A Self-paced Multiple-instance Learning Framework for Co-saliency Detection

Dingwen Zhang[1], Deyu Meng[2], Chao Li[1], Lu Jiang[3], Qian Zhao[2], and Junwei Han[1*]
[1]School of Automation, Northwestern Polytechnical University
[2]School of Mathematics and Statistics, Xi'an Jiaotong University
[3]School of Computer Science, Carnegie Mellon University
{zhangdingwen2006yyy,junweihan2010,lllcho1314,timmy.zhaoqian}@gmail.com
dymeng@mail.xjtu.edu.cn, lujiang@cs.cmu.edu

Abstract

As an interesting and emerging topic, co-saliency detection aims at simultaneously extracting common salient objects in a group of images. Traditional co-saliency detection approaches heavily rely on human knowledge for designing hand-crafted metrics to explore the intrinsic patterns underlying co-salient objects. Such strategies, however, always suffer from poor generalization capability to flexibly adapt to various scenarios in real applications, especially due to their lack of insightful understanding of the biological mechanisms of human visual co-attention. To alleviate this problem, we propose a novel framework for this task, by naturally reformulating it as a multiple-instance learning (MIL) problem and further integrating it into a self-paced learning (SPL) regime. The proposed framework on one hand is capable of fitting insightful metric measurements and discovering common patterns under co-salient regions in a self-learning way by MIL, and on the other hand tends to promise the learning reliability and stability by simulating the human learning process through SPL. Experiments on benchmark datasets have demonstrated the effectiveness of the proposed framework as compared with the state-of-the-arts.

1. Introduction

The rapid development of the imaging equipment, e.g., cameras and smartphones, and the growing popularity of social media, e.g., Flickr and Facebook, have resulted in an explosion of digital images accessible in forms of personal and internet photo-groups. Typically, such image groups are huge in size and share common objects or events. Thus, it is of great interest to identify the common and attractive objects from all images in such groups. However, in practice, the image groups are also quite complex due to diverse background, illumination conditions, and view point variations. Consequently, detecting the common and attractive objects is also of great challenge. Against this problem, co-saliency detection, as depicted in Fig. 1, has been proposed and attracting intensive research attention in the recent years.

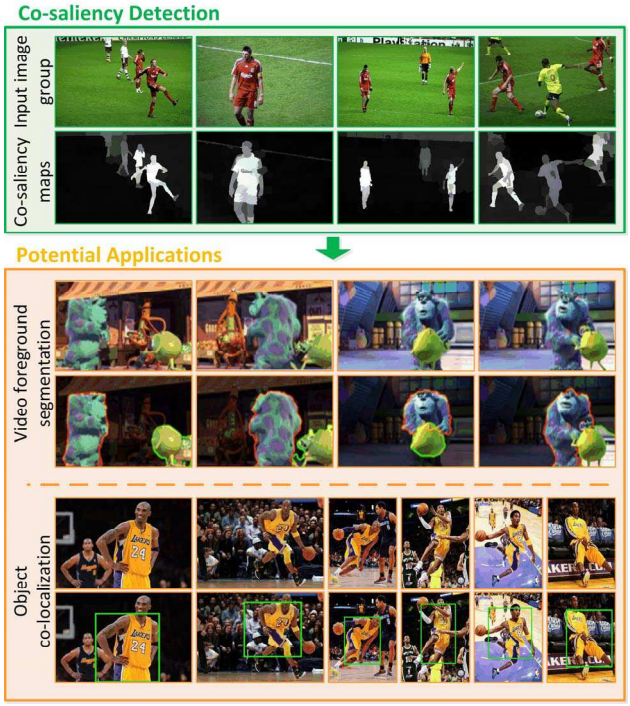

Figure 1: Examples illustrating the co-saliency detection problem (the upper row) and its potential applications (two lower rows).

Co-saliency detection aims at exploring the most important information, i.e., the common and salient foreground object regions, from the image group with implying priorities based on the human visual co-attention [1]. As one extension of the traditional saliency detection [2, 3], co-saliency detection additionally explores the global information at the group level. Thus it can not only be used in the multi-camera system [4] directly, but also provide useful common foreground prior for some real-world applications, such as video foreground co-segmentation [5] and image co-localization [6, 7]. In this paper, we mainly focus on the fundament of this problem while specific applications are beyond the scope of this paper.

Most existing works in co-saliency detection heavily rely on manually designed metrics, e.g., the intra-image contrast and the inter-image consistency, to formulate the properties of the co-salient regions for achieving satisfactory

* Corresponding author.

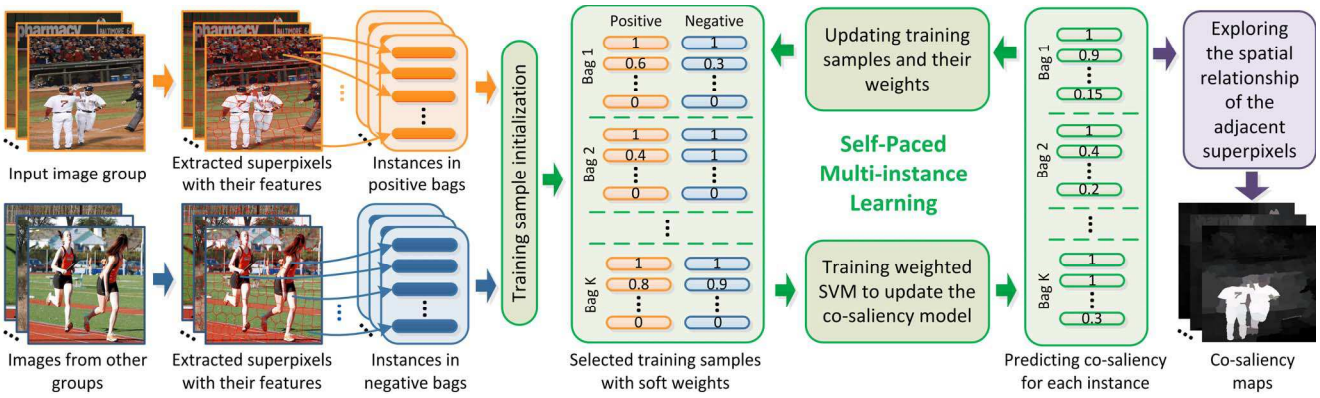

Figure 2: The framework of the proposed co-saliency detection approach.

performance. However, these hand-designed metrics are typically too subjective and cannot generalize well to flexibly adapt to various scenarios encountered in practice, especially due to the lack of thorough understanding of the biological mechanisms of human visual co-attention. It is more promising to use machine learning algorithms to automatically fit data and discover common patterns underlying the co-salient regions in a self-learning way.

