A Review of Co-saliency Detection Technique: Fundamentals, Applications, and Challenges

Dingwen Zhang, Huazhu Fu, Junwei Han, Senior Member, IEEE, Feng Wu, Fellow, IEEE

Abstract—Co-saliency detection is a newly emerging and rapidly growing research area in computer vision community. As a novel branch of visual saliency, co-saliency detection refers to discovery of the common and salient foregrounds existed in two or more relevant images, and can be more widely used in many computer vision tasks. The existing co-saliency detection algorithms mainly consist of three components: extracting effective features to represent the image regions, exploring the informative cues or factors to characterize co-saliency, and designing effective computational framework to formulate co-saliency. Although enormous methods have been developed, a deep review of the literatures concerning about the co-saliency detection technique is still lacking. In this paper, we aim to provide a comprehensive review of the fundamentals, challenges, and applications in co-saliency detection area. Specifically, this paper provides the overview of some related computer vision works, reviews the history of co-saliency detection briefly, summarizes and categorizes the major algorithms in this research area, presents the potential applications of co-saliency detection, discusses some open issues in this research area, and finally points out some unsolved challenges and promising future works. It is our hope that this review will be beneficial for both the fresh and senior researchers in this field as well as researchers working in other relevant fields to have a better understanding about what they can do with co-saliency detection in the future.

I. INTRODUCTION

In the past several years, a huge number of literatures from many different facets of science have been proposed with the aim of exploring the human visual attention mechanism and enabling the machine vision system to have such capability to selectively concentrate on a subset of interesting regions in each single image automatically [1], [2]. Rather than focusing on such relatively mature research area, in this paper, we are going to review a novel branch of visual saliency, i.e., co-saliency, which quickly gains extensive research interests in the most recent years. Different from the traditional saliency detection, co-saliency detection aims at discovering the common and salient foregrounds from an image group containing two or more relevant images, while the categories, intrinsic characteristics and locations of these objects are entirely unknown. Due to its superior scalability, co-saliency detection also has been widely used to access many other computer vision tasks, such as image/video co-segmentation [3]–[6], object co-localization [7]–[9], and weakly supervised learning [10], [11]. Thus, it is of great interest to review this newly developed co-saliency detection technique and provide the comprehensive discussion of the fundamentals existed in the major co-saliency detection methods, the existing and potential applications that can benefit from co-saliency detection technique, and the unsolved challenges that need to be addressed in the future to further promote the development of this research area.

Essentially, the problem of co-saliency detection is naturally emerged in the recent big data era. The numerous kinds of imaging equipments, like digital cameras and smart phones, have become much easier to acquire a large collection of image or video data, and some photo-sharing websites, like Flickr and Facebook, have been established to make such data extremely accessible. Consequently, people in the recent daily life are more likely to face with image collections, which typically are huge in size and share common objects or events, rather than the individual images as shown in Fig. 1. Compared to each single image, the multiple relevant images within an image group contain more rich and useful information. Due to the human visual attention mechanism, it is easy for us to feel that the frequently occurred patterns we capture during viewing the image groups influence our attention in each single...
Fig. 2. Given an image set, it contains the common foreground (CF), common background (CB), uncommon foreground (unCF), and uncommon background (unCB) regions. The task of co-saliency detection is to generate the co-saliency map highlighting the common foreground.

image and guide us to understand such images more clearly. From this phenomenon, the co-saliency task is derived, with the goal of establishing effective computational systems to endow such capability to the machines.

Although it is interesting and meaningful, the co-saliency detection task still remains to be of great challenge due to the complex scenarios in the practical image groups, which usually contain complex backgrounds, diverse illumination conditions, and view point variations. To tackle these problems, a series of works have been proposed from different motivations and solutions. It is summarized that the main problem in co-saliency detection is to explore the most important informations, i.e., the common and salient object regions (also called common foreground), from a group of images weakly pre-annotated as containing such similar objects. However, as depicted in Fig. 2, given an image set, such common foreground (CF) regions are submerged in tremendous of common background (CB) regions, uncommon foreground (unCF) regions, and uncommon background (unCB) regions. In order to discover the real CF regions from the noisy CB, unCF, and unCB regions, the existing co-saliency detection approaches mainly focus on solving three key problems: 1) extracting representative features to describe the image foregrounds; 2) exploring the informative cues or factors to characterize co-saliency; and 3) designing effective computational framework to formulate co-saliency.

Regarding the feature representation, a large number of co-saliency detection methods, e.g., [12]–[14], mainly used the low-level features such as color histogram, Gabor filter, and SIFT descriptor [15]. These methods assumed that the co-salient objects appearing in the multiple related images should share certain low-level consistency. Another group of methods, e.g., [16]–[19], also added the mid-level features to detect co-saliency. These methods usually consider the prediction results of the previous (co-)saliency detection algorithms as the mid-level features and seek to explore the consistency of the co-salient objects based on them. More recently, high-level semantic features were also proposed to be used in some co-saliency detection approaches [20]–[22] which assumed that the co-salient objects appearing in multiple images should share stronger consistency in high-level concept.

For the informative cues or factors, the most frequently used ones are the intra- and inter-image similarity. Specifically, the former is used to ensure the detected object regions are salient in each single image while the latter is used to ensure the detected object regions are commonly appeared in the given image group. Apparently, these two factors were designed corresponding to the basic definition of co-saliency as mentioned before. Except for these two factors, some other information cues, such as the objectness [18], [22]–[24], the center prior [13], [25], and the border connectivity [26], were also considered in co-saliency detection to obtain the desirable results. Moreover, Some works [20], [21], [27] have also introduced the inter-group separability to better discriminate the common foregrounds and background.

Regarding the computational framework, from the existing literatures we can observe that the trend changes from the bottom-up manner to the top-down manner. Specifically, in the early stage, most of the co-saliency detection methods, e.g., [13], [17], [23], [28], are designed in the bottom-up manner, where either the features or the computational metrics of the explored information cues are designed by heavily relying on the human knowledge and interpretation of the co-saliency detection task. With the rapid development of the machine learning technique, some co-saliency detection frameworks that are designed in the top-down manner are proposed in recent years. For example, Cao et al. [29] and Zhang et al. [21] adopted the fusion and self-learning approaches, respectively, to detect co-saliency based on the top-down priors inferred in the given image group.

While enormous methods exist, a deep review of the literatures concerning about the co-saliency detection techniques is still lacking. In this paper, we provide a comprehensive review of the recent progress in this area, which we believe could in many ways help: 1) the relative fresh researchers in exhaustively understanding the key techniques in this research area and incorporating them in this research area more easily; 2) the relative senior researchers in identifying the unsolved problems in this research area and tapping these problems towards advances in more effective and efficient systems; and 3) the researchers working on
other relevant tasks in being aware of what tasks can benefit from the co-saliency detection techniques and how they are benefited. The concrete organization of the remainder of this paper is as follows: Section II provides an overview of the works about some other computer vision tasks related to co-saliency detection. Section III reviews the history of co-saliency detection briefly. The major co-saliency detection algorithms are summarized in Section IV. The potential applications of co-saliency detection are introduced in Section V. And Section VI discusses some open issues in this research area. Finally, a conclusion with future work is proposed in Section VII.

