Co-saliency Detection via Mask-guided Fully Convolutional Networks with Multi-scale Label Smoothing

Kaihua Zhang[1], Tengpeng Li[1], Bo Liu[2]*, Qingshan Liu[1]
[1]B-DAT and CICAEET, Nanjing University of Information Science and Technology, Nanjing, China
[2]JD Digits, Mountain View, CA, USA
{zhkhua,kfliubo}@gmail.com

Abstract

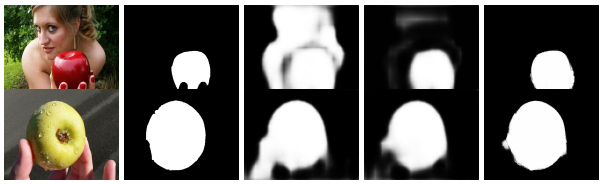
In image co-saliency detection problem, one critical issue is how to model the concurrent pattern of the co-salient parts, which appears both within each image and across all the relevant images. In this paper, we propose a hierarchical image co-saliency detection framework as a coarse to fine strategy to capture this pattern. We first propose a mask-guided fully convolutional network structure to generate the initial co-saliency detection result. The mask is used for background removal and it is learned from the high-level feature response maps of the pre-trained VGG-net output. We next propose a multi-scale label smoothing model to further refine the detection result. The proposed model jointly optimizes the label smoothness of pixels and superpixels. Experiment results on three popular image co-saliency detection benchmark datasets including iCoseg, MSRC and Cosal2015 demonstrate remarkable performance compared with the state-of-the-art methods.

1. Introduction

Image saliency detection mimics human vision system when looking at one image, through detecting the region that attracts human attention most. Given a group of images, the *image co-saliency* refers to common salient objects or regions in a group of relevant images. Discovering image co-saliency has been widely used as a pre-processing step in many applications, such as video/image foreground co-segmentaion [17, 16], object localization [42], surveillance video analysis [37] and image retrieval [38, 47].

One major theme in co-saliency detection research is image pixel or region feature representation. Traditional manually designed cues such as color histograms, Gabor filters, and SIFT descriptors, have been used in image co-saliency detection [8, 41, 15]. However, due to the limited feature discrimination, the performance of those methods are in general unsatisfactory. More recently, the deep learning methods have achieved superior accuracy because neural networks can generate more discriminative feature representations [51, 52]. The second major theme is to discover the repeated saliency across all images through a proper association strategy. The unsupervised learning methods look for the common objects or salient regions across images, by a series of models such as clustering [53, 48], multi-instance learning [54, 24] and graphical model [23]. With co-saliency label information, the supervised learning methods are promising to achieve more accurate results. The supervised single image saliency detection methods [22, 56, 55] can be used for co-saliency detection. However, they ignore the pattern concurrency of salient regions within all images, which is the essential characteristic of the co-saliency detection problem compared with single image saliency detection task. Several recent efforts are conducted to model the between-image pattern concurrency in the form of distance metric learning [20] and collaborative learning [45].

In this paper, the proposed hierarchical method is also to capture the salient pattern concurrency across images. Intuitively, image co-saliency can be derived from the single image saliency repeated in all images [8]. This motivates us to design a two-step framework. In the first step,



(a) (b) (c) (d) (e)

Figure 1. Illustration of different maps. (a) Input images; (b) Ground truth; (c) Single salient results with the proposed FCN without mask guidance; (d) Masked co-saliency maps; (e) Final refined results.

*Corresponding author. This work is supported by the NSFC (61876088, 61825601), the NSF of Jiangsu Province (BK20170040).

we generate the initial co-saliency detection results by a mask-guided fully convolutional network (FCN). The idea of mask-guided network has been successfully applied in various tasks such as object segmentation [12] and detection [57], because the mask encodes useful semantical and spatial information and hence the learned convolutional features are more discriminative with such a guidance. In our network we use the mask to remove the background information in convolutional feature learning, as shown in Figure 1 (d). The network co-saliency detection results of the mask-guided FCN are further refined by a multi-scale label smoothing model, leading to the final results in Figure 1 (e). The proposed method has the following technical novelties:

- We propose a mask-guided FCN structure for image co-saliency detection. The convolutional part of the proposed network has two channels and the mask is added at different convolutional layers in the two channels. The outputs of the two channels are merged and fed into the deconvolution layers to obtain the initial co-saliency detection results.
- To make the FCN targeted for the concurrent feature pattern learning, a mask is used as a guidance in the network. The mask is learned from the feature response map output of a pre-trained VGG net [40]. We design a learning objective that jointly maximizes the mask variance and encourages entries of the mask to be sparse. The designed learning objective is solved by an ADMM-type algorithm.
- A multi-scale label smoothing model is proposed to refine the detection results of the masked-guided FCN. The model considers the label smoothness of both image superpixels and pixels. The superpixel smoothness is modeled by manifold ranking and the pixel smoothness is modelled by a fully-connected CRF objective. An iterative optimization algorithm is designed to minimize the objective.

2. Related work

2.1. Single-image saliency detection

Exhaustively introducing the related works is beyond the scope of this paper, and some recent surveys about single-image saliency detection can be found in [10, 5]. Generally, the existing single-image salient detection methods can be categorized into unsupervised methods and supervised methods [10]. The unsupervised methods detect image saliency based on various prior knowledge as assumptions. In [9], image saliency is detected by region contrast that is evaluated by global contrast difference and spatial weighted coherence scores. Other priors such as frequency domain analysis [1], sparse learning [30, 39], background prior [61, 44] and compactness prior [60] are also considered in literature. Supervised methods, especially deep learning type models have demonstrated higher accuracy over unsupervised methods. In [33], image saliency is detected by predicting eye fixations. Extensive efforts have been conducted on network structure design, with multi-scale feature fusion network [28], hierarchical network structure [32] and skip-layer structures within the HED architecture [22] as representative work. A black box classifier is proposed in [11] for real-time image saliency detection.

2.2. Image co-saliency detection

Image co-saliency detection methods can be grouped into bottom-up, fusion based and learning based ones. Bottom-up methods score image regions based on feature priors to simulate visual attention [29, 15, 18]. In [15], three visual attention cues including contrast, spatial and corresponding are adopted. Background and foreground cues are used in a two-stage propagation framework in [18]. Fusion based methods ensemble the detection results of existing saliency or co-saliency methods. For example, Cao *et al.* obtain the self-adaptive weight via rank constraint to combine the co-saliency maps [7] and Huang *et al.* use multiscale superpixels to jointly detect salient object via low-rank analysis [25]. High-level semantic features from CNNs are extracted in [54, 52] to discover inter-image correspondence. Learning based methods have gained significant development in recent years, mainly because of the breakthrough of deep learning models [45, 58, 20, 23]. In [45], Wei *et al.* propose an end-to-end framework based on the FCN [36] to discover co-salient objects. In addition, an unsupervised CNN [23] is proposed to jointly optimize the co-saliency maps. A more comprehensive image co-saliency method survey can be found in [50].

