Co-attention CNNs for Unsupervised Object Co-segmentation

Kuang-Jui Hsu[1,2], **Yen-Yu Lin**[1], **Yung-Yu Chuang**[1,2]
[1]Academia Sinica, Taiwan
[2]National Taiwan University, Taiwan
kjhsu@iis.sinica.edu.tw, yylin@citi.sinica.edu.tw, cyy@csie.ntu.edu.tw

Abstract

Object co-segmentation aims to segment the common objects in images. This paper presents a CNN-based method that is unsupervised and end-to-end trainable to better solve this task. Our method is unsupervised in the sense that it does not require any training data in the form of object masks but merely a set of images jointly covering objects of a specific class. Our method comprises two collaborative CNN modules, *a feature extractor* and *a co-attention map generator*. The former module extracts the features of the estimated objects and backgrounds, and is derived based on the proposed *co-attention loss*, which minimizes inter-image object discrepancy while maximizing intra-image figure-ground separation. The latter module is learned to generate co-attention maps by which the estimated figure-ground segmentation can better fit the former module. Besides the co-attention loss, the *mask loss* is developed to retain the whole objects and remove noises. Experiments show that our method achieves superior results, even outperforming the state-of-the-art, supervised methods.

1 Introduction

Object co-segmentation simulates human visual systems to search for the common objects repetitively appearing in images. It was introduced in [Rother *et al.*, 2006] to address the difficulties of single-image object segmentation. It leverages not only intra-image appearance but also inter-image object co-occurrence to compensate for the absence of supervisory information. As an important component of image analysis, object co-segmentation is essential to various computer vision and AI applications, such as image matching [Chen *et al.*, 2015], semantic segmentation [Shen *et al.*, 2017], object skeletonization [Jerripothula *et al.*, 2017] and 3D reconstruction [Mustafa and Hilton, 2017].

Engineered features, such as SIFT [Lowe, 2004], HOG [Dalal and Triggs, 2005] and texton [Shotton *et al.*, 2009], are widely used in conventional co-segmentation methods, e.g., [Wang *et al.*, 2017; Joulin *et al.*, 2012; Lee *et al.*, 2015; Tao *et al.*, 2017], to cope with intra-class variations and background clutters. These features are designed in advance. They are not optimized for the given images for co-segmentation, and may lead to sub-optimal performances. *Convolutional neural networks* (CNNs) [Krizhevsky *et al.*, 2012] have demonstrated effectiveness in joint visual feature extraction and nonlinear classifier learning. Yuan et al. [Yuan *et al.*, 2017] proposed a CNN-based supervised method, which learns the mapping between images and the corresponding masks, for object co-segmentation. They achieved the state-of-the-art results by substituting the features learned by CNNs for engineered features. However, their method requires additional training data in the form of object masks for learning the CNN model. As discussed in other applications, such as semantic segmentation [Hsu *et al.*, 2014] or top-down saliency detection [Hsu *et al.*, 2017], these object masks are usually manually drawn or delineated by tools with intensive user interaction. The heavy annotation cost of training data makes their method less practical. Furthermore, the unsupervised nature of co-segmentation is also violated.

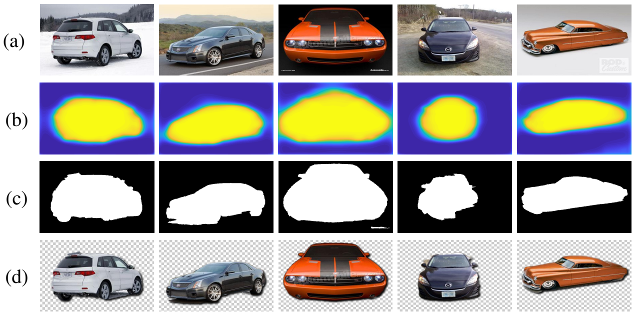

Figure 1: (a) The images for co-segmentation. (b) The estimated object maps by optimizing the co-attention loss. (c) The selected object proposals by using the mask loss. (d) Our co-segmentation results by considering the two losses simultaneously.

This paper presents an unsupervised CNN-based method for co-segmentation that makes a good compromise between the performance and data annotation cost. Specifically, we aim at co-segmenting images covering objects of a specific category without additional data annotations. This task is often referred to as unsupervised co-segmentation in the literature [Chang *et al.*, 2011; Rubio *et al.*, 2012; Rubinstein *et al.*, 2013; Jerripothula *et al.*, 2016; Tao *et al.*, 2017; Li *et al.*, 2018], though it can be also considered weakly supervised since we know all images contain objects of the same class. In this paper, we follow the previous work [Chang *et al.*, 2011; Rubio *et al.*, 2012; Rubinstein *et al.*, 2013; Jerripothula *et al.*, 2016; Tao *et al.*, 2017; Li *et al.*, 2018], and term this task unsupervised co-segmentation. Our method does not rely on training data in form of object masks, and can improve co-segmentation via using the features extracted by CNNs.

To this end, we develop the *co-attention loss* to derive a CNN model by enhancing the similarity among the estimated objects across images while enforcing the figure-ground distinctness in each image. Our model comprises two CNN modules, i.e., *a co-attention map generator* and *a feature extractor*, as shown in Figure 2. The generator compiles a heat map for the object in each image to estimate its figure-ground segmentation. The extractor computes the features of the estimated objects and backgrounds to minimize the co-attention loss. Through backpropagation, the generator is learned to compile high-quality object maps with which the resultant figure-ground segmentation can best optimize the co-attention loss. In this way, our model is end-to-end trainable and can carry out unsupervised object co-segmentation. For further enhancement, we develop the *mask loss*, which can refine the yielded object maps by preserving the whole objects and removing the noises. Figure 1 shows an example of the co-segmentation results inferred by our method.

To the best of our knowledge, this work is the first attempt to develop an unsupervised and end-to-end trainable CNN model for object co-segmentation. Compared with unsupervised conventional methods [Wang *et al.*, 2017; Joulin *et al.*, 2012; Lee *et al.*, 2015; Tao *et al.*, 2017] and the supervised CNN-based method [Yuan *et al.*, 2017], our method can enjoy the boosted performance empowered by deep CNN features and does not suffer from the high annotation cost in labeling object masks as training data. Our method is evaluated on three benchmarks for co-segmentation, *the Internet dataset* [Rubinstein *et al.*, 2013], *the iCoseg dataset* [Batra *et al.*, 2010], and *the PASCAL-VOC dataset* [Faktor and Irani, 2013]. It remarkably outperforms the state-of-the-art unsupervised and supervised methods.

2 Related Work

The literature related to our work is discussed in this section.

