Weakly Supervised Co-saliency Detection via Class Excitation Mapping

Anonymous ACCV 2018 submission

Paper ID 658

Abstract. Existing co-saliency detection algorithms mainly focus on extracting low-level features, failing to handle group of images with variations in background scenes and foreground objects. A solution for this case is to employ effective high-level features and to explore the semantic context. We observe that common image-level tags provide consistency information of common foreground salient objects for a group of images. Thus, we propose a co-classification method based on K-means clustering to decompose the complicated high-level co-salient feature extraction into identifying the common object class(es) and discovering the corresponding neural attention maps. By our proposed top-down neural attention model, the high-level consistency can be mapped to class excitation maps (CEMs), which are capable of capturing the common foreground regions. Then we integrate CEMs with bottom-up saliency maps (BUSMs), detecting co-saliency based on both high-level and low-level features combinedly. Finally, we apply a fully connected Conditional Random Field (FC-CRF) model for accurate boundary recovery. Notably, to our knowledge, this paper is the first one introducing weakly supervised learning into co-saliency detection, requiring only image-level tags as supervision to significantly reduce the costs in producing training data. Our weakly supervised co-saliency detection method is also novel in combining high-level and low-level cues. Experiment results show that our method achieves the state-of-the-art results on three benchmark datasets.

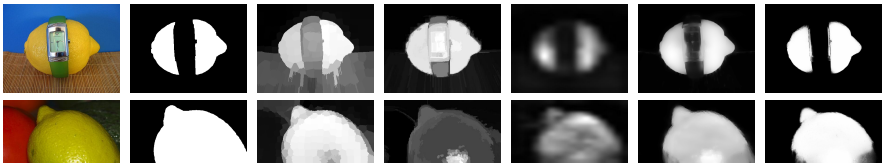
1 Introduction

As the advance in imaging equipment and the availability of online photo sharing services, there is a growing demand for people to manage large volumes of images. Being able to detect the same person or object from photos taken in the same event or under different occasions becomes important, which may facilitate effective image retrieval and indexing, or photo album organization. Co-saliency detection, an emerging research area, is considered as a prominent approach to realize such needs. Unlike saliency detection methods, which only identify distinctive features from an image based on some local contrast assumption, co-saliency detection additionally traces certain consistent occurrence of distinctive features and aims at highlighting the common and salient foreground regions across a set of images.

In order to detect co-salient regions precisely, we need to discover the intersection of the intra-image salient regions and inter-image common object regions. The former is used to guarantee that the detected object regions are salient in individual image, while the latter is to ensure that the detected object regions are commonly appearing in the image group. Many methods [1–3] address the co-saliency detection task mainly rely on low-level cues. Interestingly, despite some good results were obtained based on conventional datasets, their performance deteriorates significantly under background variation or existence of different objects across a group of images. The main reason is that low-level cues comprise only local distinctive features, which may not involve enough inter-image consistency. In Fig. 1, when the lemon is presented against several colorful fruits or with other foreground, low-level features are no longer robust attributes to discover co-saliency regions. Consequently, an available solution is to effectively capture the high-level representation of class "lemon" and locating semantically co-salient regions.

Our approach is to adopt a convolutional neural network (CNN) to support automatic high-level feature extraction. CNN has demonstrated superior performance in feature extraction for many computer vision areas [4–6], especially the ability of capturing semantic meaning. In our co-saliency detection, a reasonable assumption is that the class of a group of images which contain common foreground objects can serve as a form of weak supervision. We leverage the image-level class tags to provide inter-image consistent cues which can be mapped to semantic attention maps for detecting commonly appearing foreground. Such a learning approach has recently been making good contributions to solve the saliency detection problem. For instance, [7] proposed a two-stage weakly supervised learning saliency detection method based on image-level tags. Weakly supervised learning also achieves comparable performance with fully supervised learning in other tasks, such as segmentation [8–10] and localization [11, 12]. Compared with costly and scarce pixel-wise annotation supervision [13], the advantage of weak supervision is that image-level tags are much easier to collect.

Our approach of weak supervision is to accept images that are annotated with image-level tags indicating certain classification information without any fine-grained labeling about saliency of pixel-wise image contents. To identify the co-salient objects from a group of images, we perform K-means clustering based on the classification probabilities of each image, which can be considered as a co-classification task to identify similar objects from a set of images. Through the CNN process, we obtain the derivative of each channel of a CNN layer by back-propagation, representing the pixel importance to an image classification and using the derivative summation of each channel as the weight of its feature map. Subsequently, we fuse feature maps from three CNN layers to obtain a class excitation map (CEM), forming a top-down cue. Up to now, the inferred co-salient priors have been mapped to top-down attention maps which capture semantically co-salient regions. To illustrate how our method detects co-saliency, Fig. 1 shows intermediate and final results from our method.