II. RELATED WORKS

In this section, we discuss some related works of co-saliency detection technique: saliency detection, object co-segmentation, and weakly supervised localization. A brief summary is listed in Table I.

A. Saliency detection

Saliency detection aims at highlighting the regions which can attract the human visual attention in the single image. Generally, the saliency detection models are based on human visual attention, and can be divided into two categories: the eye fixation prediction model and the salient object detection model [1], [2]. Specifically, the former is used to predict the exact locations of the human attention when people are free-viewing one static natural scene, while the latter is used to detect the entire salient object that people are interested [1], [2]. To cope with the saliency detection, some earlier methods were mainly based on exploring the local contrast [30]–[32] and global contrast [33] with the assumption that the salient regions in each image should be distinctive from their sounding or the entire image context. Later on, some methods started to introduce some other factors, e.g., the background prior [34], [35], the high-level priors [36], [37], and depth information [38]–[40]. More recently, the deep neural networks were also adopted in saliency detection models [41]–[43], which have achieved remarkable performance.

Rather than modeling the human visual attention mechanism during the free-viewing of the single image, co-saliency makes a further step to make effort on figuring out what are the attractive things when people are viewing a group of related images. However, different from the saliency which has already been supported by a number of literatures in cognitive or psychological evidence [44], [45], little evidence about “visual co-attention” can be found in the existing literatures, even though people can feel the impact of the frequently occurring objects to their visual attention on each of those images. Recently, the researchers in the field of co-saliency detection believe that co-saliency is not only influenced by the contrast factor within each individual image but also determined by the consistent co-occurring patterns among the multiple related images, as shown in Fig. 3. Thus, co-saliency detection is a relative new trend than the saliency detection. It is worthwhile to mention that for most co-saliency detection methods, they could be also employed to detect saliency in single image with setting the image number to one.

B. Co-segmentation

Co-segmentation is used to generate the binary masks to segment out the common objects from an image group [46], as shown in Fig. 4 (C). Co-segmentation techniques aim to assign multiple labels to segment both the common things and stuffs rather than just separating the common foreground objects from the background regions [47]–[51]. More recently, by observing that in most applications of co-segmentation, the regions of interest are objects (things) rather than stuffs 1, Vicente et al. defined the object co-segmentation task more specifically and applied the objectness measure in their formulation explicitly [54]. Wang et al. proposed a semi-supervised learning algorithm to

1In this context, things usually refers to the objects like cars and cows while stuffs usually refers to the texture material like grass and rocks [52], [53].
TABLE I
COMPARISON OF CO-SALIENCY DETECTION AND THE RELATED WORKS.

|                       | Saliency detection | Image co-segmentation | Weakly supervised localization | Co-saliency detection |
|-----------------------|--------------------|-----------------------|-------------------------------|-----------------------|
| Basic model           |                    |                       |                               |                       |
| Image number          |                    |                       |                               |                       |
| Target                |                    |                       |                               |                       |
| Output                |                    |                       |                               |                       |
| Running time          |                    |                       |                               |                       |

exploit the similarity from both the limited training image foregrounds and the common objects shared among the unlabeled images [55].

According to our observation, 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 [50], [52]. 2) Co-segmentation usually needs semi- or interactive supervision [55], [56] (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. Although different, these two research areas also have strong relationship: as mentioned in [3], [4], [12], co-saliency detection can be applied to provide the useful prior of the common foregrounds for co-segmentation.

C. Weakly supervised localization

Weakly supervised localization (WSL) is another research topic closely related to co-saliency detection. The basic goal of WSL method is to jointly localize the common objects from the given weak-labeled images and learn their appearance models [57]–[59]. However, this actually leads to a chicken-and-egg scenario because to localize the common objects actually, one needs to have models capturing their appearance, and to lean such appearance models, one needs to localize the object instances in advance. In practice, most WSL works [60]–[62] choose to find ways to initialize the object instances firstly, and then jointly refine the appearance model of the common objects and the localization annotations. Specifically, Siva et al. [10] proposed to make use of the objectness, intra-image similarity, and inter-image variance to initialize the common object instances. Shi et al. [63] proposed to transfer the mapping relationship between the selection accuracy and the appearance similarity from an auxiliary fully annotated dataset to make initialization. Zhang et al. [64] and Han et al. [65] also adopted the saliency cue in their initialize stages. Song et al. [66] proposed to discover the frequent configurations of discriminative visual patterns for the initialization.

As can be seen, although no effort has been made to adopt co-saliency detection in WSL, the co-saliency detection technique still appears to have close relationship with the problems in WSL, especially for localizing the common objects from the given weak-labeled images. Eventually, co-saliency detection can be conveniently applied to improve the initialization performance of the common object instances in WSL because without additional prior knowledge or supervision, salient and common are the two useful information for initializing the WSL frameworks and they are what co-saliency detection models seek to figure out. The difference between co-saliency detection and WSL is also obvious: WSL needs to additionally lean the appearance models of the target category via incorporating more top-down factors and its final output is the bounding boxes with the corresponding confidence scores rather than the probability map outputted by the co-saliency detection algorithms.

III. BRIEF CHRONOLOGY OF CO-SALIENCY DETECTION

The history of co-saliency detection can be tracked back to the year 2010, when Jacobs et al. first defined the concept of co-saliency as calculating the importance, or saliency, of image pixels in the context of other related images [67]. After that, the first wave of interest started with the goal of discovering co-saliency from image pairs [18]. To extend co-saliency to image groups with more than two related images, the second wave of interest appeared in the year 2013 with the work in [13], which made use of multiple co-saliency cues, i.e., the contrast cue, spatial cue, and corresponding cue, to detect the co-salient regions. More recently, Zhang et al. [20]–[22] also introduced the widely used deep learning techniques in co-saliency detection.
A brief statistics of the publication history of the co-saliency detection techniques are shown in Fig. 5 and the major co-saliency detection algorithms are listed in Table II. As can be seen, even though the co-saliency detection technique only has five-year history, it has already obtained encouraging development and showed promising prospect in the future. Firstly, the quality of the co-saliency detection techniques has obtained tremendous promotion (from the ordinary publications, such as ICIP and UIST, to the top-ranked publications like TIP, IJCV, TPAMI, ICCV, and CVPR). Secondly, the quantity of the co-saliency detection techniques has also obtained obvious increment, especially for the most recent three years (more than half of the literatures were published during the year 2014 and 2015). It can also be found that the increment of the journal publications is more drastic than the conference publications. From the above analysis, we can see the development of co-saliency detection technique is in good condition and tends to become much better in the near future. More importantly, besides the development in terms of the quality and quantity of publications, the ideas and frameworks proposed in these literatures are also gained tremendous development during the past few years, which will be introduced in detail in the next section.

