This CVPR2015 paper is the Open Access version, provided by the Computer Vision Foundation.
The authoritative version of this paper is available in IEEE Xplore.

Co-saliency Detection via Looking Deep and Wide

Dingwen Zhang[1], Junwei Han[1*], Chao Li[1] and Jingdong Wang[2]
[1]Northwestern Polytechnical University, P.R. China
[2]Microsoft Research, P.R. China
{zhangdingwen2006yyy,junweihan2010,lllcho1314}@gmail.com, jingdw@microsoft.com

Abstract

With the goal of effectively identifying common and salient objects in a group of relevant images, co-saliency detection has become essential for many applications such as video foreground extraction, surveillance, image retrieval, and image annotation. In this paper, we propose a unified co-saliency detection framework by introducing two novel insights: 1) looking deep to transfer higher-level representations by using the convolutional neural network with additional adaptive layers could better reflect the properties of the co-salient objects, especially their consistency among the image group; 2) looking wide to take advantage of the visually similar neighbors beyond a certain image group could effectively suppress the influence of the common background regions when formulating the intra-group consistency. In the proposed framework, the wide and deep information are explored for the object proposal windows extracted in each image, and the co-saliency scores are calculated by integrating the intra-image contrast and intra-group consistency via a principled Bayesian formulation. Finally the window-level co-saliency scores are converted to the superpixel-level co-saliency maps through a foreground region agreement strategy. Comprehensive experiments on two benchmark datasets have demonstrated the consistent performance gain of the proposed approach.

1. Introduction

With the growing popularity of photo-sharing website like Flickr and Facebook, it has been found that people love taking photographs and there is a rich collection of related pictures sharing the common foreground regions of same object or event [1]. Humans have an extraordinary ability to rapidly scan such a collection of related images or fames and fixate their attention on the most valuable information (i.e. the co-salient regions). According to [2], the humans' fixation ability can be viewed as visual co-attention which is usually driven by bottom-up visual co-saliency. Thus, with the goal of effectively identifying common and salient objects among images within a same group, co-saliency detection has emerged to be an interesting research topic in recent years. Compared with traditional saliency detection [3-9], co-saliency detection additionally explore the mutual information among multiple images/frames. Thus it can provide more useful information and generate more precise prediction for real-world applications, such as segmenting out common objects [10, 11], capturing informative image structures for image matching [12, 13], or discovering human actions and poses in video sequences [14, 15].



Figure 1: Examples illustrating the motivation of looking deep and wide for co-saliency detection. The images in the same row belong to the same group, while the images in different rows are cross different groups. 1) The co-salient objects in an image group share different low-level features, such as colors, textures, and some local descriptors, are more consistency in higher-level concepts; 2) The images cross different groups share similar backgrounds, which is useful to suppress the common background in one image group.

In order to detect co-saliency precisely, one needs to solve two key problems: i) extract effective features to represent the image and, ii) explore useful information to formulate the properties of the co-salient regions over the extracted features. Inspired by the traditional saliency detection algorithms, a variety of hand-crafted features, such as the texture descriptors [16], color histogram [2, 17], and Gabor filter [18], are applied in the existing co-saliency detection approaches. As illustrated in [2, 16], these low-level features can only work with the assumption that the co-salient regions or objects should exhibit high similarity with respect to these features. However, this assumption is neither reasonable in the co-saliency datasets nor in the real-world images, because this kind of low-level

* Corresponding author.

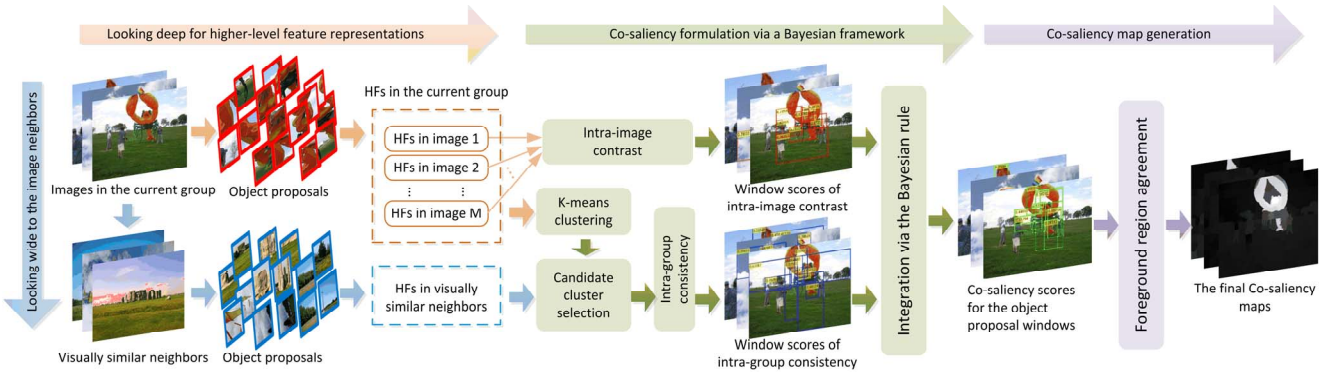


Figure 2: The paradigm of the proposed co-saliency detection approach, where HF denotes the higher-level feature.

visual stimulus appear to be quite unstable due to the variations in viewpoints, shapes, and luminance (as shown in Figure 1). Thus they could not reflect the properties of the co-salient objects in many cases. To solve this problem, we propose to look deep to adopt the high-capacity deep learning technology (i.e. convolutional neural network (CNN) with additional RBM transfer layer) to learn and transfer the higher-level representations which have been demonstrated in previous works [19, 20] to have the capability to capture the useful abstract concepts for image classification and segmentation. Thus this deep information can help us to effectively model the concept-level properties, such as consistency, of the co-salient objects.