(a) (b) (c) (d) (e) (f) (g)

Fig. 1. Sample results from our method, where brighter pixels indicate higher co-saliency values. (a) Two original images from the image group "lemon" in the Cosal2015 dataset [14]. (b) Ground truth. (c) Results of [14]. (d) Bottom-up saliency maps via [15]. (e) CEMs, indicating top-down cues. (f) Fusion results from both bottom-up and top-down cues. (g) Final co-saliency maps after FC-CRF refinement.

Our top-down neural attention technique can typically localize discriminative image regions and provide class-specific information. Due to the size of feature map in CNN, CEMs are vague (See Fig. 1(e)) after upsampling. On the contrary, bottom-up saliency maps lack semantic information, but can extract clear low-level features, e.g. intensity, color and texture (See Fig. 1(d)). We exploit the advantages from both of them, adopting a fusion strategy to integrate CEMs with bottom-up features followed by a FC-CRF refinement (See Fig. 1(f) and (g)). Experiments show that CEM possesses impressive localization ability and the combined approach we proposed achieves remarkable co-saliency detection performance on the iCoseg dataset [16], the MSRC dataset [17] and the Cosal2015 dataset [14]. Our contributions are as follows:

- We propose an effective weakly supervised co-saliency detection framework based on both low-level and high-level cues.
- We decompose the intractable top-down feature extraction task into identifying the class(es) of co-salient objects and mining the neural attention maps based on the inferred priors by proposing a simple co-classification method.
- We propose a novel class excitation mapping (CEM) method to generate top-down neural attention maps.

2 Related Work

In this section, we depict existing co-saliency detection methods and explain weakly supervised works about extracting top-down cues by neural attention.

2.1 Co-saliency Detection

Existing co-saliency detection methods can mainly be classified into three types: bottom-up methods, fusion-based methods, and learning-based methods [18]. Most bottom-up methods focus on designing hand-crafted features to assign co-saliency values to each pixel/region in an image [19]. For instance, Li *et al.* [20] utilized color and texture to extract salient regions features within an image and

detected co-saliency by comparing feature similarity hierarchically between images. Alternatively, Fu *et al.* [2] considered contrast, spatial and correspondence as features to detect co-saliency through both intra- and inter-image clustering. During the past few years, bottom-up methods have made a great progress and become a most common way for co-saliency detection. However, these methods heavily depend on manually designed bottom-up features which inevitably introduce subjective factors, failing to generalize to complicated scenarios, such as background variations across images.

Different from bottom-up methods, fusion-based methods incorporate several saliency detection algorithms to extract features from input images, fusing them or the minded mined knowledge to obtain the final co-saliency maps. For example, Cao *et al.* [21] utilized various saliency algorithms to generate multiple saliency maps, using a matrix to combine all salient regions. The rank of the matrix is low if those regions are similar, indicating they are co-salient regions. Cao *et al.* [22] applied multiple saliency detection methods to generate saliency maps. Based on these maps, they voted superpixels (group of pixels) of input images, determining how likely a superpixel should be extract as a candidate. Co-saliency is then computed based on the occurrence rate clusters of superpixels. Fusion-based methods exploit co-detected salient features from multiple methods to improve accuracy. On the other hand, the generated results may not be precise enough as a fusion method may inherit various deficiency from all incorporated saliency detection methods.

Newly developed learning-based methods aim at learning the latent patterns of co-salient objects from a group of images, which is a promising trend of co-saliency detection. A typical method was developed by Zhang *et al.* [23] to integrates both multiple instance learning (MIL) and self-paced learning (SPL). MIL learns classifiers to detect images as positive/negative depending on the presence of co-salient objects, which are formed by superpixel regions (instances). SPL assists this process by gradually learning from easy/faithful training samples to complex/confusable ones. [13] proposed an end-to-end group-wise deep co-saliency detection method based on fully convolutional network. Comparing with the aforementioned methods, this method can potentially better capture patterns about co-salient objects, as the learned features are more objective. However, the training process is usually time-consuming and hard to be converged.

2.2 Top-down Neural Attention of CNNs

Top-down neural attention is to model human visual attention considering top-down factors like goals, expectations and knowledge, which is a significant mechanism for many visual tasks. There are various methods proposed to exploit CNN classifier's prediction. Simonyan *et al.* [24] applies a CNN to visualize relevant image regions for a certain class by activated hidden neurons, which makes use of a back-propagation process. The gradient found by back-propagation indicates the importance of the activation at a spatial point leading to the classification

of an image to class c. Cao *et al.* [25] introduce a feedback loop into a CNN architecture to infer the activation status of hidden layer neurons that can capture task relevant regions successfully. Zhou *et al.* [11] develop a significant method by revisiting the global average pooling layer in CNNs which has a remarkable localization ability despite only being trained on image-level labels. Besides, a probabilistic Winner-Take-All process is used in a new back-propagation scheme to model the top-down neural attention [26].

