Unsupervised CNN-based Co-Saliency Detection with Graphical Optimization

Kuang-Jui Hsu[1,2], Chung-Chi Tsai[1,3], Yen-Yu Lin[1], Xiaoning Qian[3],
Yung-Yu Chuang[1,2]

[1]Research Center for Information Technology Innovation, Academia Sinica, Taiwan
kjhsu@citi.sinica.edu.tw, yylin@citi.sinica.edu.tw
[2]Computer Science & Information Engineering, National Taiwan University, Taiwan
cyy@csie.ntu.edu.tw
[3]Electrical & Computer Engineering, Texas A&M University, U.S.A
chungchi@ece.tamu.edu, xqian@ece.tamu.edu

Abstract. In this paper, we address co-saliency detection in a set of images jointly covering objects of a specific class by an unsupervised convolutional neural network (CNN). Our method does not require any additional training data in the form of object masks. We decompose co-saliency detection into two sub-tasks, *single-image saliency detection* and *cross-image co-occurrence region discovery* corresponding to two novel unsupervised losses, *the single-image saliency (SIS) loss* and *the co-occurrence (COOC) loss*. The two losses are modeled on a graphical model where the former and the latter act as the unary and pairwise terms, respectively. These two tasks can be jointly optimized for generating co-saliency maps of high quality. Furthermore, the quality of the generated co-saliency maps can be enhanced via two extensions: map sharpening by self-paced learning and boundary preserving by fully connected conditional random fields. Experiments show that our method achieves superior results, even outperforming many supervised methods.

Keywords: Co-saliency detection, unsupervised learning, convolutional neural networks, deep learning, graphical model.

1 Introduction

Co-saliency detection refers to searching for visually salient objects repetitively appearing in multiple given images. For its superior scalability, co-saliency has been applied to help various applications regarding image content understanding, such as image/video co-segmentation [1,2,3], object co-localization [4], content-aware compression [5], etc.

The success of co-saliency detection relies on robust feature representations of co-salient objects against appearance variations across images. Engineered features, such as color histograms, Gabor filtered texture features, and SIFT [6] are widely used in conventional co-saliency methods [7,8,9,10]. Deep learning (DL) has recently emerged and demonstrated success in many computer vision applications. DL-based features have been adopted for co-saliency detection, such

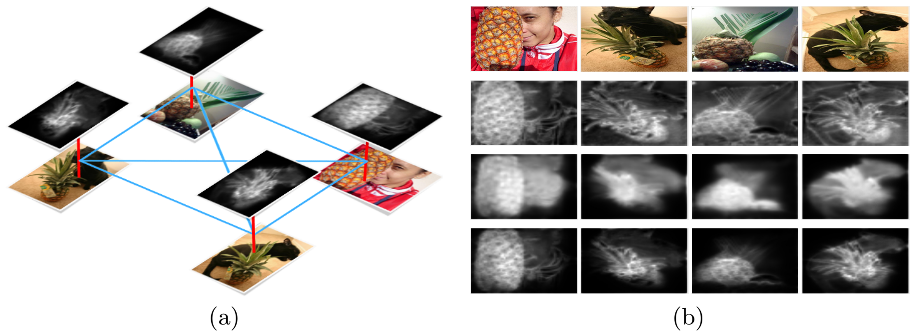


Fig. 1. Motivation of our method. (a) Our method optimizes an objective function defined on a graph where single-image saliency (SIS) detection (red edges) and cross-image co-occurrence (COOC) discovery (blue edges) are considered jointly. (b) The first row displays the images for co-saliency detection. The following three rows show the detected saliency maps by using COOC, SIS, and both of them, respectively.

as those extracted from a pre-trained convolutional neural network (CNN) [11] or from unsupervised semantic feature learning with restricted Boltzmann machines (RBMs) [12]. However, feature extraction and co-saliency detection are treated as separate steps in these approaches [7,8,9,10,11,12], leading to suboptimal performance. In contrast, the supervised methods, by metric learning [13] or DL [14], enable the integration of feature learning and co-saliency detection. However, they require additional training data in the form of object masks, often manually drawn or delineated by tools with intensive user interaction. Such heavy annotation cost makes these methods less practical as pointed out in other applications, such as semantic segmentation [15] and saliency detection [16]. Furthermore, their learned models may not perform well for unseen object categories in testing, since the models do not adapt themselves to unseen categories.

In this work, we address the aforementioned issues by proposing an unsupervised CNN-based method for joint adaptive feature learning and co-saliency detection for given images, hence making a good compromise between the performance and the annotation requirement. In the proposed method, co-saliency detection is decomposed into two complementary parts, *single-image saliency detection* and *cross-image co-occurrence region discovery*. The former detects the saliency object in a single image, which may not repetitively appear across images. The latter discovers regions repetitively appearing across images, which may not be visually salient. To this end, we design two novel losses, *the single-image saliency (SIS) loss* and *the co-occurrence (COOC) loss*, to capture the two different but complementary sources of information. These two losses measure the quality of the saliency maps by referring to individual images and the co-occurrence regions for each image pair, respectively. They are further integrated on a graphical model whose unary and pairwise terms correspond to the proposed SIS and COOC losses respectively, as illustrated in Fig. 1 (a). Through optimizing the proposed losses, our approach can generate co-saliency maps of high quality by integrating SIS and COOC cues, as shown in Fig. 1 (b).

To the best of our knowledge, our method represents the first unsupervised CNN model for co-saliency detection. Compared with unsupervised methods including those using engineered features [3,7,8,9,10] and those using DL-based features [11,12], our method achieves better performance by joint adaptive feature learning and co-saliency detection based on CNNs. Compared with the supervised method [13,17], our method can reach comparable or even slightly better performance and does not suffer from the high annotation cost of labeling object masks as training data. We comprehensively evaluate our method on three benchmarks for co-saliency detection, including *the MSRC dataset* [18], *the iCoseg dataset* [19], and *the Cosal2015 dataset* [12]. The results show that our approach remarkably outperforms the state-of-the-art unsupervised methods and even surpasses many supervised DL-based saliency detection methods.

