Hierarchical Co-salient Object Detection via Color Names

Jing Lou, Fenglei Xu, Qingyuan Xia, Mingwu Ren
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing 210094, China
Email: mingwuren@163.com

Wankou Yang
School of Automation
Southeast University
Nanjing 210096, China
Email: wkyang@seu.edu.cn

***Abstract*—In this paper, a bottom-up and data-driven model is introduced to detect co-salient objects from an image pair. Inspired by the biologically-plausible across-scale architecture, we propose a multi-layer fusion algorithm to extract conspicuous parts from an input image. At each layer, two existing saliency models are first combined to obtain an initial saliency map, which simultaneously codes for the color names based surrounded cue and the background measure based boundary connectivity. Then a global color cue with respect to color names is invoked to refine and fuse single-layer saliency results. Finally, we exploit the color names based distance metric to measure the color consistency between a pair of saliency maps and remove those non-co-salient regions. The proposed model can generate both saliency and co-saliency maps. Experimental results show that our model performs favorably against 14 saliency models and 6 co-saliency models on the Image Pair data set.**

***Keywords*-saliency, co-saliency, salient object detection, co-salient object detection, color names**

I. Introduction

Along with the rapid development of multimedia technology, saliency detection has become a hot topic in the field of computer vision. Numerous saliency models have been developed, aiming to reveal the biological visual mechanisms and explain the cognitive process of human beings. Generally speaking, saliency detection includes two different tasks: one is salient object detection [1]–[4], the other is eye fixation prediction [5]–[8]. The focus of this paper is bottom-up and data-driven saliency for detecting salient objects in images. A recent exhaustive review of salient object detection models can be found in [9].

Different from detecting salient objects in an individual image, the goal of co-saliency detection is to highlight the common and salient foreground regions from an image pair or a given image group [10]. As a new branch of visual saliency, modeling co-saliency has attracted much interest in the most recent years [11]–[16]. Essentially, co-salient object detection is still a figure-ground segmentation problem. The chief difference is that some distinctive features are required to distinguish non-co-salient objects.

The color feature based global contrast has been widely used in pure bottom-up computational saliency models. Different from the local center-surround contrast, global contrast aims to capture the uniqueness from the entire scene. In [17], the authors compute saliency maps by exploiting color names [18] and color histogram [19]. Inspired by this model, we also integrate color names into our framework to detect single-image saliency. As an effective low-level visual cue, the use of color names can facilitate co-saliency detection due to the high color consistency between two co-salient regions.

Moreover, a popular way of co-salient detection is to fuse the saliency results generated by multiple existing saliency models [13], [15]. The main advantage of the fusion-based methods is that they can be flexibly embedded with various existing saliency models [10]. However, when the adopted saliency models produce totally different saliency results, the performance of these fusion-based methods may decrease seriously.

In this paper, we also exploit a fusion technique to compute single-layer saliency maps. The proposed fusion and refinement algorithm has the ability of addressing the above issue. Furthermore, we incorporate the color names based contrast into co-salient object detection. The proposed model will be called "HCN" in the following sections, which can generate both saliency and co-saliency maps with higher accuracy.

II. Related Work

Recently, a simple and fast saliency model called the *Boolean Map based Saliency* (BMS) is proposed in [7]. The essence of BMS is a Gestalt principle based figure-ground segregation [20]. To overcome its limitation of only exploiting the surroundedness cue, Lou et al. [17] extend the BMS model to a *Color Name Space* (CNS), and invoke two global color cues to couple with the topological structure information of an input image. In CNS, the color name space is composed of eleven probabilistic channels, which are obtained by using the PLSA-bg color naming model [18]. However, the CNS model also uses a morphological algorithm [21] to mask out all the unsurrounded regions at the stage of attention map computation, so it fails when the salient regions are even slightly connected to the image borders.