3 Methods

We introduce high-level information for determining whether the given image contains a instance of the interested target and locating the area of the object. Fig. 2 illustrates the workflow of our weakly supervised co-saliency detection method.

Take the Cosal2015 dataset for example, in order to achieve co-saliency detection for a group of related images, we (1) train a CNN with a training set we established which contains the same image classes as that of the Cosal2015 dataset. We then (2) identify co-salient class label(s) by co-classifying the input images based on K-means clustering. After that, we (3) sum the derivatives as weights by back-propagating the predicted class labels and compute a weighted sum of feature maps to generate layer excitation map (LEM). The LEMs of three relatively deep convolutional layers in the CNN are being aggregated to obtain the final CEM. We also (4) employ the saliency detection method from [15] to obtain a bottom-up saliency map (BUSM), followed by fusing both CEM and BUSM. Results are refined with a fully connected CRF model.

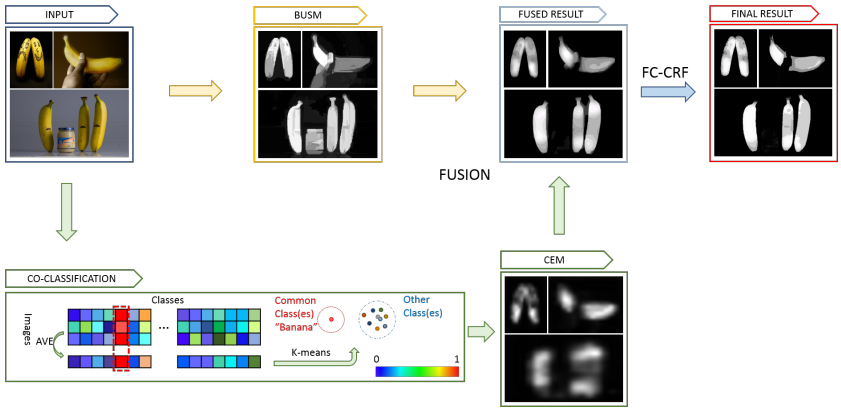


Fig. 2. Workflow of our co-saliency detection method.

It is worth mentioning that our method can also work if the input is only one image rather than an image group, which means that our method is able to

using the same framework for saliency detection and co-saliency detection. The following subsections are dedicated to a detailed description of the proposed approach.

3.1 Co-classification

It is not accurate and reliable enough to determine the image class(es) only by individual image, particularly when the image background is complex or there are several different foreground objects. We exploit the classification result of related images to address the limitation of performing classification based on only a single image. This means co-classification can be regarded as the detection of inter-image consistency to take advantage of the similarity among related images. The ultimate aim of co-classification is to obtain co-salient object class(es) through the results of classification of each image. We train a CNN for classification with image-level tags based on ResNet-50 [6]. Because it performs efficiently and produces a little better classification results than conventional deep CNNs, such as VGG-16 [27] (more details in Section 4.2). Despite ResNet-50 facilitates a high classification accuracy, a small number of images still cannot be classified correctly. We therefore propose a simple co-classification method based on K-means clustering ($k = 2$ clusters) to identify the common class(es) of an input image group. The predicted probabilities of each image in this group are averaged according to the categories before clustering the means. The smaller cluster C represents the class(es) of common foreground objects and the other represents the uncommon foreground and background, which is motivated by the observation that the co-salient class number n_c is much smaller than the other ($n_c = 1$ in the Cosal2015 dataset which contains 50 different classes).

3.2 Class Excitation Map

Top-down mechanisms drive attention to the regions that relevant to the target objects which can be learned by CNNs. We start from the intuition that the salient features leading to a class are those that best distinguish it from other visual classes of recognition interest. In this section, we depict the class excitation maps (CEMs), showing how they are formulated to become top-down cues for our method. Fig. 3 shows the architecture of CEM formulation. Given an image group $I = \{I_i\}_{i=1}^{N}$, where N is the number of images in this group. Our CEM is to discover the top-down co-saliency maps which are capable of capturing and locating semantic object regions $R = \{R_i\}_{i=1}^{N}$. We take advantage of neural attention to detect R_i based on a set of parameters Θ :

$$R_i = f(I_i; \Theta, C) \tag{1}$$

where f is a regression function that represents our class excitation mapping, $C = \{C_k\}_{k=1}^{K}$ is K co-salient class(es) obtained by co-classification.