2 Related work

2.1 Single-image saliency detection

Single-image saliency detection is to distinguish salient objects from the background by either unsupervised [20,21,22,23,24,25] or supervised [26,27,28,29,30] methods based on color appearance, spatial locations, as well as various supplementary higher-level priors, including objectness. These approaches can handle well images with single salient objects. However, they may fail when the scenes are more complex, for example when multiple salient objects are presented with intra-image variations. By exploiting co-occurrence patterns when common objects appearing in multiple images, co-saliency detection is expected to perform better. However, the appearance variations of common objects across images could also make co-saliency detection a more challenging task.

2.2 Co-saliency detection

Co-saliency detection discovers common and salient objects across multiple images using different strategies. The co-saliency detection methods have been developed within the bottom-up frameworks based on different robust features, including low-level handcrafted features [3,7,8,9,10,17,31,32] and high-level DL-based semantic features [11,12] to catch intra-image visual stimulus as well as inter-image repetitiveness. However, there are no features adopted suitable for all visual variations, and they treat the separate steps of feature extraction and co-saliency detection, leading to suboptimal performance. Data-driven methods [13,14,17] directly learn the patterns of co-salient objects to overcome the limitation of bottom-up methods. For instance, the transfer-learning-based method [17] uses the object masks to train a stacked denoising autoencoder (SDAE) to learn the intra-image contrast evidence, and propagate this knowledge to catch inter-image coherent foreground representations. Despite their impressive results, the performance might drop dramatically once the transferred knowledge on feature representations is not satisfactory as the separation of feature extraction and co-saliency detection may potentially impede the performance. Recently, Wei *et al.* [14] and Han *et al.* [13] have proposed unified learning-based methods to learn semantic features and detect co-salient objects jointly. Despite the improved performance, their methods rely on a large number of training object masks. It reduces the generalizability of their approaches to unseen images. However, our method can perform the adaptive and unified learning for given images in an unsupervised manner, and hence no aforementioned issues exist in our approach.

2.3 Graphical models with CNNs

Deep learning has demonstrated success in many computer vision applications. For better preserving spatial consistency, graphical models have been integrated with CNNs when requiring structured outputs, such as depth estimation [33], stereo matching [34], semantic segmentation [35,36,37], image denoising, and optical flow estimation [38]. Although showing promise in preserving spatial consistency and modeling pairwise relationships, these methods have three major limitations when extending to co-saliency detection. First, their graphical models are built on single images, and hence can not be directly applied to co-saliency detection with multiple images. Second, the pairwise terms in these graphical models often act as regularization terms to ensure spatial consistency but can not work alone by themselves. Finally, they require training data to train the model. For the inter-image graphical models, Hayder *et al.* [39] and Yuan *et al.* [40] respectively integrated fully-connected CRFs into CNNs for object proposal co-generation and object co-segmentation, where each node is an object proposal. However, their methods still suffer from the last two limitations. In comparison, our method integrates the merits from graphical models for co-saliency detection without aforementioned issues.

3 Our approach

We first describe the proposed formulation for co-saliency detection. Next, we propose a couple of enhancements by self-paced learning and fully connected conditional random fields. Finally, the optimization process and the implementation details are provided.

3.1 The proposed formulation

Given a set of N images $\{I_n\}_{n=1}^{N}$, co-saliency detection aims to detect the salient objects of a category commonly present in these images. We accomplish the task by decomposing it into two sub-tasks, *single-image saliency detection* and *cross-image co-occurrence discovery*. The former detects the salient regions in a single image, without considering whether the detected regions are commonly present

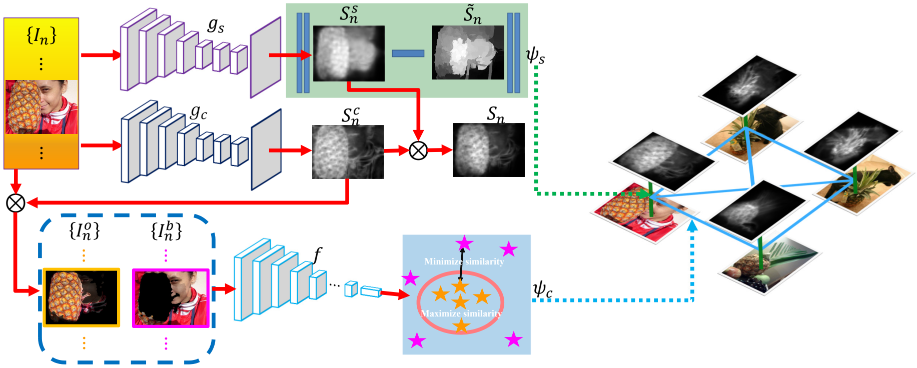


Fig. 2. Overview of our approach to co-saliency detection. It optimizes an objective function defined on a graph by learning two collaborative FCN models g_s and g_c which respectively generates single-image saliency maps and cross-image co-occurrence maps.

across images. The latter discovers the regions repetitively occurring across images, while disregarding whether the discovered regions stand out visually. Co-saliency detection, finding the salient co-occurrence regions, can then be carried out by performing and integrating the two tasks on a graph whose two types of edges respectively correspond to the two tasks, as shown in Fig. 1 (a). The proposed objective function on the graph is defined by

$$E(\mathbf{w}) = \sum_{n=1}^{N} \psi_s(I_n; \mathbf{w}) + \sum_{n=1}^{N} \sum_{m \neq n} \psi_c(I_n, I_m; \mathbf{w}), \tag{1}$$

where the unary term $\psi_s(I_n; \mathbf{w})$ focuses on saliency detection for the image I_n, the pairwise term $\psi_c(I_n, I_m; \mathbf{w})$ accounts for co-occurrence discovery for the image pair (I_n, I_m), and $\mathbf{w}$ is the set of model parameters.

As shown in Fig. 2, we learn two fully convolutional network (FCN) [41] models, g_s and g_c, to optimize the unary term ψ_s and the pairwise term ψ_c in Eq. (1), respectively. For image I_n, FCN g_s investigates intra-image clues and generates its *saliency map* S_n^s. In contrast, FCN g_c discovers cross-image evidence and produces its *co-occurrence map* S_n^c, where the repetitively occurring regions are highlighted. The resultant *co-saliency map*, highlighting the co-occurrence and salient regions, is yielded by $S_n = g_s(I_n) \otimes g_c(I_n) = S_n^s \otimes S_n^c$, where $\otimes$ denotes the element-wise multiplication operator.