In order to formulate the properties of the co-salient regions over the extracted features, the intra-group consistency is explored by the existing approaches [2, 17, 18] to discover the appearance similarity of the image regions. The purpose of these methods is to find image regions frequently occurring in the image group and put them higher probabilities to be co-salient. However, in co-saliency detection datasets as well as the real-world data, common objects usually occur in common backgrounds. That means the image regions frequently occurring in the image group are not only the co-salient objects, but also the common backgrounds (as shown in Figure 1). In this case, directly using the information within the image group would involve the background property into the co-saliency model, thus leading to unsatisfactory co-saliency detection results. To solve this problem, we propose to look wide to take advantage of the cross-group information between the images in one image group and their visually similar neighbors (SNs) in the other image groups. As shown in Figure 1, these SNs would contain similar background but different objects with the images in current image group. Thus, the wide information can be used to decide the candidate co-salient cluster in the proposed framework and finally suppress the common background regions in the generated co-saliency maps.

To explore these useful information (i.e. the deep information and wide information) for co-saliency detection, we propose a unified framework as shown in Figure 2. Firstly, the wide and deep information are explored for the object proposal windows extracted in each image. Then the co-saliency scores are calculated by integrating the intra-image contrast and intra-group consistency via a principled Bayesian formulation. Finally the window-level co-saliency scores are converted to the superpixel-level co-saliency maps through a foreground region agreement strategy. In summary, the novelties of this paper are threefold:

1) We propose to explore the properties of the co-salient objects by using CNN with additional transfer layers, which brings deep information for discovering the higher-level properties of the co-salient objects.
2) We introduce the idea to make use of the visually similar neighbors from the other image groups, which brings wide information for suppressing the common background regions in co-saliency detection.
3) We propose a simple but effective framework which uniformly embeds the deep and wide information in co-saliency detection through a Bayesian formulation and a foreground region agreement strategy.

2. Previous works

Most previous approaches for co-saliency detection explore the joint information provided by the image pair to detect co-salient object regions [16, 21-23]. Specifically, Chen [21] presented a progressive algorithm to enhance pre-attentive responses based on the distribution of the sparse representations in a pair of images. In [16], Li and Ngan combined conventional saliency detection methods to calculate the single-image saliency and applied a co-multilayer graph model to obtain the multi-image saliency. The final co-saliency was detected by linearly fusing the single-image saliency and multi-image saliency. The above models are feasible only to image pairs. However, the requirement for "pairs" means that it only seeks to detect the co-saliency of two images at a time, not accounting for the discovery of the entire mutual information that may exist when there are more than two

images. This results in a direct limitation for co-saliency detection when extending beyond pairwise relations.

In order to extend the co-saliency detection model to work on image sets containing more than two images. Li *et al*. [2] generated the intra-image saliency map and the inter-image saliency map by using multi-scale segmentation voting and pairwise similarity ranking, respectively. The co-saliency map was obtained by the weighted combination of the intra- and inter-image saliency maps. Different from [2], Fu *et al*. [18] utilized a cluster-based algorithm to capture the global correspondence information among the multiple images. They proposed to measure the cluster-level saliency by using three bottom-up saliency cues calculated based on the raw pixel values. Then, the co-saliency map was obtained by fusing these cues with multiplication. In [17], Liu *et al*. proposed a hierarchical segmentation based co-saliency model. They used color histograms of the fine-level segments to measure regional similarities and measured the object prior of each coarse-level segment based on its connectivity with image borders. Finally, the global similarity was derived based on the regional similarity measures, and then fused with the object prior to generate the co-saliency map for each image. Another way to detect co-saliency in multiple related images was proposed by Cao *et al*. [24]. Rather than devoting to discover homogeneous information from the collection of multiple related images for representing co-salient objects, they 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.

3. Looking deep and wide for co-saliency detection

As shown in Figure 2, given a group of images $\{I_m\}_{m=1}^{M}$, we first search the SNs by using Gist [25] and Color Histogram (with 256 bins in each color channel) as the global feature to represent each image. In order to capture the information of the common background, we use the averaged global feature to represent the current image group, and then follow the approach of [25] to select 20 similar images $\{I_n^-\}_{n=1}^{20}$ from other image groups based on the Euclidean distance. Then the aim of our work is to transfer higher-level features for representing object proposal windows (OPs) in the each image (section 3.1), assigning co-saliency scores to the OPs via a Bayesian formulation (section 3.2), and finally converting the obtained window-level co-saliency scores to the superpixel-level co-saliency maps (section 3.3)

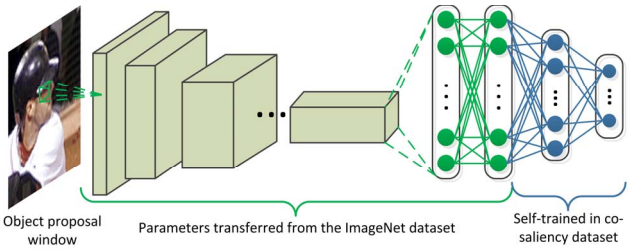


Figure 3: Transfer CNN for higher-level representation.

3.1. Transfer CNN for higher-level representation

In order to represent image regions with deep information, we extract OPs in each image firstly, and then build higher-level features for them via a CNN with additional transfer RBM layer. Specifically, we adopt BING [26] to obtain 256 object proposal windows (as shown in Figure 2) in each image, which will be used as the basic units to calculate the co-saliency scores in the following subsections.