For an input image I, a certain class C_k, and a class score function $S_{C_k}(I)$ of the CNN, we can transform the classification process into a linear function

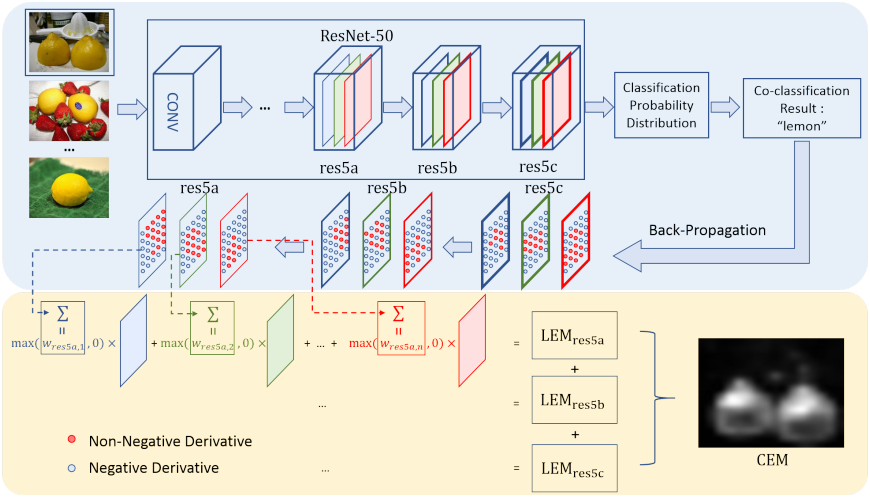


Fig. 3. Architecture overview of Class Excitation Mapping. For each layer, we obtain the derivatives of each channel by back-propagating the class label, where the derivatives are accumulated to be the weight. The class excitation maps (CEMs) are generated by a weighted sum of the feature maps in three layers.

approximately through the first-order Taylor expansion [24]:

$$S_{C_k}(I) \approx v^T I + b \tag{2}$$

where v is the derivative of S_{C_k} with respect to the image I.

Inspired by [24], we treat the derivative of a point in a channel of a layer as the ability of this point influencing the classification result, while the summation of the derivatives with normalization of a channel indicates the importance of this feature map for a class C_k. Consequently, we use the feature maps of a channel obtained by a single forward-pass to generate the layer excitation maps (LEMs), by a weighted summation.

In our paper, considering the balance of various characteristics in different LEMs (See Fig. 4), we integrate three LEMs of relatively deeper layers (res5a, rea5b, res5c in ResNet-50 and conv3_3, conv4_3, conv5_3 in VGG-16). The derivative of class score S_{C_k} with respect to the i-th layer j-th channel at spatial grid $P(i,j,x,y)$ is the class score derivative $u^{C_k}_{i,j,x,y}$ at this point of a feature map $F_{i,j}$ in layer L_i:

$$u^{C_k}_{i,j,x,y} = \frac{\partial S_{C_k}}{\partial P}\bigg|_{P_{i,j,x,y}} \tag{3}$$

To proceed, we obtain the predicted classes(C) by co-classification, and set the target class(C_k) unit to 1 every time, while setting 0 for all the other units. $u^{C_k}_{i,j,x,y}$ is captured by back-propagation, and we sum the derivatives of each channel in the three aforementioned layers. After abandoning the negative

summations, we normalize them to be the weights of each feature map:

$$w_{i,j}^{C_k} = \frac{\max(\sum_{x,y} u_{i,j,x,y}^{C_k}, 0)}{\sum_j \max(\sum_{x,y} u_{i,j,x,y}^{C_k}, 0)} \tag{4}$$

Thus, $w_{i,j}^{C_k}$ represents the importance of each feature map resulting in the classification of an image to class C_k. Intuitively, a weighted point-wise summation of feature map $F_{i,j}^{C_k}$ indicates a discriminative region of the class C_k with respect to layer L_i.

$$F_i^{C_k} = \sum_j w_{i,j}^{C_k} F_{i,j}^{C_k} \tag{5}$$

where $F_i^{C_k}$ indicates the LEM of the i-th layer of class C_k. The fusion is formulated as a linear summation of M layers with normalization to generate the final CEM CS_{CEM}:

$$CS_{CEM}^{C_k} = \frac{1}{M} \sum_i F_i^{C_k} \tag{6}$$

$$CS_{CEM}(x, y) = \max_k p_k CS_{CEM}^{C_k}(x, y) \tag{7}$$

where $p_k = \frac{P^{C_k}}{\sum_k P^{C_k}}$, and P^{C_k} represents the averaged predicted probability of class C_k of the image group. In the experiments, we fuse maps of $M = 3$ layers. After normalized summation, we obtain the final CEM, which can be regarded as a top-down class-specific saliency map. The architecture for CEM construction is shown in Fig. 3. Examples of generated CEMs can be found in Fig. 4, and the comparison with CAM [11] is shown in Section 4.2.

Fig. 4. CEMs obtained from our ResNet-50, fine-tuning with the training set we established: (a) Input image, (b) res5c, (c) res5b, (d) res5a, and (e) average of (b)-(d).