2.1 Object Co-segmentation

According to [Tao *et al.*, 2017], conventional researches on object co-segmentation can be divided into two categories, namely the *graph-based* [Chang *et al.*, 2011; Rubio *et al.*, 2012; Chang and Wang, 2015; Jerripothula *et al.*, 2016; Quan *et al.*, 2016; Wang *et al.*, 2017; Li *et al.*, 2018] and the *clustering-based* [Joulin *et al.*, 2010; Kim *et al.*, 2011; Joulin *et al.*, 2012; Lee *et al.*, 2015; Tao *et al.*, 2017] methods. The former methods adopt a structure model to capture the relationship between instances from different images, and utilize the information shared cross images to jointly select the most similar instances as the common objects. The latter methods assume that the pixels or superpixels in the common objects can be grouped together well. Thus, they formulate co-segmentation as a clustering problem to search for the common objects. In these graph-based and clustering-based methods, engineered features, e.g., SIFT, HOG, and texton, are often used for instance representation. The features are pre-designed instead of optimized for the input images. In contrast, our method adaptively learns the CNN features conditional on the given images. It can better cope with the intra-class variations and background clutters, leading to a higher performance.

To improve the performance of co-segmentation, Sun and Ponce [Sun and Ponce, 2016] further explored additional background images to help detect discriminative object parts. Yuan et al. [Yuan *et al.*, 2017] recently proposed a method integrating *conditional random fields* (CRFs) into CNNs to jointly learn the features and search the common objects. Despite the great performance, their method intensely relies on a large number of training object masks. It reduces the applicability of their method to unseen images. Instead, the proposed method does not require additional background images or any training data but merely a set of images for co-segmentation. It can adapt itself to any unseen images in an unsupervised manner. Therefore, the proposed method has better generalization than the supervised method [Yuan *et al.*, 2017], and even outperforms it based on the developed co-attention and mask losses.

2.2 Unsupervised CNN for Image Correspondence

CNNs have been applied in an unsupervised fasion to a few tasks related to image correspondence, such as optical flow [Yu *et al.*, 2016; Ren *et al.*, 2017; Meister *et al.*, 2018] and stereo matching [Godard *et al.*, 2017; Zhou *et al.*, 2017]. The common goal of these tasks is to find the cross-image correspondences of all pixels. The input images are typically adjacent video frames or stereo pairs of the same scene. The adopted objective functions are often based on brightness and cycle consistency. Namely, all matched pixels need to have similar colors or appearances, and the correspondences generated from different image perspectives should be consistent. There are three major differences between these tasks and object co-segmentation. First, co-segmentation often considers objects of the same category, instead of the same instance. Thus, the brightness consistency may not hold. Second, co-segmentation identifies the region correspondence of the common objects, instead of the pixel correspondence of the whole image. Third, cross-image large displacement of the common objects may be present in co-segmentation. Local search for correspondence detection is no longer applicable. Due to the major differences, these CNN-based methods for unsupervised image correspondence cannot be straightforwardly applied to object co-segmentation.

2.3 Weakly Supervised Semantic Segmentation

Weakly supervised semantic segmentation (WSS) [Kolesnikov and Lampert, 2016; Chaudhry *et al.*, 2017; Jin *et al.*, 2017; Shimoda and Yanai, 2016; Hou *et al.*, 2017; Wei *et al.*, 2017b; Roy and Todorovic, 2017] aims to reduce the annotation cost of semantic segmentation. Methods of this category usually train their models by using training data with image-level labels, instead of pixel-level masks. There are at least two major differences between WSS and

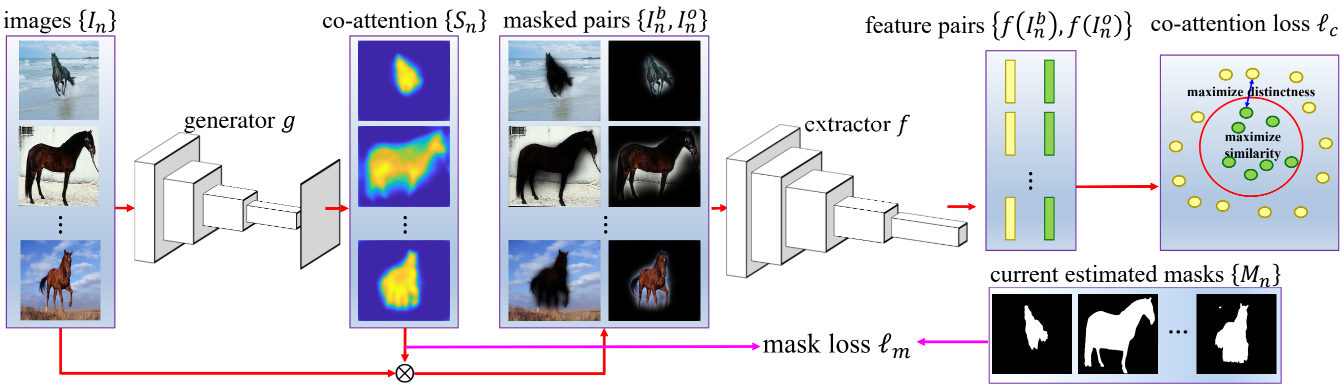
images $\{I_n\}$
generator g
co-attention $\{S_n\}$
masked pairs $\{I_n^b, I_n^o\}$
extractor f
feature pairs $\{f(I_n^b), f(I_n^o)\}$
co-attention loss ℓ_c
maximize distinctness
maximize similarity
current estimated masks $\{M_n\}$
mask loss ℓ_m

Figure 2: The overview of our method. Our network architecture is composed of two collaborative CNN modules, a map generator g and a feature extractor f, which are derived by the co-attention loss ℓ_c and the mask loss ℓ_m.

co-segmentation. First, WSS typically consists of the training and testing phases. It requires weakly annotated training data to learn the model, and applies the learned model to test images. Co-segmentation is carried out by discovering objects commonly appearing in multiple images in a single phase. Second, WSS works with training images of multiple known categories, and requires that the categories of testing images are covered by those of training images. On the contrary, co-segmentation usually works on multiple images of a single, unknown category.

3 Proposed Method

Our method is introduced in this section. First, the proposed formulation for co-segmentation is given. Then, the developed loss functions and the optimization process are described. Finally, some implementation details are provided.

3.1 Proposed Formulation

Given a set of N images, $\{I_n\}_{n=1}^N$, commonly covering objects of the same category, our goal is to segment the common objects. Figure 2 illustrates the proposed method for a quick overview. Our network architecture is composed of two collaborative CNN modules, i.e., the co-attention map generator g and the semantic feature extractor f. Two loss functions, including the co-attention loss ℓ_c and the mask loss ℓ_m, are developed to derive the network.