To this end, we make the earliest effort to introduce multi-instance learning (MIL) [8] to co-saliency detection for jointly exploring the contrast between co-salient objects and their contexts, as well as the consistency of the co-salient objects within multiple images. The aim of MIL is to learn to predict each instance based on two criteria, i.e., maximizing inter-class distances and minimizing intra-class distances. In co-saliency detection, each image in a certain image group contains at least one common object, while the images from other image groups generally do not contain such objects. The aim of co-saliency detection is to separate the co-salient object regions from the image background. Thus MIL is well-suited for the task of co-saliency detection naturally. Under this paradigm, the inter-class and intra-class distances can be automatically learned from data, which can well fit the properties of the co-salient regions.

An important component in MIL is to alternatively update labels of the training instances (located in positive bags) as well as the instance detector in iterations [9-12]. Since these instance labels can only be pseudo-annotated in a weakly supervised way, there are always bunch of false annotations involved in learning, which tends to conduct confusable or even improper detections. A MIL framework which is capable of guiding a reliable instance annotation and sound instance detection is thus needed urgently.

To this end, we further propose a self-paced MIL (SP-MIL) framework in this study. This framework is constructed by integrating the MIL regime into a self-paced learning (SPL) paradigm [13-16]. The SPL theory is inspired by the learning process of humans/animals, where samples are involved in learning from easy/faithful to gradually more complex/confusable ones [13]. It is a theoretically-sound manner to help MIL in instance selection and annotation. Especially, such a SPL manner facilitates MIL gradually achieving faithful detection knowledge from instances reliable to detect to those easily confused by current detectors.

To apply the SP-MIL to co-saliency detection, we propose a unified framework as shown in Fig. 2. Given an image group, we consider the images within this group as the positive bags and the similar images searched from other groups as the negative bags. The superpixels in each image are considered as the instances. After feature extraction, we use SP-MIL to alternatively update co-saliency object detector and annotate pseudo-labels for training instances in a SPL manner. Finally, the co-saliency maps are generated by additionally considering the spatial relationship of the adjacent superpxiels. In summary, the contributions of this paper are four-fold:

- We propose a novel co-saliency detection framework which is among the earliest efforts to infer the properties of the co-salient regions in a self-learning manner without the need of manually designed metrics.
- We first discover the natural relation between the co-saliency detection and MIL, and easily formulate the former issue into a concise MIL setting.
- We propose a new SP-MIL framework by integrating SPL paradigm into the MIL regime, which facilitates MIL to extract faithful knowledge from highly confused instance detection results in learning.
- We also advance the SPL development by proposing a new self-paced regularizer which considers easiness, diversity, and real-valued sample weighting.

2. Related Works

Co-saliency detection: Early co-saliency detection methods [17-20] were developed to discover co-saliency from image pairs. Specifically, Jacobs *et al.* [19] firstly defined visual co-saliency as the visual saliency of image pixels or regions in the context of other images. Afterwards,

Li *et al*. [17] proposed a co-multilayer graph model to explore the multi-image saliency and established the first public co-saliency dataset. Then, Chen [18] and Tan *et al*. [20] solved this problem via the sparse distribution-based representation and bipartite graph matching, respectively.

To extend to detect co-saliency from multiple images, several novel methods [1, 21-24, 46] were proposed lately. For example, Li *et al*. [1] firstly defined the intra-image saliency and inter-image saliency and then integrated them to obtain the final co-saliency of the image group. Fu *et al*. [21] proposed a cluster-based algorithm to explore the contrast cue, the spatial cue, and the corresponding cue to detect the co-salient regions. Liu *et al.* [22] proposed a hierarchical segmentation based model, where the regional contrasts, global similarity, and object prior are calculated based on multiple-level segmentation. Cao *et al*. [23] used rank constraint to exploit the relationship of multiple pre-designed saliency cues and then assigned the self-adaptive weight to generate the final co-saliency map.

As can be seen, existing co-saliency detection methods heavily rely on manually designed metrics to explore the properties of the co-salient regions. Thus, a novel learning-based model which can automatically infer the properties of the co-salient regions is desirable.

Image co-segmentation: Co-segmentation is a closely related research topic for co-saliency detection. However, they mainly have two-fold differences: 1) co-saliency detection only focuses on detecting the common salient objects while co-segmentation methods also tend to segment the similar but non-salient background regions [25, 26]; 2) co-segmentation usually needs semi- or interactive-supervision [27, 28], where some object regions need to be labeled in advance, while co-saliency detection, as a concept from human attention mechanism, is implemented in an unsupervised or super-weakly supervised manner. Thus, the latter is generally implemented under a much weaker conditions and can be used to get some informative priors for segmenting the common objects for the former.

Multi-instance learning: MIL was first proposed by [8] to classifyng molecules in the context of drug design. A remarkable progress of MIL was the presence of MI-SVM proposed by Andrews *et al.* [9], which heuristically solved the mixed integer quadratic programs in the extended support vector machine. Afterwards, numerous MIL models were developed for solving the problems in computer vision tasks [10-12]. However, as the sample labels are always learned under super-weak supervision, the iterative learning scheme still tends to instable or even unreliable solutions. In this paper, we propose a novel MIL approach based on the SPL theory, where the solid theoretical background will guide MIL to achieve more faithful knowledge from reliable instances to more confusable ones. Different from [10] which adopted saliency detection to cast the unsupervised object discovery into a MIL formulation to localize objects in all possible classes, this paper finds the natural relation between co-saliency detection and MIL and designed novel self-paced regime for learning co-salient patterns. As the self-paced regularizer in SP-MIL can finely handle the ambiguity of the unlabeled data, it might also facilitate the unsupervised object discovery in the future.