3.3 Fusing High-level and Low-level Cues

Most existing co-saliency detection approaches heavily rely on hand-crafted metrics, leading to poor generalization ability when adapting to different scenarios in terms of image content. CEMs obtained by our weakly supervised CNN can well produce semantic and localization information, but the region boundaries generated are not clear and explicit enough. In contrast, BUSM possesses rich low-level features, but lacking high-level information, leading to poor ability in

distinguishing object category. The top-down attention mechanism of CEM can effectively pruning away bottom-up co-salient regions which are irrelevant to the target objects and improve the performance of co-saliency detection. We employ a fusion method to combine both high-level and low-level cues, which is a trade-off between localization accuracy and low-level cues. In our method, BUSMs are obtained by DRFI [15]. The higher co-saliency value in the CEM indicates the higher co-saliency probability, which can be used to guide the fusion process. Besides, improving the degree of precision is comparatively more important than recall rate in (co-)saliency detection [28]. For these reasons, our fusion strategy is formulated as a point-wise multiplication of CEM and BUSM to improve the precision in detection, which is given by:

$$CS = CS_{CEM} \bigodot BUSM \tag{8}$$

where CS is the fused co-saliency map, and $\bigodot$ indicates the point-wise multiplication. We get our co-saliency map after an edge-aware smoothing [29], which is prepared for FC-CRF refinement.

3.4 Refining via Fully Connected CRF

Since the output of the fusion process is rough, we apply a fully connected CRF model to enhance both spatial coherence and refine boundary accuracy. As a classical probabilistic graphical model, the fully connected CRF [30] regards every pixel in the image as a node, and each node connects to each others. The energy function is showed as follows:

$$E(\mathbf{x}) = \sum_i \theta_i(x_i) + \sum_{i,j} \theta_{i,j}(x_i, x_j) \tag{9}$$

where $\mathbf{x}$ is the label assignment on pixels. We use as unary potential $\theta_i(c_i) = -\log(CS_{x_i})$, where CS_{x_i} is the co-saliency value at pixel i which is calculated from the co-saliency map.

The pairwise term of Eq. (9) is:

$$\begin{aligned}\theta_{i,j}(x_i, x_j) =& \mu(x_i, x_j)\Big[w_1 \exp(-\frac{\left\|p_i - p_j\right\|^2}{2\sigma_\alpha^2} - \frac{\left\|I_i - I_j\right\|^2}{2\sigma_\beta^2}) \\ &+ w_2 \exp(-\frac{\left\|p_i - p_j\right\|^2}{2\sigma_\gamma^2})\Big]\end{aligned} \tag{10}$$

where $\mu(x_i, x_j) = 1$ if $x_i \neq x_j$, and zero otherwise, which means the model penalizes nodes with different labels. The following two Gaussian kernels extract different features from pixel i and j. The first bilateral kernel depends on both pixel positions (denoted as p) and RGB color (denoted as I), and the second kernel only depends on pixel positions. The hyper parameters σ_α, σ_β and σ_γ control the scale of Gaussian kernels.

4 Experiments

In this section, we describe the setup of our experiments in Section 4.1 and present our qualitative and quantitative comparisons with other state-of-the-art methods in Section 4.2.

4.1 Experimental Setup

Dataset In the experiments, we evaluate our proposed method based on three public benchmark datasets: the iCoseg dataset [16], the MSRC dataset [17] and the Cosal2015 dataset [14]. The iCoseg is a dataset widely used in co-saliency detection which consists 38 image groups of totally 643 images along with pixel-wise groud truth data. The MSRC dataset contains 8 image groups with a total of 240 images, which are manually labeled at the pixel-wise level forming the ground truth. Similar to Zhang *et al.* [14], since the image group "grass" has no co-salient objects, we did not use it in our experiment. Cosal2015, constructed by Zhang *et al.* [14], is a newly published dataset containing 2015 images categorized into 50 image groups and provides pixel-wise annotations for evaluation. Compared with the other two datasets, the complexity of image background and the diversity of co-salient objects make Cosal2015 a more challenging benchmark dataset. Due to the lack of training data, we establish a dataset which contains the same classes as Cosal2015 with totally 128086 images by manually extracting from the ImageNet dataset and the Internet with image-level tags. The collections for the iCoseg dataset and MSRC dataset are also organized in the same way. We then split each class of images into training set and validation set randomly according to the proportion of 4:1. It is worth mentioning that in the iCoseg dataset, the common instances of football players in white and in red are in different image groups, so we regrad them as two classes and collect the corresponding images in our training set.