3. Proposed approach

Figure 2 illustrates the overall framework of the proposed method. First, a mask is learned for each image that can highlight the co-salient regions and remove the background in the FCN learning (§ 3.1). Then, the mask-guided FCN detects the co-salient objects (§ 3.2). Finally, a multi-scale label smoothing model is proposed to refine the output of the mask-guided FCN (§ 3.3).

3.1. Mask learning

We leverage the VGG-16 network [40] pre-trained on the ImageNet image classification task [13] by removing its fully connected layers as generic feature extractor, and use the convolutional feature maps (CFMs) of the last layer (i.e., conv5-3) for masking learning. Given a group of images $\mathcal{I} = \{\boldsymbol{I}^n\}_{n=1}^N$ that contain co-salient objects of a category, for each image $\boldsymbol{I}^n \in \mathcal{I}$, the VGG network generates its feature representation $\boldsymbol{X}^n = [\boldsymbol{x}_1^n, \ldots, \boldsymbol{x}_k^n]^\top \in \mathbb{R}^{k \times d}$, where $\boldsymbol{x}_i^n \in \mathbb{R}^d$ is its i-th feature, k is the dimension of

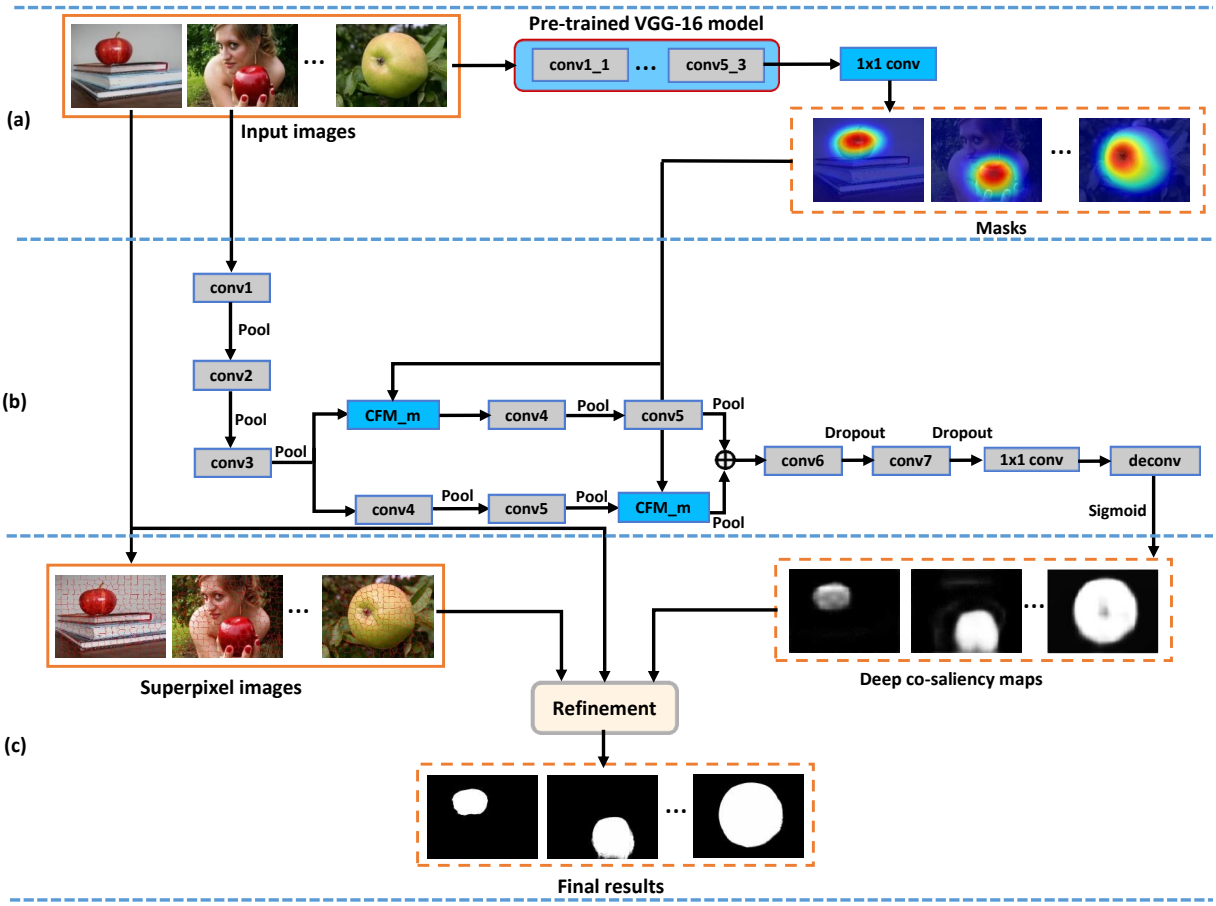


Figure 2. Proposed framework for co-saliency detection including three cascades: (a) Generate masks that can highlight the co-salient targets; (b) Develop a mask-guided FCN, wherein two branches of CFMs are masked by the masks (denoted by CFM_m), to generate deep co-saliency maps; (c) Refine the results via iteratively optimizing a multi-scale labeling smoothing model of pixels and superpixels.

the vectorized feature map of each channel and d denotes the number of channels. We aim to learn a mask function $m(\boldsymbol{x}_i^n) = \boldsymbol{w}^\top \boldsymbol{x}_i^n$ ($\boldsymbol{w}$ is equal to the top 1×1 convolutional filter in Figure 2) that classifies the feature $\boldsymbol{x}_i^n$ as foreground or background, yielding the mask responses

$$\boldsymbol{m}^n = \boldsymbol{X}^n \boldsymbol{w}, \tag{1}$$

where $\boldsymbol{m}^n = [m(\boldsymbol{x}_1^n), \ldots, m(\boldsymbol{x}_N^n)]^\top$. However, when learning the filters $\boldsymbol{w}$, supervised learning suffers from high annotation cost of labeling foreground masks as training data. Moreover, the learned model may not work well for unseen object categories in testing since it cannot be well generalized to unseen categories. To address this issue, motivated by principle component analysis (PCA) [4], we learn the filters $\boldsymbol{w}$ in an unsupervised learning manner. However, PCA selects the projection direction of maximum variance, which may lead to strong class overlap. As for our task, the direction selected by PCA may cause high classifier responses at all locations, leading to an inaccurate mask (refer to the middle row in Figure 3). To prevent this issue, we add an additional sparse regularization on the mask $\boldsymbol{m}^n$, leading to the following objective function