The generator g is a *fully convolutional network* (FCN) [Long *et al.*, 2015]. For each image I_n, the generator estimates its co-attention map, $S_n = g(I_n)$, which highlights the common object in I_n. With S_n, the estimated object image I_n^o and background image I_n^b of I_n are available. The extractor f can be one of the pre-trained CNN models for image classification, such as AlexNet [Krizhevsky *et al.*, 2012] or VGG-16 [Simonyan and Zisserman, 2015], with the softmax layer removed. It computes the semantic features of the estimated object and background images, i.e., $f(I_n^o)$ and $f(I_n^b)$. We treat the inputs to the last fully connected layer of f as the extracted features.

The co-attention loss ℓ_c is introduced to enhance both inter-image object similarity and intra-image figure-ground distinctness. The mask loss refines the co-attention maps by referring to the selected object proposals. It makes the maps retain the whole objects while removes the noises. According to our empirical studies, we pre-train the extractor f and fix it during training, although fine-tuning is possible. Suppose the generator g is parametrized by $\mathbf{w}$. The proposed unsupervised loss function for learning g is defined by

$$\ell(\mathbf{w}) = \ell_c(\{I_n\}_{n=1}^N; \mathbf{w}) + \lambda \sum_{n \in \{1,\ldots,N\}} \ell_m(I_n, M_n; \mathbf{w}), \quad (1)$$

where λ is a constant for weighting losses. M_n is the selected object proposal for I_n. For the sake of clearness, the optimization of Eq. (1), the loss ℓ_c, and the loss ℓ_m will be detailed in the following subsections.

From co-attention to co-segmentation. By applying the learned generator g to all images, the corresponding co-attention maps are obtained. Following [Yuan *et al.*, 2017], we generate the co-segmentation results via *dense CRFs* [Krähenbühl and Koltun, 2011] where the unary and the pairwise terms are set to referring to the co-attention maps and bilateral filtering, respectively.

3.2 Co-attention Loss ℓ_c

The co-attention loss ℓ_c guides the training of the generator g by referring to the object and background features computed by extractor f. This loss is designed based on the two criteria used in unsupervised object co-segmentation, namely high inter-image object similarity and high intra-image figure-ground distinctness.

As shown in Figure 2, the generator g produces the co-attention map S_n for each image I_n. Sigmoid function serves as the activation function in the last layer of g. Hence, the co-attention value at every pixel k, $S_n(k)$, ranges between 0 and 1. With S_n, the masked object and background images of I_n are respectively obtained as follows:

$$I_n^o = \otimes(S_n, I_n) \text{ and } I_n^b = \otimes(1 - S_n, I_n), \quad (2)$$

where $\otimes$ is the operator of element-wise multiplication. Images I_n^o and I_n^b highlight the estimated object and background of I_n, respectively.

The extractor f is applied to images $\{I_n^o, I_n^b\}_{n=1}^N$ for computing the features $\{f(I_n^o), f(I_n^b)\}_{n=1}^N$. With these features, the co-attention loss is then defined by

$$\ell_c(\{I_n\}_{n=1}^N; \mathbf{w}) = -\sum_{i=1}^{N}\sum_{j\neq i}\log(p_{ij}), \tag{3}$$

where p_{ij} can be considered as a score estimating two mentioned criteria of object co-segmentation, and it is defined by the following equations,

$$p_{ij} = \frac{\exp(-d_{ij}^+)}{\exp(-d_{ij}^+) + \exp(-d_{ij}^-)}, \tag{4}$$

$$d_{ij}^+ = \frac{1}{c}\|f(I_i^o) - f(I_j^o)\|^2, \text{ and} \tag{5}$$

$$d_{ij}^- = \frac{1}{2c}(\|f(I_i^o) - f(I_i^b)\|^2 + \|f(I_j^o) - f(I_j^b)\|^2). \tag{6}$$

Eq. (5) and Eq. (6) respectively measure the inter-image object distance and intra-image figure-ground discrepancy for an image pair I_i and I_j. Constant c is the dimension of the extracted features. The co-attention loss in Eq. (3) is defined over all image pairs. By minimizing this loss, the generator g will produce the co-attention maps in which low inter-image object distances and high intra-image figure-ground discrepancies can be observed. The co-attention loss is the primary part of the objective function. To the best of our knowledge, it has not been explored and is novel in the literature.

3.3 Mask Loss ℓ_m

Using the co-attention loss alone may lead to two problems. First, the resultant co-attention maps tend to highlight only the discriminative object parts, instead of the whole objects. It is not surprising, since segmenting only the discriminative parts gives even lower co-attention loss. Second, some noises, false positives here, are present in the co-attention maps.

The two problems can be alleviated by taking into account single-image objectness. To this end, we can compile a pool of object proposals, $\mathcal{O}_n$, for each image I_n, by using an unsupervised, off-the-shelf approach, e.g., [Krähenbühl and Koltun, 2014]. These proposals are designed to cover objects completely. We can pick object proposals highly consistent with co-attention maps, and use them in order to regularize co-segmentation. Unfortunately, the co-attention maps $\{S_n\}$ at the early training stage are too unstable to pick satisfactory proposals. Thus, we adopt a two-stage strategy to optimize Eq. (1). At the first stage, the mask loss is turned off. After a few epochs, the resultant co-attention maps $\{\tilde{S}_n\}$ become stable enough to pick the proposals $\{\tilde{M}_n\}$, where $\tilde{M}_n = \arg\min_{O\in\mathcal{O}_n} \|\tilde{S}_n - O\|^2$. At the second stage, the mask loss ℓ_m in Eq. (1) is turned on and it is defined by

$$\ell_m(I_n, M_n; \mathbf{w}) = \frac{-1}{|\mathcal{K}|}\sum_{k\in\mathcal{K}}(\beta\tilde{M}_n(k) + (1-\beta)M_n(k))\log(S_n(k))$$
$$+ (\beta(1-\tilde{M}_n(k)) + (1-\beta)(1-M_n(k)))\log(1-S_n(k)), \tag{7}$$

where $M_n = \text{argmin}_{O\in\mathcal{O}_n} \|S_n - O\|^2$, β is a constant, $\mathcal{K}$ is the index set of pixels, and $|\mathcal{K}|$ is the number of pixels. The mask loss in Eq. (7) is in the form of cross entropy, and enforces the co-attention map S_n to be consistent with the weighted combination of the pre-picked $\tilde{M}_n$ and the currently selected M_n.

The idea behind the mask loss is intuitive: The object proposal, covering the discriminative parts detected by S_n, likely covers the non-discriminative parts at the same time. This property is leveraged to enforce the generator g to highlight the non-discriminative parts along with the discovered discriminant parts. The loss also reduces false positives because it can suppress the unfavorable high co-attention values in the background. The mask loss is inspired by the bootstrapping loss in [Reed *et al.*, 2015], but with the difference that the estimated co-attention maps $\{S_n\}$ are updated in turn with the selected proposals instead of a hard threshold 0.5. β is set as 0.95 following [Reed *et al.*, 2015].