Evaluation Metrics To quantitatively evaluate the performance of our proposed method against the state-of-the-art co-saliency methods, we adopt three widely used evaluation metrics: the Precision Recall (PR) curve, the F-measure, and the Average Precision (AP). In order to obtain these criteria, we first binarize co-saliency maps into binary maps via a series threshold integers from 0 to 255. For each threshold, a group of true positive rates (TPR) and positive predictive values (PPV) can be calculated and they form the PR curve. Next, the AP score can be obtained by calculating the area under the PR curve. Following the approach in [31], we binarize each co-saliency map via a self-adaptive threshold $\mu + \varepsilon$, where μ is the mean value and ε is the standard deviation. Finally, we compute the average precision and recall values for all images, and the F-measure can be obtained by:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{11}$$

where $\beta^2 = 0.3$ as suggested in [28].

Parameter Settings We use ResNet-50 [6] or VGG-16 [27] training over the ImageNet dataset [4] as basic CNN, and replace the last fully connected layer with a convolutional layer, fine-tuning it with our training set. This is implemented by Caffe [32]. For ResNet-50 fine-tuning details, learning rate is 0.001 and momentum is 0.9. In addition, for VGG-16, the learning rate is 0.0001 and momentum is 0.9. The training process totally costs about 8 hours running 100,000 iterations on a PC with an Intel Xeon CPU and a TITAN X GPU. As for hyper-parameters of the fully-connected CRF, we use the following parameters: $w1 = 3$, $w2 = 5$, $\sigma_\alpha = 60$, $\sigma_\beta = 5$, $\sigma_\gamma = 3$.

4.2 Performance Comparison with State-of-art methods

In this Section, our proposed method is compared with 6 state-of-the-art methods, including DRFI [15], CSHS [1], CBCS [2], DW [14], SP-MIL [23] and SACS [3], where the first one is a remarkable single saliency method and the other five are state-of-the-art co-saliency methods. For a fair comparison, we use the source codes or results provided by those authors.

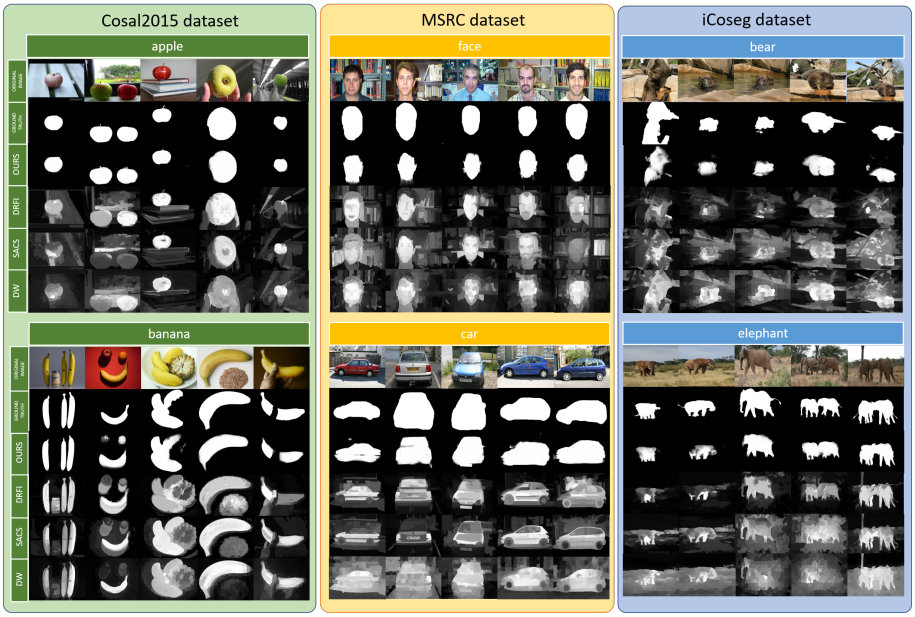


Fig. 5. Visual comparison of co-saliency maps on three benchmark datasets. The left block (in green), the middle block (in orange) and the right block (in blue) are from the Cosal2015 dataset, the MSRC dataset and the iCoseg dataset, respectively.

Subjective Comparison For subjective comparison, we apply our method on three benchmark Cosal2015, MSRC and iCoseg to generate co-saliency maps. Fig. 5 shows some experiment results based on ResNet-50. The results contain 6 image groups, including "apple", "banana", "face","car", "bear" and "elephant". The left, middle and right blocks are from the Cosal2015 dataset, the MSRC dataset, and the iCoseg dataset, respectively.

It is not sufficient to use only low-level features when objects and scenes vary significantly within a particular class of image group. Most existing methods cannot properly handle these because they do not incorporate high-level semantic information to process images. The examples in groups "face" and "car" indicate that our proposed method can efficiently highlight co-salient regions even when co-salient objects are of entirely different shapes, sizes and colors. The examples in groups "bear" and "elephant" depict that under complex scenarios, our CEM only concentrates on common object regions and it is robust enough to suppress complex background, in this case, existing methods may fail owning to the similar colors of the environment and the common objects. In the examples of groups "apple" and "banana", we can see that our method uses class-specific information to wipe irrelevant object regions and retain co-salient regions. As shown in the first column of the image group of "banana", the original image contains two classes of objects, including a bottle and three bananas, our method can effectively wipe off the region of the bottle taking into account of top-down class-specific cues. Comparing with our method, most other (co-)saliency detection methods detect foreground object aimlessly.