Self-paced learning: Inspired by the learning process of humans/animals, the theory of self-paced (or curriculum) learning [13, 14] is proposed lately. The idea is to learn the model iteratively from easy to complex samples in a self-paced fashion. By virtue of its generality, the SPL theory has been widely applied to various tasks, such as object tracking [29], image classification [30], and multi-media event detection [15, 16]. Most previous works only considered sample easiness in SPL. Jiang *et al.* [15] made the earliest effort to additionally conduct sample diversity in SPL via a nonconvex negative $l_{2,1}$ norm. Different from [15], this paper proposes a convex negative $l_{0.5,1}$ norm which is much easier to be solved. This not only more complies with the original SPL axiomic definition but also leads to real-valued sample weighting rather than the binary one as in [15], which can more faithfully reflect the importance of training samples to the classifier.

3. Self-paced Multi-instance Learning

3.1. Problem Formulation

Given K images, consisting of K_+ positive ones within a certain image group for co-saliency detection and K_- negative ones searched from other groups, consider superpixels in each image as the instances to be classified. Accumulate all instances at the k-th image obtains $\mathbf{X}_k = \{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k}$, $k = 1,2,\cdots,K$, where $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ corresponds to the feature representation of the i-th superpixel/instance of the k-th image/bag, n_k is the instance number in $\mathbf{X}_k$, and $n = \sum_{k=1}^{K} n_k$ corresponds to the number of whole instances. Correspondingly, denote the label set as $\mathbf{Y}_k = \{y_i^{(k)}\}_{i=1}^{n_k}$, where $y_i^{(k)} \in \{-1,1\}$ denotes the label of the instance $\mathbf{x}_i^{(k)}$ (if it belongs to the co-salient region or not). Without loss of generalization, we assume that the index set of all positive images is $I_+ = \{1,\cdots,K_+\}$ while the negative ones $I_- = \{K_+ + 1,\cdots,K\}$. Since the objective foreground object is known to exist in each image of the current image group, for each $k \in I_+$, at least one instance in $\mathbf{X}_k$ should be positive, i.e., at least one $y_i^{(k)}$ in $\mathbf{Y}_k$ should be +1; and for each $k \in I_-$, all $y_i^{(k)}$ s are set as -1 since all of instances are known as non-saliency object. Under such formulation, the co-saliency detection problem is naturally transformed into the MIL problem setting.

3.2. SP-MIL Model

The main idea of SP-MIL is to integrate the MIL process into a SPL framework. Specifically, SP-MIL tends to first

distinguish faithful image co-saliency regions from easy (high-confidence) instances, and then gradually transfer the learned knowledge to recognize more complex ones.

Such an idea can be formulated as a concise optimization problem as follows:

$$\min_{\mathbf{w},b,\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)},\mathbf{v}\in[0,1]^n} \mathbf{E}(\mathbf{w},b,\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)},\mathbf{v}) = \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)}\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b)) + f(\mathbf{v};\lambda,\gamma)$$
$$s.t.,\ \left\|\mathbf{y}^{(k)}+1\right\|_0 \geq 1,\ k=1,2,\cdots K_+ \quad (1)$$

where $\mathbf{v}=[v_1^{(1)},\cdots v_{n_1}^{(1)}, v_2^{(2)},\cdots v_{n_2}^{(2)},\cdots v_{n_K}^{(K)}]\in\mathbb{R}^n$ denotes the importance weights for all instances, $\mathbf{y}^{(k)}=[y_1^{(k)}, y_2^{(k)}\cdots, y_{n_k}^{(k)}]\in\mathbb{R}^{n_k}$ denotes the labels for instances in the k-th bag, $\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$ denotes the hinge loss of $\mathbf{x}_i^{(k)}$ under the linear classifier $g(\mathbf{x}_i^{(k)};\mathbf{w},b)$ with weight vector $\mathbf{w}$ and bias parameter b. The constraint $\|\mathbf{y}^{(k)}+1\|_0\geq 1$ for each $k\in I_+$ enforces at least one positive instance in each positive bag.

In the SP-MIL model (1), the self-paced capability is followed by the involvement of the SPL regularizer $f(\mathbf{v};\lambda,\gamma)$ with the following form:

$$f(\mathbf{v};\lambda,\gamma) = -\lambda\sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)} - \gamma\sum_{k=1}^{K}\sqrt{\sum_{i=1}^{n_k} v_i^{(k)}}, \quad (2)$$

where λ, γ are the parameters imposed on the easiness term (the negative l_1-norm term: $-\|\mathbf{v}\|_1 = -\sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)}$) and the diversity term (the negative $l_{2,1}$-norm-like term: $-\sum_{k=1}^{K}\sqrt{\sum_{i=1}^{n_k} v_i^{(k)}}$), respectively.

The negative l_1-norm term is inherited from the conventional SPL, which favors selecting easy over complex examples. If we omit the diversity term (i.e., let $\gamma=0$), the regularizer degenerates to the traditional hard SPL function proposed in [14], which conducts either 1 or 0 (i.e., selected in training or not) for the weight $v_i^{(k)}$ imposed on instance $\mathbf{x}_i^{(k)}$, by judging whether its loss value is smaller than the pace parameter λ or not. That is, a sample with smaller loss is taken as an "easy" sample and thus should be learned preferentially and vice versa.

Another regularization term favors selecting diverse samples residing in more bags. This can be easily understood by seeing that its negative leads to the group/bag-wise sparse representation of $\mathbf{v}$. Contrariwise, this diversity term should have a counter-effect to group-wise sparsity. That is, minimizing this diversity term tends to disperse non-zero elements of $\mathbf{v}$ over more bags, and thus favors selecting more diverse samples. Consequently, this anti-group-sparsity representation is expected to realize the desired diversity. Different from the commonly utilized $l_{2,1}$ norm, our utilized group-sparsity term is non-convex, leading to the convexity of its negative. This on one side simplifies the designation of the solving strategy, and on the other hand well fits the previous axiomic definition for the SPL regularizer [16, 31].