$$\begin{aligned} L(\boldsymbol{w}) = &- \sum_{n=1}^{N} \|\boldsymbol{X}^n \boldsymbol{w} - \bar{\boldsymbol{X}} \boldsymbol{w}\|_2^2 + \lambda_1 \|\boldsymbol{w}\|_2^2 + \lambda_2 \sum_{n=1}^{N} |\boldsymbol{m}^n|_1^1, \\ s.t.,\ &\boldsymbol{m}^n = \boldsymbol{X}^n \boldsymbol{w}, n = 1, \ldots, N, \end{aligned} \tag{2}$$

where $\bar{\boldsymbol{X}} = [\bar{\boldsymbol{x}}, \ldots, \bar{\boldsymbol{x}}]^\top$ is composed of the sample set mean $\bar{\boldsymbol{x}} = \frac{1}{Nk} \sum_{n=1}^{N} \sum_{i=1}^{k} \boldsymbol{x}_i^n$. The first term encourages the mask responses for all locations to have maximal variance, the second term leverages l_2 norm to control overfitting, and the last term penalizes high responses of $\boldsymbol{m}^n$ at background while making the first term encourage $\boldsymbol{m}^n$ to have high responses at foreground, thereby reducing class overlap compared to PCA that only resorts to maximal variance.

The objective $L(\boldsymbol{w})$ in (2) is convex with respect to the variables $\boldsymbol{w}$, and can be minimized to achieve the globally optimal solution via ADMM [6]. By introducing the step

parameter ρ, the Augmented Lagrangian form of (2) can be formulated as

$$L_\rho(\boldsymbol{w},\boldsymbol{m},\boldsymbol{s}) = g(\boldsymbol{w}) + \lambda_2 \sum_{n=1}^{N} |\boldsymbol{m}^n|_1^1 + \sum_{n=1}^{N} \boldsymbol{s}^{n\top}(\boldsymbol{X}^n\boldsymbol{w} - \boldsymbol{m}^n) + \frac{\rho}{2}\sum_{n=1}^{N} \|\boldsymbol{X}^n\boldsymbol{w} - \boldsymbol{m}^n\|_2^2, \tag{3}$$

where $g(\boldsymbol{w}) = -\sum_{n=1}^{N} \|\boldsymbol{X}^n\boldsymbol{w} - \bar{\boldsymbol{X}}\boldsymbol{w}\|_2^2 + \lambda_1\|\boldsymbol{w}\|_2^2$, and $\boldsymbol{s}^n$ is the Lagrange multiplier. By introducing $\boldsymbol{z}^n = \boldsymbol{s}^n/\rho$, (3) can be reformulated as

$$L_\rho(\boldsymbol{w},\boldsymbol{m},\boldsymbol{z}) = g(\boldsymbol{w}) + \lambda_2 \sum_{n=1}^{N} |\boldsymbol{m}^n|_1^1 + \frac{\rho}{2}\sum_{n=1}^{N} \|\boldsymbol{X}^n\boldsymbol{w} - \boldsymbol{m}^n + \boldsymbol{z}^n\|_2^2. \tag{4}$$

We then adopt the ADMM algorithm that alternatingly solves the following subproblems

$$\begin{cases} \boldsymbol{w}^{(i+1)} = \arg\min_{\boldsymbol{w}} g(\boldsymbol{w}) + \frac{\rho}{2}\sum_{n=1}^{N} \|\boldsymbol{X}^n\boldsymbol{w} - \boldsymbol{m}^n + \boldsymbol{z}^n\|_2^2, \\ \boldsymbol{m}^{n(i+1)} = \arg\min_{\boldsymbol{m}} \lambda_2|\boldsymbol{m}|_1^1 + \frac{\rho}{2}\|\boldsymbol{X}^n\boldsymbol{w} - \boldsymbol{m} + \boldsymbol{z}^n\|_2^2, \\ \boldsymbol{z}^{n(i+1)} = \boldsymbol{z}^{n(i)} + \boldsymbol{X}^n\boldsymbol{w}^{(i+1)} - \boldsymbol{m}^{n(i+1)}. \end{cases} \tag{5}$$

Update $\boldsymbol{w}$: Taking the derivative of the top equation in (5) be zero, we can obtain the closed-form solution for $\boldsymbol{w}$ as

$$\boldsymbol{w} = \left(-2\boldsymbol{S} + 2\lambda_1\boldsymbol{I} + \rho\sum_{n=1}^{N} \boldsymbol{X}^{n\top}\boldsymbol{X}^n\right)^{-1} \rho\sum_{n=1}^{N} \boldsymbol{X}^{n\top}(\boldsymbol{m}^n - \boldsymbol{z}^n), \tag{6}$$

where $\boldsymbol{S} = \sum_{n=1}^{N}(\boldsymbol{X}^n - \bar{\boldsymbol{X}})^\top(\boldsymbol{X}^n - \bar{\boldsymbol{X}})$.

Update $\boldsymbol{m}$: The middle equation in (5) can be readily solved by the soft thresholding method, and its closed-form solution is

$$\boldsymbol{m}^n(i) = \begin{cases} (\boldsymbol{X}^n\boldsymbol{w} + \boldsymbol{z}^n)(i) - \frac{\lambda_2}{\rho}, & if\ (\boldsymbol{X}^n\boldsymbol{w} + \boldsymbol{z}^n)(i) > \frac{\lambda_2}{\rho}, \\ 0, & if\ |(\boldsymbol{X}^n\boldsymbol{w} + \boldsymbol{z}^n)(i)| \le \frac{\lambda_2}{\rho}, \\ (\boldsymbol{X}^n\boldsymbol{w} + \boldsymbol{z}^n)(i) + \frac{\lambda_2}{\rho}, & else. \end{cases} \tag{7}$$

Note that the proposed objective function (2) is convex, and the ADMM algorithm has closed-form solution for each subproblem in (5). Therefore, it satisfies the Eckstein-Bertsekas condition [14] that is guaranteed to converge to global optimum. Moreover, we empirically find that the proposed ADMM can converge within 3 iterations on most images, and thus we set the iteration number to 3 for efficiency.