IV. THE MAJOR CO-SALIENCY DETECTION ALGORITHMS

The key problem of co-saliency detection is how to discover the common and salient foregrounds from the given image group. Based on this, strategies for co-saliency detection can be grouped into three main categories: bottom-up method, fusion-based method, and learning-based method.

A. Bottom-up method

The most common method to detection co-saliency is to score each pixel/region in the image group by using the manually designed co-saliency cues, which is defined as the bottom-up method \([13], [18], [27], [74], [75]\). For example, Li and Ngan proposed to detection pairwise co-saliency by exploring the single-image saliency and the multi-image co-saliency cues \([18]\). Afterwards, they further improved their previous model \([18]\) by using multi-scale segmentation voting to explore the object property of the foreground regions to generate the intra-image saliency map and extracting more powerful local descriptors to compute the inter-image saliency map \([17]\). Both of these two methods generated the final co-saliency maps by weighted combining the inter/multi-image saliency maps and the intra-image saliency maps. Differently, Fu et al. proposed a very efficient cluster-based algorithm \([13]\), which measured the cluster-level co-saliency by using three bottom-up saliency cues (i.e. the contrast cue, the spatial cue, and the corresponding cue). The global corresponding relation is built by clustering. In \([24]\), Liu et al. proposed a hierarchical segmentation based co-saliency model. They explored the local similarity and the object prior of each image region in the fine-level segments and coarse-level segments, respectively. Then, the global similarity was derived by simply summing the local similarities of several most similar segments from the other images. Finally, the co-salient objects were detected by fusing the local similarity, the global similarity, and the object prior.

Fig. 6 (A) gives the flowchart of bottom-up co-saliency detection methods, which mainly contains four important components: pre-processing, feature extraction, exploring bottom-up cues, and weighted combination. Specifically, the input images are first broken up into many computational units (pixel clusters or superpixels) in pre-processing step. Afterwards, features are extracted to explore the property of each computational unit based on the manually designed bottom-up cues. Finally, results obtained from each bottom-up cue are integrated together to generate the co-saliency maps for the input images. For the feature extraction, different kinds of features, including low-level, mid-level, and high-level features, have been adopted in co-saliency detection. Among these features, color feature is the most frequently used one while the high-level semantic feature is demonstrated in \([20]\) to be more effective during exploring the bottom-up cues. For the manually designed bottom-up cues, various cues have been explored in the existing methods. Within these cues, the contrast cue (also called single-image saliency or intra-saliency) and the corresponding cue (also called multi-image saliency, or inter-saliency) are the most widely used ones. A basic role is \([12], [13]\):

\[
\text{Co-saliency} = \text{Saliency} \times \text{Repeatedness}. \tag{1}
\]

For exploring the bottom-up cues, techniques adopted have changed from the early simple ones, e.g., matching and ranking, to the recent more powerful ones, e.g., matrix factorization and pattern mining. For the weighted combination, most bottom-up methods adopted simple strategies like multiplication or averaging, while some recent methods \([20], [28]\) have applied better ways to take advantage of both multiplication and averaging.

As the most popular way of co-saliency detection technique, bottom-up methods have obtained great development during the past few years. However, as this kind of co-saliency detection approaches heavily rely on manually designed bottom-up cues pre-defined in their frameworks, they 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 attention system.

B. Fusion-based method

Recently, fusion-based method to detect co-saliency in multiple related images was proposed \([28], [29], [76]\). Rather than devoting to discover informative cues from the collection of multiple related images for representing co-salient objects, the fusion-based method aims at mining the useful knowledge from the prediction results obtained by several existing saliency or co-saliency algorithms and then
TABLE II
A SUMMARY OF THE MAJOR CO-SALIENCY DETECTION ALGORITHMS.

| Name            | Year | Pub.   | Features                                                                 | Description                                                                 | Cues                                                                 | Img |
|-----------------|------|--------|---------------------------------------------------------------------------|-----------------------------------------------------------------------------|----------------------------------------------------------------------|-----|
| Jacob [67]      | 2010 | UIST   | Single saliency map (M), Nearest Neighbor error (L), Nearest Neighbor incoherence (L) | Learn to combine multi-features via SVM                                      | Single image saliency, change of dynamic object                       | 2   |
| Chen [16]       | 2010 | ICIP   | Sparse feature (M)                                                        | Minimize the distribution divergence on co-salient objects via KL-divergence | Inter-image similarity                                                | 2   |
| Chang [17]      | 2011 | CVPR   | SIFT (L)                                                                  | Saliency x Repeatedness                                                     | Single image saliency, inter-image similarity                        | many |
| Li [18]         | 2011 | TIP    | Single image saliency maps (M), color and texture (L)                     | Weighted combination of single-image saliency and inter-image similarity    | Single-image saliency, inter-image similarity                        | 2   |
| Fu [13]         | 2013 | TIP    | Pixel location (L), color (L), Gabor filter (L)                           | Fuse multiple cues in cluster level                                          | Contrast cue, Spatial cue, corresponding cue                        | many |
| Li [17]         | 2013 | TMM    | Pyramid color feature (L), and co-occurrence descriptor (M)               | Weighted combination of intra-image and inter-image saliency                | Objectness, co-occurrence                                           | many |
| Tan [14]        | 2013 | ICASSP | Color (L)                                                                 | Explore locality compactness and foreground cohesiveness via Bipartite graph matching and affinity propagation | Intra-image similarity (locality compactness), inter-image similarity (foreground cohesiveness) | 2   |
| Cao [29]        | 2014 | TIP    | Color (L)                                                                 | Multiple map fusion via low rank matrix approximation/recovery               | Foreground cohesiveness                                              | many |
| Cao [68]        | 2014 | ACM    | Color (L)                                                                 | Select co-saliency from multiple single-image saliency via reconstruction error of sparse coding | Foreground cohesiveness (correspondence + center prior)              | many |
| Li [23]         | 2014 | ICME   | Color (L), and spatial position (L)                                        | Region-level fusion + pixel-level refinement via low-rank matrix recovery    | Intra-saliency, object prior, global similarity                      | many |
| Liu [24]        | 2014 | SPL    | Color (L)                                                                 | Fuse intra-saliency, object prior, global similarity via Multi-level segmentation | Intra-saliency, object prior, global similarity                      | many |
| Chen [25]       | 2014 | ICPP   | Color, texture, contrast (L)                                              | Implicit Rank-Sparsity Decomposition via RPCA                                | Center prior, inter-image similarity, intra-image similarity         | 2   |
| Cheng [19]      | 2014 | VC     | Color (L), circularity (L), solidity (L), Fourier Descriptor (L), Shape Context (M) | Maximize between-image similarities and within-image distinctiveness via sketch based retrieval and GMM | Single-image saliency, appearance consistency, shape consistency     | many |
| Li [28]         | 2015 | SPL    | Color (L)                                                                 | Guide single image saliency to co-saliency via manifold ranking             | Single-image saliency, appearance consistency                        | many |
| Zhang [20]      | 2015 | CVPR   | Semantic feature (H)                                                     | Integrate deep semantic feature and wide inter-group heterogeneous via a Bayesian framework | Deep semantics, intra-image contrast, intra-group consistency       | many |
| Ye [26]         | 2015 | SPL    | Color (L), SIFT (L)                                                       | Recover co-salient object from pre-defined exemplars                        | Objectness, border connectivity, appearance consistency             | many |
| Zhang [21]      | 2015 | ICCV   | Deep feature (H)                                                         | Self-learning without handcrafted metrics based on SP-MIL                     | No explicit cues                                                    | many |
| Zhang [22]      | 2015 | TNNLS  | Color (L), steerable pyramid filter (L), Gabor filter (L)                 | Transfer intra-saliency prior and mine deep inter-saliency via SDAE          | Intra-saliency (contrast + object prior), Inter-saliency (consistency) | many |
| Du [69]         | 2015 | SPL    | Color (L)                                                                 | Determine inter-saliency from all the generated groups via nearest neighbor searching | Single saliency map, Inter-saliency                                 | many |
| Shen [70]       | 2015 | ICIMCS | Color (L)                                                                 | Extract and select the least and effective features for co-saliency via Sparse PCA | Same with CBCS                                                      | many |
| Huang [71]      | 2015 | ICME   | Position (L), color (L), texture (L)                                       | Multiple map fusion via low-rank matrix recovery                            | Multi-scale saliency                                                | many |
| Zhang [72]      | 2016 | IJCV   | Semantic feature (H)                                                     | Integrate deep semantic feature and wide inter-group heterogeneous via a Bayesian framework | Deep semantics, intra-image contrast, intra-group consistency, inter-group separability | many |
| Zhang [73]      | 2016 | TPAMI  | Deep feature (H)                                                         | Self-learning without handcrafted metrics based on SP-MIL                     | No explicit cues                                                    | many |