Algorithm 1: Algorithm of Optimizing $y_i^{(k)}$

Input: $\mathbf{x}_1^{(k)},\mathbf{x}_2^{(k)},\cdots,\mathbf{x}_{n_k}^{(k)}$, classifier parameters $\mathbf{w},b$;

Output: Pseudo-labels $\mathbf{y}^{(k)}=[y_1^{(k)},y_2^{(k)},\cdots,y_{n_k}^{(k)}]$.

1: $y_i^{(k)} = \arg\min_{y_i^{(k)}=\{+1,-1\}} \ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$ for $i=1,2,\cdots,n_k$;

2: if $\left\|\mathbf{y}^{(k)}+1\right\|_0<1$;

3: then $i^* = \arg\min_{i\in\{1,2,\cdots,n_k\}} v_i^{(k)}\ell(1, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$, $y_{i^*}^{(k)}=1$;

4: return $\mathbf{y}^{(k)}=[y_1^{(k)},y_2^{(k)},\cdots,y_{n_k}^{(k)}]$.

The alternative search algorithm can be readily employed to solve the optimization problem in (1), as introduced in the next section.

3.3. Optimization Strategy

The solution of (1) can be approximately attained by alternatively optimizing the involved parameters $\{\mathbf{w}, b\}$, $\{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}\}$ and $\mathbf{v}$ in (1).

Optimize $\{\mathbf{w}, b\}$ *under fixed* $\{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}\}$ *and* $\mathbf{v}$: This step aims to update the classifiers for detecting saliency areas. In this case, (1) degenerates to the following form:

$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)}\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b)), \quad (3)$$

which is the standard weighted SVM problem [32]. The model is convex and can be easily solved by off-the-shelf toolboxes [16].

Optimize $\{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}\}$ *under fixed* $\{\mathbf{w}, b\}$ *and* $\mathbf{v}$: The goal of this step is to learn the pseudo-labels of training instances from the current classifier. The SP-MIL model in this case is reformulated as:

$$\min_{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}} \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)}\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$$
$$\text{s.t.,}\ \left\|\mathbf{y}^{(k)}+1\right\|_0 \geq 1, k=1,2,\cdots,K_+. \quad (4)$$

This problem can be equivalently decomposed into sub-problems with respect to each $\mathbf{y}^{(k)}$, $k=1,2,\cdots,K_+$:

$$\min_{\mathbf{y}^{(k)}} \sum_{i=1}^{n_k} v_i^{(k)}\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$$
$$\text{s.t.,}\ \left\|\mathbf{y}^{(k)}+1\right\|_0 \geq 1. \quad (5)$$

The global optimum of (5) can be exactly attained by Algorithm 1, as clarified in the following theorem:

Theorem 1 *Algorithm 1 attains the global optimum to* $\min_{\mathbf{y}^{(k)}} \mathbf{E}(\mathbf{w},b,\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)},\mathbf{v})$ for each $\mathbf{y}^{(k)}$, $k=1,2,\cdots,K$ *independently under any given* $\{\mathbf{w}, b\}$ *and* $\mathbf{v}$ *in linearithmic time.*

The proof is presented in supplementary material.

Optimize $\mathbf{v}$ *under fixed* $\{\mathbf{w}, b\}$ *and* $\{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}\}$: After updating the pseudo-labels, we aim to renew the weights on all instances to reflect their different importance

Algorithm 2: Algorithm of Optimizing $\mathbf{v}$.

Input: K bag instances $\mathbf{X}_1,\cdots,\mathbf{X}_K$ with their current labels, instance detector $\{\mathbf{w}, b\}$, two parameters λ and γ;

Output: The global solution $\mathbf{v}=[\mathbf{v}^{(1)},\cdots,\mathbf{v}^{(K)}]$ of $\min_{\mathbf{v}} \mathbf{E}(\mathbf{w},b,\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)},\mathbf{v})$.

1: for k=1 **to** K **do**
2: Sort the instances in $\mathbf{X}_k$ in ascending order of their loss values, i.e., $l_1^{(k)} \le l_2^{(k)} \le \ldots \le l_{n_k}^{(k)}$; Let m=0;
3: **for** i=1 **to** n_k **do**
4: **if** $l_i^{(k)} < \lambda+\gamma/2\sqrt{i}$ **then** $v_i^{(k)}=1$;
5: **if** $l_i^{(k)} \ge \lambda+\gamma/2\sqrt{i}$ **then** count the number m of $l_j^{(k)}=l_i^{(k)}$ for $j=i, i+1, \cdots, n_k$, let $v_i^{(k)}=\cdots=v_{i+m-1}^{(k)}=\left(\left(\gamma/2\left(l_i^{(k)}-\lambda\right)\right)^2-(i-1)\right)/m$
6: and $v_{i+m}^{(k)}=\cdots=v_{n_k}^{(k)}=0$; **Break**;
7: **end for**
8: end for
9: return $\mathbf{v}$.

to learning of the current decision surface. As aforementioned, this model in this case is convex, and we propose a simple yet effective algorithm for extracting the global optimum to it, as listed in Algorithm 2, where $\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$ is simplified as $l_i^{(k)}$. The global minimum property is proved in the following theorem:

Theorem 2 *Algorithm 2 attains the global optimum to* $\min_{\mathbf{v}} \mathbf{E}(\mathbf{w},b,\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)},\mathbf{v})$ *for any given* $\{\mathbf{w}, b\}$ *in linearithmic time.*

The proof is also listed in supplementary material.

As shown, Algorithm 2 selects samples in terms of both the easiness and the diversity. Specifically:

- The "easy" instances with $l_i^{(k)} < \lambda$ will be selected into training process with v_i = 1 as in Step 4, where i is the sample's rank w.r.t. its loss value within its bag. While the "complex" instances with $l_i^{(k)} > \lambda+\gamma$ will not be involved into learning (i.e., v_i = 0) in Step 6.
- Instances with $l_i^{(k)} \le \lambda+\gamma/2\sqrt{i}$ will be selected as training samples with real-valued $v_i \in (0,1]$ in Step 4. Since the threshold decreases considerably as the rank i grows for each bag, Step 4 penalizes samples monotonously selected from the same group and thus naturally conducts diversity.