There is only one common object class in the three datasets. Considering that there are often multiple classes of objects in a image group in real life, we collect a image set that contains "apple", "banana", "bowl", etc, but "apple" and "banana" appear in every image. It is obvious that the co-salient regions should be the area of "apple" and "banana". Experimental evaluation shows that our co-classification can cluster the correct classes and our CEM is able to obtain the expected co-salient object regions without the irrelevant foreground and the background, e.g. the area of the "bowl" (See Fig. 6).

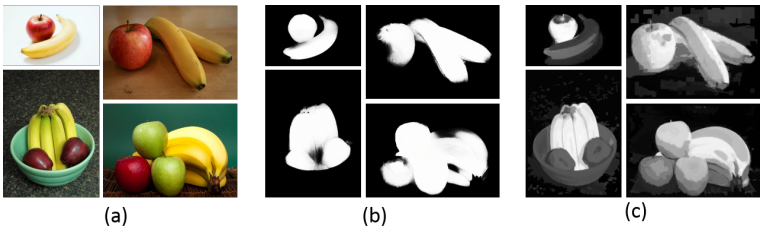
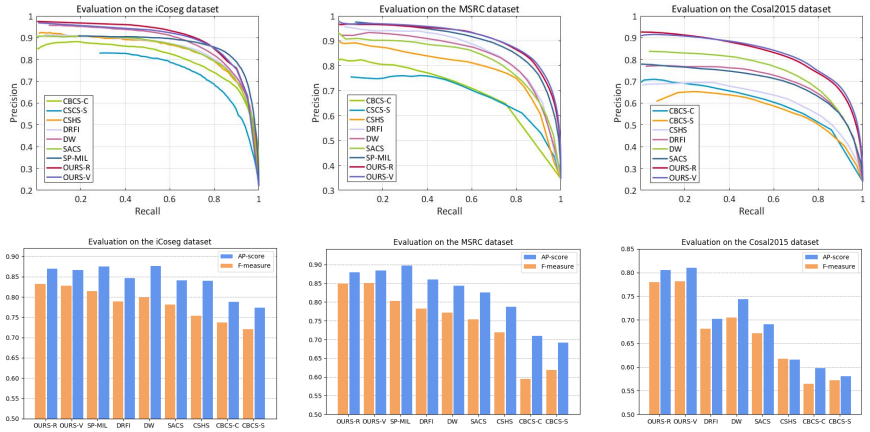


Fig. 6. (a)Original images. (b)Our results (c)Results of SACS



(a) iCoseg (b) MSRC (c) Cosal2015

Fig. 7. Quantitative comparison with state-of-the-art methods on three benchmark datasets. We define our proposed model with ResNet-50 as the "OURS-R" model and define the proposed model with VGG-16 as the "OURS-V" model.

Quantitative Comparison The Quantitative comparison consists of three parts, including PR curve, F-measure and AP-score, which are reported in Fig. 7. In the iCoseg and MSRC dataset, our method can consistently achieve promising performance in terms of F-measure and we also achieve a satisfactory performance on AP-score. Comparing with other methods, our method has a remarkable capability to identify the main object across relevant images and highlight co-salient only the foreground without background. For instance, in the "face" class, our detection result suppresses the regions of bookshelf. In contrast, as shown in Fig. 5, co-saliency maps from other methods contain a large amount of background information.

As for the more challenging dataset, the Cosal2015 dataset, which contains multiple foreground objects and complex scenes, Fig. 7 illustrates that our proposed method outperforms existing methods owing to the high-level semantic features and localization information. It achieves the highest degree of precision over the other methods along all different recall values. This implies our method has more capability to handle the balance between precision and recall. The best performance on the PR-curve also facilitates our proposed method to achieve a new state-of-the-art performance in terms of F-measure and AP score. The F-measure and AP score of our results in terms of each class are shown in our supplemental material.