Object mask refinement. An object proposal is designed to cover one single object. For an image where multiple objects are present, the aforementioned mask loss may lead to an unfavorable circumstance. Namely only one single object is detected. Thus, we develop a scheme to generate an object mask M_n by an iterative refinement procedure where multiple object proposals may be iteratively merged into M_n. Let O_n^t denote the selected proposal for image I_n at the tth iteration. At the first iteration, we pick the proposal O_n^1 from $\mathcal{O}_n$ that best matches the co-attention map S_n. The object mask M_n is initially set to O_n^1. Other proposals overlapping O_n^1 are removed from $\mathcal{O}_n$. The co-attention values in S_n are set to zero if the values are less than the average value of O_n^1. At the following iteration t, we pick proposal O_n^t that best matches the updated S_n, and merge it into M_n. Then the proposal pool $\mathcal{O}_n$ and the co-attention map S_n are similarly updated. The procedure is repeated until S_n becomes a zero matrix or no proposals remain in $\mathcal{O}_n$. This iterative scheme allows the object mask M_n to cover multiple non-overlapping and high-quality object proposals. The updated M_n is then substituted for the original M_n in Eq. (7).

The mask loss is auxiliary. It is similar to that in [Dai *et al.*, 2015], but has two major differences. First, we dynamically refine object proposals to better cover the detected salient objects. Second, we adopt the bootstrapping method via Eq. (7) to alleviate the unfavorable effect caused by the selected proposals of low quality.

3.4 Optimization Process

The objective function in Eq. (1) is differentiable and convex. We choose ADAM [Kingma and Ba, 2014] as the optimization solver for its rapid convergence. In each epoch, we perform forward propagation and get the updated co-attention maps $\{S_n\}$. Then, the most consistent object masks $\{M_n\}$ are generated based on the proposed object mask generation scheme. Once the object masks $\{M_n\}_{n=1}^N$ are determined, the objective function in Eq. (1) can be optimized by using ADAM. The gradients of each loss function with respect to the optimization variables can be derived straightforward. Therefore, we omit their derivation here.

Our method is end-to-end trainable. Feature extractor can be updated via back propagation. We keep it fixed because, for co-segmentation, there are often not sufficient images for

Method	Airplane		Car		Horse		Avg.	
	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
[Joulin *et al.*, 2012]	47.5	0.12	59.2	0.35	64.2	0.30	56.97	0.243
[Rubinstein *et al.*, 2013]	88.0	0.56	85.4	0.64	82.8	0.52	82.73	0.427
[Chen *et al.*, 2014]	90.2	0.40	87.6	0.65	89.3	0.58	89.03	0.543
[Chang and Wang, 2015]	72.6	0.27	75.9	0.36	79.7	0.36	76.07	0.330
[Lee *et al.*, 2015]	52.8	0.36	64.7	0.42	70.1	0.39	62.53	0.392
[Jerripothula *et al.*, 2016]	90.5	0.61	88.0	0.71	88.3	0.61	88.93	0.643
[Quan *et al.*, 2016]	91.0	0.56	88.5	0.67	89.3	0.58	89.60	0.603
[Hati *et al.*, 2016]	77.7	0.33	62.1	0.43	73.8	0.20	71.20	0.320
[Tao *et al.*, 2017]	79.8	0.43	84.8	0.66	85.7	0.55	83.43	0.547
[Sun and Ponce, 2016]	88.6	0.36	87.0	0.73	87.6	0.55	87.73	0.547
[Jerripothula *et al.*, 2017]	81.8	0.48	84.7	0.69	81.3	0.50	82.60	0.556
w/o ℓ_m	93.6	0.66	91.4	0.79	87.6	0.59	90.86	0.678
Ours	**94.2**	**0.67**	**93.0**	**0.82**	89.7	0.61	**92.29**	**0.698**
[Yuan *et al.*, 2017]*	92.6	0.66	90.4	0.72	**90.2**	**0.65**	91.07	0.677

Table 1: The performance of object co-segmentation on the Internet dataset. The bold and underlined numbers indicate the best and the second best results, respectively. * means the supervised method.

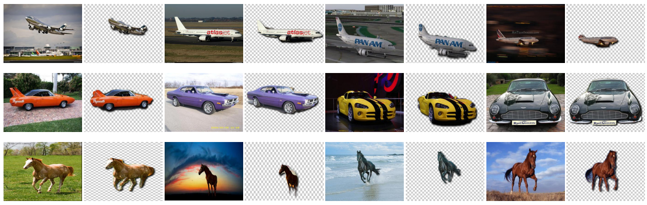

Figure 3: The co-segmentation results generated by our approach on the Internet dataset. In the three examples (rows), the common object categories are airplane, car, and horse, respectively.

stable update. In addition, object proposals are dynamically refined to better cover common objects.

3.5 Implementation Details

The proposed method is implemented based on `MatConvNet` [Vedaldi and Lenc, 2015]. The same network architecture is used in all the experiments. ResNet-50 [He *et al.*, 2016] is adopted as the feature extractor f, because AlexNet [Krizhevsky *et al.*, 2012] and VGG-16/19 [Simonyan and Zisserman, 2015] sometimes lead to the problem of gradient vanishing in our cases. The feature extractor f is the off-the-shelf model pre-trained on ImageNet [Deng *et al.*, 2009]. It is fixed during the optimization process. We have tried to fine-tune f based on the co-attention loss. The performance is not improved due to the limited number of images for co-segmentation. Thus, the feature extractor f remains fixed in the experiments. The features extracted by f are set to the inputs to the last fully connected layer of f. The feature dimension, i.e., c in Eq. (5) and Eq. (6), is set to 2,048.

The generator g is developed based on the VGG-16 [Simonyan and Zisserman, 2015] setting of FCN [Long *et al.*, 2015]. We replace the activation function *softmax* in the last layer with the *sigmoid* function. The output of the sigmoid function serves as the co-attention map. The learning rate is set to 10^{-6} and kept fixed during optimization. As mentioned previously, the generator is learned in a two-stage manner. At the first stage, we optimize the objective in Eq. (1) with the mask loss ℓ_m removed for 20 epochs. After the first stage, the co-attention maps $\{\tilde{S}_n\}$ become stable enough to pick plausible $\{\tilde{M}_n\}$. At the second stage, the mask loss ℓ_m is turned on and the objective in Eq. (1) is optimized for 40 epochs. Therefore, the total number of epoches is 60. The batch size, weight decay, and momentum are set to 5, 0.0005, and 0.9, respectively. All images for co-segmentation are resized to the resolution 384×384 in advance, since the feature extractor f is applied to only images of the same size. Then, we resize the generated co-segmentation results into their original sizes for performance measure. The parameter λ in Eq. (1) is empirically set and fixed to 9 in all experiments.

