s}_{jk}\|^2. \tag{14}$$

The second term evaluates the mutual exclusiveness of the segmentation functions in different classes on the same image, i.e., asserting they should be orthogonal to each other:

$$f_i^{\text{exclu}} = \sum_{(k,k') \in \{I_i \in \mathcal{C}_k \cap \mathcal{C}_{k'}\}} (\boldsymbol{s}_{ik}^T \boldsymbol{s}_{ik'})^2. \tag{15}$$

The final term evaluates the saliency of each segmentation. In this case, we simply evaluate the normalized cut score in terms of the normalized cut Laplacian $\boldsymbol{L}_i$:

$$f_i^{\text{cut}} = \boldsymbol{s}_{ik}^T \boldsymbol{L}_i \boldsymbol{s}_{ik}. \tag{16}$$

To avoid having all segmenting functions be the zero vector, we include the regularization constraints $\sum_{i \in \mathcal{C}_k} \|\boldsymbol{s}_{ik}\|^2 = |\mathcal{C}_k|, \quad 1 \le k \le K$. Combining these three objective terms, we arrive at the following optimization problem:

$$\begin{aligned} \underset{\boldsymbol{s}_{ik}, i \in \mathcal{C}_k}{\text{minimize}} \quad & \sum_{k=1}^{M} \sum_{(i,j) \in \mathcal{E} \cap (\mathcal{C}_k \times \mathcal{C}_k)} \|\boldsymbol{X}_{ij}\boldsymbol{s}_{ik} - \boldsymbol{s}_{jk}\|^2 \\ & + \gamma \sum_{l \neq k} \sum_{i \in \mathcal{C}_k \cap \mathcal{C}_l} (\boldsymbol{s}_{il}^T \boldsymbol{s}_{ik})^2 + \alpha \sum_{k=1}^{M} \sum_{i \in \mathcal{C}_k} \boldsymbol{s}_{ik}^T \boldsymbol{L}_i \boldsymbol{s}_{ik} \\ \text{subject to} \quad & \sum_{i \in \mathcal{C}_k} \|\boldsymbol{s}_{ik}\|^2 = |\mathcal{C}_k|, \quad 1 \le k \le M. \end{aligned} \tag{17}$$

It is hard to optimize Eq. 17 directly because the term $(\boldsymbol{s}_{ik}\boldsymbol{s}_{il})^2$ is quartic in the segmentation function coefficients. However, the objective functions becomes quadratic if we only optimize the segmentation functions associated with each class. This leads to an alternating optimization procedure. Specifically, at each step, we optimize the segmentation functions associated with class $\mathcal{C}_k$, i.e.,

$$\begin{aligned} \underset{\boldsymbol{s}_{ik}, i \in \mathcal{C}_k}{\text{minimize}} \quad & \sum_{(i,j) \in \mathcal{E} \cap (\mathcal{C}_k \times \mathcal{C}_k)} \|\boldsymbol{X}_{ij}\boldsymbol{s}_{ik} - \boldsymbol{s}_{jk}\|^2 \\ & + \gamma \sum_{i \in \mathcal{C}_l, l \neq k} (\boldsymbol{s}_{il}^T \boldsymbol{s}_{ik})^2 + \alpha \sum_{i \in \mathcal{C}_k} \boldsymbol{s}_{ik}^T \boldsymbol{L}_i \boldsymbol{s}_{ik} \\ \text{subject to} \quad & \sum_{i \in \mathcal{C}_k} \|\boldsymbol{s}_{ik}\|^2 = |\mathcal{C}_k|. \end{aligned} \tag{18}$$

This optimization is performed for each class in order. In practice, we found that the segmentation functions become stable after 4-5 complete iterations.

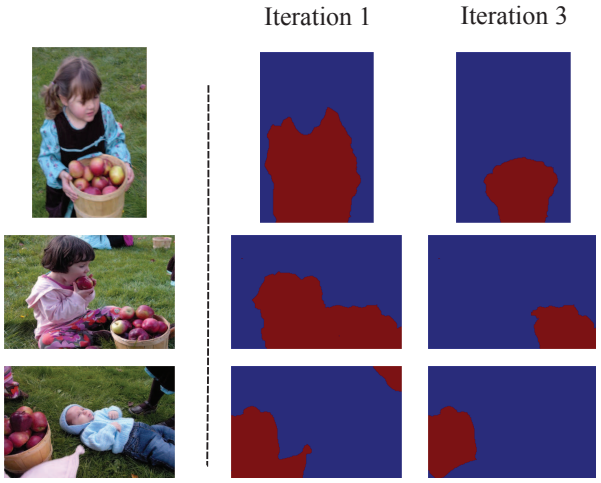


Figure 4. The generated segmentations are updated when more images are included in this object class. This figure shows how the image segmentations are improved as iterations go on.

4.3. Combinatorial Optimization

Given the current segmentation functions $\boldsymbol{s}_{ik}, I_i \in \mathcal{C}_k, 1 \le k \le M$, we proceed to expand $\mathcal{C}_k$ by propagating segmentation functions to other images using the functional maps, and detecting salient segmentations. Specifically, for each class $\mathcal{C}_k$ and for each image $I_i \notin \mathcal{C}_k$, such that there exists an image $I_j \in \mathcal{C}_k$ and $(i,j) \in \mathcal{E}$, we compute the induced segmentation $\boldsymbol{s}_{ik}$ by solving the following constrainted optimization problem

