Co-Saliency Detection via a Self-Paced Multiple-Instance Learning Framework

Dingwen Zhang, Deyu Meng, and Junwei Han

Abstract—As an interesting and emerging topic, co-saliency detection aims at simultaneously extracting common salient objects from a group of images. On one hand, traditional co-saliency detection approaches rely heavily on human knowledge for designing hand-crafted metrics to possibly reflect the faithful properties of the co-salient regions. Such strategies, however, always suffer from poor generalization capability to flexibly adapt various scenarios in real applications. On the other hand, most current methods pursue co-saliency detection in unsupervised fashions. This, however, tends to weaken their performance in real complex scenarios because they are lack of robust learning mechanism to make full use of the weak labels of each image. To alleviate these two problems, this paper proposes a new SP-MIL framework for co-saliency detection, which integrates both multiple instance learning (MIL) and self-paced learning (SPL) into a unified learning framework. Specifically, for the first problem, we formulate the co-saliency detection problem as a MIL paradigm to learn the discriminative classifiers to detect the co-saliency object in the "instance-level". The formulated MIL component facilitates our method capable of automatically producing the proper metrics to measure the intra-image contrast and the inter-image consistency for detecting co-saliency in a purely self-learning way. For the second problem, the embedded SPL paradigm is able to alleviate the data ambiguity under the weak supervision of co-saliency detection and guide a robust learning manner in complex scenarios. Experiments on benchmark datasets together with multiple extended computer vision applications demonstrate the superiority of the proposed framework beyond the state-of-the-arts.

Index Terms—Co-saliency detection, multiple-instance learning, self-paced learning

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, image groups from certain sources, e.g., similar subjects, websites, or companies, 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 for mining the intrinsic knowledge and facilitating further utilization from them. However, in practice, the image groups are achieved under complex settings and scenarios, like diverse background, illumination conditions, and view point variations. Consequently, this task is also of great challenge. To tackle this problem, co-saliency detection has been proposed and attracts intensive research attention in the recent years [1], [2], [3], [4], [5], [6], [7], [8], [9], [10].

The main aim of the co-saliency detection task is to explore the most important information, i.e., the common and salient foreground object regions from a group of images weakly pre-annotated as containing such similar objects, while their object categories, intrinsic characteristics and locations in images are entirely unknown. As one extension of the traditional saliency detection [11], [12], such a co-saliency detection task is evidently more challenging while more promising to perform in real applications. It is not only expected to be used in multi-camera systems [13] directly, but also hopeful to provide more precise common foreground prior to be served as a helpful preprocessing step for multiple related real-world applications, such as the video/image foreground co-segmentation [14] and image co-localization [15], [16]. In this paper, we not only focus on the fundament of this problem, i.e., developing insightful and effective approach for co-saliency detection, but also verify the effectiveness of using the proposed method to facilitate more practical computer vision applications.

Most previous methods for this task need to first manually design certain metrics for helping possibly explore the intra-image contrast and preserve the inter-image consistency, and then to directly integrate these metrics to demarcate the co-saliency regions among all images. There are, however, mainly two limitations in such learning framework. Firstly, the manually designed metrics are typically too subjective to generalize well and flexibly adapt various scenarios encountered in practice, especially due to the lack of thorough understanding of the biological mechanisms of human visual attention. Secondly, most current methods pursue co-saliency detection in unsupervised fashions. This, however, tends to weaken their performance in such complex scenarios, where there are only a relative small number of really informative

- *D. Zhang and J. Han are with School of Automation, Northwestern Polytechnical University, Xi'an, China. E-mail: zdw2006yyy@mail.nwpu.edu.cn, junweihan2010@gmail.com.*
- *D. Meng is with the Institute for Information and System Sciences, School of Mathematics and Statistics and Ministry of Education Key Lab for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an, China. E-mail: dymeng@xjtu.edu.cn.*

Manuscript received 13 Feb. 2016; revised 2 May 2016; accepted 4 May 2016. Date of publication 11 May 2016; date of current version 10 Apr. 2017. Recommended for acceptance by D. Xu.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2016.2567393

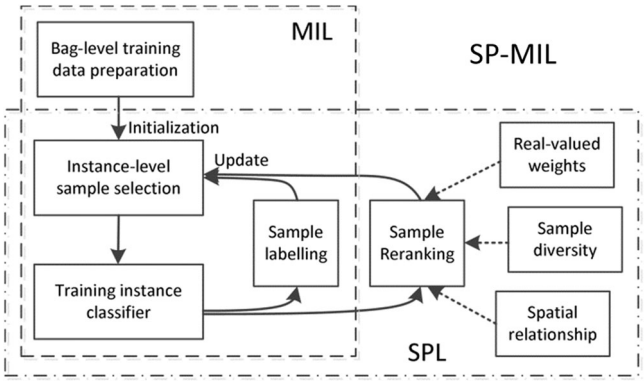


Fig. 1. Illustration of the self-paced multiple-instance learning model.

data. Actually, the co-saliency information is weakly annotated to the image groups, where we coarsely know if there are co-salient regions in images or not. However, the current methods have not made full use of such weak annotations in co-saliency detection.

For solving the first problem, we make the first effort to clarify a natural relationship between co-saliency detection and MIL problems, and accordingly utilize the latter to facilitate a self-learning strategy for producing the insightful metrics under the former problem. Specifically, in co-saliency detection, images with (without) containing a certain kind of co-salient objects can be considered as positive (negative) bags and superpixel regions in each image can be considered as instances. In this case, the co-saliency detection problem can be naturally reformulated as an MIL problem. Basically, the instance-level MIL aims at learning the classifiers which can minimize the intra-class distance among positive instances from each positive bag and maximize the inter-class distance between positive and negative instances. The classifiers so learned can then be utilized to predict the locations of the co-salient objects in the instance (super-pixel) level. Thus, by implementing MIL for co-saliency detection, the insightful metrics, which can well explore the intra-image contrast and preserve the inter-image consistency, are expected to be derived from the learned MIL classifiers, and co-saliency regions are hopeful to emerge from the recognized positive instances.

As for the second problem, we propose an effective way to model the problem in a weakly supervised fashion, which can better take advantage of the coarse-level image labels and gradually transfer them to more fine-level superpixel labels. Actually, since co-salient regions are always concealed inside many easily confused image regions in practical cases, robust learning strategy which can learn from subsets of high-confidence data and thus possibly avoid the chaos derived from the large amount of ambiguous data under such weak supervision context tends to be critical. Inspired by the recently proposed self-paced learning (SPL) theory, we embed an improved SPL paradigm into our framework to alleviate the data ambiguity and guide a robust learning manner in complex scenarios. The basic SPL model [17] mainly considers the sample easiness to gradually learn from the easy/faithful training samples to the complex/confusable ones. In this paper, we additionally explore two helpful prior knowledges on the structure of co-salient objects in SPL, which are the sample diversity, and spatial smoothness (as shown in Fig. 1), respectively. Specifically, the sample diversity encourages the learner to select training instances from a wide range of images in each given image group, which can enable the subsequent learning process to better take account of the co-salient objects in different scales, view-points, poses and shapes. The spatial smoothness encourages the spatial adjacent instances are assigned with similar impotent weights. Thus, the proposed sample diversity and spatial smoothness factors can work complementarily to enforce the real-valued weights of the training instances to have both the good inter-image property (diversity) and the good intra-image property (spatial smoothness). In addition, different from the $l_{2,1}$ -norm-based sample diversity regularizer [18], this paper ameliorates the sample diversity term as a convex negative $l_{0.5,1}$ -norm, which not only more complies with the original SPL axiomic definition, but also naturally leads to a more rational real-valued sample weighting scheme rather than the binary one as in [18].

According to the above discussion, a novel SP-MIL framework is naturally constructed for co-saliency detection as shown in Fig. 2. By intrinsically integrating the aforementioned MIL and SPL paradigms into a unified model, the proposed SP-MIL model is expected to be able to discover the faithful co-saliency patterns from the large number of ambiguous image regions. Specifically, the MIL component in SP-MIL facilitates to solve the co-saliency detection problem in a self-learning way, which is capable of automatically producing proper metrics to measure the intra-image contrast and the inter-image consistency for

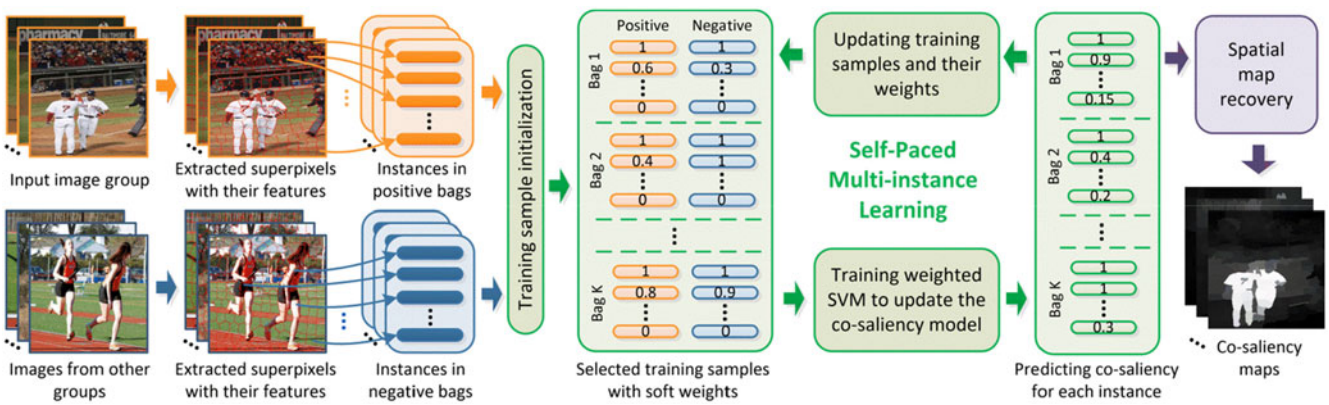


Fig. 2. The framework of the proposed SP-MIL approach for co-saliency detection.

co-saliency detection. In addition, integrated with the newly proposed self-paced regularizers, the SPL component in SP-MIL introduces an important sample selection mechanism to the learning framework and thus makes the proposed learning strategy more stable and robust in real complex scenarios. Consequently, the limitations existed in current co-saliency detection methods can thus be ameliorated under the proposed self-learning framework.