In order to address the above issue, HCN incorporates the resultant maps of the *Robust Background Detection* (RBD) based model to generate single-layer saliency maps. The boundary prior is closely related to human perception mechanism and has been suggested to compute saliency by several existing models [22], [23]. In RBD, the authors first propose a *boundary connectivity* to quantify how heavily a region is connected to the image border, and integrate the background measure into a principled optimization

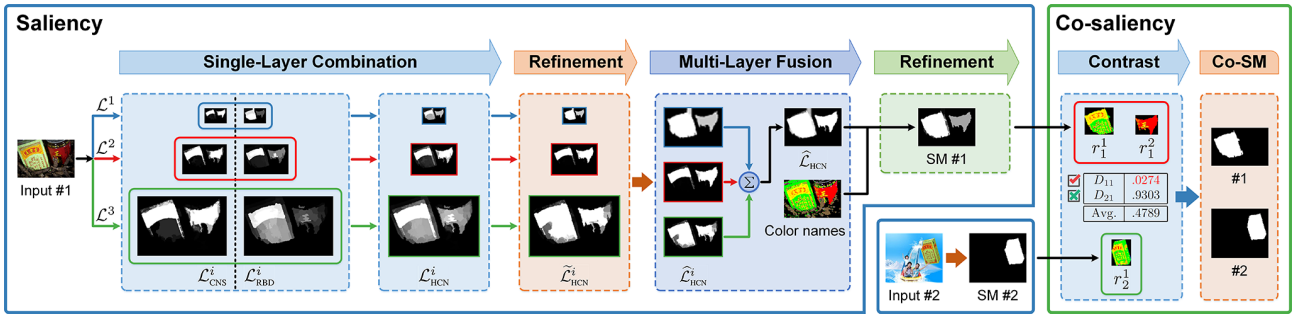


Figure 1: Pipeline of the proposed model. SM and Co-SM are abbreviations for saliency map and co-saliency map, respectively.

framework. This model is more robust to obtain uniform saliency maps, which can be used as a complementary technique to the surroundedness based saliency detection.

Moreover, many hierarchical and multi-scale saliency methods that model structure complexity have appeared in the literature [3], [24], [25]. In this paper, we also employ a multi-layer fusion mechanism to generate single-image saliency maps. We will demonstrate that a simple and bottom-up fusion approach is also effective and able to achieve promising performance improvements.

III. Color Names Based Hierarchical Model

The proposed model is as follows. First, three image layers are constructed for each input image. At each layer, we combine two individual saliency maps obtained by CNS [17] and RBD [26] separately. Then the three combination maps are fused into one single-image saliency map. Finally, we measure the color consistency of a pair of saliency maps and remove those non-co-salient regions to generate the final co-saliency maps. The pipeline is illustrated in Fig. 1.

A. Single-Layer Combination

In order to detect various sized objects in an image, we fix the layer number to 3 in our model. Each input image is first down-sampled to produce the first layer $\mathcal{L}^1$, which has a width of 100 pixels. For the second and the third layers ($\mathcal{L}^2$ and $\mathcal{L}^3$), we up-sample the input image and set the image widths to twice and four times the width of $\mathcal{L}^1$, i.e., 200 and 400 pixels, respectively. As shown in Fig. 1, such an architecture is well-suited to detecting salient regions at different scales, avoiding to obtain incorrect results by only using a single scale.

After all the three layers are produced, we generate two saliency maps $\mathcal{L}^i_{\text{CNS}}$ and $\mathcal{L}^i_{\text{RBD}}$ at the ith layer using CNS and RBD, respectively. The two maps of each layer are then combined to obtain a single-layer saliency map $\mathcal{L}^i_{\text{HCN}}$. The value at spatial coordinates (x, y) is defined as

$$\mathcal{L}^i_{\text{HCN}}(x,y) = \big(w_f \mathcal{L}^i_{\text{CNS}}(x,y) + (1-w_f)\mathcal{L}^i_{\text{RBD}}(x,y)\big) \times \big(\underbrace{2e^{-|\mathcal{L}^i_{\text{CNS}}(x,y) - \mathcal{L}^i_{\text{RBD}}(x,y)|} - 2e^{-1}}_{\text{consistency}} + 1\big), \quad (1)$$

where $|\cdot|$ indicates computing the absolute value, $w_f \in (0, 1)$ is a weighting coefficient.