$$\begin{aligned} \underset{\boldsymbol{s}_{ik}}{\text{maximize}} \quad & \frac{1}{|\mathcal{N}(i) \cap \mathcal{C}_k|} \sum_{j \in \mathcal{N}(i) \cap \mathcal{C}_k} (\boldsymbol{s}_{ik}^T \boldsymbol{X}_{ji} \boldsymbol{s}_{jk})^2 \\ & - \gamma \sum_{l \neq k, i \in \mathcal{C}_l} (\boldsymbol{s}_{ik}^T \boldsymbol{s}_{il})^2 - \alpha \boldsymbol{s}_{ik}^T \boldsymbol{L}_i \boldsymbol{s}_{ik} \end{aligned} \tag{19}$$

$$\text{subject to} \quad \|\boldsymbol{s}_{ik}\|^2 = 1. \tag{20}$$

The first term in Eq. 19 prioritizes the agreement of $\mathbf{s}_{ik}$ with the induced segmentation functions $\boldsymbol{X}_{ji}\boldsymbol{s}_{jk}$ from its neighboring images. The second term ensures that $\boldsymbol{s}_{ik}$ is orthogonal to existing segmentation functions of other classes on image i. The third term measures the saliency of $\boldsymbol{s}_{ik}$, with respect to the normalized Laplacian matrix $\boldsymbol{L}_i$. Since the objective function in quadratic in $\boldsymbol{s}_{ik}$, its optimal value can be obtained using the standard eigen-decomposition procedure.

After computing the segmentation function $\boldsymbol{s}_{ik}$, we compute the saliency score $\boldsymbol{s}_{ik}^T \boldsymbol{L}_i \boldsymbol{s}_{ik}$ (agreement with normalized cuts). We then include I_i into $\mathcal{C}_k$ if

$$\boldsymbol{s}_{ik}^T \boldsymbol{L}_i \boldsymbol{s}_{ik} < \epsilon \max_{j \in \mathcal{C}_k} \frac{\boldsymbol{s}_{jk}^T \boldsymbol{L}_j \boldsymbol{s}_{jk}}{\boldsymbol{s}_{jk}^T \boldsymbol{s}_{jk}},$$

where we choose a conservative value $\epsilon = 1/2$ to ensure that we only include the most salient images.

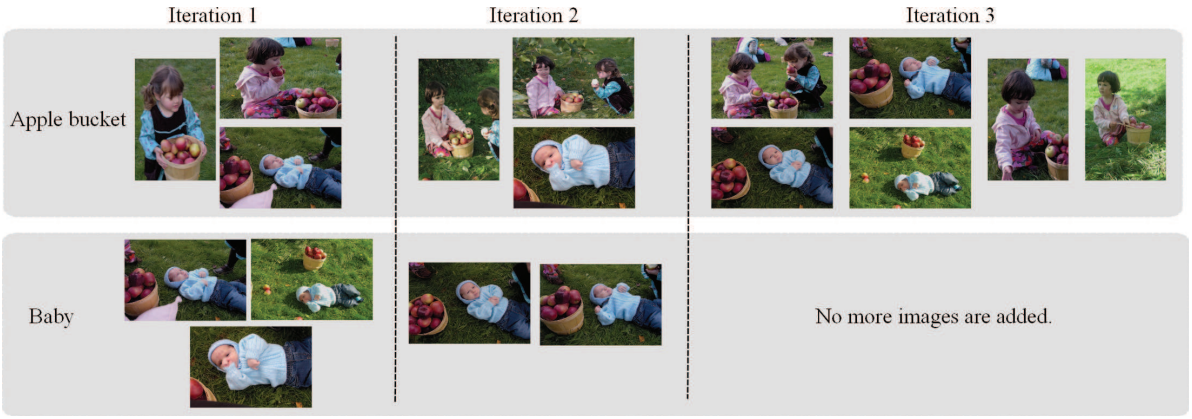


Figure 3. The segmentation propagation process on Flickr dataset. As iteration goes on, more images are included with the same foreground object ("apple bucket" or "baby" in this example).

5. Experiments

5.1. Experiments on MSRC Dataset

We first evaluate our proposed method on the co-segmentation dataset MSRC [18]. It includes 591 pixelwise labeled images in 23 object classes with one object per class. Images in each class contain a common object with the similar appearance, e.g., cow, dog, etc. This is a standard binary segmentation setting, therefore, many existing single-class co-segmentation algorithms are applicable. Table 1 gives a quantitative comparison with [7, 8, 11], and the same classes are selected as reported in [7]. [7] is designed for multi-class segmentation and [8] and [11] are state-of-the-art foreground-background cosegmentation methods. All methods are unsupervised except for knowing the total number of objects. The performance is measured by the intersection-over-union score which is standard in PASCAL challenges.

Our method is significantly better than the state-of-the-art methods in most of the cases. It is interesting to note that our method works best for natural objects, such as "Cat", "Cow", and "Sheep" despite their high appearance variability. Our algorithm performs worse for images with very cluttered background ("Face"). The lower accuracy for "Bike" and "Chair" is caused by the coarse superpixels.

class	N	[7]	[8]	[11]	Ours
Bike	30	43.3	29.9	42.8	**51.2**
Bird	30	47.7	29.9	-	**55.7**
Car	30	59.7	37.1	52.5	**72.9**
Cat	24	31.9	24.4	5.6	**65.9**
Chair	30	39.6	28.7	39.4	**46.5**
Cow	30	52.7	33.5	26.1	**68.4**
Dog	30	41.8	33.0	-	**55.8**
Face	30	70.0	33.2	40.8	60.9
Flower	30	51.9	40.2	-	**67.2**
House	30	51.0	32.2	66.4	56.6
Plane	30	21.6	25.1	33.4	**52.2**
Sheep	30	66.3	60.8	45.7	**72.2**
Sign	30	58.9	43.2	-	**59.1**
Tree	30	67.0	61.2	55.9	62.0

Table 1. Performance of binary segmentation on MSRC.