Method	[Jerripothula *et al.*, 2016]	[Quan *et al.*, 2016]	[Tao *et al.*, 2017]
$\mathcal{P}$	91.8	93.3	90.8
$\mathcal{J}$	0.72	0.76	0.74

Method	[Wang *et al.*, 2017]	Ours	[Yuan *et al.*, 2017]*
$\mathcal{P}$	93.8	**96.5**	94.4
$\mathcal{J}$	0.77	**0.84**	0.82

Table 2: The performance of object co-segmentation on the iCoseg dataset. The bold and underlined numbers indicate the best and the second best results, respectively. * means the supervised method.

For generating the pool of object proposals $\{\mathcal{O}_n\}$ used for object mask update, we adopt the fast object proposal generation algorithm, *geodesic object proposal* (GOP) [Krähenbühl and Koltun, 2014]. Following the unsupervised setting in this work, the unsupervised setting of GOP is adopted. The number of the generated proposals for an image typically ranges from 200 to 1,100.

4 Experimental Results

In this section, we evaluate the proposed method and compare it with existing methods on three benchmarks for object co-segmentation, including the Internet dataset [Rubinstein *et al.*, 2013], the iCoseg dataset [Batra *et al.*, 2010], and the PASCAL-VOC dataset [Faktor and Irani, 2013]. These datasets are composed of real-world images with large intra-class variations, occlusions and background clutters. They have been widely adopted to evaluate many existing methods for object co-segmentation, such as [Jerripothula *et al.*, 2016; Wang *et al.*, 2017; Yuan *et al.*, 2017].

4.1 Datasets and Evaluation Metrics

The three used datasets and the adopted evaluation metrics are briefly described as follows:

The Internet dataset. This dataset introduced in [Rubinstein *et al.*, 2013] contains images of three object categories including airplane, car and horse. Thousands of images in this dataset were collected from the Internet. Following the same setting of the previous work [Rubinstein *et al.*, 2013; Yuan *et al.*, 2017; Tao *et al.*, 2017], we use the same subset of the Internet dataset where 100 images per class are available.

The iCoseg dataset. There are 38 categories in the iCoseg dataset [Batra *et al.*, 2010] with total 643 images. Each category consists of several images, and these images contain either the same or different object instances of that category. Large variations of viewpoints and deformations are present in this dataset.

Method	Avg. $\mathcal{P}$	Avg. $\mathcal{J}$	A.P.	Bike.	Bird	Boat	Bottle.	Bus	Car	Cat	Chair	Cow	D.T.	Dog	Horse	M.B.	P.S.	P.P.	Sheep	Sofa	Train	TV
[Faktor and Irani, 2013]	84.0	0.46	0.65	0.14	0.49	0.47	0.44	0.61	0.55	0.49	0.20	0.59	0.22	0.39	0.52	0.51	0.31	0.27	0.51	0.32	0.55	0.35
[Lee *et al.*, 2015]	69.8	0.33	0.50	0.15	0.29	0.37	0.27	0.55	0.35	0.34	0.13	0.40	0.10	0.37	0.49	0.44	0.24	0.21	0.51	0.3	0.42	0.16
[Chang and Wang, 2015]	82.4	0.29	0.48	0.09	0.32	0.32	0.21	0.34	0.42	0.35	0.13	0.50	0.06	0.22	0.37	0.39	0.19	0.17	0.41	0.21	0.41	0.18
[Quan *et al.*, 2016]	89.0	0.52	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
[Hati *et al.*, 2016]	72.5	0.25	0.44	0.13	0.26	0.31	0.28	0.33	0.26	0.29	0.14	0.24	0.11	0.27	0.23	0.22	0.18	0.17	0.33	0.27	0.26	0.25
[Jerripothula *et al.*, 2016]	85.2	0.45	0.64	0.20	0.54	0.48	0.42	0.64	0.55	0.57	0.21	0.61	0.19	0.49	0.57	0.50	0.34	0.28	0.53	0.39	0.56	0.38
[Jerripothula *et al.*, 2017]	80.1	0.40	0.53	0.14	0.47	0.43	0.42	0.62	0.50	0.49	0.20	0.56	0.13	0.38	0.50	0.45	0.29	0.26	0.40	0.37	0.51	0.37
[Wang *et al.*, 2017]	84.3	0.52	0.75	0.26	0.53	0.59	0.51	0.70	0.59	0.70	**0.35**	0.63	0.26	0.56	0.63	0 59	0.35	0.28	0.67	0.52	0.52	0.48
Ours	**91.0**	**0.60**	**0.77**	**0.27**	**0.70**	**0.61**	**0.58**	**0.79**	**0.76**	**0.79**	0.29	**0.75**	**0.28**	**0.63**	**0.66**	**0.65**	**0.37**	**0.42**	**0.75**	**0.67**	**0.68**	**0.51**

Table 3: The performance of object co-segmentation on the PASCAL-VOC dataset under Jaccard index and Precision. The class-wise results are measured in Jaccard index. The bold and underlined numbers indicate the best and the second best results, respectively.

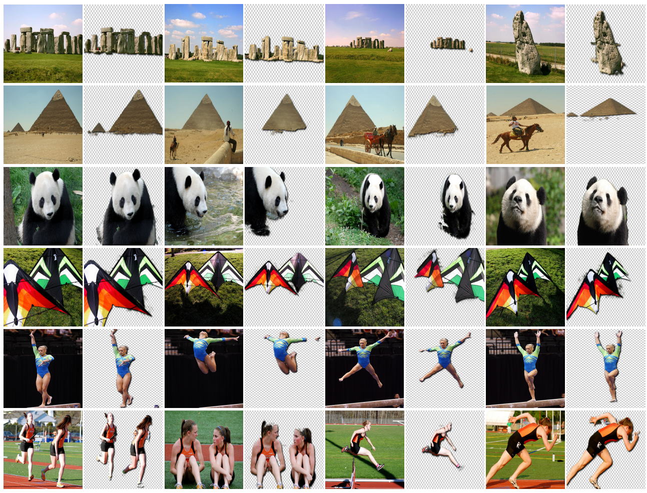

Figure 4: The co-segmentation results generated by our approach on the iCoseg dataset. In the six examples (rows), the common object categories are Stonehenge, pyramids, pandas, kite-kitekid, and track and field, respectively.