fusing those prediction results or the mined knowledge to generate the final co-saliency maps. Specifically, Cao et al. applied low-rank decomposition to exploit the relationship of the result maps of multiple existing saliency and co-saliency approaches to obtain the self-adaptive weights, and then used these weights to combine the multiple result maps for generating the final co-saliency map [29]. A reconstruction-based fusion approach is proposed in [68], which combined the knowledge mined from several existing saliency detection methods based on the reconstruction errors. Huang et al. fused the obtained multi-scale saliency maps by using low-rank analysis with introducing a GMM-
Based co-saliency prior [71].

Fig. 6 (B) illustrates the brief flowchart of the fusion-based co-saliency detection methods. Specifically, multiple coarse prediction maps generated by using several existing saliency or co-saliency detection methods in the given image set are obtained firstly. Then, the knowledge of the co-saliency regions (usually the consistency property) are mined from the obtained coarse prediction results. For example, the methods in [29], [68], [71] mined the consistency property among the salient regions through the rank constraint and the reconstruction error, respectively. Finally, the previously obtained prediction results are fused based on the guided knowledge to generate the final co-saliency maps by:

$$\text{Co-saliency} = \sum_i \text{Weight}_i \cdot \text{Submap}_i,$$

where Submap is the coarse prediction map, and Weight is the guided fusion weighting.

Generally, the fusion-based co-saliency detection methods can usually achieve promising results as they can make further improvement based on multiple existing (co-)saliency detection approaches. Another advantage of the fusion-based methods is that they can be flexibly embedded with various existing (co-)saliency detection methods, which makes them easier to adapt to different scenarios. However, as they heavily rely on the existing (co-)saliency detection methods, when most of the adopted (co-)saliency techniques lose their power, the final performance of the fusion-based co-saliency detection methods may be hurt badly.

C. Learning-based method

The third category of co-saliency detection technique is the learning-based method, which usually casts co-saliency detection as the classification problem for each image pixel/region. Actually, the earlier co-saliency detection method, i.e., Jacobs et al. [67], employed the supervised learning approach to learn the weights of the extracted cues, e.g., the contrast, semantics (faces), and motion cues, to predict co-saliency for each image pixel. More recent methods [19], [21] have relaxed the learning-based co-saliency detection technique from the (fully) supervised one to the unsupervised or weakly supervised (when considering the given image group as weakly labeled to contain one common object) ones, which can also be considered as the self-learning methods. As shown in Fig. 6 (C), these methods first adopted any off-the-shelf unsupervised saliency detection approach to provide an initial estimation. Then, the iterative self-learning schemes were designed to learn the appearance of the co-salient object (by using either the generative model like GMM in [19] or the discriminative model like SVM in [21]) and refine the obtained co-saliency maps gradually.
Among the learning-based methods, one typical model was proposed by Zhang et al. [21], which proposed a self-paced multiple-instance learning (SP-MIL) to address this problem. Given an image group containing $K_+$ images, they treated these images as the positive bags and searched $K_-$ similar images from other image groups as the negative bags. In each image, the extracted superpixels with their feature representation $x^{(k)}_i$ were treated as instances to be classified, where $k \in [1, K_+], K = K_+ + K_-$ denotes the bag index of each image and $i \in [1, n_k]$ denotes the index of the instance index of each superpixel in the $k$-th bag. The objective function of SP-MIL was formulated as:

$$\min_{w, b, y,(1), \ldots, y^{(K_+)}, v} E(w, b, y^{(1)}, \ldots, y^{(K_+)}, v) =$$

$$\frac{1}{2} ||w||^2 + \sum_{k=1}^{K} \sum_{i=1}^{n_k} v^{(k)}_i L(y^{(k)}_i, g(x^{(k)}_i; w, b)) + f(v; \lambda, \gamma)$$

s.t. $||y^{(k)} + 1||_0 \geq 1, k = 1, 2, \ldots, K_+$, (3)

where $v = [v^{(1)}_1, \ldots, v^{(1)}_{n_1}, v^{(2)}_1, \ldots, v^{(2)}_{n_2}, \ldots, v^{(K_+)}_{n_{K_+}}] \in \mathbb{R}^{n}$ denotes the importance weights for the instances, $y^{(k)} = [y^{(k)}_1, y^{(k)}_2, \ldots, y^{(k)}_{n_k}] \in \mathbb{R}^{n_k}$ denotes the labels for instances in the $k$-th bag, $L(y^{(k)}_i, g(x^{(k)}_i; w, b))$ denotes the hinge loss of $x^{(k)}_i$ under the linear SVM classifier $g(x^{(k)}_i; w, b)$ with the weight vector $w$ and bias parameter $b$. The constraint enforces at least one positive instance would be emerged in each positive bag. $f(v; \lambda, \gamma)$ is the self-paced regularizer to introduce meaningful biases, e.g., the easiness bias and diversity bias, into the learning mechanism to help selecting more confidence training instances and discovering more truthful knowledge of co-saliency. The regularizer may have many different forms.