Let $\mathbf{w}_s$ and $\mathbf{w}_c$ denote the learnable parameter sets of FCNs g_s and g_c, respectively. We learn g_s and g_c jointly by optimizing $E(\mathbf{w} = \mathbf{w}_s \cup \mathbf{w}_c)$ in Eq. (1). The unary term ψ_s and the pairwise term ψ_c in Eq. (1) are described below.

3.2 Unary term ψ_s

This term aims to identify the salient regions in a single image. It guides the training of FCN g_s, which produces saliency map S_n^s for image I_n, i.e., $S_n^s =$

$g_s(I_n)$. Inspired by Zhang *et al.* [42], we apply an existing unsupervised method for saliency detection to image I_n, and use its output saliency map $\tilde{S}_n$ as the desired target for learning FCN g_s. In this work, we adopt MILP [25] to generate $\tilde{S}_n$. Specifically, the unary term $\psi_s(I_n; \mathbf{w}_s)$ applied to image I_n is defined by

$$\psi_s(I_n; \mathbf{w}_s) = \sum_{i \in I_n} R_n(i) |S_n^s(i) - \tilde{S}_n(i)|^2, \tag{2}$$

where i is the index of the pixels in I_n, $S_n^s(i)$ and $\tilde{S}_n(i)$ are respectively the saliency values of maps S_n^s and $\tilde{S}_n$ at pixel i, and $R_n(i)$ represents the importance of pixel i. Pixels in map $\tilde{S}_n$ can be divided into the salient and non-salient groups by using the mean value of $\tilde{S}_n$ as the threshold. $R_n(i)$ is introduced here to deal with the potential size unbalance between the two groups. Let δ be ratio of salient pixels over the whole image I_n. $R_n(i)$ takes the value $1 - \delta$ if pixel i belongs to the salient group, and δ otherwise. In this way, the salient and non-salient groups contribute equally in Eq. (2).

3.3 Pairwise term ψ_c

The pairwise term ψ_c seeks the regions simultaneously appearing across images. It serves as the objective to learn FCN g_c. The regions should look similar across images but distinctive from surrounding non-detected regions. Thus, two criteria are jointly considered in the design of ψ_c, including 1) high cross-image similarity between the detected co-occurrence regions and 2) high intra-image distinctness between the detected co-occurrence regions and the rest of the image. The second criterion is auxiliary but crucial to avoid trivial solutions.

As shown in Fig. 2, FCN g_c produces the co-occurrence map S_n^c for image I_n, i.e., $S_n^c = g_c(I_n)$. The sigmoid function is used as the activation function in the last layer of g_c. Thus, the value of the co-occurrence map at each pixel i, $S_n^c(i)$, is between 0 and 1. With S_n^c, image I_n is decomposed into two masked images,

$$I_n^o = S_n^c \otimes I_n \quad \text{and} \quad I_n^b = (1 - S_n^c) \otimes I_n, \tag{3}$$

where $\otimes$ denotes element-wise multiplication. The masked image I_n^o keeps the detected co-occurrence regions of I_n, while image I_n^b contains the rest.

To measure the similarity between images, we employ a feature extractor f to compute the features of a given image. In this work, the extractor f can be a pre-trained CNN model for image classification, e.g., AlexNet [43] or VGG-19 [44], with the softmax function and the last fully connected layer removed. We apply the extractor f to all masked images $\{I_n^o, I_n^b\}_{n=1}^N$ and obtain their features $\{f(I_n^o) \in \mathbb{R}^c, f(I_n^b) \in \mathbb{R}^c\}_{n=1}^N$, where c is the feature dimension. With these extracted features, the pairwise term $\psi_c(I_n, I_m; \mathbf{w}_c)$ applied the image pair I_n and I_m is defined by

$$\psi_c(I_n, I_m; \mathbf{w}_c) = -\log(p_{nm}), \tag{4}$$

where p_{nm} is the score estimating the quality of the detected co-occurrence regions in I_n and I_m. The score p_{nm} is defined below,

$$p_{nm} = \frac{\exp(-d^+_{nm})}{\exp(-d^+_{nm}) + \exp(-d^-_{nm})}, \text{ where} \tag{5}$$

$$d^+_{nm} = \frac{1}{c}\|f(I^o_n) - f(I^o_m)\|^2 \text{ and} \tag{6}$$

$$d^-_{nm} = \frac{1}{2c}\|f(I^o_n) - f(I^b_n)\|^2 + \frac{1}{2c}\|f(I^o_m) - f(I^b_m)\|^2. \tag{7}$$

Eq. (6) measures the inter-image distance between the detected co-occurrence regions in images I_n and I_m (criterion 1). Eq. (7) evaluates the intra-image distance between the detected co-occurrence regions and the rest of the image (criterion 2). By minimizing the pairwise term $\psi_c(I_n, I_m; \mathbf{w}_c)$ in Eq. (4) for each image pair (I_n, I_m), the resultant FCN g_c will produce the co-occurrence maps where the inter-image distances between the detected co-occurrence regions are minimized while the intra-image distances between the detected co-occurrence regions and the rest of the images are maximized. After learning FCNs g_s and g_c jointly through the unary and pairwise terms in Eq. (1), the resultant co-saliency map S_n of a given image I_n is produced via $S_n = g_s(I_n) \otimes g_c(I_n)$.

Note that the pairwise term in Eq. (4) is defined by referring to the co-occurrence maps produced by FCN g_c, i.e., S^c_n and S^c_m. In practice, we found that the performance of co-saliency detection can be further improved if co-saliency maps S_n and S_m are also taking into account in the pairwise term. In our implementation, we extend the pairwise term in Eq. (4) to

$$\psi_c(I_n, I_m; \mathbf{w}_c) = -\lambda_c \log(p_{nm}) - \lambda_{\tilde{c}} \log(\tilde{p}_{nm}), \tag{8}$$

where like p_{nm}, $\tilde{p}_{nm}$ is computed in the same way but by referring to co-saliency maps S_n and S_m. Constants λ_c and $\lambda_{\tilde{c}}$ are used in Eq. (8) for weighting the corresponding terms. In the following, we will show that the quality of the co-saliency maps can be further improved via two extensions, including map enhancement by self-paced learning and postprocessing by fully connected conditional random fields (or DenseCRFs) [45].