The PASCAL-VOC dataset. This dataset was collected by Faktor and Irani [Faktor and Irani, 2013]. It contains total 1,037 images of 20 object classes from PASCAL-VOC 2010 dataset. The PASCAL-VOC dataset is more challenging and difficult than the Internet dataset due to extremely large intra-class variability and subtle figure-ground discrimination. In addition, some object categories have only a few images.

Evaluation metrics. Two widely used measures, *precision* ($\mathcal{P}$) and *Jaccard index* ($\mathcal{J}$), are adopted to evaluate the performance of object co-segmentation. Precision measures the percentage of correctly segmented pixels including both object and background pixels. Jaccard index is the ratio of the intersection area of the detected objects and the ground truth to their union area. The background pixels are taken into account in precision, so the images with larger background areas tend to have a higher performance in precision. Therefore, precision may not very faithfully reflect the quality of object co-segmentation results. Compared with precision, Jaccard index is considered more reliable to measure the quality of results. It provides more appropriate evaluation as it only focuses on objects.

4.2 Comparison with Co-segmentation Methods

We compare the proposed method with the state-of-the-art methods on the Internet, iCoseg, and PASCAL-VOC datasets, and report their performances in Table 1, Table 2, and Table 3, respectively. All methods in Table 1, Table 2, and Table 3 are unsupervised except for the supervised CNN-based method [Yuan *et al.*, 2017]. Our method achieves the state-of-the-art performance on the three datasets under both evaluation metrics.

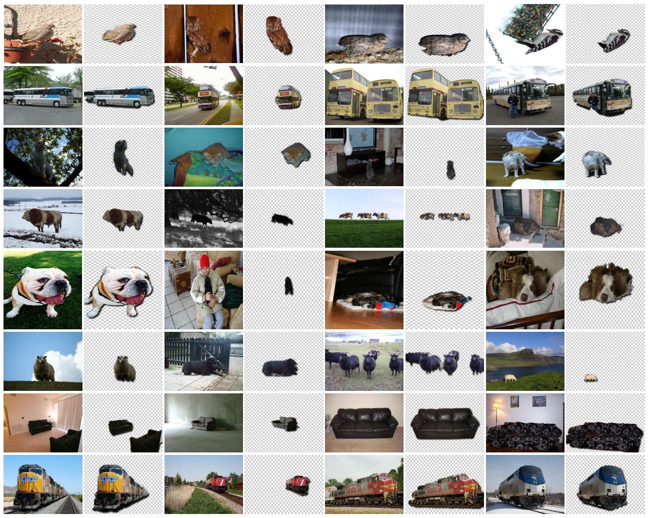

Figure 5: The co-segment results generated by our approach on the PASCAL-VOC dataset. From the first row to the last row, the classes are bird, bus, cat, cow, dog, sheep, sofa, and train, respectively.

On the Internet dataset, our method outperforms the state-of-the-art unsupervised method [Jerripothula *et al.*, 2016] by a margin around 5% in $\mathcal{J}$ and the supervised method [Yuan *et al.*, 2017] by a margin around 2% in Table 1. On the iCoseg dataset, our method performs favorably against the state-of-the-art unsupervised method [Wang *et al.*, 2017] by a margin around 7% in $\mathcal{J}$ and the supervised method [Yuan *et al.*, 2017] by a margin around 2% in Table 2. The results demonstrate that the proposed method can effectively utilize the information shared between common objects in different images without using complex graphical structures and optimization algorithms, or additional training data in the form of object masks. The effectiveness of our method mainly results from two properties. First, the co-attention loss enables CNNs to adaptively learn the robust features for unseen images, and discover the common regions. Second, the mask loss helps CNNs discover the whole objects and remove noises. Figure 3 and Figure 4 show some co-segmentation results on the Internet and iCoseg datasets, respectively. Our method can generate promising object segments under different types of intra-class variations, such as colors, sharps, views and background clutters, in the Internet and iCoseg datasets, resepctively.

In Table 3, our proposed method also outperforms the best competing method [Wang *et al.*, 2017] by a large margin around 8% in $\mathcal{J}$. Although the PASCAL-VOC dataset has higher variations than the Internet and iCoseg datasets, high performance gains over the competing methods can be obtained by our method. The results indicate that our method adapts itself well to unseen images with large variations. Some examples of the co-segmentation results on the PASCAL-VOC dataset are shown in Figure 5. Compared with the Internet dataset in Figure 3 and the iCoseg dataset Figure 4, images on this dataset contain higher intra-class variations and subtle figure-ground differences. Our method can infer the common object segments of high quality. For example, the birds in the first row are of dissimilar colors and have subtle figure-ground difference. It is difficult for hand-crafted features to handle this case well. Nevertheless, our method gets the promising segmentation results owe to its ability of adaptive feature learning. Although our method does not adopt multi-scale learning which may make the running time longer and consume more memory, it still finds objects of different scales, such as those of object classes bus, dog, sofa, and train, because CNNs can tolerate scale variations to some extent.

Ablation studies. In Table 1, w/o ℓ_m indicates the variant of our method where the mask loss ℓ_m is turned off, i.e., $\lambda = 0$ in Eq. (1). A performance drop about 2% is observed, but it still outperforms the state-of-the-arts. Therefore, the effectiveness of our method is mainly attributed to the proposed co-attention loss, instead of the mask loss with its adopted object proposals. We visualize the effect of using the mask loss in Figure 6. When the mask loss is turned off, the co-attention maps have many false positives and do not sharply cover the common objects. These co-attention maps result in the sub-optimal co-segmentation results. With the mask loss, the generator can highlight whole objects and suppress co-attention values in the background. Therefore, the attention maps result in much better co-segmentation results.

Our method employs *dense CRFs* [Krähenbühl and Koltun, 2011] for post-processing and generating the co-segmentation results. To measure the effect of using dense CRFs in our method, we evaluate the performance of our method by replacing dense CRFs with Otsu's method and GrabCut [Rother *et al.*, 2004] for post-processing. Otsu'method and GrabCut were adopted as the post-processing step to generate the co-segmentation results in previous work [Jerripothula *et al.*, 2016; Faktor and Irani, 2013; Quan *et al.*, 2016]. The results in Table 4 demonstrate that our method works well with each of the three schemes for post-processing.

In addition, our method can run with reasonable efficiency. Given 100 images for co-segmentation, model training and object mask refinement in each of 60 epochs take about 20 and 6 seconds, respectively, on an NVIDIA Titan X graphics card. Namely, it takes about $1,560$ seconds to estimate the co-segmentation results of 100 images, and the average time per image is 15.6 seconds.