In such learning-based co-saliency detection framework, most of the knowledge about the co-saliency object regions are inferred by the designed learner automatically rather than heavily relying on some manually designed metrics as in other categories of co-saliency detection methods. Essentially, with applicable learning strategy, the learning-based method has potential to obtain more truthful knowledge than the methods based on the human-designed metrics because these hand-designed metrics are typically too subjective and cannot generalize well to flexibly adapt various scenarios encountered in practice, especially due to the lack of thorough understanding of the biological mechanisms of human visual attention. The disadvantage of the learning-based co-saliency detection method is the execution time of the iterative training process, especially when the converge condition is hard to achieve.

V. APPLICATIONS OF CO-SALIENCY DETECTION

In this section we discuss potential applications of co-saliency detection, which are object co-segmentation, video foreground detection, weakly supervised object localization, image retrieval, multi-camera surveillance, and 3D object reconstruction.
in this application, the co-saliency detection approaches should be further improved to handle the influence of the noisy images which may not contain the foreground object or action.

C. Weakly supervised object localization

As discussed in Sec. II-C, co-saliency detection technique can be applied to the task of WSL because without additional prior knowledge or supervision, salient and common cues are the two major useful information for initializing the WSL frameworks and they are what co-saliency detection models seek to figure out. Even though co-saliency detection methods can be directly applied to initialize the common objects in the positive images with weak labels [85], some informative knowledge from the negative images are beyond exploration, which may lead to the un-optimal solutions.

One way to solve this problem and thus further improve the performance when applying co-saliency detection technique in WSL is to involve the prior information contained in the negative images in the formulation of co-saliency detection. For example, Zhang et al. formulated co-saliency in a Bayesian framework [20] as:

$$\text{Cosal}(x_{m,p}) = \frac{\text{Pr}(y_{m,p} = 1|x_{m,p})}{\text{Pr}(x_{m,p})} \text{Pr}(x_{m,p}|y_{m,p} = 1)$$

(4)

where $x_{m,p}$ denotes the feature representation of the $p$-th superpixel region in the $m$-th image, and $y_{m,p}$ is the label of $x_{m,p}$, indicating whether $x_{m,p}$ belongs to the co-salient region. As can be seen, when applying the above formulation in WSL, it can only consider the intra-image contrast and the intra-group consistency in positive images. To further leverage the useful information contained in the negative images, one can follow the Bayesian rule to extend the formulation as:

$$\text{Cosal}(x_{m,p}) = \frac{1}{\text{Pr}(x_{m,p})} \left\{ \text{Pr}(y_{m,p} = 1|x_{m,p}) \right\}$$

(5)

$$\frac{\text{Pr}(x_{m,p}|y_{m,p} = 1)}{\text{Intra-image contrast}} - \text{Pr}(y_{m,p} = 0) \frac{\text{Pr}(x_{m,p}|y_{m,p} = 0)}{\text{Inter-group separation}}$$

As can be seen, the extended formulation introduces the factor of inter-group separability as well as another two priors, which can be used to localize the objects in the weakly labeled images effectively.

D. Image retrieval

Recently, the object sensitive image pair-distance has been demonstrated to be beneficial for the image retrieval techniques, especially the ones based on the global feature comparison [19], [86], [87]. Thus, as a promising way to generate robust image pair-distance, co-saliency detection technique can also be applied in the image retrieval task.

For example, Fu et al. has provided an efficient and general robust image pair-distance based on the co-saliency maps [13]. Given an image pair $\{I_i\}_{i=1}^2$, each image can be segmented into the co-salient foreground regions $I^f$ and the background regions $I^b$ by thresholding the co-saliency maps. Then, the co-saliency-based image-pair distance $D(.)$ was defined as:

$$D(I_1, I_2) = w_f E\text{dist}(I_1^f, I_2^f) + w_b E\text{dist}(I_1^b, I_2^b),$$

(6)

where $w_f$ and $w_b$ are the foreground and background weights, respectively, with $w_f + w_b = 1$. $E\text{dist}(.)$ denotes the traditional image distance. As can be seen, this equation makes the co-salient object regions play the more important role in the image-pair distance computing.

E. Multi-camera surveillance

Video surveillance in a large and complex environment always requires the use of multiple cameras [88], which leads to amount of research interest in multi-camera surveillance. In multi-camera surveillance, one important task is to automatically detect and track the person with unnormal (distinct) behaviors monitored by multiple camera sources. Recently, Luo et al. posed this problem as the multi-camera saliency estimation (shown in Fig. 8), which extended the conventional saliency detection methods to identify the regions of interest with information from multiple camera sources in an integrated way [89]. As can be seen, the problem of multi-camera saliency estimation is very related to co-saliency detection. Both of these two tasks need to discover the common and salient image regions from multiple related images. Thus the co-saliency detection technique can also be applied in the task of multi-camera saliency and multi-camera surveillance. It also needs to notice that the images used in multi-camera surveillance are usually captured by the cameras positioned in various locations and thus contain huge diversity in terms of the visual perspective, which should be considered additionally when applying the co-saliency detection technique in the multi-camera surveillance.
F. 3D object reconstruction from 2D images

3D object reconstruction is a core problem in computer vision community and has been extensively researched for the last few decades. Even though it has been largely solved in the multi-view rigid case, where calibration and correspondences can be estimated and thus lead it to be a well-understood geometric optimization problem, more recent interests focus on class-based 3D object reconstruction [90], [91], where the goal is to reconstruct the objects belonging to the same category and pictured in each single 2D image of a given image group. Specifically, Vicente et al. proposed to first estimate camera viewpoint of each image using rigid structure-from-motion and then reconstruct object shapes by optimizing over visual hull proposals guided by loose within-class shape similarity assumptions [90]. Kar et al. proposed the deformable 3D models which can further implement 3D object reconstruction in a more unconstrained setting [91]. However, even for these two most recent approaches, both of them need the manually labeled ground-truth masks for the objects appearing in each image, which leads to the tedious human labeling task and limited capability to scale up for reconstructing 3D objects in arbitrary categories.

Fortunately, with the rapid development in recent years, co-saliency detection techniques can be used here to release the tedious human labeling task and scale up the 3D object reconstruction for any given object category. As shown in Fig. 9, for a given object category, we can first search for the relevant images from the internet, where many diverse views of the objects in same category are available. Some processes can be applied to purify the searched images to reduce the noisy data. Then, any off-the-shelf co-saliency detection technique can be applied to detect the co-salient objects appearing in the image collection. Afterwards, some segmentation techniques, such as direct shareholding and graph cut, can be adopted to generate the masks for the objects. Finally, by estimating the camera viewpoints of each image, the 3D object models can be reconstructed from the obtained co-saliency masks of each 2D image.

VI. DISCUSSIONS

A. Evaluation

1) Evaluation Measurement: To evaluate the performance of the proposed method, the existing literatures usually adopted six criteria, which are the precision-recall (PR) curve, average precision (AP) score, F-measure curve, mean F-score (mF), receive operator characteristic (ROC) curve, and AUC score, respectively. All of these criteria are widely used to evaluate the performance of (co-)saliency detection algorithms, especially for the former four, which are more frequently adopted for evaluating the performance of co-saliency detection methods.