The basic framework of the proposed SP-MIL strategy is 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-salient object detector and annotate pseudo-labels for training instances in a SPL manner. Finally, the co-saliency maps are generated through spatial map recovery. In summary, the contributions of this paper are mainly four-fold:

- By clarifying a natural relationship between co-saliency detection and MIL, we incorporate the MIL component into the proposed framework to learn implicit metrics for co-saliency detection.
- We propose a novel SPL formulation via introducing two useful prior knowledges for co-saliency detection, i.e., the sample diversity, and the spatial smoothness, in the learning regime, which can lead to the convex solutions and real-valued sample weighting scheme.
- We propose a novel and general SP-MIL paradigm by integrating MIL and SPL into a unified model. The proposed SP-MIL model can gradually achieve faithful knowledge of co-saliency in a pure self-learning way.
- We extend the proposed approach to three typical computer vision applications and achieve the competitive or even better performance as compared with the state-of-the-art methods specifically designed on those problems.

The rest of this paper is organized as follows. Section 2 introduces related works. Section 3 presents the proposed SP-MIL model. Section 4 describes the details of the proposed SP-MIL framework. Section 5 shows experimental results to substantiate the effectiveness of the proposed method. Section 6 extends the proposed approach to three real-world computer vision applications. Finally, conclusions are drawn in Section 7.

2 Related Works

2.1 Co-Saliency Detection

Co-saliency detection aims at finding out the common and salient regions from a group of related images just as humans review them. Different from the conventional saliency detection, co-saliency detection needs to explore the interactive information in group level rather than in image level, which makes it far more challenging beyond traditional saliency detection issue in one image.

The first wave of co-saliency detection methods [1], [2], [3], [4] was developed to discover co-saliency from image pairs. Specifically, in 2010, Jacobs et al. [3] firstly defined visual co-saliency as the visual saliency of image pixels or regions in the context of other images. Around one year later, Li et al. [1] proposed a co-multilayer graph model to explore the multi-image saliency and established a public co-saliency dataset. Then, Chen [2] and Tan et al. [4] better enhanced the performance on this problem via the sparse-distribution-based representation and bipartite graph matching, respectively.

To further detect co-saliency from multiple images, the second wave of interest [5], [6], [7], [8], [9] appeared in 2013 with the work of Li et al. [5]. Afterwards, Fu et al. [6] proposed a cluster-based algorithm to more comprehensively explore the contrast cue, the spatial cue, and the corresponding cue to detect the co-salient regions. Liu et al. [7] further 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. [8] imposed a rank constraint to exploit the relationship of multiple predesigned saliency cues and then assigned the self-adaptive weight to generate the final co-saliency map. Very recently, Zhang et al. [10] made the earliest effort to introduce the deep and wide information for co-saliency detection, where their Bayesian framework integrated the designed intra-image contrast and intra-group consistency metrics. As can be seen, most existing methods heavily rely on manually designed metrics to explore the properties of the co-salient regions.

Compared with the conference version of the work [19], this paper makes the following extensions: 1) We introduce an additional term in the self-paced regularizer to explore the spatial smoothness of the superpixel instances during the learning process. Albeit looking simple, the added smoothness term can help evidently improve the performance of the proposed method, as verified in the results of Section 5. 2) We compared with more state-of-the-art approaches to demonstrate the effectiveness of the proposed framework in co-saliency detection task. 3) More comprehensive evaluations have also been conducted to analyze capability of the models and components in the proposed framework. 4) We substantially extended the applications of the proposed model to video foreground segmentation, activity localization, and object co-localization problems. All attained state-of-the-art performance in average as compared to the state-of-the-art techniques designed for those problems.

2.2 Co-Segmentation

Co-segmentation is closely related to the co-saliency detection. The difference of such two research topics mainly lies in three-fold aspects: 1) Co-saliency detection only focuses on discovering the common and salient objects from the given image groups while co-segmentation methods additionally tend to possibly precisely segment out the similar but non-salient background regions [20], [21]. 2) Co-segmentation usually needs semi- or interactive- supervision [22], [23] (where some object regions need to be labeled in advance) while co-saliency detection is implemented in an unsupervised or super-weakly supervised manner. 3) Compared with co-segmentation, co-saliency, as a concept stemming from human vision, usually needs to introduce common pattern analysis into the contrast-based visual attention mechanism to reveal the sampling strategy of human visual system. Thus it also receives greater interest in the field of visual cognition. As can be seen, image co-segmentation is a relative higher-level computer vision task, where co-saliency detection models can be applied to

replacing the user interaction to provide the informative prior knowledge of the visually similar objects under much weaker conditions and implemented as a pre-processing step for the co-segmentation task.

2.3 Weakly Supervised Semantic Segmentation

Another topic related to co-saliency detection is weakly supervised semantic segmentation (WSSS) [24], [25], [26], [27], [28], [29], whose goal is to assign semantic labels to each image pixel under the weak supervision of image-level annotations. This topic was firstly proposed by Vezhnevets et al. [24], which casted this task as a MIL problem and further leveraged the multi-task learning to import useful knowledge from a supplementary geometry context estimation task. Afterwards, they further proposed a multi-image model in [25] to recover the pixel labels via connecting superpixels from all training images in a data-driven fashion and a Bayesian optimization framework in [26] to select the best model in a pre-defined parametric family of CRF models without using superpixel labels. More recently, Yao et al. [27] described the intrinsic representation of superpixel regions via a discriminative deep feature learning framework and exploited both the global context about co-occurrence of visual classes and the local context around each image region to mine effective supervision information. Based on the reconstruction error, Zhang et al. [28] evaluated the basis superpixels of each category and obtained the optimal classification model parameter via an iterative merging update algorithm. Zhang et al. [29] presented a joint conditional random field model leveraging various contexts to address the social images with noisy image-level labels.

To our best knowledge, the most essential difference between WSSS [24] and co-saliency is that the former aims to learn the visual semantics from the weakly labeled data while the latter aims to learn the concept of co-saliency. As there is no direct relationship between the visual semantics and the concept of co-saliency, the tactics and methodologies designed to address these two tasks also tend to be different: WSSS usually leverages the label and task correlation information to assist the learning process whereas we more focus on the learning robustness issue for the co-saliency detection task. In addition, it is also obvious that WSSS inclines to understand the global scenes of the images while co-saliency detection only focuses on the co-salient object regions in the images. Thus, co-saliency detection could benefit more to the object-centric computer vision tasks as we demonstrated in the following application section (Section 6).

2.4 Multi-Instance Learning (MIL)

MIL was first proposed in [30] for classifying molecules in the context of drug design. A typical class of these MIL models focused on learning the desired classifiers in the "bag-level". These approaches, e.g., [31] and [32], usually need to design mechanisms to map the instances of each bag into a "bag-level" training vector, where each instance is served as an individual dimension in the new feature space. To solve the finer-level problems existed in some challenging computer vision tasks, e.g., weakly supervised object localization [33] and saliency detection [34], another class of the MIL models is proposed to learn the desired classifiers in the "instance-level". An inspirational work was the mi-SVM proposed by Andrews et al. [35], which heuristically solved the mixed integer quadratic programs in the extended SVM learning approach. Afterwards, Gehler et al. [36] introduced a novel objective function with the deterministic annealing algorithm to find better local minima. To further enhance the robustness of MIL, Cao et al. [37] introduced similarity weights to the instances and built an extended SVM-based predictive classifier by following a heuristic strategy.

With a generalized soft-margin SVM, a primal form of the instance-level MIL formulation can be written as

$$\begin{aligned} &\min_{\mathbf{w},b,\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}} \frac{1}{2}||\mathbf{w}||_2^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k} \ell\big(y_i^{(k)}, g\big(\mathbf{x}_i^{(k)};\mathbf{w},b\big)\big) \\ &s.t., \|\mathbf{y}^{(k)}+1\|_0 \geq 1, k=1,2,\ldots,K_+, \end{aligned} \tag{1}$$

which aims to learn discriminative functions to separate positive and negative instances and assign accurate labels to each instance in the positive bags. The notations of the variables will be explained in Section 3.1.

2.5 Self-Paced Learning (SPL)

Inspired by the learning process of humans/animals, the theory of self-paced (or curriculum) learning [17], [38] is proposed lately. The idea is to learn the model iteratively from easy to complex samples in a self-paced fashion. The effectiveness of such a newly proposed learning regime, especially its robustness in highly corrupted data, has been validated in various computer vision tasks. For example, Supancic et al. [39] used the formalism of SPL to automatically learn robust appearance model in object tracking. For accommodating the "hidden" information of the samples into the learning procedure, Tang et al. [40] proposed to adaptively select easy samples in each iteration to learn the powerful dictionary. In multimedia event detection [41], Jiang et al. proposed self-paced reranking models in a self-paced fashion. They also proposed a nonconvex regularizer to incorporate the information of diversity in [18]. Especially, the SPL paradigm has been integrated into the system developed by CMU Informedia team, and achieved the leading performance in challenging TRECVID MED/MER competition organized by NIST in 2014 [42]. Recently, Zhao et al. [70] also applied SPL in matrix factorization by gradually including matrix elements into the training process from easy to complex.