With more than 60 million parameters, CNN [27] has been demonstrated to have the capability to capture higher-level image representations and achieve good performance in object detection and image classification. However, directly learning a whole CNN from the co-saliency dataset is problematic, since the co-saliency dataset only contains hundreds of images, which is too scarce to train such many parameters. In addition, the overall co-saliency framework proposed in this paper is in an unsupervised fashion, which means there is no available image labels to be used for CNN training. To solve this problem, we are inspired by the technologies in transfer learning and domain adaption to pre-train a CNN [27] on the source data (i.e. the images in the ImageNet dataset) firstly. Then, the obtained internal layers are considered as the generic extractor of higher-level features to represent each image region. To transfer the generic higher-level feature to the specific co-saliency dataset, we stack two additional transfer layers formed by two fully connected Restricted Boltzmann Machines (RBM) [28]. The objective of a RBM is to minimize an energy function defined as the joint distribution over the visible units and hidden units, thus maximizing probability for the training data in an unsupervised manner. As the two RBM layers are trained on the target domain, they can adapt the whole network to extract domain-specific higher-level features in the co-saliency dataset.

Figure 3 displays the architecture of the whole network, from which we can see that the final output higher-level representation of each OP is computed by forward propagating the resized 221×221 image windows through six convolutional layers (layer 1-6) and four fully connected layers (layer 7-10). Specifically, we adopt the CNN model proposed in [27] to build the first 8 layers

network. The parameters among these layers are pre-trained on the source data by [27] and kept fixed in the proposed network. For the two RBM transfer layers (i.e. the layer 9 and layer 10 of the proposed network), we set the output dimension of them to be 2048 and 512, respectively. Thus we can obtain a 512 dimensional higher-level feature outputted by the proposed CNN to represent each OP in the co-saliency dataset. The experimental results in Section 4 demonstrate the effectiveness of the transferred higher-level feature.

3.2. Co-saliency formulation

In this section, our goal is to assign scores to each OP by exploring the deep and wide information. The obtained scores are used to estimate the probability of a certain OP to be the co-salient region. We extend the Bayesian framework proposed by [29] to formulate this problem. To be specific, let $\{x_{m,p}\}_{p=1}^{256}$ denote the OPs in image I_m of the current image group, and let the binary random variable $y_{m,p}$ denote whether or not a certain $x_{m,p}$ belongs to a co-salient region. Then, we define the co-saliency of $x_{m,p}$ as:

$$\begin{aligned}\mathrm{Cosal}(x_{m,p}) &= \Pr(y_{m,p}=1 \mid x_{m,p}) \\ &= \frac{\Pr(x_{m,p} \mid y_{m,p}=1)\Pr(y_{m,p}=1)}{\Pr(x_{m,p})} \\ &\propto \underbrace{\frac{1}{\Pr(x_{m,p})}}_{\substack{\text{Intra-image} \\ \text{contrast}}} \underbrace{\Pr(x_{m,p} \mid y_{m,p}=1)}_{\substack{\text{Intra-group} \\ \text{consistency}}} \end{aligned} \quad (1)$$

In the information theory, $-\log \Pr(x_{m,p})$, which is the log form of $1/\Pr(x_{m,p})$, is known as the self-information of the random variable [29]. Self-information increases when the probability of a OP decreases. In other words, OPs that are discriminative from the others in one image are more informative and thus more likely to be the foreground object. Therefore, the first term in the right side of Eq.(1) is associated with the intra-image contrast. Note that BING would extract a number of overlapping OPs in one image (as shown in Figure 2). These overlapping and adjacent windows would cause unfair bias in this formulation because they always have similar appearance. To discourage this influence, we follow [25] and [30] to introduce a spatial distance term into our formulation as:

$$\Pr(x_{m,p}) = \frac{1}{Z_m^{IMC}} \sum_{q=1,q\neq p}^{256} \exp\left(\frac{-D(x_{m,p},x_{m,q})}{\sigma^2}\right) \quad (2)$$

$$D(x_{m,p},x_{m,q}) = \left(\frac{Ed(f_{m,p},f_{m,q})^2}{1+Ed(l_{m,p},l_{m,q})^2}\right) \quad (3)$$

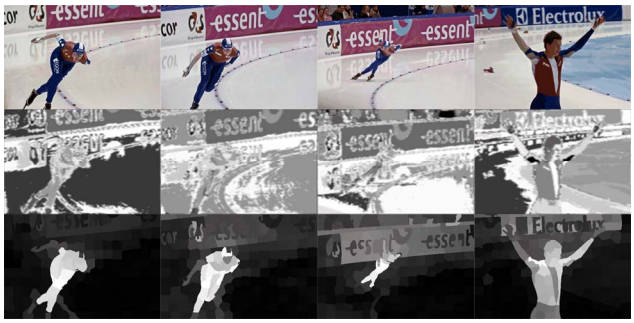


Figure 4: Some examples of the images containing common backgrounds. The co-saliency maps in the second row are generated based on the corresponding cue proposed in [18]. The co-saliency maps in the third row are generated by the proposed intra-group consistency in this paper.

$$Z_m^{IMC} = \sum_{p=1}^{256}\sum_{q=1,q\neq p}^{256} \exp\left(\frac{-D(x_{m,p},x_{m,q})}{\sigma^2}\right) \quad (4)$$

where σ is a constant, $Ed(\cdot)$ indicates the Euclidean distance. $f_{m,p}$ and $l_{m,p}$ denote the extracted higher-level feature and location (in the normalized image coordinate) of $x_{m,p}$, respectively.