PR curve and AP score are generated by separating the pixels in a saliency map into salience or non-salience by varying the quantization threshold within the range [0, 1]. The resulting true positive rate (or the recall rate) versus precision rate at each threshold value forms the PR curve. The area under the PR curve is calculated as the AP curve, and AUC score, respectively. All of these criteria are widely used to evaluate the performance of (co-)saliency detection, especially for the former four, which are more frequently adopted for evaluating the performance of co-saliency detection methods.

\[ PRE = \frac{|SF \cap GF|}{|SF|}, \]
\[ TPR = \frac{|SF \cap GF|}{|GF|}, \]
\[ FPR = \frac{|SF \cap GB|}{|GB|}, \]

where SF, GF and GB denote the set of segmented foreground pixels after a binary segmentation using a certain threshold, the set of ground truth foreground pixels and the set of ground truth background pixels, respectively. Under each threshold, we also calculate the corresponding F-score as:

\[ F_{\beta} = \frac{(1 + \beta^2) \cdot PRE \times TPR}{\beta^2 \times PRE + TPR}, \]

where \( \beta^2 = 0.3 \) as suggested in [92]. The threshold value versus the corresponding F-score forms the F-measure curve and the mean value of the F-scores under all the thresholds is the mean F-score (mF).

2) Evaluation Datasets: Co-saliency detection approaches are mainly evaluated on three public benchmark datasets: the Image Pair dataset [18], the iCoseg dataset [56], and the MSRC dataset [93]. The brief summary of these datasets are shown in Table III, and more concrete properties of these datasets are described and discussed as follows:

The Image Pair dataset [18] (see Fig. 10 (a)) contains 105 image pairs (i.e. 210 images) with resolution around 128\times100. It also has the manually labeled pixel-level ground truth data and is the earliest benchmark dataset built for evaluating the performance of co-saliency detection, in which each image pair contains one or more similar

---

**Fig. 9.** Applying co-saliency detection for 2D image based 3D reconstruction.
objects with different backgrounds. After being proposed by Li et al. [18], this dataset was frequently utilized by the co-saliency detection approaches designed in early years like [12], [13], [16], [18]. As the content of images in the Image Pair dataset is not as cluttered as other benchmark datasets, the state-of-the-art co-saliency detection methods have achieved promising performance (0.973 and 0.933 in terms of AUC and AP, respectively, according to Zhang et al. [22]) on it.

The iCoseg dataset [56] (see Fig. 10 (b)) is a large-scale publicly available dataset that is widely used for co-saliency detection. It consists of 38 image groups of totally 643 images along with pixel ground-truth hand annotations. Notice that each image group in the iCoseg dataset contains 4 to 42 images rather than two images in the Image Pair dataset. Besides, most images in the iCoseg dataset contain complex background and multiple co-salient objects. Thus, the iCoseg dataset is considered as a more challenging dataset for co-saliency detection and is widely used to evaluate the modern co-saliency detection technique. As reported by Zhang et al. [21], the performances of the state-of-the-arts on this dataset are around 0.87 and 0.81 in terms of AP and F-measure, respectively. Thus, there is still a room for further improvement of the co-saliency detection technique on this dataset.

Another dataset which is widely used for evaluating co-saliency detection approaches in recent time is the MSRC dataset [93]. The MSRC dataset consists of 8 image groups (240 images) with manually labeled pixel-wise ground truth data. As shown in Fig. 10 (c), image groups like airplane, car, and bicycle are contained in this dataset. Notice that the image group of ‘grass’ is not usually used to evaluate co-saliency detection performance as it does not have co-salient objects in each image. Compared with the iCoseg dataset, there are different colors and shapes for the co-salient objects in image groups of the MSRC dataset, making it more challenging. Thus, most of the existing co-saliency detection approaches can only obtain performance less than 0.85 and 0.77 in terms of AP and F-measure, respectively, which is still not satisfactory.

The Cosal2015 dataset [72] (see Fig. 10 (d)) is a newly published dataset that is particularly established for the research of co-saliency detection. This dataset contains 50 image groups and totally 2015 images which are collected from challenging scenarios in the ILSVRC2014 detection benchmark [94] and the YouTube video set [95]. As shown in Fig. 10 (d), the images in this dataset are highly cluttered. Moreover, in some cases, there are multiple salient objects in each single image, which, however, do not commonly appear in the image group. This involves lots of ambiguity for the algorithm to determine which image regions are the co-salient ones. Consequently, the Cosal2015 dataset tends to be much bigger and challenger than any other existed datasets. In addition, it also provides the pixel-level annotations manually labeled by 20 subjects. The state-of-the-art performance on this dataset can only reach to 0.74 and 0.71 in terms of AP and F-measure, respectively.
B. Low-level feature vs. High-level feature

As one basic yet critical component, the features adopted to represent the image pixels or regions can also affect the performance of co-saliency detection to a large extent. From the existing literatures, we can observe that the early co-saliency detection methods are mainly based on the low-level features such as the texture descriptors [18], color histogram [24], and Gabor filter [13]. These methods assumed that the common objects appearing in the given image group should have high consistency in terms of these low-level features and can be distinct from the image backgrounds according to these features. However, as recently proposed by Zhang et al [20], high-level features which embeds semantic information can overcome the unstable issue caused by the variations in viewpoints, shapes, and luminance and thus can model the concept-level properties, such as consistency, of the co-salient objects more effectively.

Essentially, each of the low-level feature and high-level feature cannot handle all the cases in co-saliency detection. As shown in Fig. 11 (c), the elephants in the image groups have the similar color and texture with the background regions containing the grass. However, in higher semantic level, the image regions of elephants are related to the semantics like animals whereas the grass regions are related to the semantics like ground. Thus, high-level features could be used to detect co-saliency more efficiently in this case. On the contrary, as shown in Fig. 11 (d), the red football players and the white ones share the similar semantic as human. However, as only the white players appear in every image of the given image group, the desirable co-salient regions should be the white players rather than the red ones. In this case, we need to adopt the low-level color feature to separate the real co-salient regions from the confusing ones.

Consequently, the proper features that should be used in co-saliency detection are hard to select and design because they are closely related to the content of the given image group. As there is no existing literature working on the idea of learning to select proper features according to the specific content of the given image group, we think this could be another promising yet challenging direction in the future.

C. Contrast vs. Consistency

As we know, contrast and consistency are the two most critical factors considered to explore the properties of the co-salient object regions. To reveal the insight about which factor is more important in co-saliency detection, some literatures have proved evidences in their model settings or experiments. For example, Li et al. used equal weights to add these two factors to obtain the final estimation[18]. Similarly, Du et al. integrated these two factors via multiplying them with equal weights [69]. These works indicate that the contrast factor and consistency factor play the same importance during co-saliency detection. Whereas Zhang et al. verified the parameter settings in their experiments and concluded the consistency factor (with the weight of 0.7) is more important than the contrast factor (with the weight of 0.3) [22].