5.2. Experiments on FLickr Dataset

We then evaluated our proposed method on the public multi-class image dataset Flickr [9]. This dataset consists of 14 groups, where each group contains between 10 and 20 images along with groundtruth pixel-level annotations. We compare our method with other state-of-the-art methods, including [9, 8, 6, 16] and summarize the comparison in Table 2. For [9], an unsupervised version is applied for a fair comparison. [8], [6] and [16] are applied to each subgroup of images which share the same foregrounds. On the other hand, our algorithm is applied to the entire dataset in a completely unsupervised way. In the unsupervised setting, after obtaining the segmentation functions for M different clusters, we need to find the correspondences between each cluster and each ground truth object. We pick the one-to-one matching which maximizes the average accuracy. As can be seen in Table 2, for image collections with irregularly appearing objects, our algorithm can significantly improve the performance in most of the classes.

5.3. Experiments on PASCAL-multi Dataset

Besides the standard benchmark datasets, we create a more challenging multi-class dataset ("PASCAL-multi") based on PASCAL VOC 2012 dataset [3]. Given a pre-selected set of class labels, a group of images is retrieved from the PASCAL dataset such that each image only contains a subset of the pre-selected labels. This can ensure the pre-selected classes are the only re-occurring object classes in the images. Images with foreground object smaller than 0.5% of the total image area are discarded as these objects are not salient. This dataset is much more challenging than the Flickr dataset in §5.2 due to its larger size and the larger

class	N	M	[9]	[8]	[6]	[16]	Ours
Apple	20	6	40.9	32.6	24.8	25.6	**46.6**
baseball	18	5	31.0	31.3	19.2	16.1	**50.3**
Butterfly	18	8	29.8	32.4	29.5	10.7	**54.7**
Cheetah	20	5	32.1	40.1	50.9	41.9	**62.1**
Cow	20	5	35.6	43.8	25.0	27.2	38.5
Dog	20	4	34.5	35.0	32.0	30.6	**53.8**
Dolphin	18	3	34.0	47.4	37.2	30.1	**61.2**
Fishing	18	5	20.3	27.2	19.8	18.3	**46.8**
Gorilla	18	4	41.0	38.8	41.1	28.1	**47.8**
Liberty	18	4	31.5	41.2	44.6	32.1	**58.2**
Parrot	18	5	29.9	36.5	35.0	26.6	**54.1**
Stonehenge	20	5	35.3	49.3	47.0	32.6	**54.6**
Swan	20	3	17.1	18.4	14.3	16.3	**46.5**
Thinker	17	4	25.6	34.4	27.6	15.7	**68.6**
Average	-	-	31.3	36.3	32.0	25.1	**53.1**

Table 2. Performance comparison on the Flickr data set.

class	imgNum	Ncut [17]	[2]	Ours
Bike + person	248	27.3	30.5	**40.1**
Boat + person	260	29.3	32.6	**44.6**
bottle + dining table	90	37.8	39.5	**47.6**
bus + car	195	36.3	39.4	**49.2**
bus + person	243	38.9	41.3	**55.5**
chair + dining table	134	32.3	30.8	**40.3**
chair + potted plant	115	19.7	19.7	**22.3**
cow + person	263	30.5	33.5	**45.0**
dog + sofa	217	44.6	42.2	**49.6**
horse + person	276	27.3	30.8	**42.1**
potted plant + sofa	119	37.4	37.5	**40.7**

Table 3. Performance comparison on the PASCAL-multi data set.

object appearance variability.

We compare our framework with baseline methods [17] and [2]. The number of foreground objects in each image is provided as a prior for these two baseline methods. The results are shown in Table 3; we can see that our method is very robust when dealing with larger dataset and when the foreground objects are not quite similar.

6. Conclusion

In this paper we have proposed a framework for multi-class joint image segmentation. Unlike the traditional image co-segmentation task which only has one foreground object, we deal with images containing a large number of objects, with a variable number of objects from multiple classes appearing in each image. We have shown an approach to this problem using the framework of functional maps and demonstrated how it can be adapted to reflect partial similarity between images. Based on the optimized maps, segmentation functions for multiple groups emerge from the image network, and the group assignment is updated through a combination of continuous and discrete optimization steps.

This framework is completely unsupervised, and the object existence and segmentation are obtained simultaneously. It is straightforward to add supervision information, such as image labels or ground truth segmentations of a few images, but we leave that as future work.

7. Acknowledgement

Support from ONR MURI N00014-13-1-0341, NSF grants IIS 1016324, CNS 0832820, Marie Curie CIG-334283-HRGP, two Google faculty research awards, a gift from HTC Corporation, a grant from the Max Planck Center for Visual Computing and Communications, and a CNRS chaire d'excellence is gratefully acknowledged.

References

[1] K.-Y. Chang, T.-L. Liu, and S.-H. Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*, 2011. 2

[2] T. Cour, F. Benezit, and J. Shi. Spectral segmentation with multiscale graph decomposition. In *CVPR*, 2005. 8

[3] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman. The PASCAL visual object classes (VOC) challenge. *IJCV*, 88:303–338, 2010. 7

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *International Journal of Computer Vision*, 59, 2004. 1

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

[6] A. Joulin, F. Bach, and J. Ponce. Discriminative clustering for image co-segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2010. 1, 2, 7, 8

[7] A. Joulin, F. Bach, and J. Ponce. Multi-Class cosegmentation. In *CVPR*, 2012. 2, 7

[8] G. Kim, E. Xing, L. Fei-fei, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011. 2, 7, 8

[9] G. Kim and E. P. Xing. On multiple foreground cosegmentation. In *CVPR*, 2012. 2, 7, 8