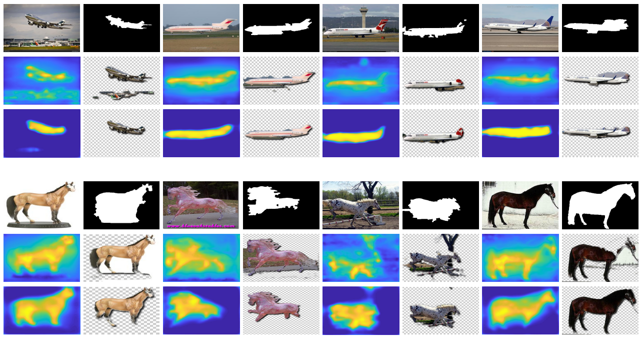

Figure 6: The effect of using the mask loss ℓ_m. The co-segmentation results on two object classes, including airplane (**top**) and horse (**bottom**). For each class, the first row shows four images and the corresponding estimated object masks, i.e., M_n in Eq. (7). When the mask loss ℓ_m is turned off, the second row gives the co-attention maps and the corresponding co-segmentation results. When the mask loss ℓ_m is turned on, the co-attention maps and the corresponding co-segmentation results displayed in the third row become better.

Method	Internet		iCoseg		PASCAL-VOC	
	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$	$\mathcal{P}$	$\mathcal{J}$
Otsu's method	91.17	0.680	**96.53**	**0.837**	90.1	0.58
GrabCut	91.60	0.692	96.35	0.835	90.6	**0.61**
dense CRFs	**92.29**	**0.698**	96.46	0.835	**91.0**	0.60

Table 4: The performance of our approach with three different schemes for post-processing.

4.3 Comparison with WSS Methods

The setting of weakly supervised semantic segmentation (WSS) is similar to that of co-segmentation in the sense that images of specific categories are given for segmentation. Therefore, we compare our method with three state-of-the-art WSS methods, including [Kolesnikov and Lampert, 2016; Jin *et al.*, 2017; Chaudhry *et al.*, 2017] in Table 5. Note that we follow the previous methods for co-segmentation, i.e., those in Table 3, and use the PASCAL-VOC dataset collected in [Faktor and Irani, 2013] as the test bed, instead of the standard PASCAL-VOC 2012 dataset [Everingham *et al.*, 2015], for evaluating the performance of object co-segmentation. In Table 5, our method outperforms the methods in [Kolesnikov and Lampert, 2016; Jin *et al.*, 2017] and achieves a similar performance to that in [Chaudhry *et al.*, 2017]. Nevertheless, our method has two advantages over these WSS methods. First, our method does not require a training phase and does not rely on any background information. Second, our method can be applied to images of an arbitrary and unknown category. On the contrary, the models learned by WSS methods can segment only objects whose categories are covered in the training data.

5 Conclusions

In this paper, we presented an unsupervised approach for object co-segmentation task with CNNs, and to best of our

Method	Avg. $\mathcal{P}$	Avg. $\mathcal{J}$.	A.P.	Bike.	Bird	Boat	Bottle.	Bus	Car	Cat	Chair	Cow	D.T.	Dog	Horse	M.B.	P.S.	P.P.	Sheep	Sofa	Train	TV
[Kolesnikov and Lampert, 2016]	90.4	0.57	0.68	0.28	0.61	0.41	**0.62**	0.79	0.67	0.71	**0.32**	0.67	**0.31**	0.65	0.60	**0.66**	**0.53**	**0.44**	0.68	0.65	0.49	**0.57**
[Jin *et al.*, 2017]	89.0	0.56	0.71	**0.29**	0.60	0.55	0.57	0.74	0.71	0.76	0.21	**0.80**	0.15	**0.72**	**0.74**	**0.66**	0.52	**0.44**	**0.80**	0.41	0.49	0.43
[Chaudhry *et al.*, 2017]	**92.0**	0.59	**0.78**	**0.29**	0.64	**0.63**	0.59	**0.82**	0.74	0.68	0.31	0.75	0.21	0.63	0.67	**0.66**	0.49	0.34	0.74	0.62	**0.70**	0.53
Ours	91.0	**0.60**	0.77	0.27	**0.70**	0.61	0.58	0.79	**0.76**	**0.79**	0.29	0.75	0.28	0.63	0.66	0.65	0.37	0.42	0.75	**0.67**	0.68	0.51

Table 5: The comparison of our method and three WSS methods on the PASCAL-VOC dataset under Jaccard index and Precision. The class-wise results are measured in Jaccard index. The bold and underlined numbers indicate the best and the second best results, respectively.

knowledge, we are the first one to solve this task with an unsupervised CNNs. The proposed CNN architecture is composed of two CNN modules, *a feature extractor* and *a co-attention map generator*, along with two unsupervised losses, *a co-attention loss* and *a mask loss*. During the optimization process, the similarity of estimated objects and background is calculated in the co-attention loss, and the information can be propagated to guide the optimization of the generator. Thus, the co-attention loss can enable the generator to produce maps correctly localizing the common objects. The optimization is further regularized by the mask loss. The mask loss can regularize the generator to remove false negatives and positives on objects and background, respectively. The experimental results on three challenging benchmarks are promising, and the proposed method outperforms the existing state-of-the-art unsupervised and supervised methods. In the future, we will generalize our approach to other tasks which also require multiple images as inputs, such as semantic correspondence [Hsu *et al.*, 2015], scene co-parsing [Zhong *et al.*, 2016] or image co-localization [Wei *et al.*, 2017a].

Acknowledgments

This work was supported by Ministry of Science and Technology (MOST) under grant 105-2221-E-001-030-MY2 and MOST Joint Research Center for AI Technology and All Vista Healthcare under 107-2634-F-002-007.

References

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

[Chang and Wang, 2015] H.-S. Chang and Y.-C. Wang. Optimizing the decomposition for multiple foreground cosegmentation. *CVIU*, 2015.

[Chang *et al.*, 2011] 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.

[Chaudhry *et al.*, 2017] A. Chaudhry, P. Dokania, and P. Torr. Discovering class-specific pixels for weakly-supervised semantic segmentation. In *BMVC*, 2017.

[Chen *et al.*, 2014] X. Chen, A. Shrivastava, and A. Gupta. Enriching visual knowledge bases via object discovery and segmentation. In *CVPR*, 2014.

[Chen *et al.*, 2015] H.-Y. Chen, Y.-Y. Lin, and B.-Y Chen. Co-segmentation guided hough transform for robust feature matching. *TPAMI*, 2015.

[Dai *et al.*, 2015] J. Dai, K. He, and J. Sun. BoxSup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentati. In *ICCV*, 2015.

[Dalal and Triggs, 2005] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A preview of a large-scale hierarchical database. In *CVPR*, 2009.

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

[Faktor and Irani, 2013] A. Faktor and M. Irani. Co-segmentation by composition. In *ICCV*, 2013.