Essentially, the importance of these two factors is on one hand related to the designed co-saliency detection algorithms and, more importantly, is highly related to the content of the image group. For example, as shown in Fig. 11 (a), the consistency factor is more important in discovering the co-salient objects in the image group of ‘Egypt Pyramids’ as the contrast factor would mislead the detector to focus on the regions of human and horse. On the contrary, for the image group of ‘Sox Players’ (see Fig. 11 (b)), the consistency factor tends to be less important than the contrast factor as both the foreground and the background regions would obtain high consistency in such image group, which reduces the capability of the consistency factor to discover the co-salient object regions.

Consequently, one promising solution for assigning proper importance weights to these investigated factors in one algorithm should be in a content-aware manner, which means the designed algorithms should have the capability to automatically infer the proper weights for the factors based on the content of the specific image group. Along this direction, novel co-saliency detection methods are needed to further improve the detection performance.

D. Bottom-up vs. Top-down

Inspired by the conventional saliency detection methods, most of the existing co-saliency detection techniques are designed in a bottom-up manner, where the frameworks are established highly based on the human designed features.
and metrics to explore the visual stimulus in each image without any task-related guidance. Although it has been successfully used in the conventional saliency detection, performing co-saliency detection in such bottom-up manner cannot obtain satisfactory results in many scenarios. The fundamental reason is that the conventional saliency detection only needs to simulate the early stage of the visual attention system when human are receiving the visual stimulus from one single image, while co-saliency detection task needs to additionally formulate the abstract knowledge obtained during the human reviewing of the images in a given group. Thus, compared with the conventional saliency detection, co-saliency detection tends to be more complicated because only exploring the bottom-up cues is not enough.

Essentially, as discussed in the above two subsections, co-saliency detection is highly related to the case-by-case content of the given image groups. In other words, co-saliency detection results would be largely influenced by the global-level information among the entire image group. Thus, one more proper way to formulate co-saliency should be in the top-down manner, where the global-level information could provide critical top-down guidance and priors for exploring the co-salient object regions in a specific image group.

To introduce the top-down priors in co-saliency detection, some of recent techniques have made efforts in two ways. The first way is to fuse multiple bottom-up (co-)saliency detection results under the guidance of the global-level information in a specific image group, i.e., the fusion-based approaches, and the second way is to learn to model co-saliency under the weak supervision provided by the image-level labels among different image groups, i.e., the learning-based methods. By leveraging the top-down priors, the fusion-based methods and learning-based methods can outperform the bottom-up methods in most cases.

To further demonstrate the superiority of incorporating the top-down priors, we compared three state-of-the-art co-saliency detection methods on the iCoseg and MSRC datasets. These methods are Fu’s method [13], Cao’s method [29], and Zhang’s method [21], which can be seen as the most typical methods from the bottom-up method, fusion-based method, and learning-based method, respectively. To make the experiments as fair as possible, we use the same feature representation as adopted in [13] for each method. Fig. 12 shows the quantitative evaluation results in terms of the PR curve, AP score, F-measure curve, mF score, ROC curve, and AUC score, respectively. As can be seen, with better involving the group-specific top-down priors into the co-saliency detection framework, the fusion-based method and learning-based method could generally outperform the bottom-up method. More specifically, as displayed in Fig. 13 we also compared the AP scores among these co-saliency detection methods in each image group of the iCoseg and MSRC datasets. It is easy to observe that the fusion-based method [29] and the learning-based method [21] can obtain better performance on much more image groups than the bottom-up method [13], which demonstrates the top-down priors explored by the former two categories of co-saliency detection technique can evidently improve the co-saliency detection performance in various cases.

E. Challenges

Although the co-saliency detection technique as a community has achieved substantial progress in the past years, there is still a long road to make co-saliency to be widely applied to more practical applications. This is more related to the limitations of existing methods in terms of robustness and efficiency. To be more specifically, we summarize the challenges that need to be addressed as follows.

The first challenge is the complexity of the image. For example, when the non-salient background involves the similar appearance (e.g. color, texture) as the salient areas, as shown in Fig. 14 (A), it is difficult to distinguish the foreground from the complex background only based on human visual fixation. A solution for this case is that employing more high-level features, and exploring the semantic context. The discriminative feature could help to identify the foreground from the complex background.

Another challenge is that if the foreground is composed of multiple components, as shown in Fig. 14 (B), the co-saliency detection method could not provide the entire object-level foreground. The main reason is that the most co-saliency detection methods are based on bottom-up role without heavy learning, which could not provide the high-level constraint. A solution for this case is that employing more high-level features, and exploring the semantic context. A useful solution is to introduce the objectness constraint, which measures the likelihood of a region being an object. The objectness is based on general object property, and widely is employed to generate object bounding box and proposal. The objectness preserves the wholeness of the general object, and could be used to handle the multiple component foreground in saliency detection.

The third challenge is the large scale data. The co-saliency detection aims to discover the common foreground from the multiple images, which is easily considered to apply on the pattern discovery from the large scale data. The most co-saliency detection algorithms works well in small and middle size datasets (within 100 ∼ 1000 images), but they are hardly to deal with large scale data. There are some unsolved issues, such as outlier and noise image, variation of intra-class, and fast optimization. How to deal with the large scale data is a direction for future work of co-saliency detection.

VII. CONCLUSION

In this paper, we reviewed the co-saliency detection technique. We described the recent progress of the co-saliency detection technique, analyzed the major algorithms in this research area, and discussed the potential applications and issues of co-saliency detection. In summary, co-saliency detection is a a newly emerging and rapidly growing research area in computer vision community, which is
Fig. 12. Comparison of the three categories of co-saliency detection technique in terms of Precision-recall curve, AP score, F-measure curve, mF score, RPC curve, and AUC score, respectively.

Fig. 13. Comparison of the three categories of co-saliency detection technique on the image groups of the iCoseg and MSRC datasets, respectively. The fraction behind each category of method reflects the successful cases, i.e., achieving superior performance than other methods, of the corresponding method.

Fig. 14. Some challenging cases for the co-saliency detection.

derived from the conventional saliency detection but with more challenging task to explore the relationship beyond one single image. As the images that can be accessed easily from the internet today are huge in scale and contain rich relationship with each other, we believe the co-saliency detection technique has been endowed with considerable potential to be applied in more real-world tasks. This motivates us to review the co-saliency detection technique and discuss its fundamentals, applications, and challenges in this paper. We hope this review could be beneficial for both the fresh and senior researchers in this field as well as researchers working in other relevant fields to have better understanding about what they can do with co-saliency detection in the future.