After obtaining the optimal mask $\boldsymbol{m}^n$, we reshape it to the size of the feature map of the conv5-3 layer, and then use bicubic interpolation to resize it to the desired size of each masked feature map, yielding a group of masks $\mathcal{M} = \{\boldsymbol{M}^n\}_{n=1}^N$ for the input image set $\mathcal{I}$.

Figure 3. Illustration of the masks generated by PCA (middle row) and our method (bottom row). PCA yields noisy background responses while our method can uniformly highlight the foreground common targets.

Figure 3 shows some examples of masks generated by our method and PCA. The input images consist of multiple targets such as cat, girl, helmet, baseball, *etc*, making it challenging to accurately localize the common baseballs. PCA suffers from noisy background, failing to uniformly localize the common targets. On the contrary, by introducing sparse representation to suppress the background responses, the proposed approach enables to better highlight the common targets than PCA.

3.2. Mask-guided FCN

The recently proposed SPP-Net [21] shows that CFMs encode both the semantics (by strengths of their activations) and spatial layouts of objects (by their positions), and hence SPP-Net masks the CFMs by a rectangular region, and directly pools the masked CFMs for recogniton. Afterwards, Dai *et al*. [12] further show that using a fine segment with an irregular shape to mask the CFMs enables to achieve top-level performance for semantic segmentation. Motivated by these works, we further propose the mask-guided FCN that masks out the CNN features of concurrent patterns across images for co-saliency detection.

We use the backbone architecture as the FCN proposed by [36], which consists of 16 convolutional layers interleaved by ReLU non-linearity, 5 max pooling and 2 dropout layers, and one deconvolutional layer. Given the training samples $\{\boldsymbol{I}^n, \boldsymbol{G}^n\}$, where $\boldsymbol{G}^n$ is the binary ground-truth mask of the input image $\boldsymbol{I}^n$. $\boldsymbol{I}^n$ is then fed forward through the FCN to generate the CFMs $\mathcal{F}^n = \{\boldsymbol{F}_1^n, \boldsymbol{F}_2^n, \ldots\}$. Then, we mask each feature map $\boldsymbol{F}_k^n$ in $\mathcal{F}^n$ with the learned mask $\boldsymbol{M}^n$ introduced by § 3.1

$$\hat{\boldsymbol{F}_k^n} = \boldsymbol{F}_k^n \odot \boldsymbol{M}^n, \tag{8}$$

where $\odot$ denotes the element-wise multiplication.

As shown in Figure 2 (b), after the conv3 layer, we achieve two branch features. One is only masking the pool3 layer while the other is only masking the pool5 layer. A-

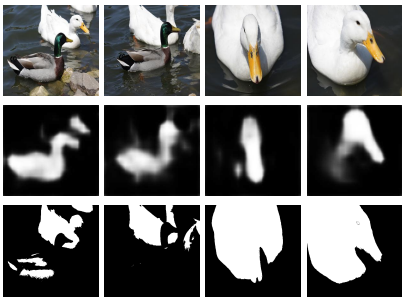


Figure 4. Top row: input images; middle row: co-saliency detection results of the mask-guided FCN, among which the gray geese with the same category are mistakenly detected; bottom row: through multi-scale label smoothing, the common salient targets are accurately detected while the distractors are suppressed.

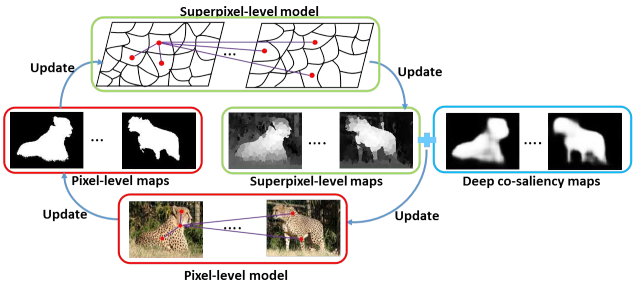


Figure 5. Alternatively optimizing the pixel-level and superpixel-level models.

mong them, the CFMs of pool3 layer mainly encode mid-level patterns such as triangular structures, red blobs, specific textures, *etc*., which are generic to describe all categories [19, 43]. Therefore, masking the pool3 layer captures salient regions of category-agnostic objects and can better generalize to unseen categories. In addition, the CFMs of pool5 layer encode rich high-level semantic information that is robust to significant appearance variations across the images, and masking these high-level features can further boost performance, which is verified by our ablative study in § 4.4. Then, to make full use of the complementary advantages of the two branch features, we fuse them by adding them together to feed forward through the following layers. Finally, we apply a 1×1 convolutional layer to compute the saliency map, and apply a deconvolutional layer to make the output map have the same size as the input image. The output layer is a sigmoid layer, which converts the saliency score into $[0, 1]$. For each input image $\boldsymbol{I}^n$, the FCN finally outputs a probability map $\boldsymbol{S}^n(\boldsymbol{\theta})$ with $\boldsymbol{\theta}$ denoting the network parameters, and is trained by minimizing the following loss function

$$L(\boldsymbol{\theta}) = \sum_n \|\boldsymbol{S}^n(\boldsymbol{\theta}) - \boldsymbol{G}^n\|_F^2 + \lambda_3 \|\boldsymbol{\theta}\|_2^2, \tag{9}$$

where $\|\cdot\|_F$ denotes the Frobenius norm, the last term denotes weight decay and $\lambda_3 > 0$ is a pre-defined trade-off parameter. Minimizing $L(\boldsymbol{\theta})$ via the stochastic gradient descent (SGD) method yields the optimal solution $\hat{\boldsymbol{\theta}}$, and the FCN outputs the deep co-saliency maps for image set $\{\boldsymbol{I}^n\}$ as $\mathcal{S} = \{\boldsymbol{S}^n(\hat{\boldsymbol{\theta}})\}$.

3.3. Refinement with multi-scale label smoothing

The presented mask-guided FCN uses high-level semantic features for co-saliency detection, which are not only robust to appearance changes, but also can well tell the salient objects from cluttered background. However, these features are not discriminative enough to different objects of the same category. As shown in Figure 4, there exist several geese with different colors, among which only the white ones are the co-salient targets, but our mask-guided FCN also mistakenly detects the gray goose distractors as co-salient targets (see middle row of Figure 4). To address this issue, we complement the semantic features with pixel-level and superpixel-level cues, and propose a multi-scale label smoothing model that alternatively optimizes the two models with these cues (refer to Figure 5).