The second term in the right side of Eq.(1), $\Pr(x_{m,p} \mid y_{m,p}=1)$, indicates a likelihood that favors OPs sharing the similar characteristics with the co-salient objects in the current image group. Hence it can be considered as the metric of intra-group consistency in co-saliency detection. Although the existing approaches have proposed some methods to explore the intra-group consistency among the images within an image group (e.g. the corresponding cue used in [18] and the global similarity used in [17]), the obtained results may badly suffer from the similar background regions that appear frequently in the multiple related images as shown in Figure 4. In this case, the information contained within the certain image group is too limited to solve the problem. Consequently, we propose to look wider in this paper by taking advantage of the visually similar neighbors in other image groups. Intuitively, if these visually similar neighbors contain similar background regions with the images in the current image group, we can use them to select image regions that are more likely to be the co-salient objects and thus suppress the co-saliency values of the background regions in the generated co-saliency maps (as shown in Figure 4).

To formulate the intra-group consistency with considering the wide information, we first apply K-means to divide $\{x_{m,p}\}_{m=1,p=1}^{M,256}$ to K clusters $\{C^k\}_{k=1}^{K}$ with cluster centers $\{c^k\}_{k=1}^{K}$ based on the extracted higher-level representations. Ideally, most of the co-salient regions would be grouped into one cluster, and the common backgrounds would be grouped into the other clusters (as shown in Figure 5). Thus, if we can separate the co-saliency

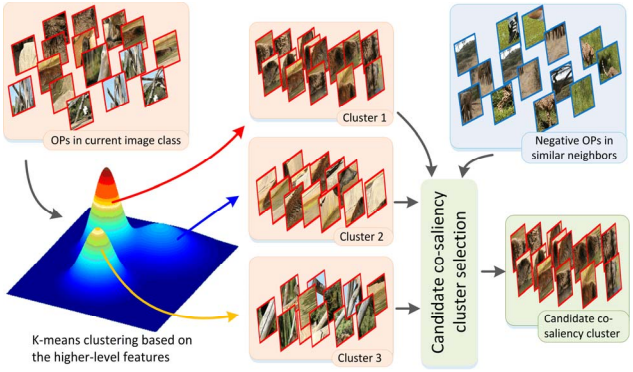


Figure 5: Illustration of the candidate co-saliency cluster selection. The OPs come from the image group of bears. After clustering, the three clusters mainly contain the parts of bear, rocks and trees, respectively. Since rocks and trees are also contained in the negative OPs, the cluster 1 will be selected as the candidate co-saliency cluster via Eqs.(5-7).

cluster from the background clusters, the obtained co-saliency cluster can be used to provide precise information of the co-salient regions and suppress the influence of backgrounds in our co-saliency formulation. To this end, the negative OPs $\{x_{n,p}^{-}\}_{n=1\,p=1}^{20\ \ 256}$ in $\{I_n^{-}\}_{n=1}^{20}$ are used to separate the candidate co-saliency cluster $\bar{C}$ by:

$$\bar{C} = \arg\max_{C^k} D^{-}(C^k) \tag{5}$$

$$\bar{c} = \arg\max_{c^k} D^{-}(C^k) \tag{6}$$

$$D^{-}(C^k) = \min_{n\in[1,20],\, p\in[1,256]} Ed(c^k, f_{n,p}^{-}) \tag{7}$$

where $\bar{c}$ is the cluster center of $\bar{C}$, Eqs.(5-7) are inspired by the negative mining technology used in [25, 31]. After obtaining the candidate co-saliency cluster center $\bar{c}$, we calculate the intra-group consistency via:

$$\Pr(x_{m,p} \mid y_{m,p} = 1) = \frac{1}{Z^{IGC}} \exp\left(-\frac{Ed(f_{m,p}, \bar{c})^2}{\sigma^2}\right) \tag{8}$$

$$Z^{IGC} = \sum_{m=1}^{M}\sum_{p=1}^{256} \exp\left(-\frac{Ed(f_{m,p}, \bar{c})^2}{\sigma^2}\right) \tag{9}$$

Finally, the co-saliency scores for each OP can be obtained by Eq.(1).

3.3. Co-saliency map generation

In order to capture the co-salient objects with finely defined boundaries, we need to convert the window-level co-saliency score to the superpixel-level saliency map. Some existing literatures [32, 33] have proposed effective ways to solve this problem. However, they must rely on the (category-specific) segmentation masks labelled by the manual annotation, which badly limits the application scope of these approaches. In this paper, inspired by [34] we propose a foreground region agreement (FRA) strategy

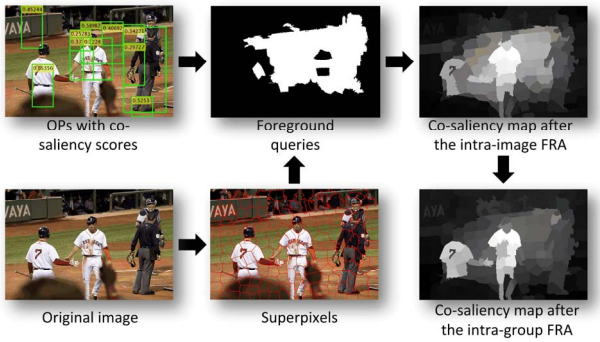


Figure 6: Illustration of the two-phase FRA.

to tackle this problem as shown in Figure 6. The proposed FRA strategy takes into consideration the overall agreement between salient regions over the whole image and the whole group with containing two phases, i.e. the intra-image FRA and the intra-group FRA.