The above equation has an intuitive explanation. It aims at mining useful information from two saliency maps $\mathcal{L}^i_{\text{CNS}}$ and $\mathcal{L}^i_{\text{RBD}}$ at each layer. We use the *consistency* term to encourage the two combined models to have similar saliency maps. At each point (x, y), $\mathcal{L}^i_{\text{HCN}}(x, y)$ will be assigned with a higher saliency value if the two maps have the same saliency, otherwise it will be zero. However, considering that the combined models may produce two totally different saliency maps, we add a factor of 1 to avoid obtaining a combination result without any salient region. An example of single-layer saliency combination is illustrated in Fig. 2, where the $\mathcal{L}^1$ layer of the original image *amira1* is shown in Fig. 2(a). Note that the proposed combination algorithm takes the advantages of two models and provides a more precise saliency result.

(a) $\mathcal{L}^1$ (b) $\mathcal{L}^1_{\text{CNS}}$ (c) $\mathcal{L}^1_{\text{RBD}}$ (d) $\mathcal{L}^1_{\text{HCN}}$

Figure 2: Illustration of single-layer combination.

Border Effect. In the testing data set, some of the input images have thin artificial borders, which may affect the output of CNS. To address this issue, we exploit an image's edge map to automatically determine the border width [26]. In our experiments, the border width is assumed to be fixed and no more than 15 pixels. The edge map is computed using the Canny method [27] with the edge density threshold of 0.7.[1] Then we trim each test image before the stage of layer generation. For the RBD model, we set the option *doFrameRemoving* to "false", and directly feed the three layers to its superpixel segmentation module. In the whole data set, sixteen images have thin image borders. After trimming them automatically, the average MAE [9] of the three layers decreases by 57.56%.

B. Single-Layer Refinement

The essence of salient object detection is a figure-ground segmentation problem, which aims at segmenting the salient

[1] We have noted that different versions of MATLAB have substantial influences on the edge detection results. In our experiments, both CNS, RBD, and HCN are all run in MATLAB R2017a (version 9.2).

Algorithm 1 refinement for the saliency map $\mathcal{L}^i_{\text{HCN}}$

Input: C^i and J^i
Output: refined saliency map $\widetilde{\mathcal{L}}^i_{\text{HCN}}$
1: $\widetilde{\mathcal{L}}^i_{\text{HCN}} = \texttt{RECONSTRUCT}(C^i,\ J^i)$
2: $\widetilde{\mathcal{L}}^i_{\text{HCN}} = (\widetilde{\mathcal{L}}^i_{\text{HCN}})^{\circ 2}$ ▷ background suppression
3: $\widetilde{\mathcal{L}}^i_{\text{HCN}} = \texttt{ADJUST}(\widetilde{\mathcal{L}}^i_{\text{HCN}},\ t_a)$ ▷ foreground highlighting
4: $\widetilde{\mathcal{L}}^i_{\text{HCN}} = \texttt{HOLE-FILL}(\widetilde{\mathcal{L}}^i_{\text{HCN}})$
5: $\widetilde{\mathcal{L}}^i_{\text{HCN}} = \texttt{NORMALIZE}(\widetilde{\mathcal{L}}^i_{\text{HCN}})$

foreground object from the background [9]. Under this definition, the ideal output of salient object detection should be a binary mask image, where each salient foreground object has the uniform value of 1. However, most previous saliency models have not been developed toward this goal. In this work, we propose a *color names based refinement* algorithm that directly aims for this goal. To use color names for the refinement of the obtained saliency map $\mathcal{L}^i_{\text{HCN}}$, we extend Eq. (1) by introducing a *color names based consistency* term as follows:

$$J^i = \underbrace{\left(W^i \circ (\mathcal{L}^i_{\text{CNS}})^{\circ 2}\right) \circ \left(W^i \circ (\mathcal{L}^i_{\text{RBD}})^{\circ 2}\right)}_{\text{color names based consistency}} + \left(\mathcal{L}^i_{\text{HCN}}\right)^{\circ 2}, \tag{2}$$

where W^i is a weighting matrix with the same dimension as $\mathcal{L}^i$, the two symbols $\circ$ and $\circ 2$ denote the Hadamard product and Hadamard power, respectively.[2]