Superpixel-level model: Given the input image set $\mathcal{I}$ and its corresponding deep co-saliency map set $\mathcal{S}$, we first use SLIC method [2] to separate each image $\boldsymbol{I}^n \in \mathcal{I}$ into a set of superpixels $\mathcal{Y}^n = \{y_i^n\}_{i=1}^{\bar{n}}$, where y_i^n denotes the mean of superpixel i in the LAB color space, and $\bar{n}$ is the number of superpixels. Then, we transform the deep co-saliency map $\boldsymbol{S}^n(\hat{\boldsymbol{\theta}}) \in \mathcal{S}$ into an initial indicator vector $\bar{\boldsymbol{l}}^n = \{\bar{l}_i^n\}_{i=1}^{\bar{n}}$ with $\bar{l}_i^n = 1$, if the mean of values in superpixel i is larger than the mean of all values in the deep co-saliency map n, otherwise, $\bar{l}_i^n = 0$. We then define a graph $G = (V, E)$, where the nodes $V = \{\mathcal{Y}^n\}_{n=1}^N$ are the superpixels on the image set $\mathcal{I}$ and the edges E are weighted by an affinity matrix $\boldsymbol{W} = [w_{ij}]_{\bar{N}\times\bar{N}}$ with the number of nodes $\bar{N} = \sum_{n=1}^N \bar{n}$, and

$$w_{ij} = \begin{cases} \exp^{-\frac{|y_i^m - y_j^n|}{\sigma^2}}, \; if \; m \neq n \; or \; i \in \mathcal{N}(j) \& m = n, \\ 0, \; if \; i \notin \mathcal{N}(j) \& m = n, \end{cases} \tag{10}$$

where $\mathcal{N}(j)$ is the 8-neighbors of the node j. Given G, its degree matrix $\boldsymbol{D} = \mathtt{diag}\{d_{11}, \ldots, d_{\bar{N}\bar{N}}\}$, where $d_{ii} = \sum_j w_{ij}$. Then, similar to the manifold ranking algorithm [59], the optimal ranking is computed by minimizing the following objective function

$$E_{sup}(\mathcal{R}|\mathcal{L}) = \sum_{i,j=1}^{\bar{N}} w_{ij}|r_i - r_j|^2 + \lambda_4 \sum_{i=1}^{\bar{N}} |r_i - l_i|^2, \tag{11}$$

where $\mathcal{R} = \{r_i\}_{i=1}^{\bar{N}}$ denotes the ranking scores of all nodes

in V and its corresponding given indicator set $\mathcal{L} = \{l_i\}_{i=1}^{\bar{N}}$, which are initialized by the indicators transformed by the deep co-saliency maps in $\mathcal{S}$.

Pixel-level model: Let $\mathcal{X} = \{x_i \in \{0, 1\}\}$ denote random variables that are the labels associated with all the pixels in the set $\mathcal{I}$, and given the superpixel ranking scores $\mathcal{R}$, we define an energy functional in a dense CRF form [27]

$$E_{pix}(\mathcal{X}|\mathcal{R}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j), \tag{12}$$

where the unary term is defined as

$$\psi_u(x_i) = -(\beta \boldsymbol{s} + (1-\beta)\boldsymbol{P}\boldsymbol{r})(i), \tag{13}$$

where $\boldsymbol{s}$ denotes the deep co-saliency map vector for all images in $\mathcal{I}$, $\boldsymbol{r} = [r_1, \ldots, r_{\bar{N}}]^\top$ denotes the ranking score vector for all superpixels, and $\boldsymbol{P}$ is a position indicator matrix which projects the superpixel scores to the corresponding pixel scores. The pairwise term is formulated as

$$\begin{aligned}\psi_p(x_i, x_j) = \Bigg[w_1 \exp\left(-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2}{2\theta_\alpha^2} - \frac{\|\boldsymbol{c}_i - \boldsymbol{c}_j\|^2}{2\theta_\beta^2}\right) \\ + w_2 \exp\left(-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2}{2\theta_\gamma^2}\right)\Bigg] \phi(x_i, x_j),\end{aligned} \tag{14}$$

where $\phi(x_i, x_j) = 1$ if $x_i \neq x_j$, and zero otherwise. $\boldsymbol{c}_i$ and $\boldsymbol{p}_i$ are RGB feature and position of pixel i respectively. Parameters w_1, w_2, θ_α, θ_β and θ_γ balance the importance of each Gaussian kernel. These parameters are set following [27].

Multi-scale model: The multi-scale model is formulated as

$$\min_{\mathcal{R},\mathcal{X},\mathcal{L}} E(\mathcal{R}, \mathcal{X}, \mathcal{L}) = E_{sup}(\mathcal{R}|\mathcal{L}) + E_{pix}(\mathcal{X}|\mathcal{R}), \tag{15}$$

where $E_{sup}(\mathcal{R}|\mathcal{L})$ is the superpixel-level model in (11) and $E_{pix}(\mathcal{X}|\mathcal{R})$ is the pixel-level model in (12).

As shown by Figure 5, (15) is alternatively optimized with respect to each variables:

Update $\mathcal{R}$: Fixing $\mathcal{X}$ and $\mathcal{L}$, we minimize $E(\mathcal{R}, \mathcal{X}, \mathcal{L})$ with respect to $\mathcal{R}$ by setting the derivative of $E(\mathcal{R}, \mathcal{X}, \mathcal{L})$ to zero:

$$\frac{\partial E(\mathcal{R}, \mathcal{L}, \mathcal{X})}{\partial \mathcal{R}} = 2(\boldsymbol{r} - \boldsymbol{S}\boldsymbol{r} + \lambda_4(\boldsymbol{r} - \boldsymbol{l})) - (1-\beta)\boldsymbol{P}^\top \boldsymbol{1} = 0, \tag{16}$$

where $\boldsymbol{S} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{W}\boldsymbol{D}^{-\frac{1}{2}}$, $\boldsymbol{l} = [l_1, \ldots, l_{\bar{N}}]^\top$ and $\boldsymbol{1}$ denotes all-ones vector. From (16), we have

$$\boldsymbol{r} = ((1+\lambda_4)\boldsymbol{I} - \boldsymbol{S})^{-1}\left(\lambda_4 \boldsymbol{l} + \frac{(1-\beta)\boldsymbol{P}^\top \boldsymbol{1}}{2}\right). \tag{17}$$

Update $\mathcal{X}$: Fixing $\mathcal{R}$, $\mathcal{L}$ and putting $\boldsymbol{r}$ in (17) into (13), minimizing $E(\mathcal{R}, \mathcal{L}, \mathcal{X})$ with respect to $\mathcal{X}$ is equal to minimizing $E_{pix}(\mathcal{X}|\mathcal{R})$ that is the objective of dense CRF, and hence we use the algorithm in [27] to efficiently obtain the optimal solution of $\mathcal{X}$.