The whole alternative search process can then be summarized as Algorithm 3. According to [14], such an alternative search algorithm converges as the objective function $\min \mathbf{E}(\mathbf{w},b,\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)},\mathbf{v})$ is monotonically decreasing and is bounded from below.

Algorithm 3: Algorithm of SP-MIL.

Input: K bag instances $\mathbf{X}_1,\cdots,\mathbf{X}_K$, two parameters λ, γ;
Output: Instance detector $\mathbf{w}$, b.

1: Initialize $\{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}\}$, $\mathbf{v}$;
2: while *not converge* **do**
3: Update $\{\mathbf{w},b\}$ via weighted SVM;
Update $\{\mathbf{y}^{(1)},\cdots,\mathbf{y}^{(K_+)}\}$ via **Algorithm 1**;
4: Update $\mathbf{v}$ via **Algorithm 2**;
Renew the pace parameter λ, γ;
5: end while
6: return $\{\mathbf{w},b\}$.

4. Co-saliency Detection via SP-MIL

4.1. Feature Extraction

In order to extract useful information from the low-level contrast to the high-level semantics, we take advantage of all the convolutional layers in the convolutional neural network (CNN) to establish the hypercolumn feature representation [33] for each superpixel. Specifically, for each image, we first extract the features via a CNN pre-trained on the ImageNet with the same architecture as the "CNN S" model proposed in [34]. Then we up-sample all the five convolutional layers to the scale of the original image. Thus, we obtain a set of feature maps that can represent each pixel of the input image. To represent the superpixel regions extracted by [35], we max-pool the feature vectors located within each superpixel region. The obtained 1888 dimensional feature vectors are the hypercolumn representations.

4.2. Co-saliency Inference

In this section, we propose details of applying the SP-MIL algorithm, i.e., Algorithm 3, to automatically inferring the co-saliency of each superpixel region. First we discuss the initialization issue. As shown in Fig. 2, we need to initialize pseudo labels and SPL weights for all training instances firstly, and then iteratively train the co-saliency detector until convergence.

Since the co-salient object regions usually have the distinguishable appearances from the image background, we follow [36] to assume that the salient regions within a single image are most likely to contain parts of the co-salient object regions. Consequently, any off-the-shelf singe-image saliency detection approach can be adopted to roughly initialize the training samples. In this paper, we adopt the graph-based manifold ranking method [37] due to its computational efficiency. After obtaining the initial score of each superpixel, we select the top 10% superpixels in each positive bag, i.e., the image from the current image group, as the initial positive samples. For negative samples, we extract the Gist [38] and Color Histogram as the image

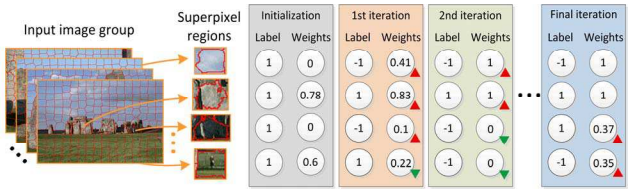

Figure 3: An example to illustrate the update of the samples' labels and SPL weights in each iteration.

feature and use the averaged image feature to represent the current image group. Then we follow [38] to search 20 similar images from other image groups based on the Euclidean distance. Finally, the bottom 10% superpixels in the searched images are selected as the negative samples. The weights of the initial positive samples are the initial saliency scores of these superpixels given by [37], while the weights of all the initial negative samples are equal to 1.

We then discuss the termination condition setting issue. Updating the co-saliency detector as well as the labels and weights of the training samples alternatively as the above description (see Fig. 3) could progressively lead to a strong co-saliency detector (see Fig. 4). To judge when the algorithm reach convergence, we calculate the Kullback-Leibler (KL) divergence of two Gaussian distributions which are inferred by the positive samples in the previous and the current iteration, respectively. Then, the convergence condition is reached when

$$D_{KL}^{*} < \tau D_{KL}, \tag{6}$$

where D_{KL}^{*} and D_{KL} are the KL divergences calculated in the current and the previous iteration, respectively, and τ is a free parameter. When the convergence condition is reached, we stop the iterative training process and predict the co-saliency of each sample, i.e., the superpixel region, via:

$$\mathrm{Cosal}(\mathbf{x}_i^{(k)}) = \mathbf{w}^T \mathbf{x}_i^{(k)} + b, \tag{7}$$

where $\mathbf{w}$ and b are the final converged solution of Algorithm 3.

4.3. Spatial Map Recovery

In SP-MIL, each superpixel is considered as an individual instance during the learning process. Thus, the spatial relationship among these superpixel regions, which is another important factor for the desirable co-saliency map, remains unexplored. In order to explore the spatial relationship of adjacent superpixels which are likely to be assigned to close co-saliency values, we adopt a graph model [37] to smooth the co-saliency values of each superpixel and thus obtain the spatial co-saliency maps.