[10] H. Li, F. Meng, Q. Wu, and B. Luo. Unsupervised multi-class region co-segmentation via ensemble clustering and energy minimization. *Circuits and Systems for Video Technology, IEEE Transactions on*, 2013. 2

[11] L. Mukherjee, V. Singh, and C. R. Dryer. Scale invariant cosegmentation for image groups. In *CVPR*, 2011. 2, 7

[12] A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001. 3

[13] M. Ovsjanikov, M. Ben-Chen, J. Solomon, A. Butscher, and L. Guibas. Functional maps: A flexible representation of maps between shapes. In *SIGGRAPH*, 2012. 1

[14] C. Rother, V. Kolmogorov, T. Minka, and A. Blake. Cosegmentation of image pairs by histogram matching - incorporating a global constraint into MRFs. In *CVPR*, 2006. 1, 2

[15] J. C. Rubio, J. Serrat, A. Lopez, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, 2012. 1, 2

[16] B. Russel, A. Efros, J. Sivic, W. T. Freeman, and A. Zisserman. Using multiple segmentations to discover objects and their extent in image collection. In *CVPR*, 2006. 7, 8

[17] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE TPAMI*, 22:888–905, 2000. 1, 2, 8

[18] J. Shotton, J. Winn, C. Rother, and A. Criminisi. TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In *ECCV*, 2006. 7

[19] S. Vicente, C. Rother, and V. Kolmogorov. Object cosegmentation. In *CVPR*, 2011. 1, 2

[20] F. Wang, Q. Huang, and L. Guibas. Image co-segmentation via consistent functional maps. In *ICCV*, 2013. 1, 2, 3, 4, 5

Supplemental Material for "Unsupervised Multi-Class Joint Image Segmentation"

Fan Wang
Stanford University
fanw@stanford.edu

Qixing Huang
Stanford University
huangqx@stanford.edu

Maks Ovsjanikov
LIX, Ecole Polytechnique
maks@lix.polytechnique.fr

Leonidas J. Guibas
Stanford University
guibas@cs.stanford.edu

1. More Segmentation Examples

Five examples are shown for each class in MSRC and Flickr, and ten examples are shown for each class in PASCAL data set.

1.1. Examples in MSRC data Set

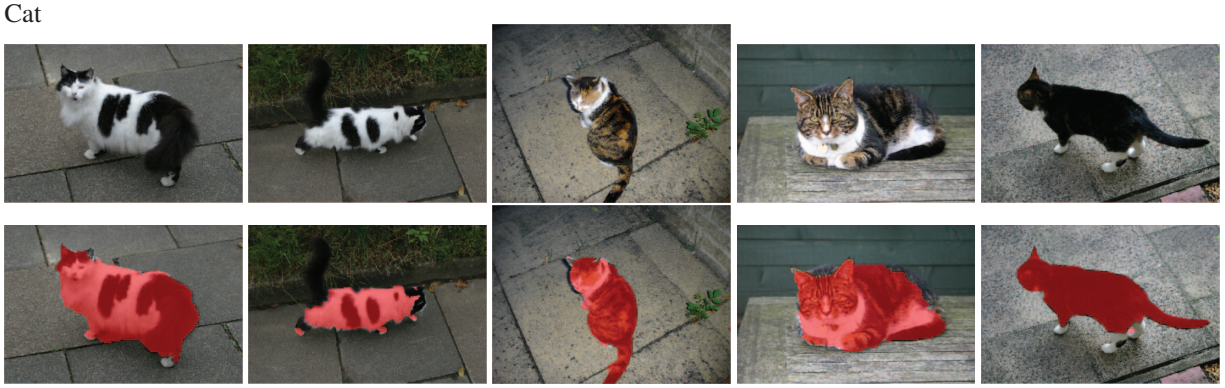
All foreground objects are colored in red.

Bike

Bird

Car

Cat



Chair

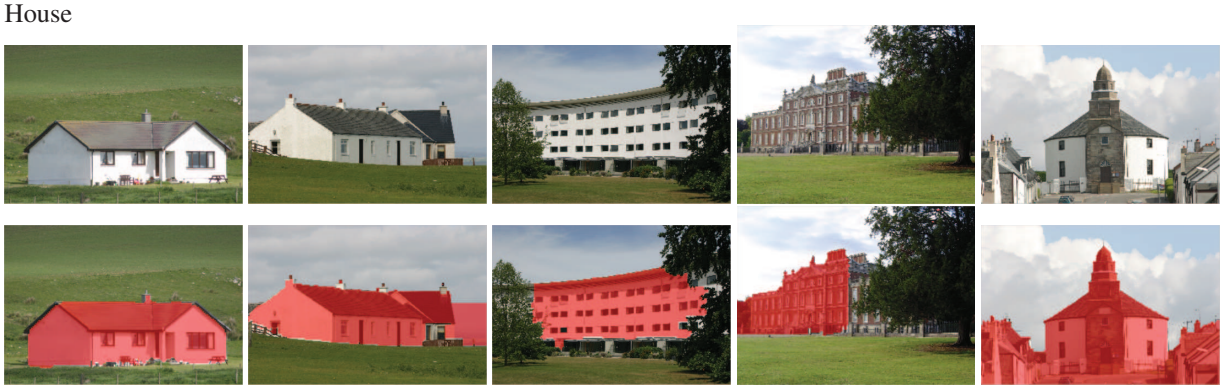
Cow

Dog

Face

Flower

House



Plane

Sheep

Sign

Tree

1.2. Examples in Flickr Data Set

Apple+picking (red: apple bucket; magenta: girl in red; yellow: girl in blue; green: baby; cyan: pumpkin.)