The intra-image FRA. We implement the intra-image FPA by first selecting some foreground queries (the superpixel regions with higher rough co-saliency scores) in each image, and then use the Manifold Ranking algorithm [35] to compute the co-saliency of each mode in a graph model based on their agreement (i.e, ranking) to those queries.

For an image I_m, we adopt the SLIC algorithm [36] to produce superpixels $\{sp_i\}_{i=1}^{N_m}$, where N_m is the number of superpixels in I_m. Then, we apply a pooling-like approach, which is usually used in image classification, by computing the rough co-saliency scores of each superpixel sp_i as the sum of the co-saliency scores of all the OPs that covers it:

$$\text{Cosal}^{rgh}(sp_i) = \sum_{p=1}^{256} \text{Cosal}(x_{m,p}) B(sp_i, x_{m,p}) \tag{10}$$

$$B(sp_i, x_{m,p}) = \begin{cases} 1, & Area(sp_i \cap x_{m,p}) > \frac{1}{2} Area(sp_i) \\ 0, & otherwise \end{cases} \tag{11}$$

where $\text{Cosal}^{rgh}(sp_i)$ denotes the rough co-saliency scores of each superpixel sp_i, $Area(\cdot)$ indicates the area of a certain region. Then an adaptive threshold (i.e. the mean value of the rough co-saliency scores over all superpixels in one image) is set to select the foreground queries. Once we obtain the foreground queries, we follow [37] to establish a graph model $G = (V, E)$ and rank all the superpixels in the image, where the nodes V are the superpixels. The weight w_{ij} between two connected nodes is determined by the difference between their features (i.e. the mean values in the CIE LAB color space) as defined in [37]. Given G, the edges E are weighted by the affinity matrix $\mathbf{W}=[w_{ij}]_{N_m \times N_m}$, and the degree matrix is $\mathbf{D}=\text{diag}\{d_{11}, \cdots, d_{N_m N_m}\}$, where

$d_{ii}=\sum_j w_{ij}$. With the foreground queries, the co-saliency of each node sp_i can be defined as its ranking score r_i assigned by a ranking function:

$$\Re=(\mathbf{D}-\alpha\mathbf{W})^{-1}\ell \quad (12)$$

where $\Re=[r_1,\cdots r_{N_m}]^T$, $\ell=[\ell_1,\cdots \ell_{N_m}]^T$, ℓ_i is a binary variable indicating whether a node sp_i is a foreground query, and α is a free parameter. Thus, after the intra-image FPA, the co-saliency of each superpixel is assigned as: $\text{Cosal}^{FRA_1}(sp_i)=r_i$.

The intra-group FRA. We implement the intra-group FRA by modifying the co-saliency of each superpixel based on the agreement to its similar regions among other images in the image group. Specifically, for each sp_i in image I_m, its H most similar superpixels in each of the other images are searched based on the Euclidean distance of their features (i.e. the mean values in the CIE LAB color space) to form the collection $\{sp_h\}_{h=1}^{(m-1)H}$. Then we process the intra-group FRA by:

$$\text{Cosal}^{FRA_2}(sp_i)=\Psi(sp_i)\cdot\exp(-\Phi(sp_i)) \quad (13)$$

where

$$\Psi(sp_i)=\sum_{h=1}^{(M-1)H}\text{Cosal}^{FRA_1}(sp_h) \quad (14)$$

$$\Phi(sp_i)=\sum_{h=1}^{(M-1)H}Ed(col_i,col_h) \quad (15)$$

where $\Psi(sp_i)$ reflects the overall co-saliency of superpixels similar to sp_i. $\exp(-\Phi(sp_i))$ is the term to take into consideration the similarity between the searched similar superpixels. Thus Eq.(13) calculates the agreement between superpixels within the whole image group and the obtained co-saliency maps are considered as the final co-saliency detection results.

4. Experiment

4.1. Experimental settings

Datasets. We evaluated the proposed algorithm on two benchmark datasets: the iCoseg dataset [1] and the MSRC dataset [38]. The former one is the largest publicly available dataset so far used for co-saliency detection. It consists of 38 image groups of totally 643 images along with manually labeled pixel-wise ground truth data. The latter one consists of 8 image groups (240 images) with manually labeled pixel-wise ground truth data. Image groups like airplane, car, and bicycle are contained in this dataset. Since the *grass* image group has no co-salient objects in each image, we did not use it in our experiment. Compared with the iCoseg dataset, different colors and shapes are allowed for the co-salient objects appearing in image groups of the MSRC dataset, making it more challenging for co-saliency detection.

Evaluation metrics. To evaluate the performance of the proposed method, we adopted three widely used criteria, i.e. the precision recall (PR) curve, the average precision (AP) score, and the F-measure. PR curves and AP scores are generated by thresholding pixels in a co-saliency map into binary co-salient object masks with a series of fixed integers from 0 to 255. The resulting true positive rate versus the precision rate at each threshold value forms the PR curve.

To further evaluate the performance of the proposed method for co-salient object segmentation, we reported the performance in segmenting the saliency map using a self-adaptive threshold $T=\mu+\varepsilon$ as suggested in [34], where μ and ε are the mean value and the standard deviation of the co-saliency map, respectively. Then, we followed the benchmark introduced in [39] to obtain the average precision and recall values over the images, as well as the F-measure defined by:

$$F_\beta=\frac{(1+\beta^2)\,Precision\times Recall}{\beta^2\times Precision+Recall} \quad (16)$$

where $\beta^2=0.3$ as suggested in [2, 37, 39].