3.4 Co-saliency map enhancement

The self-paced learning with CNNs is proposed to make salience map sharper. Then, fully connected conditional random fields are adopted to preserve co-salient objects' boundaries. The details of these two extensions are given below.
Co-saliency map enhancement by self-paced learning. The co-saliency maps obtained by optimizing Eq. (1) are sometimes over smooth, because both FCNs g_s and g_c do not take into account the information regarding object boundaries. To address this issue, we oversegment each image I_n into superpixels $Q_n = \{q^k_n\}^K_{k=1}$, where q^k_n is the kth superpixel and K is the number of superpixels. Pixels in a superpixel tend to belong to either a salient object or

the background all together. This property can be leveraged to propagate information from pixels of high confidence to those of low confidence within the same superpixel. We divide superpixels into three groups, i.e., $Q_n = O_n \cup B_n \cup T_n$. The first two groups, O_n and B_n, contain superpixels that likely belong to the object and the background, respectively. The third group T_n covers the rest. Given the co-saliency map S_n, the three groups are yielded by

$$q_n^k \in \begin{cases} O_n, & \text{if } \mu_n^k > \mu_n + \sigma_n, \\ B_n, & \text{if } \mu_n^k < \mu_n - 0.25 * \sigma_n, \\ T_n, & \text{otherwise,} \end{cases} \text{ for } k = 1, 2, ..., K, \tag{9}$$

where μ_n^k is the mean saliency value of superpixel q_n^k, while μ_n and σ_n are the mean and the standard deviation of $\{\mu_n^k\}_{k=1}^K$. In addition, we follow the background seed sampling strategy used in previous work [20,46], and add superpixels on the image boundary to the set B_n. Superpixels in O_n and B_n are confident to be assigned to either the salient regions or the background. Those in T_n are ambiguous, so they are not taken into account here. With O_n and B_n of image I_n, another FCN g_e for co-saliency map **e**nhancement is trained by optimizing

$$\psi_e(I_n; \mathbf{w}_e) = w_o \sum_{q \in O_n} \sum_{i \in q} |S_n^e(i) - 1|^2 + w_b \sum_{q \in B_n} \sum_{i \in q} |S_n^e(i) - 0|^2, \tag{10}$$

where map $S_n^e = g_e(I_n)$ is generated by FCN g_e, and i the index of pixels in I_n. Constants $w_o = \frac{|B_n|}{|O_n|+|B_n|}$ and $w_b = \frac{|O_n|}{|O_n|+|B_n|}$ are the weights used to balance the contributions of O_n and B_n, where $|O_n|$ and $|B_n|$ are the numbers of pixels in O_n and B_n, respectively.

The term in Eq. (10) enhances the consensus within superpixels of high confidence. If it is turned on, the objective is extended from that in Eq. (1) to

$$E(\mathbf{w}) = \sum_{n=1}^{N} \psi_s(I_n; \mathbf{w}_s) + \lambda_e \psi_e(I_n; \mathbf{w}_e) + \sum_{n=1}^{N-1} \sum_{m=n+1}^{N} \psi_c(I_n, I_m; \mathbf{w}_c), \tag{11}$$

where λ_e is a weight, and $\mathbf{w} = \mathbf{w}_s \cup \mathbf{w}_e \cup \mathbf{w}_c$ is the union of the learnable parameter sets of FCNs g_s, g_e, and g_c. After optimizing the objective function in Eq. (11), the co-saliency map S_n of image I_n is generated by $S_n = g_s(I_n) \otimes g_e(I_n) \otimes g_c(I_n) = S_n^s \otimes S_n^e \otimes S_n^c$.

Postprocessing using DenseCRFs. The co-saliency maps obtained by optimizing the objective in Eq. (11) can be further improved by enforcing spatial coherence and preserving object boundaries. To this end, we follow previous work [47,28] and adopt DenseCRFs [45] to postprocess the co-saliency map S_n of a given image I_n. We use the DenseCRFs code implemented by Li and Yu [47] in this work.

3.5 Optimization

To reduce memory consumption and speed up the training, the proposed method is optimized by using a two-stage procedure. At the first stage, we respectively learn FCNs g_s and g_c by using the objective functions in Eq. (2) and Eq. (4) with all images for 20 epochs. The co-saliency maps $\{S_n = g_s(I_n) \otimes g_c(I_n)\}_{n=1}^N$ become stable enough. Thus, we divide the superpixels of each image into three groups via Eq. (9). FCN g_e is then trained with the objective in Eq. (10) with all images for 20 epochs. At the second stage, we turn on all the three terms in Eq. (11) where the extended pairwise term in Eq. (8) is adopted. The three FCNs, g_s, g_e, and g_c, are optimized jointly for 20 epochs. Note that at the second stage, we optimize only the parameters in the last two convolutional layers and the skip connections of each FCN model.

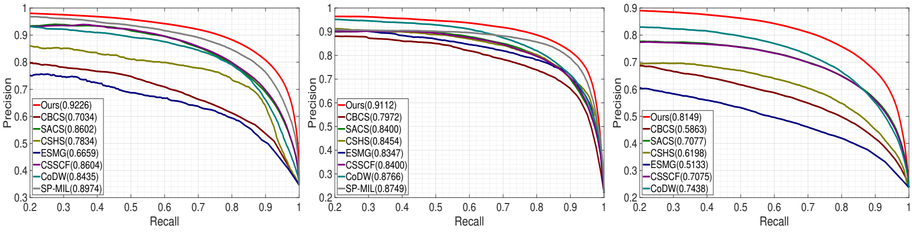
The objectives in Eq. (1) and Eq. (11) are defined on a fully-connected graph. It is difficult to directly optimize either objective with all images at the same time due to the limited memory size. Thereby, we adopt the *piecewise training* scheme [48]. Namely, we consider only the sub-graph yielded by a subset of images at each time. The learning rate is set to 10^{-6} at the first stage and is reduced to 10^{-8} at the second stage. The weight decay and momentum are set to 0.0005 and 0.9, respectively. The objective function in Eq. (11) is differentiable. We choose ADAM [49] as the optimization solver for its rapid convergence. The gradients with respect to the optimization variables can be derived straightforward, so we omit their derivation here.