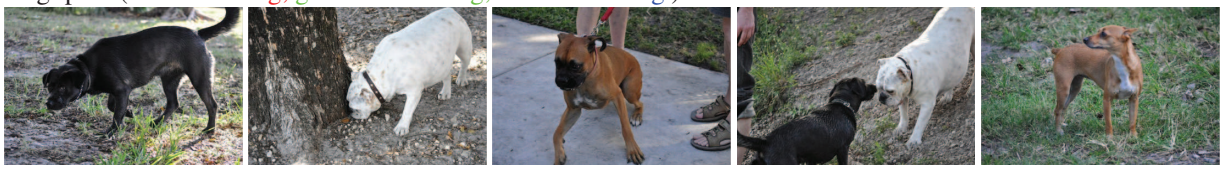
Baseball+kids (green: boy in black; blue: boy in grey; yellow: coach.)

Butterfly+blossom (green: butterfly in orange; yellow: butterfly in yellow; cyan: red flower.)

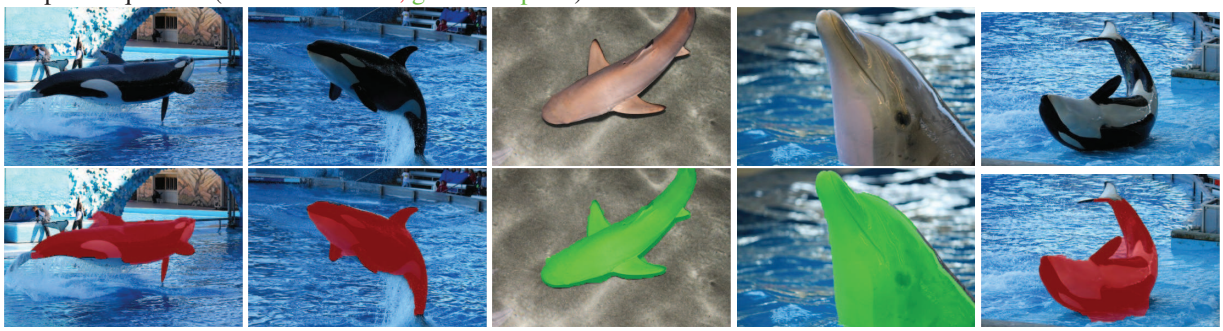
Cheetah+safari (red: cheetah; yellow: lion; magenta: monkey.)

Cow+pasture (red: black cow; green: brown cow; blue: man in blue.)

Dog+park (red: black dog; green: brown dog; blue: white dog.)

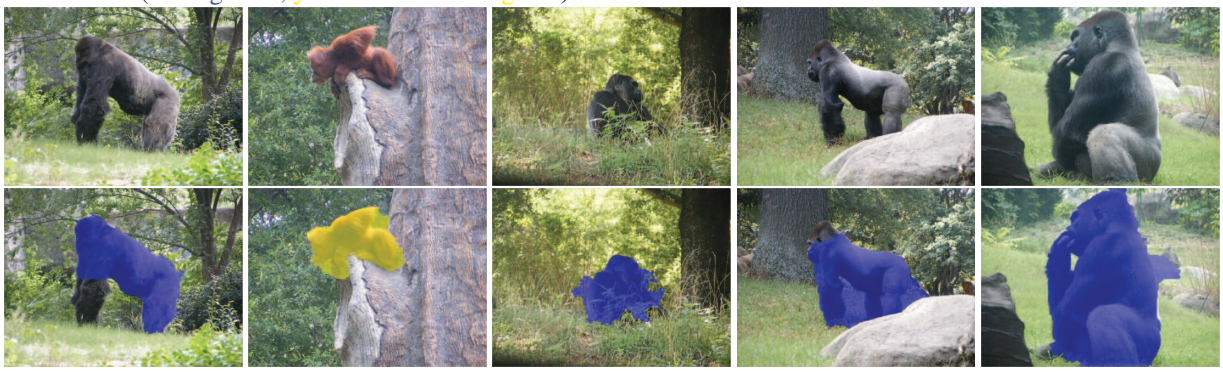


Dolphin+aquarium (red: killer whale; green: dolphin.)