Update $\mathcal{L}$: We denote the optimal solution of $\mathcal{X}$ as $\boldsymbol{x} = [x_1, \ldots]^\top$, and then use $\boldsymbol{x}$ to mask the deep co-saliency map vector $\boldsymbol{s}$, yielding $\hat{\boldsymbol{s}} = \boldsymbol{x} \odot \boldsymbol{s}$. Then, we reshape and split $\hat{\boldsymbol{s}}$ to generate a group of masked deep co-saliency maps $\{\hat{\boldsymbol{S}^n}\}_{n=1}^N$, which are used to update $\mathcal{L}$ by means of generating the initial indicator vector introduced by the section of **Superpixel-level model**.

4. Results and analysis

In this section, we first introduce implementation details of our algorithm (§ 4.1) and then introduce the benchmark datasets and evaluation metrics (§ 4.2). Afterwards, we show the qualitative and quantitative comparison results of our method with the state-of-the-arts (§ 4.3). Finally, we conduct ablative study to show the effectiveness of each component in the proposed method (§ 4.4).

4.1. Implementation details

We leverage MSRA-B [34] as the training set to train the mask-guided FCN, and all training images are resized to 500× 500 pixels. The parameters in our method are set by experience as $\lambda_1 = \lambda_3 = 0.001$, $\lambda_2 = 100$, $\lambda_4 = 1$, $\rho = 10$ and $\beta = 0.9$. We minimize the objective function (9) using mini-batch SGD with a batch size of 64, and momentum of 0.99. The learning rate is set to 1e-10, and the weight decay is set to 0.0005. The iteration number is set to 12,000. The CNNs are implemented in Caffe [26] and a Titan X GPU is used for acceleration.

4.2. Datasets and evaluation metrics

Datasets: We evaluate the proposed algorithm on three co-saliency benchmark datasets including iCoseg [3], MSRC [46] and Cosal2015 [52]. ICoseg has 38 groups of total 643 images, and each group has 4∼42 images. The images in iCoseg have similar objects with various poses and sizes. MSRC consists of 8 groups of total 240 images, and the co-salient object appearances in one group exhibit significant difference, increasing the difficulty for co-saliency detection. Cosal2015 is the largest dataset with 2015 images of 50 categories, and it is also the most challenging dataset since its images from one group suffer from the challenging factors such as different colors, sizes, poses, appearance variations and background clutters, *etc*.

Evaluation metrics: We compare our approach with the other state-of-the-art methods in terms of five criterions including the precision-recall (PR) curve, the receive operator characteristic (ROC) curve, the average precision (AP)

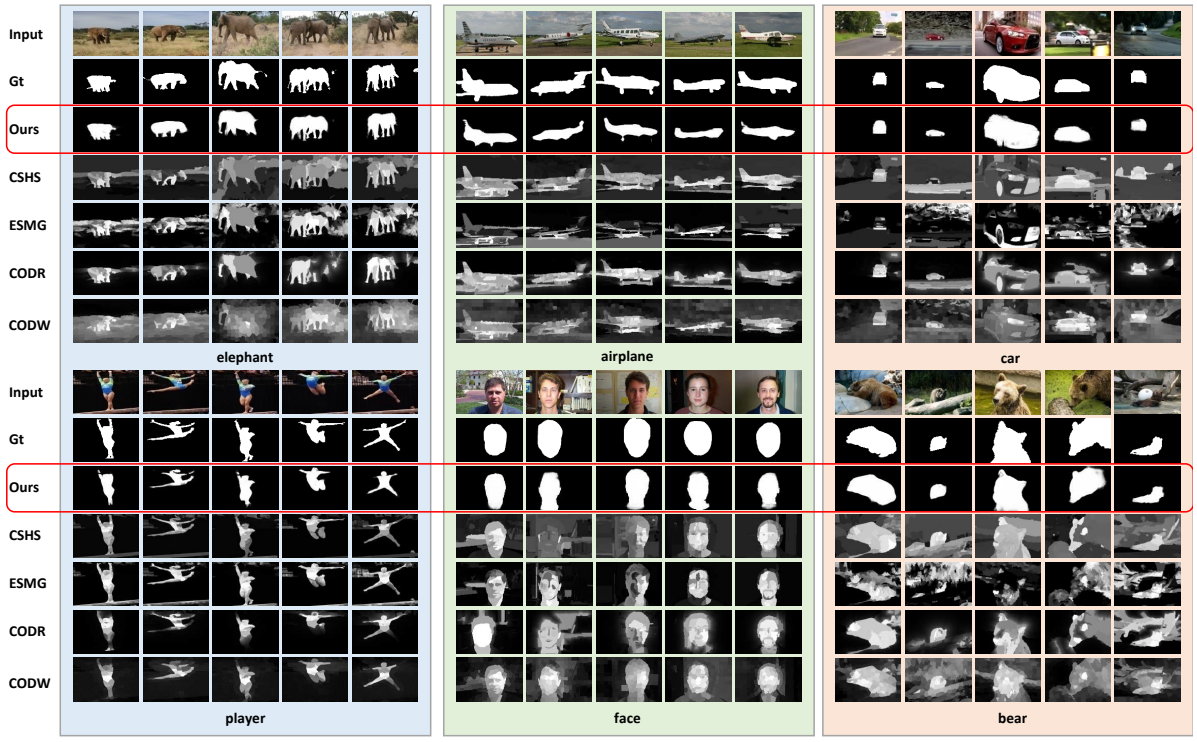


Figure 6. Example co-saliency maps generated by our method and the state-of-arts including CSHS [35], ESMG [31], CODR [49], and CODW [52].

score, the AUC score, and the F-measure score. F-measure is computed by an adaptive threshold $T = \mu + \epsilon$ to segment the co-saliency maps, where μ and ϵ represent the mean and standard deviation respectively. Here we define the F-measure based on the obtained average precision and recall

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}, \qquad (18)$$

where β^2 is set to 0.3 to enhance the importance of recall as suggested in [54, 52, 5].

4.3. Comparisons with state-of-the-art methods

We compare our algorithm with 6 state-of-art co-saliency detection methods including CBCS [15], CSHS [35], ESMG [31], CODR [49], CODW [52], and SPMIL [54]. For fair comparisons, we directly report the results released by the authors.