Implementation details. In the proposed 10-layer CNN, the first 8 layers use the Overfeat implementation described by [27]. Through a greedy layer-wise hyper-parameter search [40], the learning rate and sparsity target for the last two layers were set to be 0.002 and 0.005, respectively. The sparsity costs for layer 9 and layer 10 were set to be 0.005 and 0.003, respectively. As suggested in [37], we set the number of superpixels to be 200 for each image, and α=0.99 in Eq.(12). For other parameters, we empirically set σ=5 in Eqs.(2, 8, 9), K=3 in section 3.2 and H=3 in Eq.(14) and Eq.(15), respectively. The proposed algorithm is implemented via MATLAB on a PC with 2.8GHz CPU, 64G RAM and GeForce GTX Titan Black. It costs 5.23s per image, which is faster than the previous best method SACS [24] who needs 7.36s per image.

4.2. Evaluation on the iCoseg dataset

We first evaluated the proposed approach on the widely-used iCoseg dataset. For subjective evaluation, we show some experimental results in Figure 7(a), which contains examples in two image groups, i.e., the *Pyramid* group and the *Cheetah* group. As can be seen, compared with the state-of-the-art approaches CSHS [17] and CBCS [18], our proposed approach can yield co-saliency maps more correctly and robustly.

For quantitative evaluation, we compared the proposed approach with 6 state-of-the-art methods, i.e., CSHS [17], CBCS [18], SACS [24], CBCS-S[1][18], BLSM [4], and LR

[1]CBCS-S is the single image saliency detection method in [18].

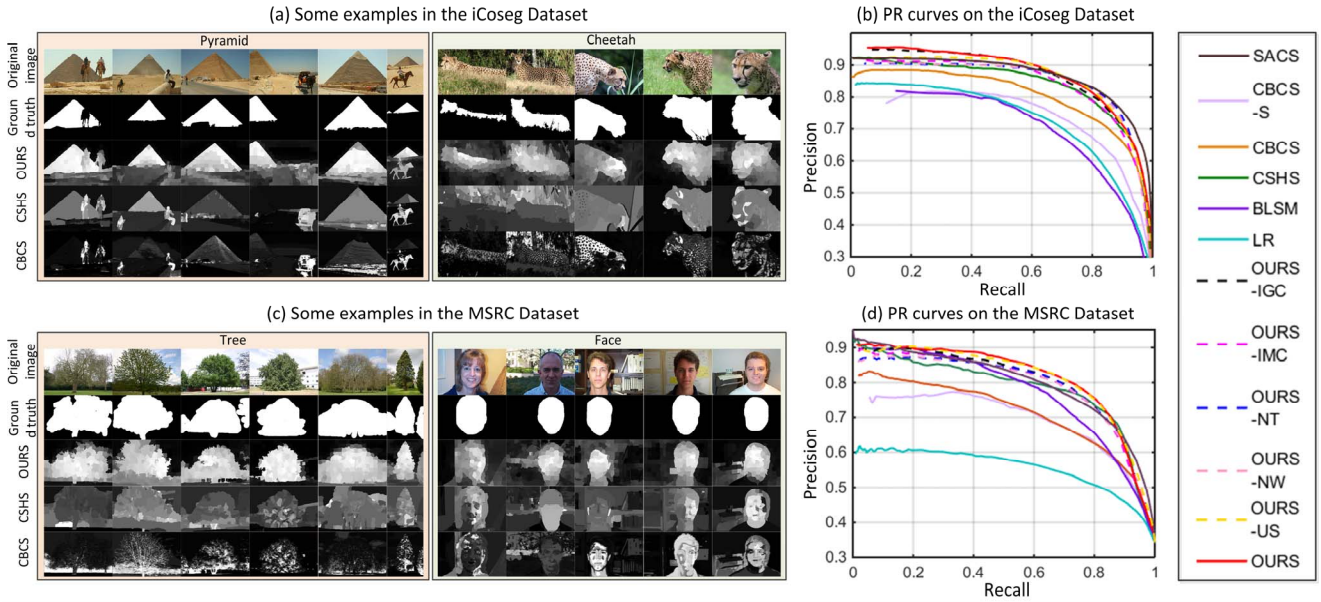


Figure 7: Subjective and quantitative evaluation of the proposed approach in the iCoseg and MSRC datasets.

Methods	LR [3]	BLSM [4]	CSHS [17]	CBCS [18]	CBCS -S[18]	SACS [24]
AP	0.727	0.719	0.839	0.805	0.769	0.865
F-measure	0.698	0.674	0.754	0.740	0.715	0.793
Methods	OURS -IMC	OURS -IGC	OURS -NW	OURS -NT	OURS -US	OURS
AP	0.845	0.867	0.848	0.850	0.862	**0.873**
F-measure	0.777	0.787	0.782	0.791	0.795	**0.798**

Table 1: AP and F-measure scores on the iCoseg dataset.

Methods	LR [3]	BLSM [4]	CSHS [17]	CBCS [18]	CBCS -S[18]	SACS [24]
AP	0.557	0.774	0.783	0.713	0.699	0.799
F-measure	0.481	0.725	0.711	0.588	0.620	0.711
Methods	OURS -IMC	OURS -IGC	OURS -NW	OURS -NT	OURS -US	OURS
AP	0.794	0.802	0.791	0.792	0.812	**0.814**
F-measure	0.732	0.737	0.730	0.729	0.747	**0.751**

Table 2: AP and F-measure scores on the MSRC dataset.