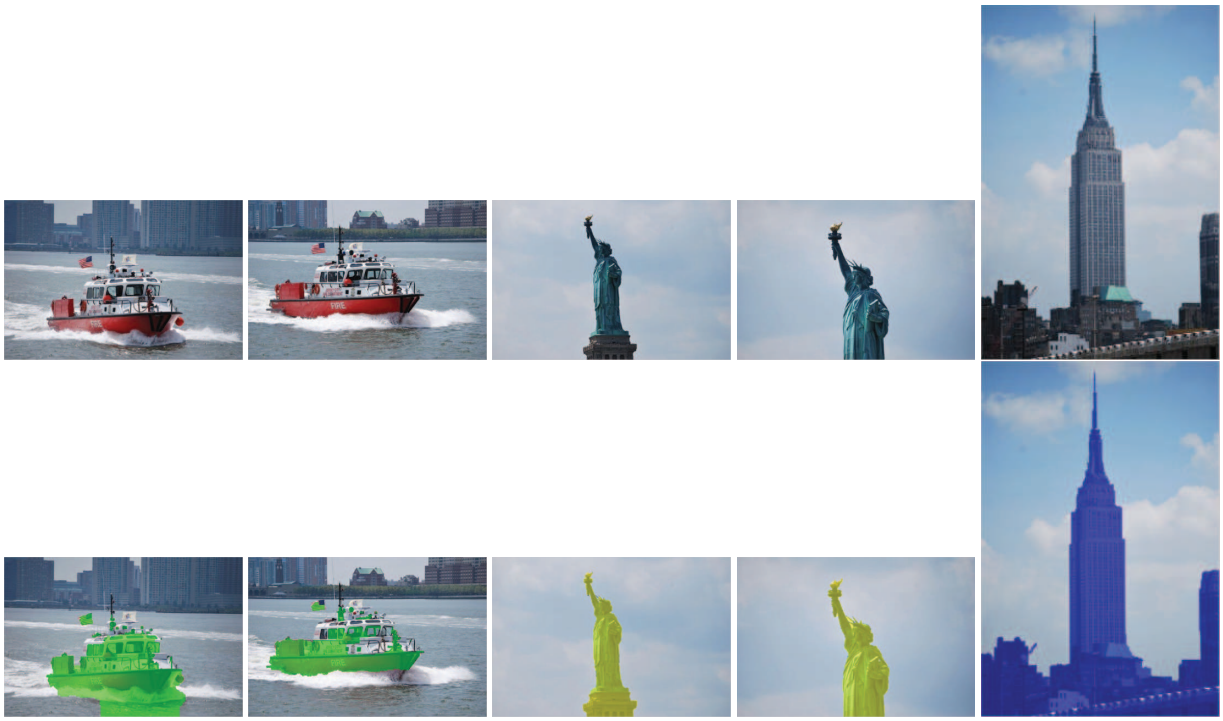


Fishing+alaska (blue: man in white; green: man in gray;magenta: woman in gray; yellow: salmon.)

Gorilla+zoo (blue: gorilla; yellow: brown orangutan)



Liberty+statue (blue: empire state building; green: red boat; yellow: liberty statue.)

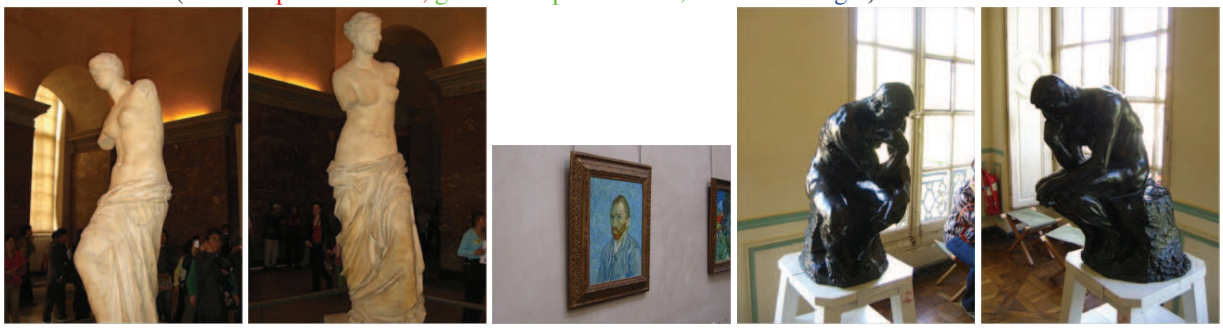


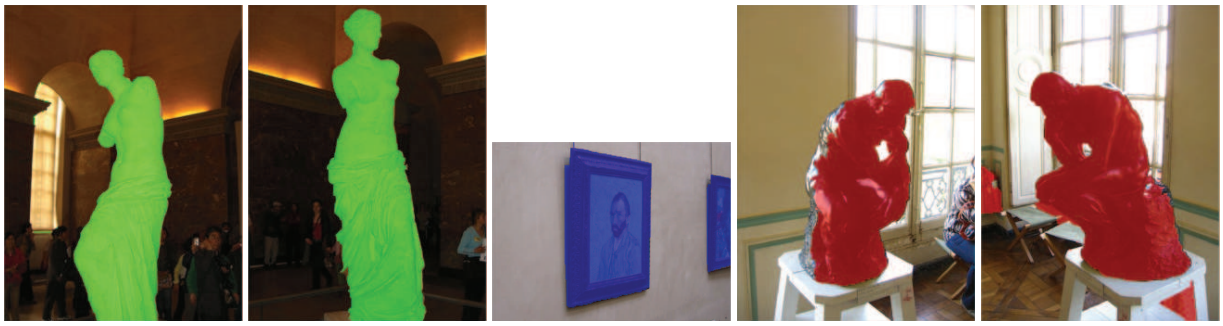
Parrot+zoo (red: hand; green: parrot in green; blue: parrot in red.)

Stonehenge (blue: cow in white; yellow: person; magenta: stonehenge.)

Swan+zoo (blue: gray swan; green: black swan.)

Thinker+Rodin (red: sculpture Thinker; green: sculpture Venus; blue: Van Gogh.)