In SPL, the goal is to jointly learn the model parameters $\mathbf{w}$ and the latent weight variable $\mathbf{v}$ by minimizing:

$$\min_{\mathbf{w},b,\mathbf{v}} \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)} \ell\big(\mathrm{y}_i^{(k)}, g\big(\mathbf{x}_i^{(k)};\mathbf{w},b\big)\big) + f(\mathbf{v};\lambda), \tag{2}$$

which indicates the loss of a sample is discounted by a weight. The notation of the variables can be referred to in Section 3.1. The objective of SPL is to minimize the weighted training loss together with the self-paced regularizier $f(\mathbf{v};\lambda)$, where λ represents the age parameter for SPL. The axiom proposed in [18] has defined that the self-paced regularizer should be convex with respect to $\mathbf{v} \in [0,1]^n$, and the optimal weight of each sample should be monotonically decreasing with respect to its loss and increasing with respect to the age λ. By gradually increasing λ and implementing such self-paced learning process, the model can then be

gradually learned from easy to complex samples, which well simulates how human design curriculums to learn from young to old [17], [38].

3 Self-Paced Multi-Instance Learning

3.1 Notification

Let $\{\mathbf{X}_k\}_{k=1}^{K}$ denote the given K bags which include K_+ positive ones and K_- negative ones. By accumulating all instances at the kth image, we obtain $\mathbf{X}_k = \{\mathbf{x}_i^{(k)}\}_{i=1}^{n_k}$, where $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$ corresponds to the feature representation of the ith instance of the k-th bag, n_k is the instance number in $\mathbf{X}_k$, and $n = \sum_{k=1}^{K} n_k$ corresponds to the number of the entire instance set. The label set is defined as $\mathbf{y}^{(k)} = [y_1^{(k)}, y_2^{(k)}, \ldots, y_{n_k}^{(k)}] \in \mathbb{R}^{n_k}$, where $y_i^{(k)} \in \{-1, 1\}$ denotes the label of the instance $\mathbf{x}_i^{(k)}$. Without loss of generalization, we assume that the index set of all positive bags is $I_+ = \{1, \ldots, K_+\}$ while the negative ones $I_- = \{K_+ + 1, \ldots, K\}$. 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. Let $\mathbf{v} = [v_1^{(1)}, \ldots, v_{n_1}^{(1)}, v_1^{(2)}, \ldots, v_{n_2}^{(2)}, \ldots, v_{n_K}^{(K)}] \in \mathbb{R}^n$ denote the importance weights for all instances and $\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)}; \mathbf{w}, b))$ denote the hinge loss of $\mathbf{x}_i^{(k)}$ under the linear SVM classifier $g(\mathbf{x}_i^{(k)}; \mathbf{w}, b)$.

3.2 SP-MIL Model

The main idea of the proposed SP-MIL is 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:

$$\begin{aligned} &\min_{\mathbf{w}, b, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K_+)}, \mathbf{v} \in [0,1]^n} \mathbf{E}\big(\mathbf{w}, b, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K_+)}, \mathbf{v}\big) \\ &= \frac{1}{2}||\mathbf{w}||_2^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)} \ell\big(y_i^{(k)}, g\big(\mathbf{x}_i^{(k)}; \mathbf{w}, b\big)\big) + f(\mathbf{v}; \lambda) \\ &s.t., \|\mathbf{y}^{(k)} + 1\|_0 \geq 1, k = 1, 2, \ldots, K_+ \end{aligned} \tag{3}$$

where the constraint $\mathbf{y}^{(k)} + 1_0 \geq 1$ for each $k \in I_+$ enforces at least one positive instance in each positive bag.

The simplest SPL regularizer was proposed in [38], and is with the following form:

$$f(\mathbf{v}; \lambda) = -\lambda \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)} = -\lambda ||\mathbf{v}||_1, \tag{4}$$

where λ is the parameter imposed on the negative l_1 -norm term ($-||\mathbf{v}||_1$) which favors selecting easy over complex examples [38]. That is, a sample with smaller loss is taken as an "easy" sample and thus should be learned preferentially and vice versa. In addition, this regularizer 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 [17], [38]. However, when dealing with the task in more complicated real-world tasks, such a simple learning regularizer (with hard weights) is not strong enough for guiding a satisfactory sample selection. To this end, we introduce two novel learning criterions, each corresponding to some helpful prior knowledge underlying co-saliency images, into this traditional SPL regularizer to improve its learning capability.

The first learning criterion is the sample diversity. As we know, a rational curriculum for a pupil should include not only the examples of suitable easiness matching her learning pace but also some diverse examples for her to possibly develop more comprehensive knowledge. Likewise, a rational SPL should consider not only the sample easiness but also the sample diversity. This principle is more evident and natural in the co-saliency detection context. Essentially, the common objects appearing in each image of a certain image group are in different scales, view-points and exhibit different poses, shapes, structures and appearances (as shown in Fig. 5). Without the diversity regularizer, the learner tends to be stuck to a local region of training image space, which would lead to the poor sample selection performance during the learning process. Therefore, an additional diversity regularizer is needed here to encourage selecting instances from different images and thus enforce the learner to take consideration of all the aforementioned diverse factors. To this end, we propose a novel SPL regularizer:

$$f(\mathbf{v}; \lambda, \gamma) = -\lambda \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)} - \gamma \sum_{k=1}^{K} \sqrt{\sum_{i=1}^{n_k} v_i^{(k)}}, \tag{5}$$

where γ is the parameter imposed on the negative $l_{0.5,1}$ term which 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. Technically, any $l_{\mathrm{p},1}$ ($0 \leq \mathrm{p} < 1$) regularizer can be rationally utilized for this anti-group sparsity task and all of them also can guarantee the convexity of the SPL regime. We adopt $l_{0.5,1}$ just because under this term, a closed-form solution for our SPL model with respect to $\mathbf{v}$ can be easily deduced (Algorithm 2 in the main text). Such closed-form solution to other $l_{\mathrm{p},1}$ terms, especially the non-smooth $l_{0,1}$ term, however, is relatively more difficult to achieve. We thus directly utilize this one for efficiency and convenience.

The second learning criterion is the spatially smooth weights of the instances in each bag/image. As a complementary regularizer term with the diversity term which aims to select instances from different images in the image group, the proposed smoothness term is to encourage the selected instances in each image to be located with the smooth connections, i.e., the spatial adjacent instances are assigned with similar impotent weights. We show an example to study this problem more clearly in Fig. 3. As we can see, without using the smoothness term, the selected instances (the instances with $v_i^{(k)} > 0$) are located without a smooth connection and only a limited number of instances are selected in each image. On the contrary, by using the proposed smoothness term, the self-paced regularizer (6) tends to help select and assign similar important weights $v_i^{(k)}$ to the spatially adjacent superpixel instances. Thus, smoothing the important weight

vector $\mathbf{v}$ would benefit the learning process. In order to take into account the spatial smoothness over the superpixels, we supplementally adopt the Laplacian prior term [43],[44] into the self-paced regularizer as:

$$f(\mathbf{v};\lambda,\gamma,\eta) = -\lambda\sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)} - \gamma\sum_{k=1}^{K}\sqrt{\sum_{i=1}^{n_k} v_i^{(k)}} + \eta\sum_{k=1}^{K}\mathbf{v}^{(k)T}\mathbf{L}^{(k)}\mathbf{v}^{(k)}, \tag{6}$$

where $\mathbf{v}^{(k)T}\mathbf{L}^{(k)}\mathbf{v}^{(k)} = \frac{1}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{n_k}(v_i^{(k)} - v_j^{(k)})^2 W_{i,j}^{(k)}$, $\mathbf{L}^{(k)} = \mathbf{D}^{(k)} - \mathbf{W}^{(k)}$ denotes the Laplacian matrix for each bag, $\mathbf{D}^{(k)}$ is the diagonal weight matrix whose entries are column sums of $\mathbf{W}^{(k)}$, i.e., $D_{i,i}^{(k)} = \sum_j W_{j,i}^{(k)}$. As defined in [43], [45], $W_{i,j}^{(k)}$ is an element in the symmetry matrix $\mathbf{W}^{(k)}$, which is equal to $\exp(-||\mathbf{x}_i^{(k)} - \mathbf{x}_j^{(k)}||_2^2/\sigma^2)$ if the ith node and the jth node are adjacent in the k-th image and 0 otherwise. $\mathbf{x}_i^{(k)}$ is the extracted hypercolumn representation capturing the rich appearance information of the ith node/superpixel as described in Section 4.1 and σ^2 is set to be 0.1 as suggested by [45]. η is the parameter imposed on the spatial regularization term which encourages the solution with desired spatial smoothness. As can be seen, it is a simple but effect way to encourage the instances with close spatial distance to share similar importance weights. More importantly, by involving this regularization term, the obtained new regularizer still remains to be convex. Thus, it can be effectively solved by adopting some off-the-shelf convex optimization toolkits. As introduced in the next section, the gradient descend algorithm can be readily employed to solve the optimization problem in (6).

Algorithm 1. Algorithm of SP-MIL

Input: K bag instances $\mathbf{X}_1,\ldots,\mathbf{X}_K$, the parameters λ,γ,η;
Output: Instance detector $\{\mathbf{w}, b\}$.
1: Initialize $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}\}$, $\mathbf{v}^*$;
2: while *not converge* **do**
3: Update $\{\mathbf{w}, b\}$ via the weighted SVM;
4: Update $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}\}$ via Algorithm 2;
5: Update $\mathbf{v}$ via Algorithm 3 (if use the self-paced regularizer of (5)) or the CVX toolbox (if use the self-paced regularizer of (6));
6: end while
7: return $\{\mathbf{w}, b\}$.