3.6 Implementation details

The proposed method is implemented using `MatConvNet` [50]. The same network architecture is used in all the experiments. ResNet-50 [51] is adopted as the feature extractor f for the pairwise term, because AlexNet [43] and VGG-16/19 [44] sometimes lead to the problem of vanishing gradients in our application. The feature extractor f is pre-trained on ImageNet [52] and fixed during the optimization process. The features extracted by f are the inputs to the last fully connected layer of f. The feature dimension, i.e., c in Eq. (6) and Eq. (7), is set to 2,048. All FCNs, including g_s, g_e and g_c, are developed based on the VGG-16 [44] setting of FCN [41]. We replace the activation function *softmax* in the last layer with the *sigmoid* function. SLIC [53] is adopted to generate superpixels because of its computational efficiency, better compactness and regularity. The models pre-trained on the ImageNet [54] dataset for classification are required. Following previous co-saliency detection methods [11,12], we determine the values of hyperparameters empirically and keep them fixed in all the experiments.

4 Experimental results

In this section, we first describe the datasets and evaluation metrics. Next, we compare our method with a set of state-of-the-art methods. Finally, we investigate contributions of different components by reporting ablation studies.



MSRC iCoseg Cosal2015

Fig. 3. Comparison with the state-of-the-art methods with the same setting in terms of PR curves on three benchmark datasets. The numbers in parentheses are AP values.

4.1 Datasets and evaluation metrics

Datasets. We evaluated the proposed approach on three public benchmark datasets: *iCoseg* [19], *MSRC* [18] and *Cosal2015* [12]. *iCoseg* consists of 38 groups of total 643 images, and each group has $4 \sim 42$ images. The images of *iCoseg* contain single or multiple similar objects in various poses and sizes with complex backgrounds. *MSRC* contains 7 groups of total 240 images, and each group has $30 \sim 53$ images. Compared to *iCoseg*, objects in *MSRC* exhibit greater appearance variation. *Cosal2015* is a more recent and more challenging dataset than the other two. It has 50 groups and a total of 2015 images. Each group contains 26 to 52 images, with various poses and sizes, appearance variations and even more complex backgrounds. Because the images of *iCoseg* and *Cosal2015* have larger sizes than the ones of *MSRC*, different batch sizes and resolutions were used. The batch size is 3 and the resolution is 512×512 for *iCoseg* and *Cosal2015*, while the batch size is 5 and the resolution is 320×320 for *MSRC*.
Evaluation metrics. To evaluate the performance of co-saliency detection, we consider three metrics, *average precision* (AP), *F-measure* (F_β), and *structure measure* (S_α). AP is computed from the area under the Precision-Recall (PR) curve, which is produced by binarizing saliency maps with every integer threshold in the range of $[0, 255]$. *F-measure* denotes the harmonic mean of the precision and recall values obtained by a self-adaptive threshold $T = \mu + \sigma$, where μ and σ are respectively the mean and standard deviation of the saliency map. With the precision and recall values, the *F-measure* is computed by $F_\beta = \frac{(1+\beta^2)\times precision \times recall}{\beta^2 \times precision + recall}$, where $\beta^2 = 0.3$ to emphasize more on recall as suggested in previous work [11,12,56]. The *structure measure* (S_α) [57] is adopted to evaluate the spatial structure similarities of saliency maps based on both region-aware structural similarity S_r and object-aware structural similarity S_o, defined as $S_\alpha = \alpha * S_r + (1 - \alpha) * S_o$, where $\alpha = 0.5$ following [57].

4.2 Comparison with state-of-the-art methods

To have a thorough comparison with state-of-the-art methods, we divide them into four groups, i.e., the unsupervised saliency [20,22,23,24,25,42] and co-

Table 1. The performance of co-saliency detection on three benchmark datasets. SI and CS denote the single-image saliency and co-saliency methods, respectively. US and S indicate the unsupervised and supervised methods, respectively. The numbers in red and green respectively indicate the best and the second best results of the unsupervised co-saliency methods (CS+US), the group which the proposed method belongs to.

Method	Setting	MSRC			iCoseg			Cosal2015		
		AP	F_β	S_α	AP	F_β	S_α	AP	F_β	S_α
DIM [17]	CS+S	-	-	-	0.8773	0.7918	0.7583	-	-	-
UMLBF [13]	CS+S	0.9160	0.8410	-	-	-	-	0.8210	0.7120	-
CBCS [7]	CS+US	0.7034	0.5910	0.4801	0.7972	0.7408	0.6580	0.5863	0.5579	0.5439
SACS [31]	CS+US	0.8602	0.7877	0.7074	0.8400	0.7973	0.7523	0.7077	0.6923	0.6938
CSHS [8]	CS+US	0.7834	0.7118	0.6661	0.8454	0.7549	0.7502	0.6198	0.6181	0.5909
ESMG [32]	CS+US	0.6659	0.6245	0.5804	0.8347	0.7766	0.7677	0.5133	0.5114	0.5446
CSSCF [3]	CS+US	0.8604	0.8005	0.7383	0.8400	0.7811	0.7404	0.7075	0.6815	0.6710
CoDW [12]	CS+US	0.8435	0.7724	0.7129	0.8766	0.7985	0.7500	0.7438	0.7046	0.6473
SP-MIL [11]	CS+US	0.8974	0.8029	0.7687	0.8749	0.8143	0.7715	-	-	-
MVSRC [55]	CS+US	0.8530	0.7840	-	0.8680	0.8100	-	-	-	-
Ours	CS+US	0.9226	0.8404	0.7948	0.9112	0.8497	0.8200	0.8149	0.7580	0.7506
LEGS [26]	SI+S	0.8479	0.7701	0.6997	0.7924	0.7473	0.7529	0.7339	0.6926	0.7068
DCL [47]	SI+S	0.9065	0.8259	0.7742	0.9003	0.8444	0.8606	0.7815	0.7386	0.7591
DSS [28]	SI+S	0.8700	0.8313	0.7435	0.8802	0.8386	0.8483	0.7745	0.7509	0.7579
UCF [29]	SI+S	0.9217	0.8114	0.8175	0.9292	0.8261	0.8754	0.8081	0.7194	0.7790
Amulet [30]	SI+S	0.9219	0.8159	0.8162	0.9395	0.8381	0.8937	0.8201	0.7384	0.7856
GMR [20]	SI+US	0.8092	0.7460	0.6547	0.7990	0.7805	0.7068	0.6649	0.6605	0.6599
GP [22]	SI+US	0.8200	0.7422	0.6844	0.7821	0.7495	0.7198	0.6847	0.6576	0.6714
MB+ [23]	SI+US	0.8367	0.7817	0.7200	0.7868	0.7706	0.7272	0.6710	0.6689	0.6724
MST [24]	SI+US	0.8057	0.7491	0.6460	0.8019	0.7659	0.7292	0.7096	0.6669	0.6676
MILP [25]	SI+US	0.8334	0.7776	0.6871	0.8182	0.7883	0.7514	0.6797	0.6734	0.6752
SVFSal [42]	SI+US	0.8669	0.7934	0.7688	0.8376	0.8056	0.8271	0.7468	0.7120	0.7604