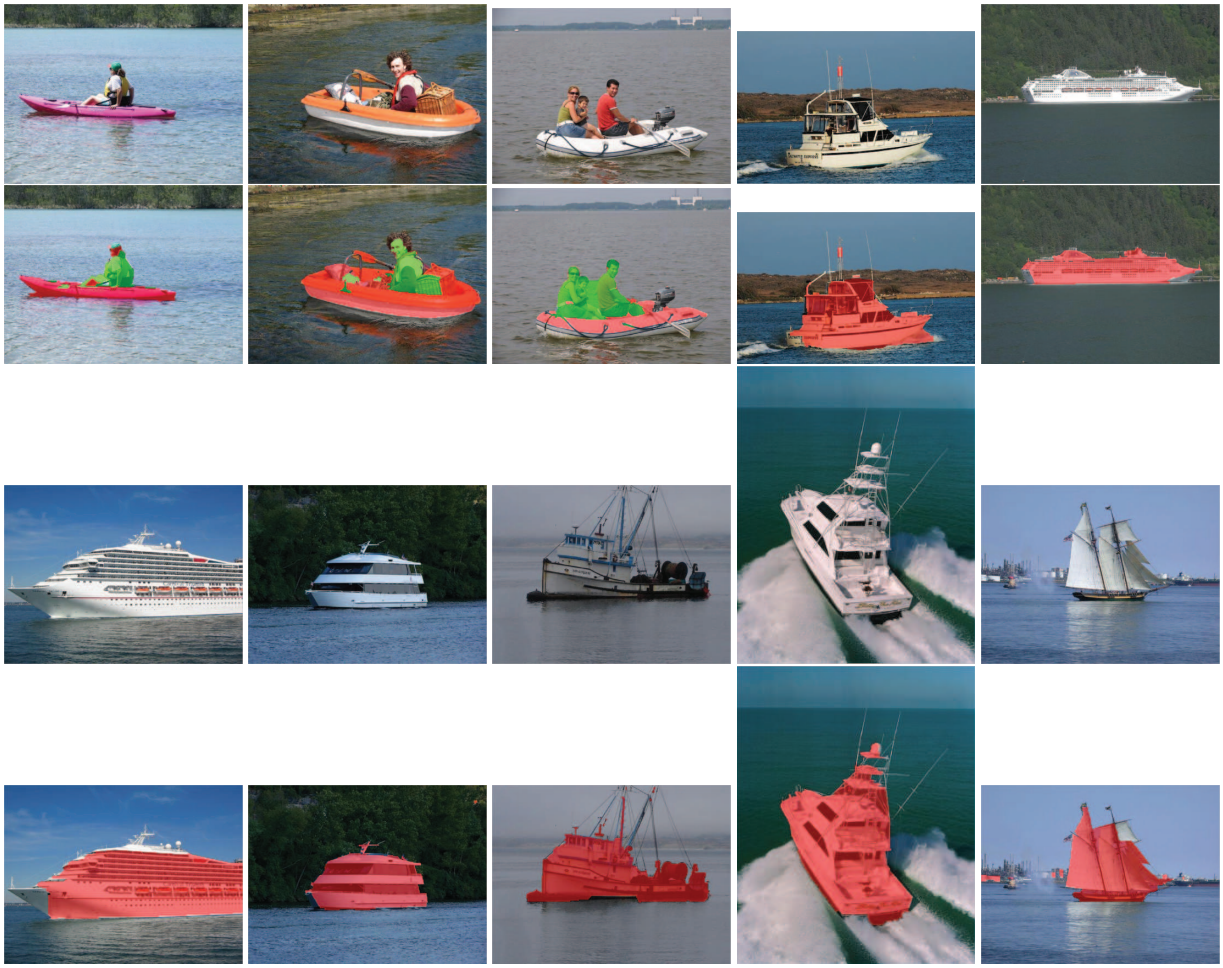




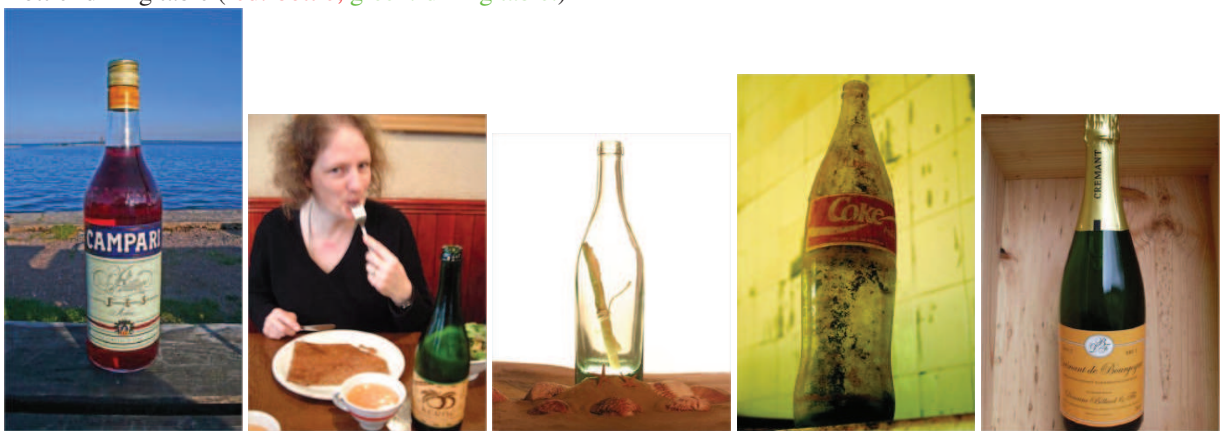
1.3. Examples in PASCAL Data Set

Bike+person (red: bike; green: person.)

Boat+person (red: boat; green: person.)



Bottle+dining table (red: bottle; green: dining table.)

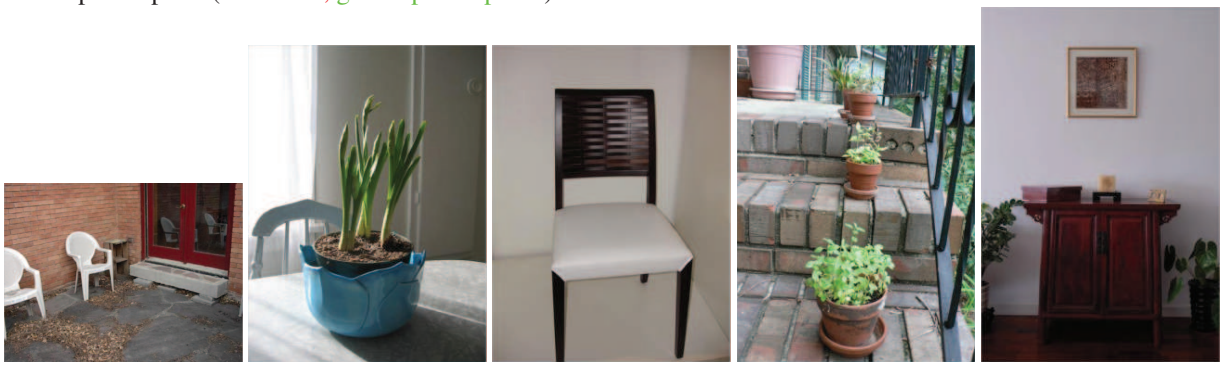


Bus+car (red: bus; green: car.)