[3], where the first three methods are the state-of-the-art co-saliency detection methods and the last three methods are the state-of-the-art saliency detection methods. The experimental results are shown in Figure 7(b) and Table 1. As can be seen, both the state-of-the-art saliency detection methods and co-saliency detection methods could not solve the problems in co-saliency detection satisfyingly. However exploring the deep and wide information via the proposed framework (OURS) could improve the performance of existing methods obviously.

We also conducted experiments to evaluate the effectiveness of each part in the proposed framework. From Figure 7(b) and Table 1, we can see that: 1) Without using the wide information (OURS-NW), the AP score of proposed framework decreases from 0.873 to 0.848, demonstrating the significance of the wide information in co-saliency detection. Since OURS-NW is still better than all the existing methods which are all using low-level features, the importance of the deep information is also demonstrated; 2) Directly using the CNN proposed in [27] without transfer layers (OURS-NT) obtains worse performance than the proposed method (OURS), which demonstrates the effectiveness of the additional RBM layers proposed in this paper for building domain-specific higher-level feature. 3) Exploring both the wide and deep information (OURS) by fusing intra-image contrast (OURS-IMC) and intra-group consistency (OURS-IGC) via the proposed framework yields the best performance for co-saliency detection. In addition, we use an unsupervised way to extract higher-level features for fair comparison. To be specific, we train two RBM layers on the extracted HOGgles [41]. Based on this feature, our approach (OURS-US) is reasonably worse than it with the transferred feature but still obtains the satisfying performance.

4.3. Evaluation on the MSRC dataset

In this section, we conducted the comparisons on a more challenging dataset, i.e., the MSRC dataset. For subjective evaluation, we showed some experimental results in Figure 7(c), which contains examples in two image groups, i.e., the *Tree* group and the *Face* group. From the former group, we can see that the proposed approach can effectively capture the homogeneity of the co-salient objects and uniformly highlight them even though they exhibit large variations in shape and texture. From the latter group, we can observe that the proposed approach can better suppress the complex backgrounds and yield the co-saliency maps robustly.

For quantitative evaluation, we also compared the

Image group	AP			F-measure		
	CBCS	CSHS	OURS	CBCS	CSHS	OURS
Cattle	**0.876**	0.855	0.856	0.782	**0.865**	0.840
Tree	0.753	0.772	**0.888**	0.478	0.526	**0.715**
House	0.689	**0.923**	0.892	0.525	**0.820**	0.784
Airplane	0.527	**0.624**	0.578	0.462	**0.581**	0.577
Face	0.614	0.775	**0.860**	0.592	0.739	**0.831**
Bike	0.636	0.646	**0.695**	0.458	0.519	**0.621**
Car	0.794	0.819	**0.892**	0.570	0.690	**0.751**
Overall	0.713	0.783	**0.814**	0.588	0.711	**0.751**

Table 3: Comparison of AP and F-measure scores between the proposed approach and the other state-of-the-art co-saliency detection methods for each image group in the MSRC dataset.

proposed approach with 5 state-of-the-art methods, i.e., CSHS, CBCS, SACS[2], BLSM, and LR. Figure 7(d) shows the PR curves of these approaches, from which we can see that the proposed approach (OURS) performs the best at recall rates among [0, 0.85] approximately and obtains the highest AP and F-measure scores as shown in Table 2. Being consist with it on the iCoseg dataset, the performance of OURS-NW is lower than OURS but higher than the other existing methods, which demonstrates the importance of the wide and deep information, respectively. Exploring both the deep and wide information via the proposed framework (OURS) yields the best co-saliency detection result which obviously outperforms other state-of-the-art methods. In addition, the comparision between OURS and OURS-NT also demonstrates the effectiveness of the proposed CNN architecture for transfering higher-level features.

To perform further verification, we compared the scores between the proposed approach and other state-of-the-art co-saliency detection methods for each image group in the MSRC dataset. The comparison results are reported in Table 3. As can be seen, CBCS [18], CSHS [17], and the proposed approach can obtain the superior performance than other methods in 1, 2, and 4 image groups, respectively. Obvious performance gain of the proposed approach can be found in the image groups of *Tree*, *Face* and *Car*. In the other image groups, the failure of our method is mainly caused by the limited quality of the object proposals. The overall performance of the proposed approach is much better than other state-of-the-art methods.

4.4. Comparing with co-segmentation approaches

Co-saliency is relevant to co-segmentation. But they are different: co-saliency detection only focuses on detecting common salient objects while the similar but non-salient background might be also segmented out in co-segmentation. Some methods use the prior knowledge to suppress the common background and thus can segment out the common foreground objects. In order to compare with these works, we do experiment using the evaluation metrics and the Internet-100 Dataset as in [42]. Although our segmentation mask is just obtained by a simple threshold and no object detector is trained in our system, the obtained result shown in Figure 10 is competitive with the most state-of-the-art co-segmentation approaches EVK [42] and UJD [10].

[2]The performance of SACS in the MSRC dataset was obtained by using the algorithm of [24] to integrate MR [37], RC [45], SP [18], and CO [18].

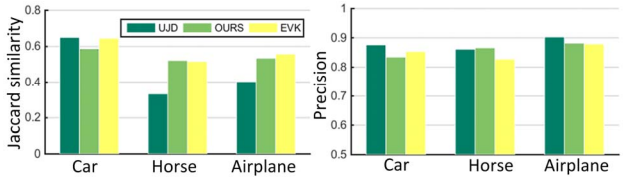


Figure 8: Co-segmentation results on the Internet-100 Dataset.