Specifically, the graph is established by connecting the superpixels adjacent with each other as well as the superpixels at the four image boundaries. Then, we set an adaptive threshold (i.e. the mean value of the co-saliency values over the superpixels in one image) to select the foreground superpixels and use them to calculate the

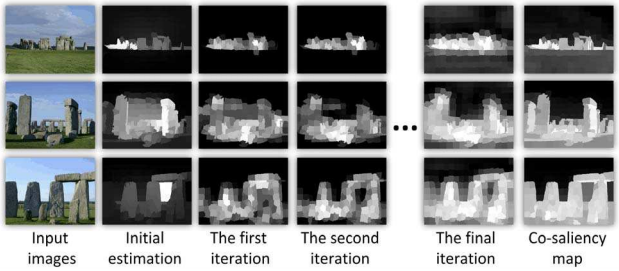

Figure 4: An example to illustrate the iterative co-saliency inference of our approach. It is seen that our approach can gradually converge to satisfactory results under conditions that the initialized estimation is incomplete (first row), imprecise (second row), and even totally wrong (third row).

co-saliency values of other superpixels in each image via a ranking function:

$$\Re = (\mathbf{D} - \alpha \mathbf{W})^{-1} \mathbf{q}, \tag{8}$$

where $\mathbf{D}$ and $\mathbf{W}$ are the affinity matrix and the degree matrix, respectively, $\mathbf{q}$ is the binary vector indicating which superpixels are the foreground query in one image, $\Re = \{r_i^{(k)}\}_{i=1}^{n_k}$ indicates the smoothed co-saliency values of the superpixels in such image, and α is a free parameter which is set to 0.99 according to [37]. Finally, the co-saliency map of the k-th image in a certain image group is obtained by:

$$\mathrm{Cosal}_{map}(\mathbf{x}_i^{(k)}) = r_i^{(k)}. \tag{9}$$

5. Experimental Results

5.1. Experimental Settings

We evaluated the proposed algorithm on two public benchmark datasets: the iCoseg dataset [28] and the MSRC dataset [39]. To the best of our knowledge, the former is one of the largest, publicly available, datasets so far that can be used for co-saliency detection. It contains 38 image groups of totally 643 images with manually labeled pixel-wise ground-truth masks. The later contains 7 image groups of totally 240 images with manually labeled pixel-wise ground truth masks. The complex background of the MSRC dataset makes it more challenging.

To evaluate the performance of the proposed method, we conducted comprehensive experiments based on three widely used criteria, which are the precision recall (PR) curve, the average precision (AP), and the F-measure. For each co-saliency map, we first segmented it via a series of fixed thresholds from 0 to 255. Then, the PR curve was drawn by using the precision rate versus the true positive rate (or the recall rate) at each threshold and the AP score was obtained by calculating the area under the PR curve. F-measure was obtained by using a self-adaptive threshold $T = \mu + \varepsilon$ as suggested in [40] to segment the co-saliency maps, where μ and ε are the mean value and the standard deviation of the co-saliency map, respectively. After

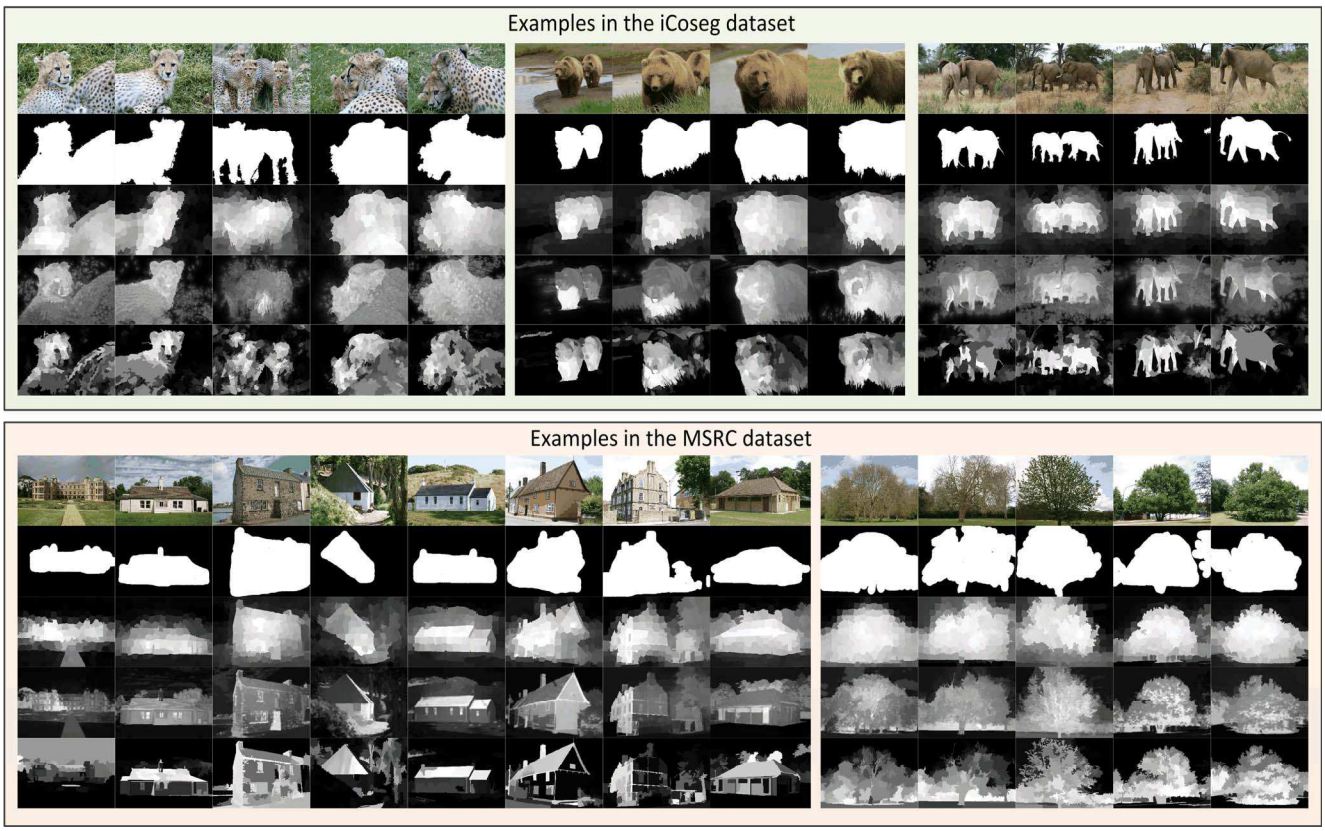

Figure 5: Subjective comparisons of the proposed approach and two state-of-the-art methods. For the examples in each dataset, the first row is the input image groups, the second row is the ground truth masks, and the 3-5 rows are the co-saliency maps generated by the proposed approach, SACS, and ESMG, respectively.