saliency [3,7,8,11,12,31,32,55] detection methods as well as supervised saliency [26,47,28,29,30,58] and co-saliency [13,17] detection methods. The overall performance statistics are compared in Table 1 and Fig. 3. Please note that all compared supervised single-image saliency detection methods are CNN-based. Among unsupervised single-image saliency methods, SVFSal [42] is CNN-based. When available, we used the publicly released source code with default parameters provided by the authors to reproduce the experimental results. For methods without releasing source code, we either evaluated metrics on their pre-generated co-saliency maps (SP-MIL [11], CoDW [12] and DIM [17]), or directly copied the numbers reported in their papers (UMLBF [13] and MVSRC [55]).

From Table 1, our method outperforms all methods with the same unsupervised co-saliency detection setting by a significant margin. Most approaches of this category take feature extraction and co-salient object detection as separating steps. Our approach excels them by performing these steps simultaneously and adopting CNN models. Comparing with the group of the supervised co-saliency method, UMLBF [13] and DIM [17], our method yields comparable or

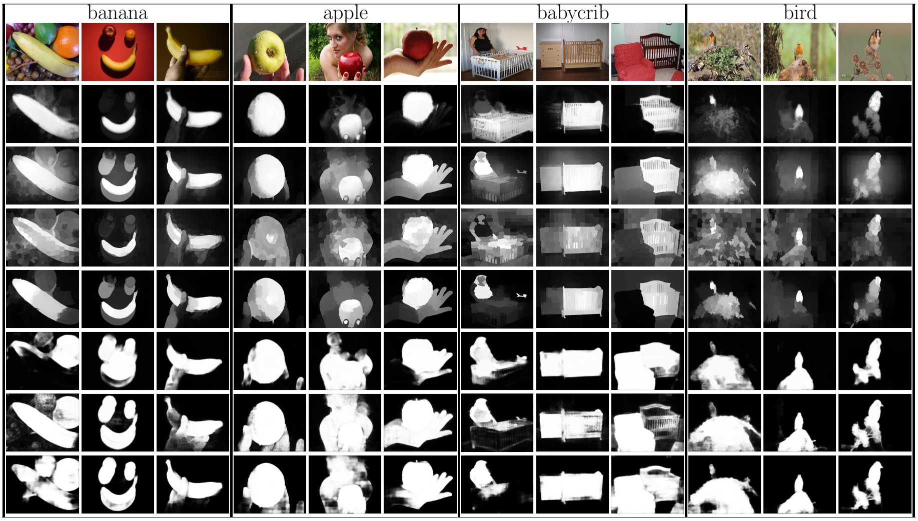
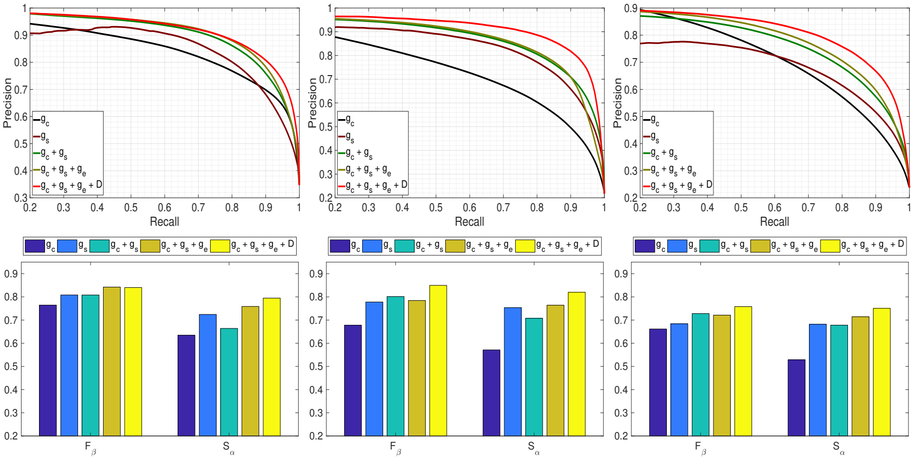


Fig. 4. Example saliency maps generated by our method and some state-of-the-art methods. From the top to the bottom, they are the given images, ours, `CSSCF` [3], `CoDW` [12], `MILP` [25], `SVFSal` [42], `UCF` [29] and `Amulet` [30].

even slightly better performance without expensive object annotations. Although both with the unsupervised setting, by taking advantage of additional information within an image set, our method clearly outperforms the group of unsupervised single-image saliency detection methods. It's worth mentioning that our method also outperforms the unsupervised CNN-based single-saliency method, `SVFSal` [42] that requires saliency proposal fusion for generating high-quality pseudo ground-truth as training data. In general, the supervised CNN-based single-image saliency methods perform the best among four groups of methods as they better utilize the object annotations. Even so, our method still outperforms many of the methods in this group by exploiting cross-image referencing and adaptive feature learning. From the PR curves in Fig. 3, the proposed method outperforms the state-of-the-arts by a large margin.