5. Conclusion

We have proposed a novel framework by introducing the deep and wide information for co-saliency detection. By looking deep, our approach is able to capture the concept-level properties of the co-salient objects. By looking wide, our approach is able to suppress the common backgrounds in the image group by modeling the background using cross-group information. For the future work, we tend to further explore the relationship between co-saliency detection and weakly supervised learning [43, 44], and extend our method to real-world video processing tasks.

Acknowledgements: This work was partially supported by the National Science Foundation of China under Grant 61473231.

References

[1] D. Batra, A. Kowdle, D. Parikh, L. Jie, and T. Chen. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR,* 2010.

[2] H. Li, F. Meng, and K. N. Ngan. Co-Salient Object Detection From Multiple Images. *IEEE Trans. Multimedia,* 15(8): 1896-1909, 2013.

[3] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *CVPR, 2012*.

[4] Y. Xie, H. Lu, and M.-H. Yang. Bayesian saliency via low and mid level cues. *IEEE Trans. Image Process.,* 22(5): 1689-1698, 2013.

[5] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.

[6] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li. Automatic salient object segmentation based on context and shape prior. In *BMVC*, 2011.

[7] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu. An object-oriented visual saliency detection framework based on sparse coding representations. *IEEE Trans. Circuits Syst. Video Technol.,* 23(12):2009 -2021, 2013.

[8] J. Han, D. Zhang, X. Hu, L. Guo, J. Ren, and F. Wu. Background Prior Based Salient Object Detection via Deep Reconstruction Residual. *IEEE Trans. Circuits Syst. Video Technol.,* 2014. DOI: 10.1109/TCSVT.2014.2381471.

[9] J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, and X. Li. Two-Stage Learning to Predict Human Eye Fixations via SDAEs. *IEEE Trans. on Cybernetics,*2015. DOI: 10.1109/TCYB.2015.2404432.

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

[11] L. Wang, G. Hua, R. Sukthankar, J. Xue, and N. Zheng. Video object discovery and co-segmentation with extremely weak supervision. In *ECCV,* 2014.

[12] A. Toshev, J. Shi, and K. Daniilidis. Image matching via saliency region correspondences. In *CVPR,* 2007.

[13] J. Xue, L. Wang, N. Zheng, and G. Hua. Automatic salient object extraction with contextual cue and its applications to recognition and alpha matting. *Pattern Recognit.*, 46(11): 2874-2889, 2013.

[14] J. Guo, Z. Li, L.-F. Cheong, and S. Z. Zhou. Video Co-segmentation for Meaningful Action Extraction. In *ICCV,* 2013.

[15] M. Eichner, and V. Ferrari. Human pose co-estimation and applications. *IEEE Trans. Pattern Anal. Mach. Intell.,* 34(11):2282-2288, 2012.

[16] H. Li and K. N. Ngan. A Co-Saliency Model of Image Pairs. *IEEE Trans. Image Process.,* 20(12):3365-3375, 2011.

[17] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur. Co-Saliency Detection Based on Hierarchical Segmentation. *IEEE Signal Process. Lett.,* 21(1):88-92, 2014.

[18] H. Fu, X. Cao, and Z. Tu. Cluster-Based Co-Saliency Detection. *IEEE Trans. Image Process.,* 22(10):3766-3778, 2013.

[19] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. *arXiv preprint arXiv:1311.2524*, 2013.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic. Learning and transferring mid-level image representations using convolutional neural networks. In *CVPR*, 2014.

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

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

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

[24] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng. Self-Adaptively Weighted Co-Saliency Detection via Rank Constraint. *IEEE Trans. Image Process.,* 23(9):4175-4186, 2014.

[25] P. Siva, C. Russell, T. Xiang, and L. Agapito. Looking Beyond the Image: Unsupervised Learning for Object Saliency and Detection. In *CVPR*, 2013.

[26] M.-M. Cheng, Z. Zhang, W.-Y. Lin, and P. Torr. BING: Binarized normed gradients for objectness estimation at 300fps. In *CVPR*, 2014.

[27] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. *arXiv preprint arXiv:1312.6229*, 2013.

[28] Y. Bengio. Learning deep architectures for AI. *Foundations and Trends in Machine Learning,* 2(1):1-127, 2009.

[29] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *J. Vision,* 8(7):32, 2008.

[30] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.,* 34(10):1915-1926, 2012.

[31] P. Siva, C. Russell, and T. Xiang. In defence of negative mining for annotating weakly labelled data. In *ECCV,* 2012.

[32] J. Tighe and S. Lazebnik. Finding things: Image parsing with regions and per-exemplar detectors. In *CVPR,* 2013.

[33] D. Kuettel and V. Ferrari. Figure-ground segmentation by transferring window masks. In *CVPR,* 2012.

[34] Y. Jia and M. Han. Category-Independent Object-level Saliency Detection. In *ICCV,* 2013.

[35] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. In *NIPS*, 2004.

[36] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, Su, x, and S. sstrunk. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.,* 34(11):2274-2282, 2012.

[37] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency Detection via Graph-Based Manifold Ranking. In *CVPR*, 2013.

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

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

[40] Y. Bengio. Practical recommendations for gradient-based training of deep architectures. *Neural Networks: Tricks of the Trade*, 2012.

[41] C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *ICCV*, 2013.

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

[43] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren. Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning. *IEEE Trans. Geosci. Remote Sens.,* 53(6): 3325-3337, 2015.

[44] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo. Weakly Supervised Learning for Target Detection in Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.,* 12(4): 701-705, 2015.

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