3.3 Optimization Strategy

The solution of (3) can be approximately attained via the alternative search strategy which alternatively optimizes the involved parameters $\{\mathbf{w}, b\}$, $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}\}$ and $\mathbf{v}$ as shown in Algorithm 1. According to [38], such an alternative search algorithm converges as the objective function $\mathbf{E}(\mathbf{w}, b, \mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}, \mathbf{v})$ is monotonically decreasing and is bounded from below. More specifically, the optimization strategy contains the following steps:

Optimize $\{\mathbf{w}, b\}$ *under fixed* $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}\}$ *and* $\mathbf{v}$: This step aims to update the classifiers for detecting salient areas. In this case, (2) degenerates to the following form:

$$\min_{\mathbf{w},b}\frac{1}{2}||\mathbf{w}||_2^2 + \sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)}\ell\big(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b)\big) \tag{7}$$

which is the standard weighted SVM problem [46]. The model is convex and can be easily solved by off-the-shelf toolboxes [41].

Optimize $\{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}\}$ *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:

$$\min_{\mathbf{y}^{(1)},\ldots,\mathbf{y}^{(K_+)}}\sum_{k=1}^{K}\sum_{i=1}^{n_k} v_i^{(k)}\ell\big(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b)\big) \quad s.t.,\ \|\mathbf{y}^{(k)}+1\|_0 \geq 1, k = 1,2,\ldots,K_+. \tag{8}$$

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

$$\min_{\mathbf{y}^{(k)}}\sum_{i=1}^{n_k} v_i^{(k)}\ell\big(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b)\big) \quad s.t.,\ \|\mathbf{y}^{(k)}+1\|_0 \geq 1. \tag{9}$$

As can be seen, for a positive bag (i.e., an image containing co-salient object), it should at least contains one instance (i.e., superpixel) with positive label. Thus when we use the current classifier to categorize all instances from a positive bag, if $\|\mathbf{y}^{(k)}+1\|_0 \geq 1$, then at least one instance is predicted as positive and the iteration can be continued; otherwise, if $\|\mathbf{y}^{(k)}+1\|_0 < 1$, i.e., all instances in this positive bag are categorized as negative, then we need to pick up one instance which brings the minimal cost increase when we predict it as positive. This can intuitively explain the meaning of the solution to (9) obtained by Algorithm 1. The global optimum of (9) can be exactly attained by Algorithm 1, as clarified in the following theorem:

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

The proof is presented in the supplementary material, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2016.2567393.

Algorithm 2. Algorithm of Optimizing $y_i^{(k)}$

Input: $\mathbf{x}_1^{(k)}, \mathbf{x}_2^{(k)},\ldots,\mathbf{x}_{n_k}^{(k)}$, classifier parameters $\{\mathbf{w}, b\}$;
Output: Pseudo-labels $\mathbf{y}^{(k)} = [y_1^{(k)}, y_2^{(k)},\ldots,y_{n_k}^{(k)}]$.
1: $y_i^{(k)} = arg\min_{y_i^{(k)}\in\{+1,-1\}}\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$ for $i = 1,2,\ldots,n_k$;
2: if $\|\mathbf{y}^{(k)}+1\|_0 < 1$;
3: then $i^* = arg\min_{i\in\{1,2,\ldots,n_k\}} v_i^{(k)}\ell(1, g(\mathbf{x}_i^{(k)};\mathbf{w},b))$, $y_{i^*}^{(k)} = 1$;
4: return $\mathbf{y}^{(k)} = [y_1^{(k)}, y_2^{(k)},\ldots y_{n_k}^{(k)}]$.

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

model with either self-paced regularizer of (5) or (6) is convex. For solving the optimization problem, we use two effective algorithms for extracting the global optimum of the model with the self-paced regularizer of (5) and (6), respectively. To be specific, when considering the sample easiness and sample diversity, we design a novel algorithm to find the explicit solution of the model with the self-paced regularizer of (5). The detailed process is listed in Algorithm 2, where $\ell(y_i^{(k)}, g(\mathbf{x}_i^{(k)}; \mathbf{w}, b))$ is simplified as $l_i^{(k)}$. By satisfying the KKT (Karush–Kuhn–Tucker) conditions of the Lagrangian condition, the global optimum of (4) can be efficiently calculated, as proved in the following theorem:

Theorem 2. *Algorithm 3 attains the global optimum to* $\min_{\mathbf{v}} \mathbf{E}(\mathbf{w}, b, \mathbf{y}^{(1)}, \ldots \mathbf{y}^{(K_+)}, \mathbf{v})$ *with* $f(\mathbf{v}; \lambda, \gamma)$ *for any given* $\{\mathbf{w}, b\}$ *in linearithmic time.*

The proof is also listed in the supplementary material, available online.

Alternatively, when additionally considering the spatial relationship of the training samples, we need to find the optimal solution of the model with the self-paced regularizer of (6). In this relatively more difficult case, the explicit solution of the problem cannot be obtained in theory. The problem, however, is still convex and thus can also be efficiently solved by utilizing certain off-the-shelf optimization techniques, e.g., the CVX toolbox [47], to finely approach its global solution.

Algorithm 3. Algorithm of Optimizing $\mathbf{v}$ in the Model with the Self-Paced Regularizer (4)

Input: K bag instances $\mathbf{X}_1, \ldots, \mathbf{X}_K$ with their current labels, instance detector $\{\mathbf{w}, b\}$, two parameters λ and γ;
Output: Solution $\mathbf{v}$ of $\min_{\mathbf{v}} \mathbf{E}(\mathbf{w}, b, \mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(K_+)}, \mathbf{v})$.
1: for $k = 1$ **to** K **do**
2: Sort the instances in $\mathbf{X}_K$ in ascending order of their loss values, i.e., $l_1^{(k)} \leq l_2^{(k)} \leq \cdots \leq l_{n_k}^{(k)}$; 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** $l_i^{(k)} \geq \lambda + \gamma/(2\sqrt{i})$ **then** count the number m of $l_j^{(k)} = l_i^{(k)}$ for $j = i, i+1, \ldots, n_k$, let $v_i^{(k)} = \cdots = v_{i+m-1}^{(k)} = ((\gamma/2(l_i^{(k)} - \lambda))^2 - (i-1))/m$ and $v_{i+m}^{(k)} = \cdots = v_{n_k}^{(k)} = 0$; **Break**;
6: **end for**
7: end for
8: return $\mathbf{v}$.

4 Co-Saliency Detection via Self-Paced Multiple-Instance Learning

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 [48], [49], [50] for each superpixel. Specifically, for each image, we first resize it to 224×224 and then extract the feature maps via a CNN pre-trained on the ImageNet. With the same architecture as the "CNN-S" model proposed in [51], the CNN used in this paper consists of 13 convolutional layers, 5 pooling layers, and 1 fully connected layer, where the 5 convolutional layers before each polling layer are considered as the feature maps representing the image content from low level to high level. As the convolution and polling operations in CNN lead to feature maps in different scales, we up-sample each of the feature maps to the scale of the original input image. Then, the obtained feature maps can represent each pixel of the input image. To represent the superpixel regions extracted by [52], we max-pool the feature vectors located within the corresponding superpixel regions. The obtained 1,888 dimensional feature vectors are the hypercolumn representations for the instances needed to be inferred in our SP-MIL model.

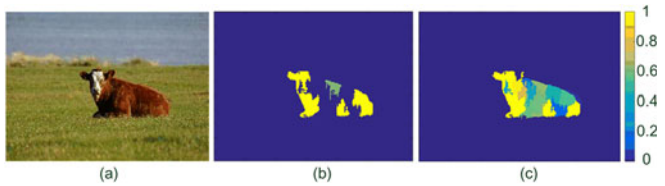


Fig. 3. An example to visualize the selected training instances with their important weights. (a) One image in the given image group. (b) The selected training instances with their important weights obtained without using the spatial smoothness term. (c) The selected training instances with their important weights obtained using the spatial smoothness term.

4.2 Co-Saliency Inference

In this section, we propose details of applying the SP-MIL algorithm, i.e., Algorithm 1, to automatically inferring the co-saliency of each superpixel region. First we discuss the initialization issue. As shown in Fig. 3, we need to initialize pseudo labels and SPL weights for all training instances firstly, and then iteratively train the co-saliency detector until convergence.

For initialization, we can just easily take any off-the-shelf single-image saliency detection approach. In this paper, we adopt the graph-based manifold ranking method [45] due to its computational efficiency. After obtaining the initial score of each superpixel, we select the top 10 percent superpixels in each positive bag, i.e., the image from the current image group, as the initial positive samples. For negative samples, we follow [53], [54], [55] to extract the Gist and Color Histogram as the image feature and use the averaged image feature to represent the current image group. Then we follow [53] to search 20 similar images from other image groups based on the Euclidean distance. Finally, the bottom 10 percent 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 [45], 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 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

$$D_{KL}^* < \tau D_{KL}, \tag{10}$$

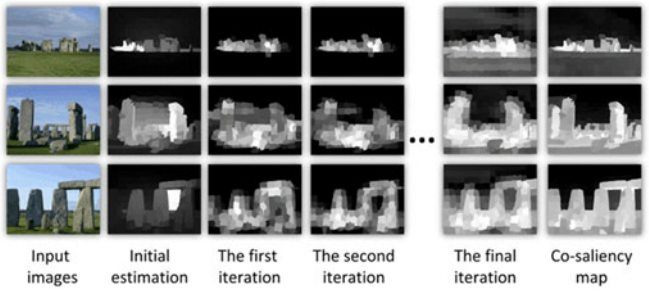


Fig. 4. An example to show that our approach can gradually converge to satisfactory results under conditions that the initial estimation is incomplete (row 1), imprecise (row 2), and even totally wrong (row 3).

where D_{KL}^{*} and D_{KL} are the KL divergences calculated in the current and the previous iteration, respectively, and τ is a constant. 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

$$Cosal\big(\mathbf{x}_i^{(k)}\big) = \mathbf{w}^T\mathbf{x}_i^{(k)} + b, \tag{11}$$

where $\mathbf{w}$ and b are the final converged solution of Algorithm 1.