Ablation Study Our method takes advantage of the summation of non-negative derivatives as the weight and fuses multiple LEMs to generate CEMs, hence it can improve the performance of top-down neural attention. We can note that our method with CEM surpasses the method with CAM in Fig. 8, which confirms

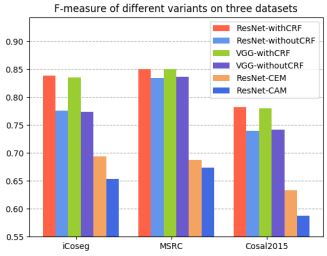


Fig. 8. F-measure of different variants on three benchmark datasets.

the ability of our CEM to cover the common object regions and learn discriminative and accurately localized features compared with CAM. To further evaluate the proposed framework, we compare our method "CNN+CEM+BUSM+CRF" with "CNN+CEM+BUSM", Fig. 7(c) and Fig. 8 show that our method without FC-CRF can achieve the state-of-the-art results on the MSRC and Cosal2015 datasets, which means that the remarkable performance of our method is independent of post-processing. After FC-CRF, our approach can reach a higher F-measure and outperforms other methods on three datasets.

5 Conclusions

In this paper, we propose a new technique called Class Excitation Mapping to model the top-down neural attention and develop a novel weakly supervised co-saliency detection framework by integrating high-level with low-level cues. We co-classification the class(es) for common foreground which can be mapped to attention maps by CEM. Based on our formulation, CEM has the ability to highlight discriminative object regions across a set of relevant images without requiring pixel-level or any bounding box annotations. We estimate CEM and BUSM, fusing both cues and refining the co-saliency map with FC-CRF to generate the final result. In our evaluation on the iCoseg dataset, the MSRC dataset and the Cosal2015 dataset, we demonstrate the effectiveness of our method qualitatively and quantitatively.

For future work, we intend to explore ways further improving the contrastive attention for making top-down attention maps more discriminative, which is key to our co-saliency detection framework.

References

1. Liu, Z., Zou, W., Li, L., Shen, L., Le Meur, O.: Co-saliency detection based on hierarchical segmentation. IEEE Signal Processing Letters **21** (2014) 88–92
2. Fu, H., Cao, X., Tu, Z.: Cluster-based co-saliency detection. IEEE Transactions on Image Processing **22** (2013) 3766–3778
3. Cao, X., Tao, Z., Zhang, B., Fu, H., Feng, W.: Self-adaptively weighted co-saliency detection via rank constraint. IEEE Transactions on Image Processing **23** (2014) 4175–4186
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
5. Girshick, R., Donahue, J., Darrell, T., Malik, J.: Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 580–587
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 770–778
7. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 136–145
8. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
9. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B.: Simple does it: Weakly supervised instance and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017)
10. Shimoda, W., Yanai, K.: Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: European Conference on Computer Vision, Springer (2016) 218–234
11. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2016) 2921–2929
12. Bency, A.J., Kwon, H., Lee, H., Karthikeyan, S., Manjunath, B.: Weakly supervised localization using deep feature maps. In: European Conference on Computer Vision, Springer (2016) 714–731
13. Wei, L., Zhao, S., El Farouk Bourahla, O., Li, X., Wu, F.: Group-wise deep co-saliency detection. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence, AAAI Press (2017) 3041–3047
14. Zhang, D., Han, J., Li, C., Wang, J., Li, X.: Detection of co-salient objects by looking deep and wide. International Journal of Computer Vision **120** (2016) 215–232
15. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2013) 2083–2090
16. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: icoseg: Interactive co-segmentation with intelligent scribble guidance. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 3169–3176

17. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. Volume 2., IEEE (2005) 1800–1807
18. Zhang, D., Fu, H., Han, J., Borji, A., Li, X.: A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. ACM Transactions on Intelligent Systems and Technology (TIST) **9** (2018) 38
19. Chang, K.Y., Liu, T.L., Lai, S.H.: From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2011) 2129–2136
20. Li, H., Ngan, K.N.: A co-saliency model of image pairs. IEEE Transactions on Image Processing **20** (2011) 3365–3375
21. Cao, X., Tao, Z., Zhang, B., Fu, H., Li, X.: Saliency map fusion based on rank-one constraint. In: Multimedia and Expo (ICME), 2013 IEEE International Conference on, IEEE (2013) 1–6
22. Cao, X., Cheng, Y., Tao, Z., Fu, H.: Co-saliency detection via base reconstruction. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 997–1000
23. Zhang, D., Meng, D., Han, J.: Co-saliency detection via a self-paced multiple-instance learning framework. IEEE transactions on pattern analysis and machine intelligence **39** (2017) 865–878
24. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: Proceedings of the International Conference on Learning Representations (ICLR). (2014)
25. Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al.: Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision. (2015) 2956–2964
26. Zhang, J., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. In: European Conference on Computer Vision, Springer (2016) 543–559
27. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014)
28. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, IEEE (2009) 1597–1604
29. Gastal, E.S., Oliveira, M.M.: Domain transform for edge-aware image and video processing. In: ACM Transactions on Graphics (ToG). Volume 30., ACM (2011) 69
30. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in neural information processing systems. (2011) 109–117
31. Jia, Y., Han, M.: Category-independent object-level saliency detection. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 1761–1768
32. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the 22nd ACM international conference on Multimedia, ACM (2014) 675–678