[Godard *et al.*, 2017] C. Godard, O Aodha, and G. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

[Hati *et al.*, 2016] A. Hati, S. Chaudhuri, and R. Velmurugan. Image co-segmentation using maximum common subgraph matching and region co-growing. In *ECCV*, 2016.

[He *et al.*, 2016] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Hou *et al.*, 2017] Q. Hou, P. Dokania, D. Massiceti, Y. Wei, M.-M. Cheng, and P. Torr. Bottom-up top-down cues for weakly-supervised semantic segmentation. In *EMM-CVPR*, 2017.

[Hsu *et al.*, 2014] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang. Augmented multiple instance regression for inferring object contours in bounding boxes. *TIP*, 2014.

[Hsu *et al.*, 2015] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang. Robust image alignment with multiple feature descriptors and matching-guided neighborhoods. In *CVPR*, 2015.

[Hsu *et al.*, 2017] K.-J. Hsu, Y.-Y. Lin, and Y.-Y. Chuang. Weakly supervised saliency detection with a category-driven map generator. In *BMVC*, 2017.

[Jerripothula *et al.*, 2016] K. Jerripothula, J. Cai, and J. Yuan. Image co-segmentation via saliency co-fusion. *TMM*, 2016.

[Jerripothula *et al.*, 2017] K. Jerripothula, J. Cai, J. Lu, and J. Yuan. Object co-skeletonization with co-segmentation. In *CVPR*, 2017.

[Jin *et al.*, 2017] Bin Jin, M. Segovia, and S. Susstrunk. Webly supervised semantic segmentation. In *CVPR*, 2017.

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

[Joulin *et al.*, 2012] A. Joulin, F. Bach, and J. Ponce. Multi-class cosegmentation. In *CVPR*, 2012.

[Kim *et al.*, 2011] G. Kim, E. Xing, F.-F. Li, and T. Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, 2011.

[Kingma and Ba, 2014] D. Kingma and J. Ba. ADAM: A method for stochastic optimization. In *ICLR*, 2014.

[Kolesnikov and Lampert, 2016] A. Kolesnikov and C. Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *ECCV*, 2016.

[Krähenbühl and Koltun, 2011] P. Krähenbühl and V. Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[Krähenbühl and Koltun, 2014] P. Krähenbühl and V. Koltun. Geodesic object proposals. In *ECCV*, 2014.

[Krizhevsky *et al.*, 2012] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[Lee *et al.*, 2015] C. Lee, W.-D. Jang, J.-Y. Sim, and C.-S. Kim. Multiple random walkers and their application to image cosegmentation. In *CVPR*, 2015.

[Li *et al.*, 2018] L. Li, Z. Liua, and J. Zhang. Unsupervised image co-segmentation via guidance of simple images. *CVIU*, 2018.

[Long *et al.*, 2015] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional models for semantic segmentation. In *CVPR*, 2015.

[Lowe, 2004] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[Meister *et al.*, 2018] S. Meister, J. Hur, and S. Roth. UnFlow: Unsupervised learning of optical flow with a bidirectional census loss. In *AAAI*, 2018.

[Mustafa and Hilton, 2017] A. Mustafa and A. Hilton. Semantically coherent co-segmentation and reconstruction of dynamic scenes. In *CVPR*, 2017.

[Quan *et al.*, 2016] R. Quan, J. Han, D. Zhang, and F. Nie. Object co-segmentation via graph optimized-flexible manifold ranking. In *CVPR*, 2016.

[Reed *et al.*, 2015] S. Reed, H. Lee, D. Anguelov, C. Szegedy, D. Erhan, and A. Rabinovich. Training deep neural networks on noisy labels with bootstrapping. In *ICLRW*, 2015.

[Ren *et al.*, 2017] Z Ren, J. Yan, B. Ni, B. Liu, X. Yang, and H. Zha. Unsupervised deep learning for optical flow estimation. In *AAAI*, 2017.

[Rother *et al.*, 2004] C. Rother, V. Kolmogorov, and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *TOG*, 2004.

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

[Roy and Todorovic, 2017] A. Roy and S. Todorovic. Combining bottom-up, top-down, and smoothness cues for weakly supervised image segmentation. In *CVPR*, 2017.

[Rubinstein *et al.*, 2013] M. Rubinstein, A. Joulin, J. Kopf, and C. Liu. Unsupervised joint object discovery and segmentation in internet images. In *CVPR*, 2013.

[Rubio *et al.*, 2012] J. Rubio, J. Serrat, A. Lopez, and N. Paragios. Unsupervised co-segmentation through region matching. In *CVPR*, 2012.

[Shen *et al.*, 2017] T. Shen, G. Lin, L. Liu, and I. Reid. Weakly supervised semantic segmentation based on co-segmentation. In *BMVC*, 2017.

[Shimoda and Yanai, 2016] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, 2016.

[Shotton *et al.*, 2009] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *IJCV*, 2009.

[Simonyan and Zisserman, 2015] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[Sun and Ponce, 2016] J. Sun and J. Ponce. Learning dictionary of discriminative part detectors for image categorization and cosegmentation. *IJCV*, 2016.

[Tao *et al.*, 2017] Z. Tao, H. Liu, H. Fu, and Y. Fu. Image cosegmentation via saliency-guided constrained clustering with cosine similarity. In *AAAI*, 2017.

[Vedaldi and Lenc, 2015] A. Vedaldi and K. Lenc. MatConvNet – convolutional neural networks for matlab. In *ACMMM*, 2015.

[Wang *et al.*, 2017] C. Wang, H. Zhang, L. Yang, X. Cao, and H. Xiong. Multiple semantic matching on augmented n-partite graph for object co-segmentation. *TIP*, 2017.

[Wei *et al.*, 2017a] X.-S. Wei, C.-L. Zhang, Y Li, C.-W. Xie, J. Wu, C. Shen, and Z.-H. Zhou. Deep descriptor transforming for image co-localization. In *IJCAI*, 2017.

[Wei *et al.*, 2017b] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In *CVPR*, 2017.

[Yu *et al.*, 2016] J. Yu, A. Harley, and K. Derpanis. Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness. In *ECCVW*, 2016.

[Yuan *et al.*, 2017] Z. Yuan, T. Lu, and Y. Wu. Deep-dense conditional random fields for object co-segmentation. In *IJCAI*, 2017.

[Zhong *et al.*, 2016] G. Zhong, Y.-H. Tsai, and M.-H. Yang. Weakly-supervised video scene co-parsing. In *ACCV*, 2016.

[Zhou *et al.*, 2017] C. Zhou, H. Zhang, X. Shen, and J. Jia. Unsupervised learning of stereo matching. In *ICCV*, 2017.