REFERENCES

[1] A. Borji and L. Itti, “State-of-the-art in visual attention modeling,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 185–207, 2013.
[2] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient Object Detection: A Benchmark,” IEEE Transactions on Image Processing, vol. 24, no. 12, pp. 5706–5722, 2015.
[3] H. Fu, D. Xu, B. Zhang, S. Lin, and R. K. Ward, “Object-based multiple foreground video co-segmentation via multi-state selection graph,” IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3415–3424, 2015.
[4] H. Fu, D. Xu, S. Lin, and J. Liu, “Object-based RGBD Image Co-segmentation with Mutex Constraint,” in CVPR, 2015, pp. 4428–4436.
[5] W. Wang, J. Shen, and F. Porikli, “Saliency-aware geodesic video object segmentation,” in CVPR, 2015, pp. 3395–3402.
[6] W. Wang, J. Shen, X. Li, and F. Porikli, “Robust video object cosegmentation,” IEEE Transactions on Image Processing, vol. 24, no. 10, pp. 3137–3148, 2015.
[7] J. Tang, A. Joulin, L.-J. Li, and L. Fei-Fei, “Co-localization in real-world images,” in CVPR, 2014, pp. 1464–1471.
[8] A. Joulin, K. Tang, and L. Fei-Fei, “Efficient image and video co-localization with frank-wolfe algorithm,” in ECCV, 2014, pp. 253–268.
[9] M. Cho, S. Kwak, C. Schmid, and J. Ponce, “Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals,” in CVPR, June 2015.
[10] P. Siva and T. Xiang, “Weakly supervised object detector learning with model drift detection,” in ICCV, 2011, pp. 343–350.
[64] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, “Weakly supervised learning for target detection in remote sensing images,” *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 4, pp. 701–705, 2015.

[65] 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 features learning,” *IEEE Transactions on Geoscience and Remote Sensing*, vol. 53, no. 6, pp. 3325–3337, 2015.

[66] H. O. Song, Y. J. Lee, S. Jegelka, and T. Darrell, “Weakly-supervised discovery of visual pattern configurations,” in *NIPS*, 2014, pp. 1637–1645.

[67] D. E. Jacobs, D. B. Goldman, and E. Shechtman, “Cosaliency: Where people look when comparing images,” in *ACM Symposium on User Interface Software and Technology*, 2010, pp. 219–228.

[68] X. Cao, Y. Cheng, Z. Tao, and H. Fu, “Co-saliency detection via base reconstruction,” in *ACM MM*, 2014, pp. 997–1000.

[69] S. Du and S. Chen, “Detecting co-salient objects in large image sets,” *Signal Processing Letters*, vol. 22, no. 2, pp. 145–148, 2015.

[70] X. Cao, Z. Tao, B. Zhang, H. Fu, and X. Li, “Saliency map fusion for video objects discovery,” in *International Conference on Internet Multimedia Computing and Service*, 2015, pp. 1–4.

[71] R. Huang, W. Feng, and J. Sun, “Saliency and co-saliency detection by low-rank multiscale fusion,” in *ICME*, 2015, pp. 1–6.

[72] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, “Detection of co-salient objects by looking deep and wide,” in *International Journal of Computer Vision*, 2016.

[73] D. Zhang, D. Meng, and J. Han, “Co-saliency detection via a self-paced multiple-instance learning framework,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. In press, 2016.

[74] J. Jing, X. He, Q. Han, and X. Niu, “Co-saliency detection linearly combining single-view saliency and foreground correspondence,” *IEICE Transactions on Information and System*, vol. 98, no. 4, pp. 985–998, 2015.

[75] N.-M. Shen, J. Li, P.-Y. Zhou, Y. Huo, and Y. Zhuang, “BSFCoS: Block and Sparse Principal Component Analysis-Based Fast Co-Saliency Detection Method,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 30, no. 1, 2016.

[76] X. Cao, Z. Tao, B. Zhang, H. Fu, and X. Li, “Saliency map fusion based on rank-one constraint,” in *ICME*, 2013, pp. 1–8.

[77] J. Yang, G. Zhao, J. Yuan, X. Shen, Z. Lin, B. Price, and J. Brandt, “Discovering primary objects in videos by saliency fusion and iterative appearance estimation,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. In press, 2016.

[78] W.-T. Li, H.-S. Chang, K.-C. Lien, H.-T. Chang, and Y. Wang, “Exploring visual and motion saliency for automatic video object extraction,” *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2600–2610, 2013.

[79] Y. Luo, J. Yuan, and J. Lu, “Finding spatio-temporal salient paths for video objects discovery,” *Journal of Visual Communication and Image Representation*, vol. 38, pp. 45–54, 2016.

[80] X. Chang, Y. Yang, E. P. Xing, and Y.-L. Yu, “Complex event detection using semantic saliency and nearly-isotonic svm,” in *ICML*, 2015.

[81] V. Mahadevan and N. Vasconcelos, “Spatiotemporal saliency in dynamic scenes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 1, pp. 171–177, 2010.

[82] W. Kim, C. Jung, and C. Kim, “Spatiotemporal saliency detection and its applications in static and dynamic scenes,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 4, pp. 446–456, 2011.

[83] N. Riche, M. Mancas, D. Culibrk, V. Cnobejvic, B. Bosselin, and T. Dutoit, “Dynamic saliency models and human attention: A comparative study on videos,” in *ACCV*, 2013, pp. 586–598.

[84] X. Cao, F. Wang, B. Zhang, H. Fu, and C. Li, “Unsupervised pixel-level video foreground object segmentation via shortest path algorithm,” *Neurocomputing*, vol. 172, pp. 235–243, 2016.

[85] X. Cao, C. Zhang, H. Fu, X. Guo, and Q. Tian, “Saliency-aware nonparametric foreground annotation based on weakly labeled data,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. In press, 2016.

[86] L. Yang, B. Geng, Y. Cai, A. Hanjallic, and X.-S. Hua, “Object retrieval using visual query context,” *IEEE Transactions on Multimedia*, vol. 13, no. 6, pp. 1295–1307, 2011.

[87] A. Papusho and A. G. Bors, “Image retrieval based on query by saliency content,” *Digital Signal Processing*, vol. 36, pp. 156–175, 2015.

[88] V. Kettner and R. Zabih, “Bayesian multi-camera surveillance,” in *CVPR*, 1999.

[89] Y. Luo, M. Jiang, Y. Wong, and Q. Zhao, “Multi-camera saliency,” *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 10, pp. 2057–2070, 2015.

[90] S. Vicente, J. Carreira, L. Agapito, and J. Batista, “Reconstructing pascal VOC,” in *CVPR*, 2014, pp. 41–48.

[91] A. Kar, S. Tulsiani, J. Carreira, and J. Malik, “Category-specific object reconstruction from a single image,” in *CVPR*, 2015.

[92] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in *CVPR*, 2009, pp. 1597–1604.

[93] J. Winn, A. Criminisi, and T. Minka, “Object categorization by learned universal visual dictionary,” in *ICCV*, vol. 2, 2005, pp. 1800–1807.

[94] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein et al., “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[95] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari, “Learning object class detectors from weakly annotated video,” in *CVPR*, 2012, pp. 3282–3289.