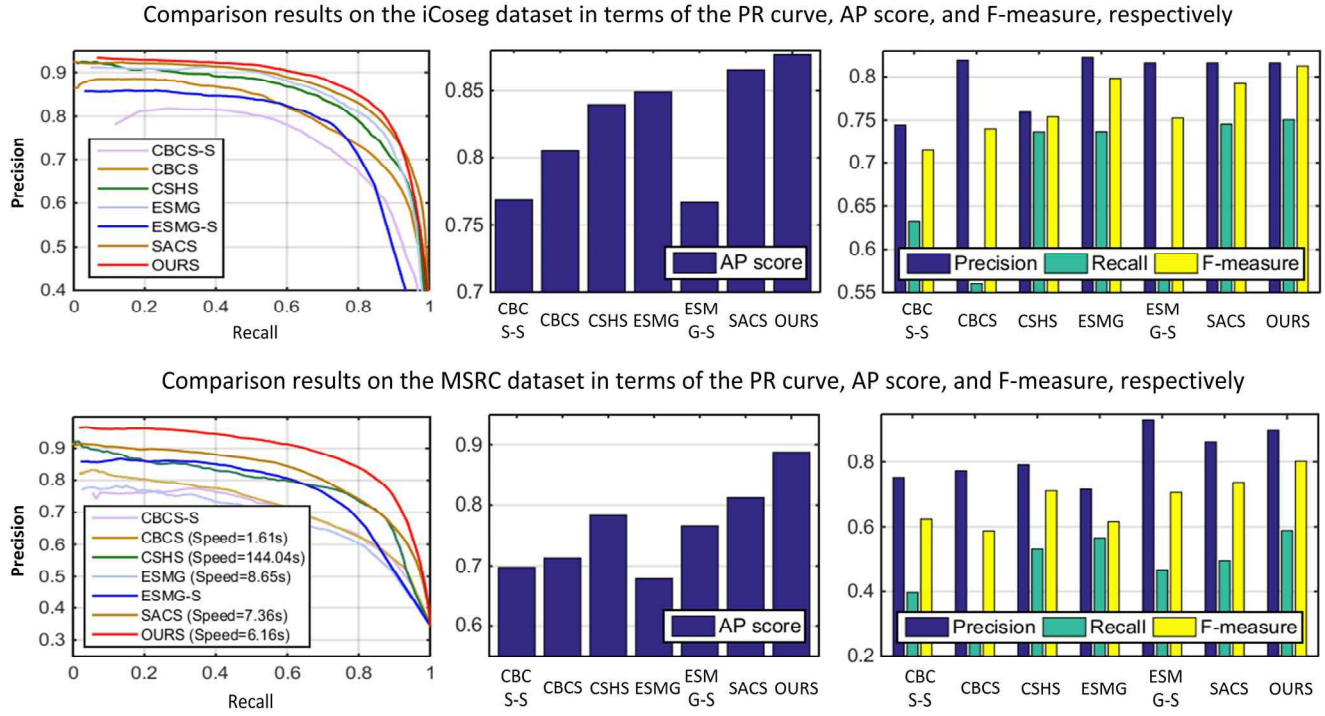

Figure 6: Quantitative comparisons of the proposed approach and other state-of-the-art methods.

obtaining the average precise and recall via the adaptive threshold T, we can obtain the F-measure as in [1, 37, 41].

In our experiments, the CNN was implemented via the MatConvNet toolbox [42]. As suggested in [15, 16], the λ and γ in (2) were chosen according to the number of the selected samples which was set to be 10% of the total superpixels in each image group. For the parameter in convergence condition, we empirically set τ=0.1 in (6).

5.2. Comparison with State-of-the-art Methods

In this section, we evaluated the proposed co-saliency detection approach by comparing with 6 state-of-the-art methods, i.e., CSHS [22], SACS [23], CBCS [21], CBCS-S [21], ESMG [24], and ESMG-S [24]. For qualitative evaluation, we show some co-saliency maps generated by the different methods in Fig. 5. The examples demonstrate that the proposed approach can uniformly highlight the co-salient regions even if they exhibit different poses, shapes, and points of view.

For quantitative comparison, we report the evaluation results in terms of the PR curve, the AP score, and the F-measure in Fig. 6. As can be seen, in the iCoseg dataset, the proposed approach obtains the highest precision when the recall rate is between 0 and 0.95. Thus, it outperforms other state-of-the-art methods both in terms of the AP score and the F-measure. The MSRC dataset is a more challenging dataset, where all state-of-the-art methods which are based on the hand-designed metrics cannot obtain as good performance as in the iCoseg dataset. However, the proposed approach obtains better performance than it does in the iCoseg dataset. Compared with other state-of-the-art methods, the proposed approach obtains obviously better performance in terms of all the three evaluation criteria. More specifically, the proposed approach outperforms the previous best method, i.e., the SACS, by 7.5% and 6.5% in terms of the AP score and F-measure, respectively. In addition, as shown in Fig. 6, the speed of our algorithm is faster than most previous works.

Furthermore, we applied our method for object co-segmentation in a large scale dataset [43] which contains thousands of noisy internet images. The experimental results (in terms of Jaccard similarity) also demonstrate our method (0.63) outperforms the state-of-the-art methods, e.g., Rubinstein's [43] (0.59) and Chen's [5] (0.61).

5.3. Model Analysis

Firstly, we compared the proposed SP-MIL with some baseline approaches, which are the initial estimation [37], the learning model without considering the sample diversity and the real-valued weights, and the learning model only considering the sample diversity. The quantitative comparison results on the two benchmark datasets are shown in Fig. 7 (a), from which we can observe

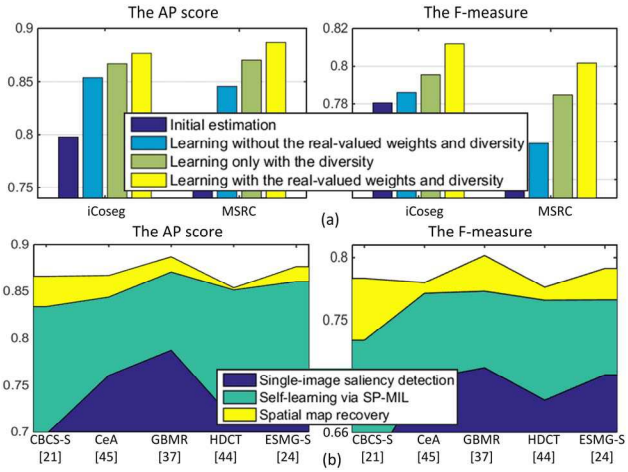

Figure 7: Analysis of the proposed SP-MIL model.

that: 1) even without considering the sample importance and diversity, the proposed SP-MIL framework can always work effectively to improve the initial estimation; 2) learning with considering the sample diversity could obviously improve the performance; 3) learning with considering the real-valued weights and sample diversity could always obtain the best performance.