4.3 Spatial Map Recovery

In order to obtain co-saliency maps with satisfactory spatial recovery, we explore the spatial relationship of the adjacent superpixels in each image by adopting a graph-based manifold ranking model. Different from the smoothness term used in (6), such spatial map recovery step is to enforce the adjacent superpixels to have similar co-saliency values to present the smoothly high-lighted co-salient object regions in the obtained co-saliency maps rather than encouraging the self-paced regularizer to select and assign similar important weights to the spatially adjacent superpixel instances during the learning process.

Specifically, we adopt a graph model to smooth the co-saliency values of each superpixel by considering its adjacent ones. 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 co-saliency values of other superpixels in each image via a ranking function [45]:

$$\Re^{(k)} = (\mathbf{D}^{(k)} - \alpha\mathbf{W}^{(k)})^{-1}\mathbf{q}^{(k)}, \tag{12}$$

where $\mathbf{q}^{(k)}$ is the binary vector indicating which superpixels are the foreground query in the k-th image, and $\Re^{(k)} = \{r_i^{(k)}\}_{i=1}^{n_k}$ indicates the smoothed co-saliency values of the superpixels in the k-th image. Finally, the co-saliency map of the k-th image in a certain image group is obtained by:

$$\mathrm{Cosal}_{map}\big(\mathbf{x}_i^{(k)}\big) = r_i^{(k)}. \tag{13}$$

5 Experimental Evaluation

5.1 Experimental Settings

We evaluated the proposed algorithm on two public benchmark datasets: the iCoseg dataset [23] and the MSRC dataset [56]. To the best of our knowledge the iCoseg dataset is the largest publicly available dataset so far widely used for co-saliency detection. It contains 38 image groups of totally 643 images with manually labeled pixel-wise ground-truth masks. The MSRC dataset 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 for co-saliency detection.

To evaluate the performance of the proposed method, we compared our approach with other state-of-the-art co-saliency detection methods based on five criteria: the precision recall (PR) curve, the average precision (AP), the ROC curve, the AUC, and the F-measure. In addition, we also conducted comprehensive experiments to analyze the properties of the model and components in the proposed framework based on the two most widely used measures AP and F-measure. To calculate these criteria 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 [57] 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 obtaining the average precision and recall via the adaptive threshold T, we defined the F-measure as:

$$\mathrm{F}_\beta = \frac{(1+\beta^2)\mathrm{Precision} \times \mathrm{Recall}}{\beta^2 \times \mathrm{Precision} + \mathrm{Recall}}, \tag{14}$$

where β^2=0.3 as suggested in [5], [45], [58].

In our experiments, the CNN was implemented via the MatConvNet toolbox [59] and the parameter τ in convergence condition (10) was empirically set to be 0.05. For the parameters in the proposed SP-MIL model in (5) and (6), we first chose the value of λ according to the number of the selected samples as suggested in [18], [41]. The number of the selected samples was set as 10 percent of the total superpixels in each image group. When determining the values of γ and η, to avoid the possible overfitting issue of selecting parameters according to the domain knowledge in the test data, we implemented parameter tuning on MSRC experiments by taking iCoseg as the validation set, and vice versa.

TABLE 1
Average Running Time of Each Programming Component for One Image

Operation	Cost (seconds)
SLIC superpixel extraction	2.49
CNN feature extraction	1.79
Optimization with regularizer (5)	1.37
Optimization with regularizer (6)	7.46
Saliency Map Recovery	0.51

5.2 Running Time Analysis

Table 1 lists the average execution time of each programming component in processing an image. The experiment was run on a workstation with two 2.8 GHz 6-core CPUs, 64 GB memory, additionally with a GTX Titan black GPU

TABLE 2
Average Running Time per Image

	CBCS	CSHS*	SACS	ESMG*	OURS-fast	OURS
Time(s)	1.61	144.04	7.36	8.65	6.16	12.25

*Running time of CSHS and ESGM are obtained from [9] which took the experiments on a laptop with an Intel 1.8 GHz CPU and 8 GB RAM.

for acceleration. The code was implemented in CUDA, MATLAB, and C without optimization. From Table 1 we can observe that the feature extraction component (including the SLIC superpixel extraction and the CNN feature extraction) is the most timing consuming part of the algorithm. In addition, co-saliency inference with regularizer (5) is faster than regularizer (6). Consequently, we define the proposed SP-MIL model with the regularizer (5) as the "OURS-fast" model and define the proposed SP-MIL model with the regularizer (6) as the "OURS" model.

We further implemented experiments to compare the average execution time of our models with other state-of-the-art methods. From Table 2 we can observe that the computational cost of OURS-fast model is less than most of the state-of-the-art methods and OURS model also has competitive computational complexity which is much less than the CSHS [6] method. More encouragingly, both OURS-fast and OURS model can obtain the best performance as evaluated in the next section.

5.3 Comparison with State-of-the-Art Methods

In this section, we compared the proposed co-saliency detection approach with 9 state-of-the-art methods, including CSHS [7], SACS [8], CBCS [6], CBCS-S[1][6], ESMG [9], ESMG-S[1][9], EQCUT [44], HDCT [60], and RBD [61]. Qualitative and quantitative comparison results on two benchmark datasets, i.e., the iCoseg dataset and the MSRC dataset, are shown in Figs. 5 and 6, respectively.[1]

Subjective comparison results: For subjective evaluation, we show some co-saliency maps generated by the different methods in Fig. 5, which consists of the examples from three image groups in the iCoseg dataset, i.e., the image groups of *Panda*, *Cheetah*, and *Bear*, as well as two image groups in the MSRC dataset, i.e., the image groups of *Face* and *Building*. The examples in the *Panda* and *Cheetah* groups indicate that the proposed approach can uniformly highlight the co-salient regions even if they exhibit different poses and shapes. The examples in the *Bear* and *Building* groups depict that the proposed approach can precisely detect the co-salient regions even if they are in different scales and points of view. The examples in *Face* group indicate that the proposed approach can robustly suppress the background regions even when they are very complex and interferential.

Quantitative comparison results: For quantitative comparison, we report the evaluation results in Fig. 6. As can be seen, in the iCoseg dataset, the proposed approaches obviously improve the precision even when the recall is high (from about 0.6 to 0.95), which implies that our methods have better capability to handle the tradeoff between the precision and recall. This also facilitates our methods to obtain the highest F-measure when using the adaptive threshold T to segment the co-saliency maps. In addition, the comparison of the AP and AUC scores in this dataset also demonstrates that the PR and ROC curves of the proposed methods are better than other state-of-the-art methods, respectively. Although SACS can also obtain competitive performance with the proposed methods in this dataset, the hand-designed metrics used by it cannot scope well in more complex scenes in the MSRC dataset as demonstrated below.

As for the MSRC dataset, all state-of-the-art methods cannot perform as well as in the iCoseg dataset as shown in Fig. 6. In such challenging case, the proposed approach obtains promising performance than other competing methods. It is easy to see that, as compared with other state-of-the-arts, the proposed approaches obtain obviously better performance in terms of all of the five evaluation criteria. In PR curve, the proposed methods obtain the highest precision along all the recall values. Similarly, the true positive rates of the proposed methods are also higher than other methods along all the false positive rates in ROC curve.

5.4 Model Analysis

In this section, we comprehensively evaluated the proposed SP-MIL model by comparing it with two traditional MIL models and analyzing some baseline performance. The traditional MIL models used in this experiment are the AL-SVM [36] and the mi-SVM [35]. In this experiment, we use the two traditional MIL models to replace the proposed SP-MIL model for co-saliency inference, while other parts of the algorithm in the same settings. The experimental results are shown in Fig. 7, which depict that the traditional MIL models cannot obtain satisfactory performance in co-saliency detection. The main reason is that the traditional MIL methods lack powerful sample selection theory to support them to discovery the most informative samples from a mass of ambiguous instances, especially when using them in the tasks with complex image content.