Qualitative comparison results: Figure 6 shows some qualitative comparison results. Among them, it is obvious that the proposed method can better extract the co-saliency regions in cases of different colors, poses, appearance variance and complex background. In Figure 6, the left two groups of images are from iCoseg. Among them, in group of elephant, the grass in the background shares the same color with the elephants, making the other compared methods fail to fully detect the elephants, but our method can achieve satisfying results because it uses the high-level semantic information that can well discriminate target from background. The middle groups are from MSRC that are mainly for semantic segmentation, and our method can also get favorable results under the interface of significant appearance variations. The right groups are from Cosal2015, and in both groups, our method can better detect the targets even they suffer from different complex scenes and significant appearance changes.

Quantitative comparison results: Figure 7 shows the PR and ROC curves of the compared methods on three benchmark datasets. The performance statistics are summarized in Table 1. From Figure 7, we can observe that our algorithm outperforms the other state-of-the-art methods in terms of both PR and ROC curves on all benchmarks. Especially on Cosal2015, the curves generated by the proposed method are much higher than the other methods. Furthermore, as listed by Table 1, CODW provides the best performance on the Cosal2015 among the state-of-the-arts, with AP score of 0.7437, AUC score of 0.9127, and F_β score of 0.7046. Meanwhile, our approach achieves AP score of 0.8527, AUC score of 0.9578, and F_β score of 0.8142, and significantly outperforms CODW by 10.9%, 4.51%, and 10.96%, respectively. These results verify the effectiveness of our mask-guided FCN with multi-scale label smoothing for co-saliency detection.

Table 1. Statistic comparisons of our method with the other state-of-the-arts. Here, -R, -P3 and -P5 represent our method in absence of refinement, pool3 masking and pool5 masking respectively. **Red**, **blue** and **green** bold fonts indicate the best, second best and third best performance respectively.

Dataset		CBCS [15]	ESMG [31]	CSHS [35]	CODR [49]	CODW [52]	SPMIL [54]	-R	-P3	-P5	Ours
iCoseg	AP	0.8021	0.8532	0.8397	0.8847	0.8766	0.8749	0.8395	**0.8959**	**0.9007**	**0.9057**
	AUC	0.9326	0.9559	0.9546	**0.9689**	0.9574	0.9649	0.9557	0.9688	**0.9705**	**0.9741**
	F_β	0.7432	0.7968	0.7540	0.8171	0.7985	0.8143	0.8110	**0.8457**	**0.8434**	**0.8553**
MSRC	AP	0.6998	0.6842	0.7868	0.8636	0.8435	**0.8974**	0.8686	0.8875	**0.8922**	**0.9096**
	AUC	0.8023	0.8228	0.8679	0.9167	0.9048	**0.9395**	0.9332	0.9351	**0.9408**	**0.9455**
	F_β	0.5986	0.6301	0.7184	0.7675	0.7724	0.8029	0.7971	**0.8142**	**0.8075**	**0.8250**
Cosal2015	AP	0.5972	0.5181	0.6212	0.6908	0.7437	-	**0.8372**	0.8297	**0.8335**	**0.8527**
	AUC	0.8166	0.7691	0.8512	0.9084	0.9127	-	**0.9499**	0.9341	**0.9537**	**0.9578**
	F_β	0.5644	0.5201	0.6225	0.6603	0.7046	-	0.7963	**0.7999**	**0.8076**	**0.8142**

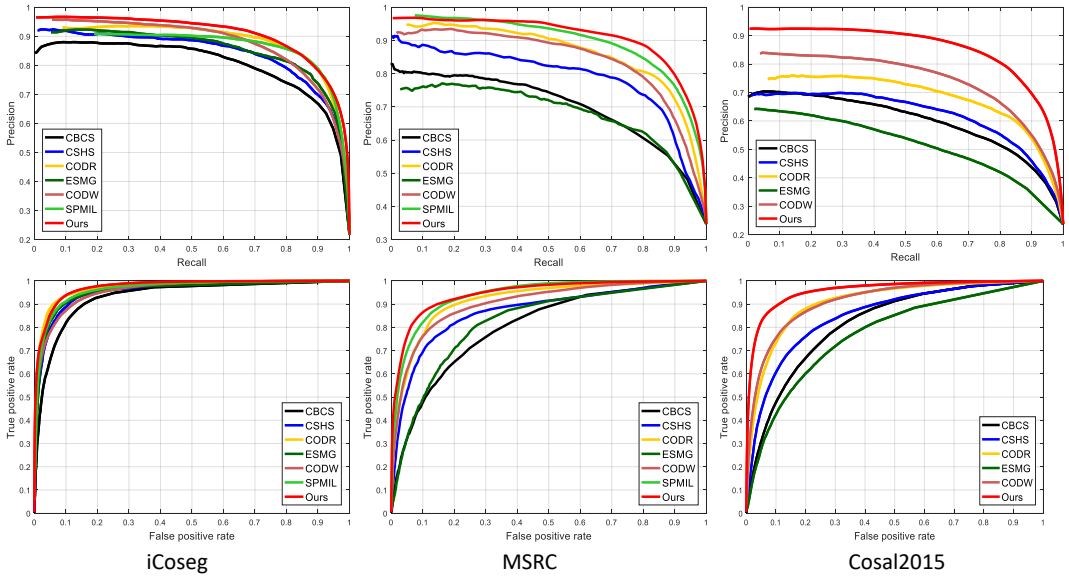


Figure 7. Comparisons with the state-of-art methods in terms of PR curves and ROC curves on three benchmark datasets

4.4. Ablative study

To further verify our main contributions, we compare different variants of our method including those without refinement (-R), pool3 masking (-P3), and pool5 masking (-P5) respectively.

From Table 1, we can observe that *without refinement,* both the AP and F_β scores drop obviously on all benchmark datasets. Especially on the iCoseg, the AP score drops significantly from 0.9057 to 0.8395 by 6.62% while the F_β score drops from 0.8553 to 0.8110 by 4.43%. This verifies the effectiveness of our multi-scale refinement strategy that can significantly boost the co-saliency detection accuracy. Furthermore, *without pool3 masking,* most scores of AP, AUC, and F_β are lower than those *without pool5 masking*. We think this is due to that the pool3 masking can capture salient regions of category-agnostic objects, leading to better generalization to unseen categories. In addition, *without pool5 masking,* all three scores drop, verifying the effectiveness of masking high-level semantic features that can further boost performance.

5. Conclusion

This paper has presented a masked-guided FCN for co-saliency detection including three cascades: in the first cascade, an unsupervised learning method has been proposed to learn co-salient object masks. In the second cascade, the learned masks are leveraged to mask the convolutional feature maps of the FCN for salient object extraction. In the third cascade, the output of the FCN has been further refined by iteratively optimizing a novel multi-scale label smoothing model. Extensive evaluations on three benchmarks have verified the effectiveness of our method.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, Sabine Süsstrunk, et al. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 2012.