In order to obtain the weighting matrix W^i, we convert $\mathcal{L}^i$ to a *color name image* and compute the probability f_j of the jth color name ($j = 1, 2, \cdots, 11$). Supposing the pixel $\mathcal{L}^i(x, y)$ belongs to the kth color name, the value of W^i at spatial coordinates (x, y) is defined as

$$W^i(x, y) = \sum_{j=1}^{11} f_j \|c_k - c_j\|_2^2, \tag{3}$$

where $\|\cdot\|_2$ denotes the ℓ_2-norm, c_k and c_j are the RGB color values of the kth and jth color names, respectively. For convenience, we define $\mathcal{L}^i_{\text{CNS}} \circ \mathcal{L}^i_{\text{RBD}}$ as C^i, and rewrite J^i as follows:

$$J^i = \left(W^i \circ C^i\right)^{\circ 2} + \left(\mathcal{L}^i_{\text{HCN}}\right)^{\circ 2}. \tag{4}$$

Finally, we sequentially perform a morphological reconstruction [28] and a post-processing step to obtain the refinement result $\widetilde{\mathcal{L}}^i_{\text{HCN}}$. The whole algorithm is summarized in Algorithm 1. To highlight foreground pixels, an adaptive threshold t_a is employed to linearly expand the gray-scale interval $[0, t_a]$ to the full $[0, 1]$ range. In our experiments, the value of t_a is set as the mean value of $\widetilde{\mathcal{L}}^i_{\text{HCN}}$.

The advantage of the refinement algorithm is threefold. First, the Hadamard power is used to suppress background pixels. Second, by exploiting a global contrast strategy with respect to color names, we further emphasize the common and salient pixels shared between the two combined saliency models. Third, the post-processing step is able to uniformly highlight salient foreground pixels, while facilitating the problem of multi-layer fusion. An example of single-layer refinement is illustrated in Fig. 3, where the refined saliency map is shown in Fig. 3(e). We can see it has comparable capability to highlight the salient region more uniformly than the combination map (Fig. 2(d)).

[2]For two matrices A, B of the same dimension, the Hadamard product for A with B is $(A \circ B)_{xy} = A_{xy}B_{xy}$, where x and y are spatial coordinates. The Hadamard power of A is defined as $(A^{\circ 2})_{xy} = A^2_{xy}$.

(a) CN (b) W^1 (c) C^1 (d) J^1 (e) $\widetilde{\mathcal{L}}^1_{\text{HCN}}$

Figure 3: Saliency refinement. (a) Color name image of Fig. 2(a).

C. Multi-Layer Fusion and Refinement

After three single-layer saliency maps are obtained, a multi-layer fusion step is then performed. Considering the possible diversity of saliency information among different layers, we propose a cross-layer consistency based fusion algorithm, rather than the use of linear averaging of them.

We resize each single-layer saliency map $\widetilde{\mathcal{L}}^i_{\text{HCN}}$ to the image size determined before the layer generation, such that three new maps have the same resolution. If the original input image has an artificial border, we add an outer frame with the same width as that of the previously trimmed image border, and set the saliency value of each pixel in it to zero. For each new map $\widehat{\mathcal{L}}^i_{\text{HCN}}$, we measure its bias to the average map $\overline{\mathcal{L}}_{\text{HCN}}$ of the three layers by using a cross-layer consistency metric d_i, which is defined as

$$d_i = \frac{1}{M \times N} \sum_{x=1}^{M} \sum_{y=1}^{N} \left| \overline{\mathcal{L}}_{\text{HCN}}(x, y) - \widehat{\mathcal{L}}^i_{\text{HCN}}(x, y) \right|, \tag{5}$$

where the average map $\overline{\mathcal{L}}_{\text{HCN}} = \frac{1}{3}\sum_{i=1}^{3} \widehat{\mathcal{L}}^i_{\text{HCN}}$.

By exploiting the d_i ($i = 1, 2, 3$) as the guided fusion weighting coefficient of the ith layer, we first perform a weighted linear fusion to produce a coarse single-image saliency map

$$\widehat{\mathcal{L}}_{\text{HCN}} = \sum_{i=1}^{3} \left(\exp(-\frac{d_i}{2\overline{d}}) \cdot \widehat{\mathcal{L}}^i_{\text{HCN}} \right), \tag{6}$$

where $\overline{d} = \sum_{i=1}^{3} d_i$. In this fashion, the multi-layer fusion result is more weighted toward the similar single-layer result compared with $\overline{\mathcal{L}}_{\text{HCN}}$. Then we refine $\widehat{\mathcal{L}}_{\text{HCN}}$ by performing similar steps as discussed in Section III-B, and obtain the final single-image saliency map S_{s} as follows:

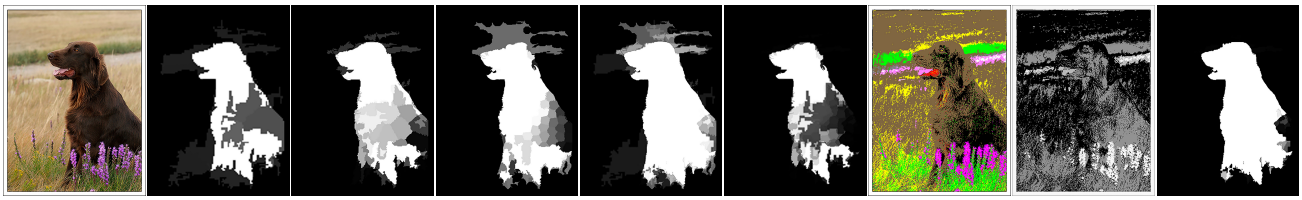
$$\widehat{C} = \widehat{\mathcal{L}}^1_{\text{HCN}} \circ \widehat{\mathcal{L}}^2_{\text{HCN}} \circ \widehat{\mathcal{L}}^3_{\text{HCN}}, \tag{7}$$

$$\widehat{J} = \left(\widehat{W} \circ \widehat{C}\right)^{\circ 3} + \left(\widehat{\mathcal{L}}_{\text{HCN}}\right)^{\circ 3}, \tag{8}$$

$$S_{\text{s}} = \left(\texttt{RECONSTRUCT}(\widehat{C},\ \widehat{J})\right)^{\circ 3}, \tag{9}$$

where $\widehat{W}$ is the weighting matrix of the input image.

An example of multi-layer saliency fusion and refinement is illustrated in Fig. 4, where the final single-image saliency map is shown in Fig. 4(i). Compared with Figs. 4(b)–4(d),



(a) original (b) $\widehat{\mathcal{L}}^1_{\text{HCN}}$ (c) $\widehat{\mathcal{L}}^2_{\text{HCN}}$ (d) $\widehat{\mathcal{L}}^3_{\text{HCN}}$ (e) $\widehat{\mathcal{L}}_{\text{HCN}}$ (f) $\widehat{C}$ (g) CN (h) $\widehat{W}$ (i) S_{s}

Figure 4: Illustration of multi-layer saliency fusion and refinement. (g) CN: Color name image of (a).

the proposed multi-layer fusion algorithm makes further improvement and achieves better accuracy.

D. Color Names Based Co-saliency Detection

To discovery the common and salient foreground objects in multiple images, a widely used cue is [12], [29]:

$$\text{Co-saliency} = \text{Saliency} \times \text{Repeatedness}\,. \tag{10}$$

That is to say, we can mine useful information from the similar pattern of a given image pair. In the previous stages, the contrast cue with respect to color names is exploited to perform single-layer refinement and multi-layer fusion. This cue will be used again to detect co-saliency.

For each single-image saliency map of an image pair, we segment it to a binary image using an adaptive threshold [2], which is twice the mean value of the saliency map. By exploiting each connected component r in the binary image, we extract the corresponding image region from the original input image, and convert it to a *color name image region*. The average color $A(r)$ of r is computed as $\sum_{j=1}^{11} f_j c_j$, where f_j and c_j are the probability and the RGB color value of the jth color name. The computation process of $A(r)$ is similar to that used in Section III-B. Then we use $A(r)$ as a contrast cue to measure the average color difference between two different regions as follows:

$$D_{ij} = \texttt{Diff}(r_1^i, r_2^j) = \left\| A(r_1^i) - A(r_2^j) \right\|_2^2, \tag{11}$$

where the subscript (1 or 2) of a region r denotes the corresponding image id in a given image pair, and the superscript denotes the region id in the binary image.

Then we compute the average value $\overline{D}$ of all the D_{ij}. The final co-saliency maps can be obtained by discarding those non-co-salient regions when their average color values are greater than $\overline{D}$, as has been presented in Fig. 1.