Bus+person (red: bus; green: person.)

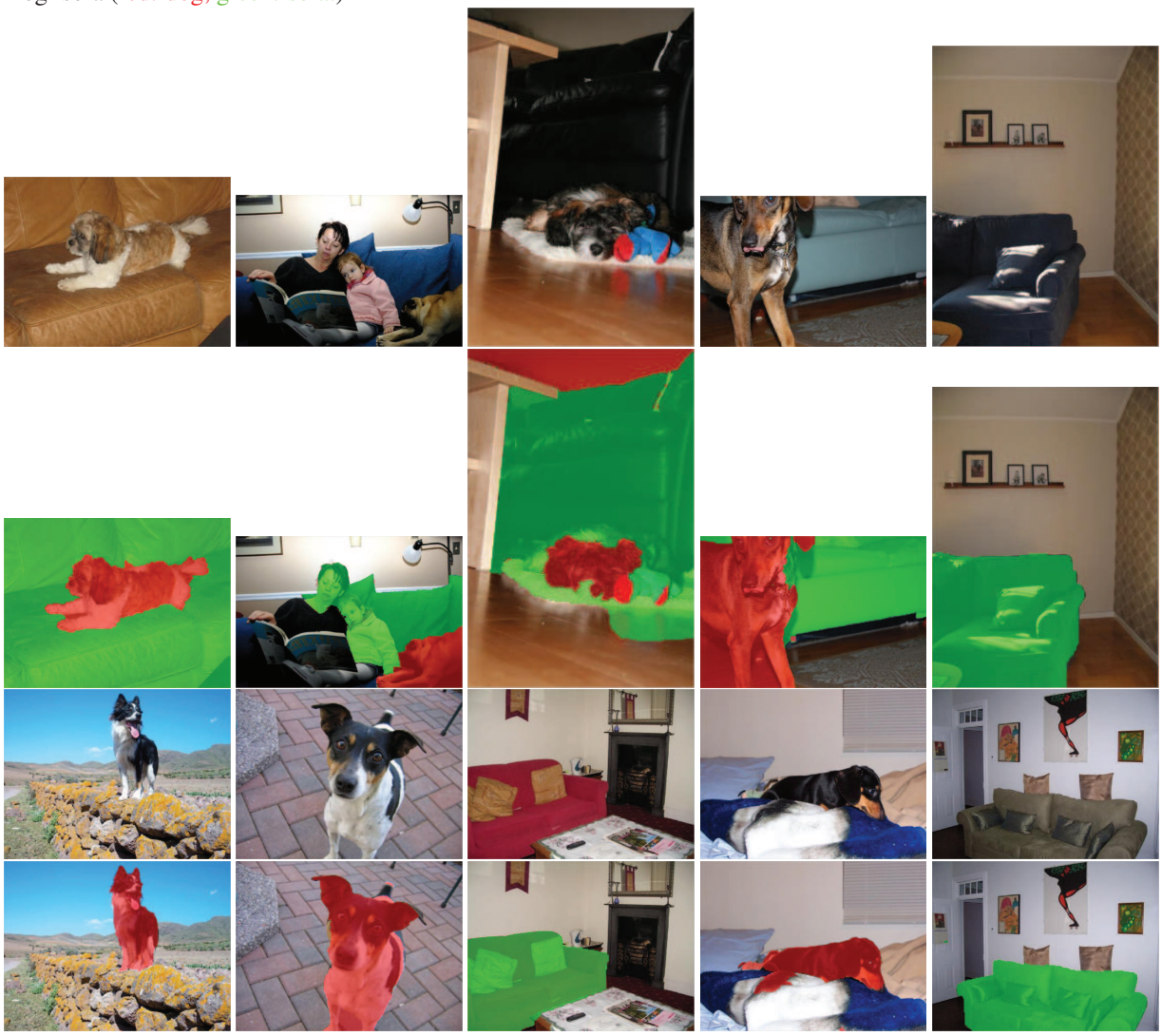
Chair+dining table (red: chair; green: dining table.)

Chair+potted plant (red: chair; green: potted plant.)

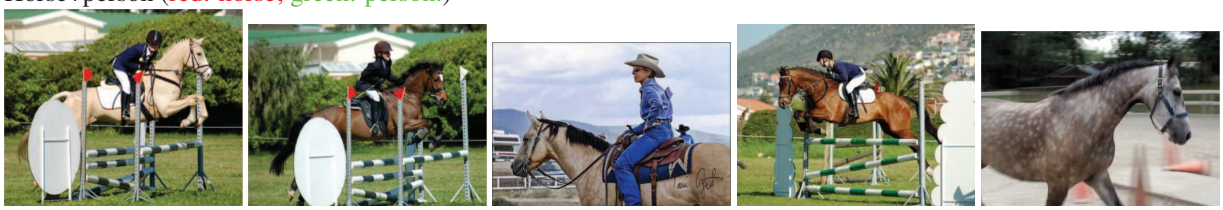


Cow+person (red: cow; green: person.)

Dog+sofa (red: dog; green: sofa.)



Horse+person (red: horse; green: person.)



Sofa+potted plant (red: sofa; green: potted plant.)