For further analyzing the factors considered in the proposed SP-MIL model, we reported the performance of some baseline models, including the SP-MIL model without considering the real-valued sample weights, the sample diversity, and the sample smoothness (OURS-NW-ND-NS), the SP-MIL model without considering the sample diversity and the sample smoothness (OURS-ND-NS), the SP-MIL model without considering the sample smoothness (OURS-NS), and the proposed SP-MIL model considering all the three factors. The quantitative comparison results on the two benchmark datasets are shown in Fig. 7, from which we can observe that: 1) Even only adopting the most basic self-paced regularizer, the proposed model, i.e., OURS-NW-ND-NS, can still largely improve the performance of the traditional MIL models. The reason is that OUS-NW-ND-NS tends to select the "easy" instances (as defined in Section 3.2) reflected by $\mathbf{v}$ during the learning process rather than directly using instances labelled by $\mathbf{y}$ as in traditional MIL models. Essentially, as the training instances in our task have large ambiguity, learning with all these instances without sample selection can hardly extract the truthful knowledge, which leads to significant performance drop of the AL-SVM and mi-SVM. 2) Each of the three factors proposed

1. CBCS-S is the single saliency model proposed in [6], and ESMG-S is the single saliency model proposed in [9].

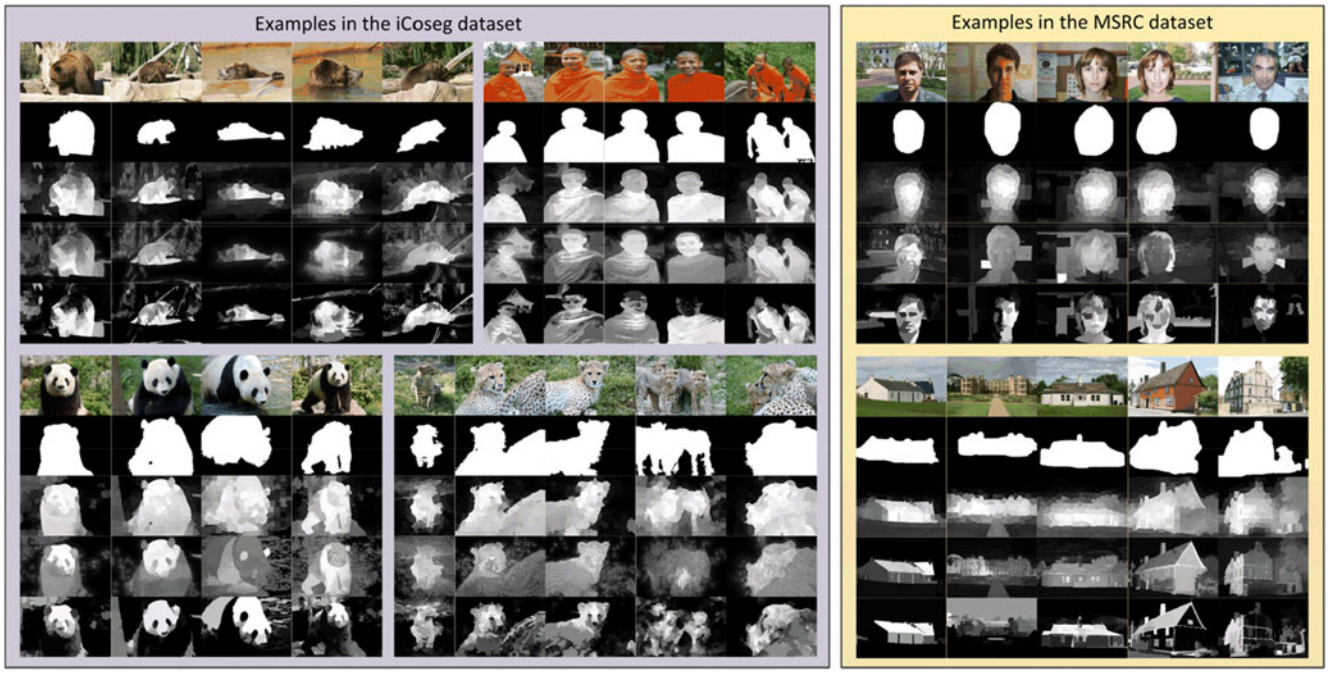


Fig. 5. Visual comparisons of the proposed SP-MIL approach and two state-of-the-art methods. For the examples in each dataset, the first row is the input image groups, the second row is the ground truth masks, and the 3-5 rows are the co-saliency maps generated by the proposed approach, SACS, and ESMG,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

in this paper could evidently contribute to the learning capability of the proposed SP-MIL model, which is invariant of the evaluation criterions and test datasets.

In addition, we also compared the proposed approach with a random selection baseline, where we selected the subset of the positive and negative training instances in each iteration randomly rather than based on the proposed self-paced regularizer. As the parameters $\lambda, \gamma,$ and η in our original SPL scheme are no longer used in this case, we introduced a new parameter φ to control the learning pace via

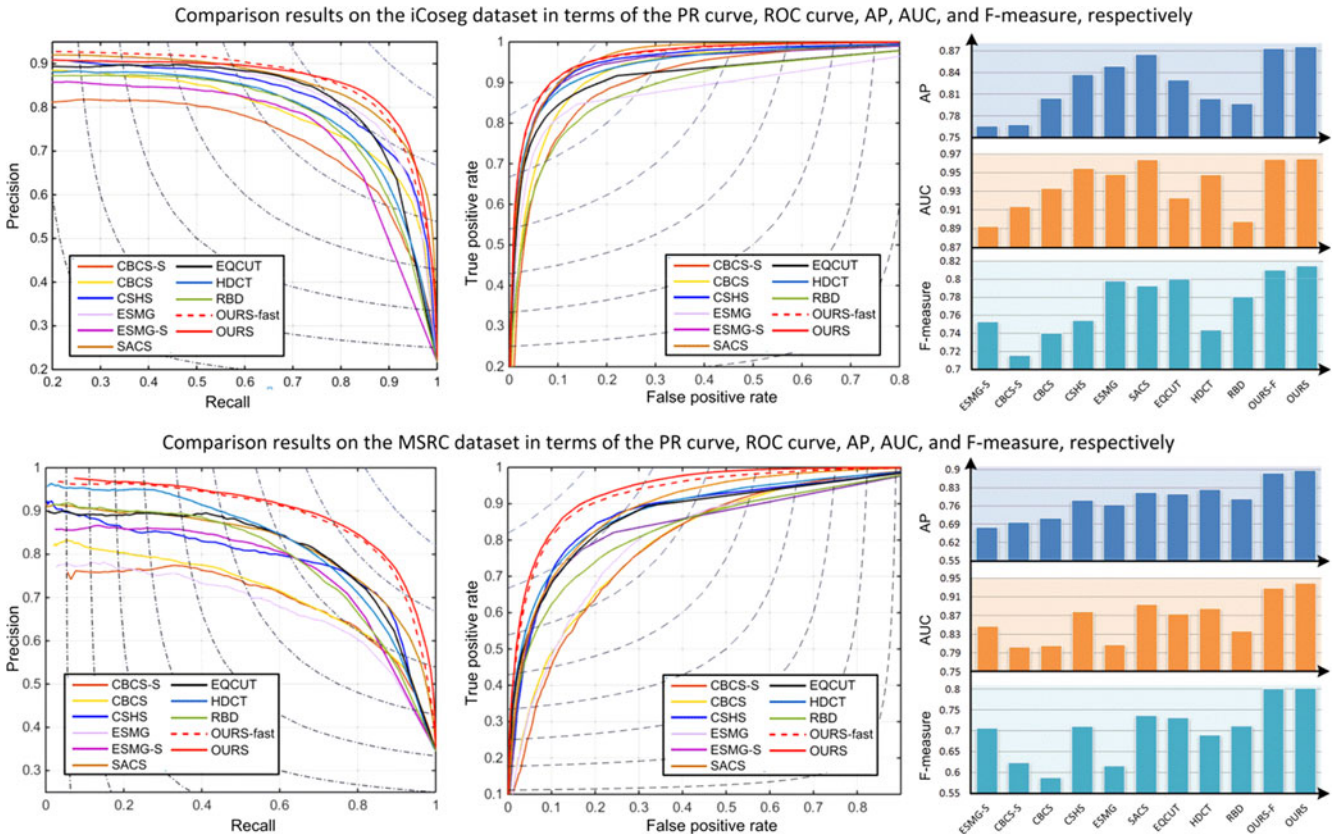


Fig. 6. Quantitative comparisons of the proposed approach and other state-of-the-art methods.

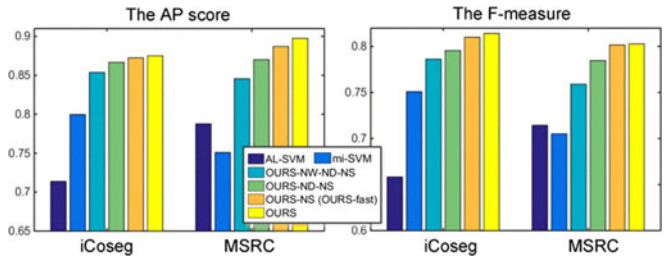


Fig. 7. Analysis of the proposed SP-MIL model, where NW, ND, NS indicate the SP-MIL model without considering the real-valued sample weights, the sample diversity, and the sample smoothness, respectively.

$\Phi_t = \varphi \cdot \Phi_{t-1}$, where Φ_t indicates the number of the selected training instances in the t-th iteration. We tuned this parameter with different values to guarantee that the sample number involved in learning in this strategy is also increased at an adequate pace for fair comparison and reported the experimental results in Fig. 8. As can be seen, such random selection baseline cannot achieve acceptable performance due to its poor training instance selection strategy, which reflects the importance of the well-organized sample selection scheme lead by the SPL strategy in this problem.

5.5 Component Evaluation

In this section, we further evaluated the contribution of the components, i.e., the training sample initialization, the co-saliency inference based on SP-MIL, the feature representation, and the spatial map recovery process, in the proposed framework. The experiments were implemented on the MSRC dataset. To demonstrate the robustness of the proposed method, we also reported results based on five different initialization methods. From Fig. 9 we can see that: 1) The initialization results obtained by 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 even with conventional features (the 1024 dimensional color SIFT-based BOW features used in this experiment), while learning with deep features could obtain better results. 3) Without using the spatial map recovery process, our approach obtains a marginal performance drop, which indicates such post-processing step is helpful to refine the co-saliency maps on the foundation of the SP-MIL prediction. 4) The proposed framework is robust to different initiation methods. In summary, all the components in the proposed framework can benefit for the final results, while the most fundamental improvement comes from the proposed SP-MIL model.