IV. Experiments

In this section, we evaluate the proposed model with 6 co-saliency models including CoIRS [11], CBCS [12], IPCS [13], CSHS [14], SACS [15], and IPTDIM [16] on the Image Pair data set [13]. Moreover, we compare it with 14 saliency models including BMS [7], CNS [17], DSR [30], GC [31], GMR [23], GU [31], HFT [8], HS [3], IRS [11], MC [32], PCA [33], RBD [26], RC [19], and TLLT [4]. The developed MATLAB code will be published in the project page: http://www.loujing.com/hcn-co-sod/.

A. Data set

The Image Pair data set [13] is designed for co-salient object detection research, where the object classes involve flowers, human faces, various vehicles and animals, etc. The authors collect 105 image pairs (i.e., 210 images) and provide accurate pixel-level annotations for an objective comparison of co-saliency detection. The whole image set includes 242 human labeled salient regions, but most of the images (191 images in total) contain only one salient region. There are 45 human labeled salient regions connected to the image borders. On average, the image resolution of this data set is around 131×105 pixels, while the ground truth salient part contains 23.87% image pixels.

B. Evaluation Metrics

To evaluate the effectiveness of the proposed model, we employ the standard and widely adopted Precision-Recall (PR) and F-measure (F_β). We use both 256 fixed thresholds (i.e., $T_f \in [0, 255]$) and an adaptive threshold (i.e., T_a proposed in [2]) to segment each resultant salient map, and compute the precision, recall, and F_β as follows:

$$\begin{gathered} \text{Precision} = \frac{|M \cap G|}{|M|}\,, \quad \text{Recall} = \frac{|M \cap G|}{|G|}\,, \\ F_\beta = \frac{(1+\beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}\,, \end{gathered} \tag{12}$$

where M is a binary segmentation, and G is the corresponding ground truth mask. The β^2 is also set to 0.3 for weighing precision more than recall [2]. Moreover, to quantitatively evaluate and compare different saliency/co-saliency models, we also report three evaluation metrics including AvgF, MaxF, and AdaptF as suggested in [17].

C. Parameter Analysis

The implementation of HCN includes 6 parameters where five of them are same as that used in CNS, i.e., sample step δ, kernel radii ω_c and ω_r, saturation ratio ϑ_r, and gamma ϑ_g. We use the same parameter ranges suggested by the authors. Considering that the two kernel radii have direct impacts on the performance, we fix the settings of δ, ϑ_r, and ϑ_g for all the three layers, but assign each layer with different values of ω_c and ω_r. In our experiments, we determine each optimal parameter value by finding the peak of the corresponding MaxF curve. The influences of the five parameters are shown in Figs. 5(a)–5(e). Moreover, HCN needs an additional parameter w_f to control single-layer saliency combination. We empirically

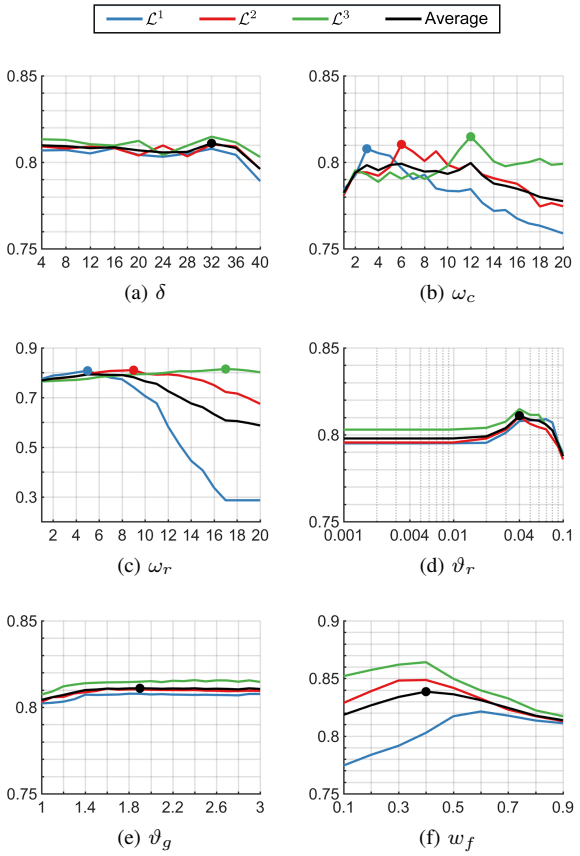


Figure 5: Parameter analysis of the proposed model.

set its initial value in the range $[0.1 : 0.1 : 0.9]$. The influence of w_f is shown in Fig. 5(f).