Fig. 4 shows example saliency maps produced by our method and some state-of-the-art methods, including unsupervised co-saliency detection methods (`CSSCF` [3], `CoDW` [12]), unsupervised single-image saliency methods (`MILP` [25] and `SVFSal` [42]), and supervised CNN-based methods (`UCF` [29] and `Amulet` [30]). Without referring to other images in the given image set, single-image saliency methods could detect the visually salient objects that do not repetitively appear in other images, such as the orange and the apple in the second image of the banana set or the woman in the first image of the babycrib set. Co-saliency detection methods perform better in this regard. The competing co-saliency methods, `CSSCF` [3] and `CoDW` [12], cannot perform well for images with low figure-ground discrepancies or highly-textured backgrounds, such as the second and third im-



MSRC iCoseg Cosal2015

Fig. 5. Ablation studies on three benchmarks. The top row plots the PR curves, while the bottom row shows the performance in F_β and S_α.

ages of the babycrib set or the first and second images of the bird set. The major drawback of their approaches is to treat feature extraction as a separate step. Thus, they cannot find the most discriminative features across images. Our method addresses the problem by performing adaptive feature learning and co-saliency detection jointly.

4.3 Ablation studies

We have performed ablation studies to investigate the contributions of individual components, g_c, g_s, g_e and DenseCRFs. Fig. 5 reports the results with different metrics. +D denotes the results refined by DenseCRFs. For both AP and F_β, the integration of g_c and g_s outperforms either alone. It is not the case for S_α measuring the structure of the detected objects. It will be explained later. Both self-paced learning and DenseCRFs further improve the results.

Fig. 6 gives example co-saliency maps for ablation studies. They demonstrate that g_c and g_s can be complementary to each other. Taking the butterfly set as an example, g_s highlights both butterflies and flowers in the first, third and fourth images. After integrating the co-occurrence information discovered by g_c, the flowers are mostly removed and lightened in g_c+g_s. As mentioned above, g_c+g_s could perform worse in terms of S_α. It is because g_c tends to have less certainty, particularly inside objects or ambiguous background regions, as illustrated in the second row of Fig. 6. Thus, g_c+g_s usually generates fuzzier maps than g_s alone. For example, the cattle have lower saliency values in g_c+g_s (the fourth row of Fig. 6) than g_s (the third row of Fig. 6). By propagating information from regions with high confidence, g_e improves the certainty of the results of g_c+g_s.

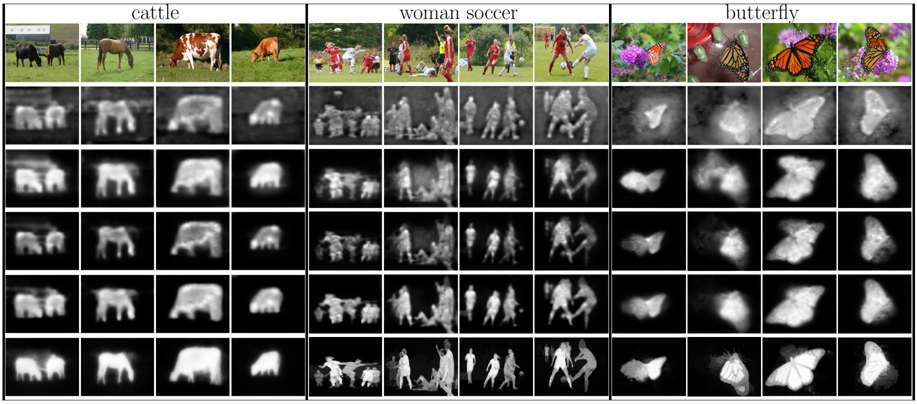


Fig. 6. Example co-saliency maps generated by combinations of different components. From the top to the bottom, they are the given images, g_c, g_s, g_c+g_s, g_c+g_s+g_e and g_c+g_s+g_e+D, respectively.

Although with less gain in AP and F_β, it brings large improvement in S_α since objects are more highlighted and the backgrounds are further lightened as shown in the fifth row of Fig. 6. Finally, the DenseCRF enhances spatial coherence and boundary preservation, thus improving both quantitative and qualitative results.

5 Conclusions

In this paper, we have presented an unsupervised method for co-saliency detection using CNNs. To the best of our knowledge, it is the first one to address this problem with an unsupervised CNN. Our method decomposes the problem into two sub-tasks, *single-image saliency detection* and *cross-image co-occurrence region discovery*, by modeling the corresponding novel losses: *single-image saliency (SIS) loss* and *co-occurrence (COOC) loss*. The graphical model is adopted to integrate these two losses with unary and pairwise terms corresponding to the SIS and COOC losses, respectively. By optimizing the energy function associated with the graph, two networks are learnt jointly. The quality of co-saliency maps is further improved by self-paced learning and postprocessing by DenseCRFs. Experiments on three challenging benchmarks show that the proposed method outperforms the state-of-the-art unsupervised methods. In the future, we plan to generalize our method to other applications, such as semantic correspondence [59], image co-localization [14] and object co-segmentation [60] that also require learning among multiple images.

Acknowledgments. This work was supported in part by Ministry of Science and Technology (MOST) under grant 105-2221-E-001-030-MY2 and MOST Joint Research Center for AI Technology and All Vista Healthcare under grant 107-2634-F-002-007. This work was partially supported by Awards #1547557 and #1553281 from the National Science Foundation.