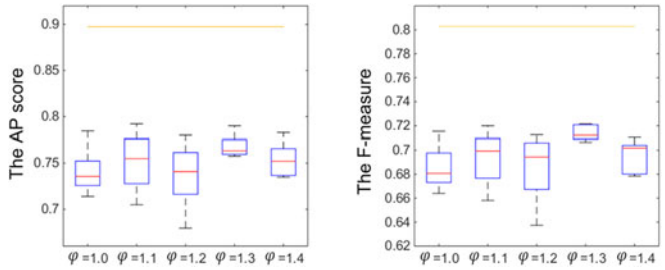


Fig. 8. The box plots display the experimental results of the random selection baseline in the MSRC dataset. The yellow lines indicate the performance of the proposed approach obtained based on the well-organized sample selection scheme.

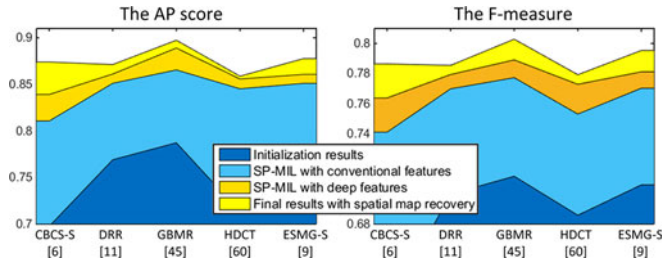


Fig. 9. Component evaluation of the proposed framework. Notice that the results of SP-MIL with conventional/deep features are obtained without using the spatial map recovery.

6 Applications

6.1 Video Foreground Segmentation

Video foreground segmentation aims at automatically discovering and segmenting the foreground objects in a given video [62]. For this task, the proposed co-saliency detection approach can be used to provide such regions of interest because the identical foreground objects in each frame usually share certain coherence. Different from the video saliency detection methods which attempt to discover the foreground objects based on the shape and motion cues, our co-saliency detection approach provides an alternative solution to discover such objects in videos.

To demonstrate the effectiveness of the proposed co-saliency detection approach in such an application, we trained the SP-MIL model in the Segtrack dataset [65] and adopted the self-adaptive threshold T to segment the obtained co-saliency maps. We also compared the results with two state-of-the-art video saliency detection approaches, i.e., VSMD [63] and SBSS [64], on this dataset. For the evaluation metric, we adopt the average per-frame pixel error rate $\epsilon(S) = |\mathbf{XOR}(S, GT)|/(F \cdot P)$, where S is each method's segmentation result, GT is the ground-truth segmentation, F is the total number of frames, and P is the total number of pixels in each frame. The evaluation results are shown in Table 3. Compared with the state-of-the-art video saliency detection methods, our approach can attain competitive performance. Notice that our failure in the "birdfall" is caused by the fact that the falling bird is not salient in each frame at all, which goes against with the assumption of the co-saliency detection aim.

6.2 Activity Localization

Activity localization aims at localizing every occurrence of a given action within a long video. The action or activity required to be localized in each frame is usually displayed

TABLE 3
Comparison between the Proposed Approach and Other State-of-the-Art Methods in Terms of the Average Per-Frame Pixel Error Rate on Segtrack

	VSMD	SBSS	OURS
BIRDFALL	**1.29%**	7.84%	8.54%
CHEETAH	5.71%	6.46%	**4.36%**
GIRL	5.03%	4.57%	**4.35%**
MONKEYDOG	15.17%	12.87%	**4.55%**
PARACHUTE	9.29%	2.90%	**0.37%**
PENGUIN	31.34%	22.47%	**20.51%**
Overall	11.31%	9.52%	**7.11%**

Fig. 10. Some examples of the activity localization results based on the proposed co-saliency detection framework.

by the same one (or more) person. Thus, applying co-saliency detection approach to discover such distinct and frequently occurring person would help to localize the activity in the video. To this end, we also apply the proposed approach to activity localization. Specifically, we follow Tran et al. [66] to make evaluation on 39 videos from the UCF-Sport dataset [67], which contain three different actions: the "running", "diving", and "horse-riding". To use the proposed approach, we first combine all the frames labelled as containing the same action category into one group. Then, we trained the SP-MIL model on these video frames and applied it to generating the co-saliency maps for each frame. Afterwards, the threshold T was easily used to segment the generated co-saliency maps to obtain the binary maps. Finally, the activities were localized by the bounding boxes generated by using the smallest rectangle to enclose the highlight regions in the binary maps of each frame. The experimental results are shown in Fig. 10 and Table 4. As can be seen, our method can obtain an average better performance as compared with two state-of-the-art techniques specifically designed on this task, i.e., OSTPD [68] and M2SOR [66]. Notice that OSTPD and M2SOR are the supervised methods needing human labelled action localizations during the training scheme, whereas our approach performs without using such information.

6.3 Object Co-Localization

Nowadays, it is of great interest to automatically co-localize the common objects in images obtained from the internet [15], [16], [69]. Similar to co-saliency detection, object co-localization aims at simultaneously localizing the common objects across a set of images which are usually obtained from social media. Instead of co-saliency maps, the output of object co-localization is the bounding boxes which include the common objects. Object co-localization is still of great challenging because the objects within them often display a large degree of intra-class variation and inter-class diversity. Since object co-localization shares the similar problem setting and challenges with co-saliency detection, it is expected that the proposed method can be beneficial for this task.

TABLE 4
Comparison between the Proposed Approach and Other State-of-the-Art Methods in Terms of the Localization Precision on UCF-Sport

	OSTPD	M2SOR	OURS
Running	0.2264	0.6186	**0.6746**
Diving	0.4809	0.3703	**0.7403**
Horse- Riding	0.6306	**0.6401**	0.4028
Overall	0.4460	0.5430	**0.6059**

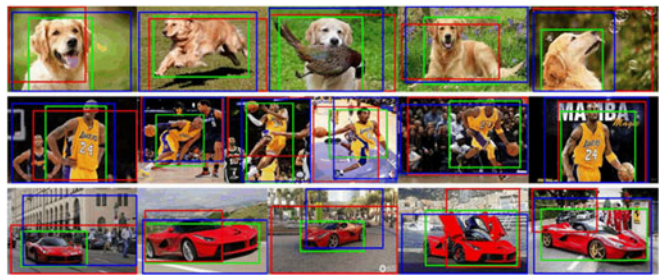


Fig. 11. Visual comparisons of our approach and two state-of-the-art methods for object co-localization in website images. The green boxes are the results of our method. The red and blue boxes are the results of Joulin's method [16] and Tang's method [15], respectively.

To demonstrate the effectiveness of the proposed approach in this application, we crawled the first 50 images returned by the Google image search with keywords "Golden Retriever", "Kobe, Lakers", and "LaFerrari". Since the noisy images containing uncommon objects are beyond the scope of the proposed method, we followed [69] to remove these images before the experiments. As shown in Fig. 10, it is easy to observe that the images collected in each image group are of high diversity in pose, background, and points of view. For obtaining the localization bounding boxes, we first applied the proposed approach to train the SP-MIL model in the crawled images and generate the co-saliency maps. Then the self-adaptive threshold T was easily used to segment the co-saliency maps. After obtaining the binary maps, we generated the bounding boxes by using the smallest rectangle to enclose the highlight regions. Following the previous work, we use the CorLoc evaluation metric. Fig. 11 and Table 5 display several co-localization results. It is interesting to see that even using such simple post-processing strategy, the proposed SP-MIL approach demonstrates evident benefit to this application, as compared with the previous state-of-the-art techniques designed for this task.

7 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 useful prior knowledges, including sample diversity and sample smoothness were rationally embedded into such formulated SP-MIL model.

TABLE 5
Comparison between the Proposed Approach and Other State-of-the-Art Co-Localization Methods in Terms of the Average CorLoc

	Tang's [15]*	Joulin's [16]*	OURS
Golden Retriever	0.65	0.50	**1.00**
Lakers' Kobe	0.70	0.70	**0.90**
LaFerrari	0.45	0.35	**0.95**
Overall	0.60	0.52	**0.95**

*The experimental results of Tang's method [15] are obtained by using the "box model" proposed in [15] while the experimental results of Joulin's method [16] are obtained by using the "image model" proposed in [16].

Comprehensive experiments on two benchmark datasets have demonstrated the effectiveness of the proposed co-saliency detection approach as well as the quality of the proposed SP-MIL model. Good performance of the proposed method in applications of video foreground segmentation, activity localization, and object co-localization has shown its potential usefulness in more extensive computer vision tasks.

For future work, we plan to 1) design a multi-class SPL model for learning of multiple categories of objects jointly; 2) establish effective framework to perform co-saliency detection in more practical yet challenging scenarios where the collected image groups may contain noisy images; and 3) perform large-scale co-saliency detection under a weaker assumption where the collected image group is not constrained to contain only one class of common objects.

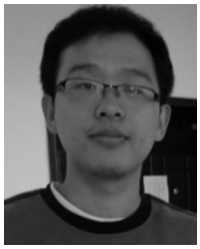
ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation of China under Grants 61522207, 61473231, and 61373114, the Doctorate Foundation, and the Excellent Doctorate Foundation of Northwestern Polytechnical University. J. Han is the corresponding author.