To demonstrate the robustness of our method, we used some state-of-the-art single-image saliency methods [21, 24, 37, 44, 45] for initialization and reported the performance after the initialization, the learning, and the smoothing processes, respectively. The experimental results in Fig. 7 (b) demonstrate that 1) the single-image saliency detection methods can only provide coarse estimation for co-saliency detection; 2), SP-MIL makes significant contribution to the final performance; and 3) it is robust to different initialization methods as well.

6. Conclusion

In this paper, we have proposed a novel co-saliency detection approach which formulates the co-saliency detection under a MIL framework and introduces the SPL theory into the MIL framework for selecting training samples in a theoretically sound manner. In addition, two novel factors, i.e., the real-valued weights and sample diversity, were considered in the SPL model. The comprehensive experiments on two benchmark datasets have demonstrated the effectiveness of the proposed co-saliency detection approach as well as the robustness of the proposed SP-MIL model. For future work, we plan to apply the proposed method to more extensive real applications, such as image/video co-segmentation [5], co-localization [6], and weakly supervised learning [47].

Acknowledgements: This work was partially supported by the National Science Foundation of China under Grant 61473231, 61522207, 61373114, and 11131006.

References

[1] H. Li, F. Meng, and K. N. Ngan. Co-Salient object detection from multiple images. *TMM,* 15(8): 1896-1909, 2013.

[2] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu. Background prior based salient object detection via deep reconstruction residual. *TCSVT,* 25(8): 1309-1321, 2015.

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

[4] Y. Luo, M. Jiang, Y. Wong, and Q. Zhao. Multi-camera saliency. *TPAMI,* 37(10): 2057 - 2070, 2015.

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

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

[7] A. Joulin, K. Tang, and L. Fei-Fei. Efficient image and video co-localization with frank-wolfe algorithm. In *ECCV* 2014.

[8] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intell.*, 89(1): 31-71, 1997.

[9] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, 2002.

[10] J.-Y. Zhu, J. Wu, Y. Wei, E. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. In *CVPR*, 2012.

[11] P. Siva, and T. Xiang. Weakly supervised object detector learning with model drift detection. In *ICCV*, 2011.

[12] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren. Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning. *TGRS*, 53(6): 3325-3337, 2015.

[13] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[14] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[15] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014.

[16] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: self-paced reranking for zero-example multimedia search. In *ACM-MM,* 2014.

[17] H. Li and K. N. Ngan. A co-saliency model of image pairs. *TIP,* 20(12):3365-3375, 2011.

[18] H. Chen. Preattentive co-saliency detection. In *ICIP*, 2010.

[19] D. E. Jacobs, D. B. Goldman, and E. Shechtman. Cosaliency: where people look when comparing images. In *UIST*, 2010.

[20] Z. Tan, L. Wan, W. Feng, C. Pun. Image co-saliency detection by propagating superpixel affinities. In *ICASSP*, 2013.

[21] H. Fu, X. Cao, and Z. Tu. Cluster-based co-saliency detection. *TIP,* 22(10):3766-3778, 2013.

[22] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur. Co-saliency detection based on hierarchical segmentation. *SPL,* 21(1):88-92, 2014.

[23] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng. Self-adaptively weighted co-Saliency detection via rank constraint. *TIP,* 23(9): 4175-4186, 2014.

[24] Y. Li, K. Fu, Z. Liu, and J. Yang. Efficient saliency-model-guided visual co-saliency detection. *SPL,* 22(5): 588-592, 2014.

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

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

[27] Z. Wang, and R. Liu. Semi-supervised learning for large scale image cosegmentation. In *ICCV*, 2013.

[28] D. Batra, A. Kowdle, D. Parikh, J. Luo, and C. Tsuhan. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010.

[29] J. Supancic, and D. Ramanan. Self-paced learning for long-term tracking. In *CVPR*, 2013.

[30] Y. Tang, Y.-B. Yang, and Y. Gao. Self-paced dictionary learning for image classification. In *ACM-MM*, 2012.

[31] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[32] X. Yang, Q. Song, and Y. Wang. A weighted support vector machine for data classification. *Int. J. Pattern Recognit. Artificial Intell.*, 21(5): 961-976, 2007.

[33] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik. Hypercolumns for object segmentation and fine-grained localization. *arXiv preprint arXiv*:1411.5752, 2014.

[34] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: delving deep into convolutional nets. *arXiv preprint arXiv*:1405.3531, 2014.

[35] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, Su, x, and S. sstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *TPAMI,* 34(11): 2274-2282, 2012.

[36] 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.

[37] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[38] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking beyond the image: unsupervised learning for object saliency and detection. In *CVPR*, 2013.

[39] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[40] Y. Jia, and M. Han. Category-independent object-level saliency detection. In *ICCV,* 2013.

[41] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[42] A. Vedaldi, and K. Lenc. MatConvNet-convolutional neural networks for matlab. *arXiv preprint arXiv*:1412.4564, 2014.

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

[44] J. Kim, D. Han, Y. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *CVPR*, 2014.

[45] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *CVPR*, 2015.

[46] D. Zhang, J. Han, C. Li, and J. Wang. Co-saliency detection via looking deep and wide. In *CVPR*, 2015.

[47] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo. Weakly supervised learning for target detection in remote sensing images. *GRSL,* 12(4): 701-705, 2015.