References

1. Fu, H., Xu, D., Zhang, B., Lin, S., Ward, R.: Object-based multiple foreground video co-segmentation via multi-state selection graph. TIP (2015)
2. Fu, H., Xu, D., Lin, S., Liu, J.: Object-based rgbd image co-segmentation with mutex constraint. In: CVPR. (2015)
3. Jerripothula, K., Cai, J., Yuan, J.: Image co-segmentation via saliency co-fusion. TMM (2016)
4. Jerripothula, K., Cai, J., Yuan, J.: CATS: Co-saliency activated tracklet selection for video co-localization. In: ECCV. (2016)
5. Xue, J., Li, C., Zheng, N.: Proto-object based rate control for jpeg2000: An approach to content-based scalability. (2011)
6. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
7. Fu, H., Cao, X., Tu, Z.: Cluster-based co-saliency detection. TIP (2013)
8. Liu, Z., Zou, W., Li, L., Shen, L., Meur, O.L.: Co-saliency detection based on hierarchical segmentation. SPL (2014)
9. Tsai, C.C., Qian, X., Lin, Y.Y.: Segmentation guided local proposal fusion for co-saliency detection. In: ICME. (2017)
10. Tsai, C.C., Qian, X., Lin, Y.Y.: Image co-saliency detection via locally adaptive saliency map fusion. In: ICASSP. (2017)
11. Zhang, D., Meng, D., Han, J.: Co-saliency detection via a self-paced multiple-instance learning framework. TPAMI (2017)
12. Zhang, D., Han, J., Li, C., Wang, J., Li, X.: Detection of co-salient objects by looking deep and wide. IJCV (2016)
13. Han, J., Cheng, G., Li, Z., Zhang, D.: A unified metric learning-based framework for co-saliency detection. TCSVT (2017)
14. Wei, L., Zhao, S., Bourahla, O., Li, X., Wu, F.: Group-wise deep co-saliency detection. In: IJCAI. (2017)
15. Hsu, K.J., Lin, Y.Y., Chuang, Y.Y.: Augmented multiple instance regression for inferring object contours in bounding boxes. TIP (2014)
16. Hsu, K.J., Lin, Y.Y., Chuang, Y.Y.: Weakly supervised saliency detection with a category-driven map generator. In: BMVC. (2017)
17. Zhang, D., Han, J., Han, J., Shao, L.: Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. TNNLS (2016)
18. Winn, J., Criminisi, A., T. Minka, T..
19. Batra, D., Kowdle, A., Parikh, D., Luo, J., Chen, T.: iCoseg: Interactive co-segmentation with intelligent scribble guidance. In: CVPR. (2010)
20. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.M.: Saliency detection via graph-based manifold ranking. In: CVPR. (2013)
21. Huan, C.R., Chang, Y.J., Yang, Z.X., Lin, Y.Y.: Video saliency map detection by dominant camera motion removal. TCSVT (2014)
22. Jiang, P., Vasconcelos, N., Peng, J.: Generic promotion of diffusion-based salient object detection. In: ICCV. (2015)
23. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 fps. In: ICCV. (2015)
24. Tu, W.C., He, S., Yang, Q., Chien, S.Y.: Real-time salient object detection with a minimum spanning tree. In: CVPR. (2016)
25. Huang, F., Qi, J., Lu, H., Zhang, L., Ruan, X.: Salient object detection via multiple instance learning. TIP (2017)

26. Wang, L., Lu, H., Ruan, X., M.-M.-Yang: Deep networks for saliency detection via local estimation and global search. In: CVPR. (2015)
27. Liu, N., Han, J.: DHSNet: Deep hierarchical saliency network for salient object detection. In: CVPR. (2016)
28. Hou, Q., M.-M.-Cheng, Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: CVPR. (2017)
29. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: ICCV. (2017)
30. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: ICCV. (2017)
31. Cao, X., Tao, Z., Zhang, B., Fu, H., Feng, W.: Self-adaptively weighted co-saliency detection via rank constraint. TIP (2014)
32. Li, Y., Fu, K., Liu, Z., Yang, J.: Efficient saliency-model-guided visual co-saliency detection. SPL (2015)
33. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: CVPR. (2015)
34. Knobelreiter, P., Reinbacher, C., Shekhovtsov, A., Pock, T.: End-to-end training of hybrid cnn-crf models for stereo. In: CVPR. (2017)
35. Chandra, S., Usunier, N., Kokkinos, I.: Dense and low-rank gaussian crfs using deep embeddings. In: ICCV. (2017)
36. Shen, F., Gan, R., Yan, S., Zeng, G.: Semantic segmentation via structured patch prediction, context crf and guidance crfs. In: ICCV. (2017)
37. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineety, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: ICCV. (2015)
38. Wang, S., Fidler, S., Urtasun, R.: Proximal deep structured models. In: NIPS. (2016)
39. Hayder, Z., He, X., Salzmann, M.: Learning to co-generate object proposals with a deep structured network. In: CVPR. (2016)
40. Yuan, Z., Lu, T., Wu, Y.: Deep-dense conditional random fields for object co-segmentation. In: IJCAI. (2017)
41. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional models for semantic segmentation. In: CVPR. (2015)
42. Zhang, D., Han, J., Zhang, Y.: Supervision by fusion: Towards unsupervised learning of deep salient object detector. In: ICCV. (2017)
43. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
44. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
45. Krahenbuhl, P., Koltun, V.: Efficient inference in fully connected CRFs with gaussian edge potentials. In: NIPS. (2011)
46. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.M.: Saliency detection via absorbing markov chain. In: ICCV. (2013)
47. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: CVPR. (2016)
48. Sutton, C., McCallum, A.: Piecewise training for structured prediction. (2009)
49. Kingma, D., Ba, J.: ADAM: A method for stochastic optimization. In: ICLR. (2014)
50. Vedaldi, A., Lenc, K.: MatConvNet – convolutional neural networks for matlab. In: ACMMM. (2015)

51. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. (2016)
52. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: CVPR. (2009)
53. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. TPAMI (2012)
54. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A., Li, F.F.: Imagenet large scale visual recognition challenge. IJCV (2015)
55. Yao, X., Han, J., Zhang, D., Nie, F.: Revisiting co-saliency detection: A novel approach based on two-stage multiview spectral rotation co-clustering. TIP (2017)
56. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. TIP (2015)
57. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV. (2017)
58. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: CVPR. (2015)
59. Hsu, K.J., Lin, Y.Y., Chuang, Y.Y.: Robust image alignment with multiple feature descriptors and matching-guided neighborhoods. In: CVPR. (2015)
60. Hsu, K.J., Lin, Y.Y., Chuang, Y.Y.: Co-attention cnns for unsupervised object co-segmentation. In: IJCAI. (2018)