[3] Dhruv Batra, Adarsh Kowdle, Devi Parikh, Jiebo Luo, and Tsuhan Chen. icoseg: Interactive co-segmentation with intelligent scribble guidance. In *CVPR*, 2010.

[4] Christopher Bishop. Pattern recognition and machine learning.

[5] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *TIP*, 2015.

[6] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 2011.

[7] Xiaochun Cao, Zhiqiang Tao, Bao Zhang, Huazhu Fu, and Wei Feng. Self-adaptively weighted co-saliency detection via rank constraint. *TIP*, 2014.

[8] Kai-Yueh Chang, Tyng-Luh Liu, and Shang-Hong Lai. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In *CVPR*.

[9] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 2015.

[10] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *TCSVT*, 2018.

[11] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *NIPS*, 2017.

[12] Jifeng Dai, Kaiming He, and Jian Sun. Convolutional feature masking for joint object and stuff segmentation. In *CVPR*, 2015.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[14] Jonathan Eckstein and Dimitri P Bertsekas. On the douglasłrachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 1992.

[15] Huazhu Fu, Xiaochun Cao, and Zhuowen Tu. Cluster-based co-saliency detection. *TIP*, 2013.

[16] Huazhu Fu, Dong Xu, Stephen Lin, and Jiang Liu. Object-based rgbd image co-segmentation with mutex constraint. In *CVPR*, 2015.

[17] Huazhu Fu, Dong Xu, Bao Zhang, and Stephen Lin. Object-based multiple foreground video co-segmentation. In *CVPR*, 2014.

[18] Chenjie Ge, Keren Fu, Fanghui Liu, Li Bai, and Jie Yang. Co-saliency detection via inter and intra saliency propagation. *SPIC*, 2016.

[19] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *CVPR*, 2014.

[20] Junwei Han, Gong Cheng, Zhenpeng Li, and Dingwen Zhang. A unified metric learning-based framework for co-saliency detection. *TCSVT*, 2017.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, 2014.

[22] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. In *CVPR*, 2017.

[23] Kuang-Jui Hsu, Chung-Chi Tsai, Yen-Yu Lin, Xiaoning Qian, and Yung-Yu Chuang. Unsupervised cnn-based co-saliency detection with graphical optimization. In *ECCV*, 2018.

[24] Fang Huang, Jinqing Qi, Huchuan Lu, Lihe Zhang, and Xiang Ruan. Salient object detection via multiple instance learning. *TIP*, 2017.

[25] Rui Huang, Wei Feng, and Jizhou Sun. Saliency and co-saliency detection by low-rank multiscale fusion. In *ICME*, 2015.

[26] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*, 2014.

[27] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011.

[28] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015.

[29] Hongliang Li and King Ngi Ngan. A co-saliency model of image pairs. *TIP*, 2011.

[30] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *CVPR*, 2013.

[31] Yijun Li, Keren Fu, Zhi Liu, and Jie Yang. Efficient saliency-model-guided visual co-saliency detection. *SPL*, 2015.

[32] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.

[33] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. In *CVPR*, 2015.

[34] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *TPAMI*, 2011.

[35] Zhi Liu, Wenbin Zou, Lina Li, Liquan Shen, and Olivier Le Meur. Co-saliency detection based on hierarchical segmentation. *SPL*, 2014.

[36] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[37] Yan Luo, Ming Jiang, Yongkang Wong, and Qi Zhao. Multi-camera saliency. *TPAMI*, 2015.

[38] Alex Papushoy and Adrian G Bors. Image retrieval based on query by saliency content. *DSP*, 2015.

[39] Houwen Peng, Bing Li, Haibin Ling, Weiming Hu, Weihua Xiong, and Stephen J Maybank. Salient object detection via structured matrix decomposition. *TPAMI*, 2017.

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ICLR*, 2015.

[41] Zhiyu Tan, Liang Wan, Wei Feng, and Chi-Man Pun. Image co-saliency detection by propagating superpixel affinities. In *ICASSP*, 2013.

[42] Kevin Tang, Armand Joulin, Li-Jia Li, and Li Fei-Fei. Co-localization in real-world images. In *CVPR*, 2014.

[43] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[44] Zilei Wang, Dao Xiang, Saihui Hou, and Feng Wu. Background-driven salient object detection. *TMM*, 2017.

[45] Lina Wei, Shanshan Zhao, Omar El Farouk Bourahla, Xi Li, and Fei Wu. Group-wise deep co-saliency detection. *IJCAI*, 2017.

[46] John Winn, Antonio Criminisi, and Thomas Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[47] Linjun Yang, Bo Geng, Yang Cai, Alan Hanjalic, and Xian-Sheng Hua. Object retrieval using visual query context. *TMM*, 2011.

[48] Xiwen Yao, Junwei Han, Dingwen Zhang, and Feiping Nie. Revisiting co-saliency detection: a novel approach based on two-stage multi-view spectral rotation co-clustering. *TIP*, 2017.

[49] Linwei Ye, Zhi Liu, Junhao Li, Wan-Lei Zhao, and Liquan Shen. Co-saliency detection via co-salient object discovery and recovery. *SPL*, 2015.

[50] Dingwen Zhang, Huazhu Fu, Junwei Han, Ali Borji, and Xuelong Li. A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. *TIST*, 2018.

[51] Dingwen Zhang, Junwei Han, Jungong Han, and Ling Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *TNNLS*, 2016.

[52] Dingwen Zhang, Junwei Han, Chao Li, and Jingdong Wang. Co-saliency detection via looking deep and wide. In *CVPR*, 2015.

[53] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. In *CVPR*, 2015.

[54] Dingwen Zhang, Deyu Meng, and Junwei Han. Co-saliency detection via a self-paced multiple-instance learning framework. *TPAMI*, 2017.

[55] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017.

[56] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, 2017.

[57] Xiangyun Zhao, Shuang Liang, and Yichen Wei. Pseudo mask augmented object detection. In *CVPR*, 2018.

[58] Xiaoju Zheng, Zheng-Jun Zha, and Liansheng Zhuang. A feature-adaptive semi-supervised framework for co-saliency detection. In *MM*, 2018.

[59] Denny Zhou, Olivier Bousquet, Thomas N Lal, Jason Weston, and Bernhard Schölkopf. Learning with local and global consistency. In *NIPS*, 2004.

[60] Li Zhou, Zhaohui Yang, Qing Yuan, Zongtan Zhou, and Dewen Hu. Salient region detection via integrating diffusion-based compactness and local contrast. *TIP*, 2015.

[61] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.