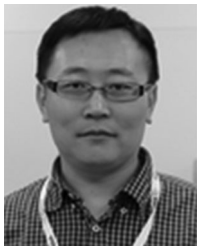
REFERENCES

[1] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20 no. 12, pp. 3365–3375, Dec. 2011.
[2] H. Chen, "Preattentive co-saliency detection," in *Proc. IEEE Int. Conf. Image Process.*, 2010, pp. 1117–1120.
[3] D. E. Jacobs, D. B. Goldman, and E. Shechtman, "Cosaliency: Where people look when comparing images," in *Proc. 23rd Annu. ACM Symp. User Interface Softw. Technol.*, 2010, pp. 219–228.
[4] Z. Tian, L. Wang, W. Feng, C. Pun, "Image co-saliency detection by propagating superpixel affinities," in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.*, 2013, pp. 2114–2118.
[5] H. Li, F. Meng, and K. N. Ngan, "Co-Salient object detection from multiple images," *IEEE Trans. Multimedia*, vol. 15 no. 8, pp. 1896–1909, Dec. 2013.
[6] H. Fu, X. Cao, and Z. Tu , "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22 no. 10, pp. 3766–3778, Oct. 2013.
[7] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21 no. 1, pp. 88–92, Jan. 2014.
[8] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23 no. 9, pp. 4175–4186, Sep. 2014.
[9] Y. Li, K. Fu, Z. Liu, and J. Yang, "Efficient saliency-model-guided visual co-saliency detection," *IEEE Signal Process. Lett.*, vol. 22, no. 5, pp. 588–592, May 2015.
[10] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2994–3002.
[11] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu, "Background prior-based salient object detection via deep reconstruction residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, pp. 1309–1321, Aug. 2015.
[12] M. -M. Cheng, G. -X. Zhang, N. J. Mitra, X. Huang, and S. -M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 409–416.
[13] Y. Luo, M. Jiang, Y. Wong, and Q. Zhao, "Multi-camera saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 10, pp. 2057–2070, Oct. 2015. DOI:10.1109/TPAMI. 2015.2392783.
[14] H. Fu, D. Xu, B. Zhang, and S. Lin, "Object-based multiple foreground video co-segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.
[15] K. Tang, A. Joulin, L. -J. Li, and L. Fei-Fei, "Co-localization in real-world images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 1464–1471.
[16] A. Joulin, K. Tang, and L. Fei-Fei, "Efficient image and video co-localization with Frank-Wolfe algorithm," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 253–268.
[17] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 41–48.
[18] L. Jiang, D. Meng, S. -I. Yu, Z. Lan, S. Shan, and A. Hauptmann, "Self-paced learning with diversity," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2078–2086.
[19] D. Zhang, D. Meng, C. Li, L. Jiang, Q. Zhao, and J. Han, "A self-paced multiple-instance learning framework for co-saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 594–602.
[20] A. Joulin, F. Bach, and J. Ponce, "Multi-class cosegmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 542–549.
[21] G. Kim, E. P. Xing, L. Fei-Fei, and T. Kanade, "Distributed cosegmentation via submodular optimization on anisotropic diffusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 169–176.
[22] Z. Wang, and R. Liu, "Semi-supervised learning for large scale image cosegmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 393–400.
[23] D. Batra, A. Kowdle, D. Parikh, J. Luo, and C. Tsuhan, "iCoseg: Interactive co-segmentation with intelligent scribble guidance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3169–3176.
[24] A. Vezhnevets, and J. M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multi-task learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2010, pp. 3249–3256.
[25] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised semantic segmentation with a multi-image model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 643–650.
[26] A. Vezhnevets, V. Ferrari, and J. M. Buhmann, "Weakly supervised structured output learning for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 845–852.
[27] X. Yao, J. Han, G. Cheng, and L. Guo, "Semantic segmentation based on stacked discriminative autoencoders and context-constrained weakly supervised learning," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 1211–1214.
[28] K. Zhang, W. Zhang, Y. Zheng, and X. Xue, "Sparse reconstruction for weakly supervised semantic segmentation," in *Proc. Int. Joint Conf. Artif. Intell.*, 2013, pp. 1889–1895.
[29] W. Zhang, S. Zeng, D. Wang, and X. Xue, "Weakly supervised semantic segmentation for social images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 2718–2726.
[30] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the multiple instance problem with axis-parallel rectangles," *Artif. Intell.*, vol. 89, no. 1,pp. 31–71, 1997.
[31] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp 1931–1947, Dec. 2006.
[32] Z. Fu, A. Robles-Kelly, and J. Zhou, "Milis: Multiple instance learning with instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 958–977, May 2011.
[33] P. Siva and T. Xiang, "Weakly supervised object detector learning with model drift detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 343–350.
[34] Q. Wang, Y. Yuan, P. Yan, and X. Li, "Saliency detection by multiple-instance learning," *IEEE Trans. Cybern.*, vol. 43, no. 2, pp. 660–672, Apr. 2013.
[35] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 561–568.
[36] P. V. Gehler, and O. Chapelle, "Deterministic annealing for multiple-instance learning," in *Proc. 11th Int. Conf. Artif. Intell. Statist.*, 2007, pp. 123–130.
[37] Y. Xiao, B. Liu, L. Cao, J. Yin, and X. Wu, "SMILE: A similarity-based approach for multiple instance learning," in *Proc. Int. Conf. Data Mining*, 2010, pp. 589–598.
[38] M. P. Kumar, B. Packer, and D. Koller, "Self-paced learning for latent variable models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 1189–1197.
[39] J. Supancic, and D. Ramanan, "Self-paced learning for long-term tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 2379–2386.
[40] Y. Tang, Y. -B. Yang, and Y. Gao, "Self-paced dictionary learning for image classification," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 833–836.

[41] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann, "Easy samples first: Self-paced reranking for zero-example multimedia search," in *Proc. 22nd ACM Int. Conf. Multimedia*, 2014, pp. 547–556.
[42] S. Yu and L. Jiang and Z. Maoet al. "CMU-Informedia@ TRECVID 2014 Multimedia Event Detection (MED)," in *Proc. TRECVID Video Retrieval Evaluation Workshop* 2014, http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.14.org.html.
[43] K. Fu, I. Y. Gu, C. Gong, and J. Yang, "Robust manifold-preserving diffusion-based saliency detection by adaptive weight construction," *Neurocomputing*, vol. 175, pp. 336–347, 2016.
[44] C. Aytekin, S. Kiranyaz, and M. Gabbouj, "Automatic object segmentation by quantum cuts," in *Proc. Int. Conf. Pattern Recog.*, 2014, pp. 112–117.
[45] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3166–3173.
[46] X. Yang, Q. Song, and Y. Wang, "A weighted support vector machine for data classification," *Int. J. Pattern Recog. Artif. Intell.*, vol. 21, no. 5, pp. 961–976, 2007.
[47] CVX Research, Inc, "{CVX}: Matlab Software for Disciplined Convex Programming, version 2.0.," Aug, 2012. [Online]. Available: http://cvxr.com/cvx.
[48] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 447–456.
[49] P. O. Pinheiro, R. Collobert, and P. Dollar, "Learning to segment object candidates," In *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1981–1989.
[50] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, "Hierarchical convolutional features for visual tracking," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3074–3082.
[51] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *Proc. Brit. Mach. Vis. Conf.*, 2014.
[52] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[53] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3238–3245.
[54] C. Lang, J. Feng, G. Liu, J. Tang, S. Yan, and J. Luo, "Improving bottom-up saliency detection by looking into neighbors," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 1016–1028, Jun. 2013.
[55] L. Mai, Y. Niu and F. Liu, "Saliency aggregation: A data-driven approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1131–1138.
[56] J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2005, pp. 1800–1807.
[57] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1761–1768.
[58] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2009, pp. 1597–1604.
[59] A. Vedaldi and K. Lenc, "MatConvNet-convolutional neural networks for MATLAB," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 689–692.
[60] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 883–890.
[61] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 2814–2821.
[62] Y. J. Lee, J. Kim, and K. Grauman, "Key-segments for video object segmentation," In *Proc. IEEE Int. Conf. Comput. Vis.*, 2011, pp. 1995–2002.
[63] C.-R. Huang, Y.-J. Chang, Z.-X. Yang, and Y.-Y. Lin, "Video saliency map detection by dominant camera motion removal," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 8, pp. 1336–1349, Aug. 2014.
[64] Z. Liu, X. Zhang, S. Luo, and O. Le Meur, "Superpixel-based spatiotemporal saliency detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 9, pp. 1522–1540, Sep. 2014.
[65] D. Tsai, M. Flagg, A. Nakazawa, and J. M. Rehg, "Motion coherent tracking using multi-label MRF optimization," *Int. J. Comput. Vis.*, vol. 100, no. 2, pp. 190–202, 2012.
[66] D. Tran, and J. Yuan, "Max-margin structured output regression for spatio-temporal action localization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 359–367.
[67] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2008, pp. 1–8.
[68] D. Tran, and J. Yuan, "Optimal spatio-temporal path discovery for video event detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2011, pp. 3321–3328.
[69] J. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr. 2015.
[70] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. Hauptmann, "Self-paced learning for matrix factorization," in *Proc. 29th AAAI Conf. Artificial Intell.*, 2015, pp. 3196–3202.



Dingwen Zhang received the BE degree from the Northwestern Polytechnical University, Xi'an, China, in 2012. He is currently working toward the PhD degree at Northwestern Polytechnical University. His research interests include computer vision and multimedia processing, especially on saliency detection, co-saliency detection, and weakly supervised learning.



Deyu Meng received the BSc, MSc, and PhD degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He is currently an associate professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University. From 2012 to 2014, he took his two-year sabbatical leave in Carnegie Mellon University. His current research interests include self-paced learning, noise modeling, and tensor sparsity.

Junwei Han received the PhD degree from Northwestern Polytechnical University in 2003. He is a professor with Northwestern Polytechnical University, Xi'an, China. He was a research fellow in Nanyang Technological University, The Chinese University of Hong Kong, and University of Dundee. He was a visiting researcher in University of Surrey and Microsoft Research Asia. His research interests include computer vision and brain imaging analysis. He is an associate editor of *IEEE Trans. on Human-Machine Systems*, *Neurocomputing*, *and Multidimensional Systems and Signal Processing*.