By and large, our model is not very sensitive to the parameters δ, ϑ_r, and ϑ_g. Based on the peak of the average MaxF curve of the three layers (see black curves), we set $\delta = 32$, $\vartheta_r = 0.04$, and $\vartheta_g = 1.9$. We use the same way to determine the optimal value of w_f, which is set as 0.4. For the other two parameters ω_c and ω_r, the MaxF curves clearly show that the proposed model performs well when smaller kernel radii are used for the $\mathcal{L}^1$ layer, while achieving better performance using larger kernel radii at the $\mathcal{L}^3$ layer. So we set the parameter values of ω_c and ω_r for the three layers as $\{(3, 5), (6, 9), (12, 17)\}$, respectively.

D. Evaluation of Saliency Fusion

We evaluate the proposed single-layer saliency combination and refinement algorithms, as well as our multi-layer fusion algorithm. The evaluation results are reported in Tables I and II. The best score under each evaluation metric is highlighted in red.

With respect to single-layer saliency combination and refinement, Table I shows that the single-layer combination result $\mathcal{L}^i_{\text{HCN}}$ achieves better accuracy in detecting salient objects than any of the combined models (i.e., CNS and RBD) at all the three layers. Although the color names based single-layer refinement result $\widetilde{\mathcal{L}}^i_{\text{HCN}}$ improves the performance slightly, we have demonstrated that it can facilitate the subsequent multi-layer saliency fusion.

Table I: MaxF statistics of single-layer combination

Model	$i = 1$	$i = 2$	$i = 3$	Average
$\mathcal{L}^i_{\text{CNS}}$	.8078	.8103	.8148	.8110
$\mathcal{L}^i_{\text{RBD}}$	.7597	.8288	.8526	.8137
$\mathcal{L}^i_{\text{HCN}}$	.8029	.8487	.8641	.8386
$\widetilde{\mathcal{L}}^i_{\text{HCN}}$	.8148	.8510	.8657	.8438

Table II: F_β statistics of multi-layer fusion

Layer	AvgF	MaxF	AdaptF	Average
$\widehat{\mathcal{L}}^1_{\text{HCN}}$	.7995	.8148	.8027	.8056
$\widehat{\mathcal{L}}^2_{\text{HCN}}$	.8391	.8510	.8435	.8445
$\widehat{\mathcal{L}}^3_{\text{HCN}}$	.8568	.8657	.8591	.8605
S_{S}	.8587	.8663	.8611	.8621

Table II shows that the multi-layer fusion result S_{S} further refines single-layer saliency maps and performs better than them. In addition, it can be seen that the evaluation results are more close to that of the third layer. This means we can achieve better performance in a larger scale when the original input images are somewhat small.

E. Comparisons with Other Models

Figure 6 shows the evaluation results of HCN compared with 14 saliency models and 6 co-saliency models on the Image Pair data set. We use the subscripts "s" and "co" to denote our saliency and co-saliency models, respectively. First, each PR curve is concentrated in a very narrow range when the fixed segmentation threshold $T_f \geqslant 1$. For HCN_{s} the standard deviations of the precision and recall are 0.0133 and 0.0127, while for HCN_{co} the two values are 0.0029 and 0.0031. Second, our F-measure curvs are more flat, this means the proposed model can facilitate the figure-ground segmentation. Third, when $T_f = 0$, all the models have the same precision, recall, and F_β values (precision 0.2387, recall 1, and F_β 0.2848), indicating that there are 23.87% image pixels belonging to the ground truth co-salient objects.

Some visual results are displayed in Figure 7. We can see that HCN generates more accurate co-saliency maps with uniformly highlighted foreground and well suppressed background. In addition, Tables III and IV show the F_β statistics of all the evaluated saliency and co-saliency models. The top three scores under each metric are highlighted in red, green, and blue, respectively. Overall, our model ranks the best in terms of all the three metrics. Although performing slightly worse than [16] with respect to the MaxF score (about 0.692‰), HCN_{co} outperforms it with large margins using the AvgF and AdaptF metrics.

V. Conclusion

By exploiting two existing saliency models and a color naming model, this paper presents a hierarchical co-saliency detection model for an image pair. We first demonstrate the

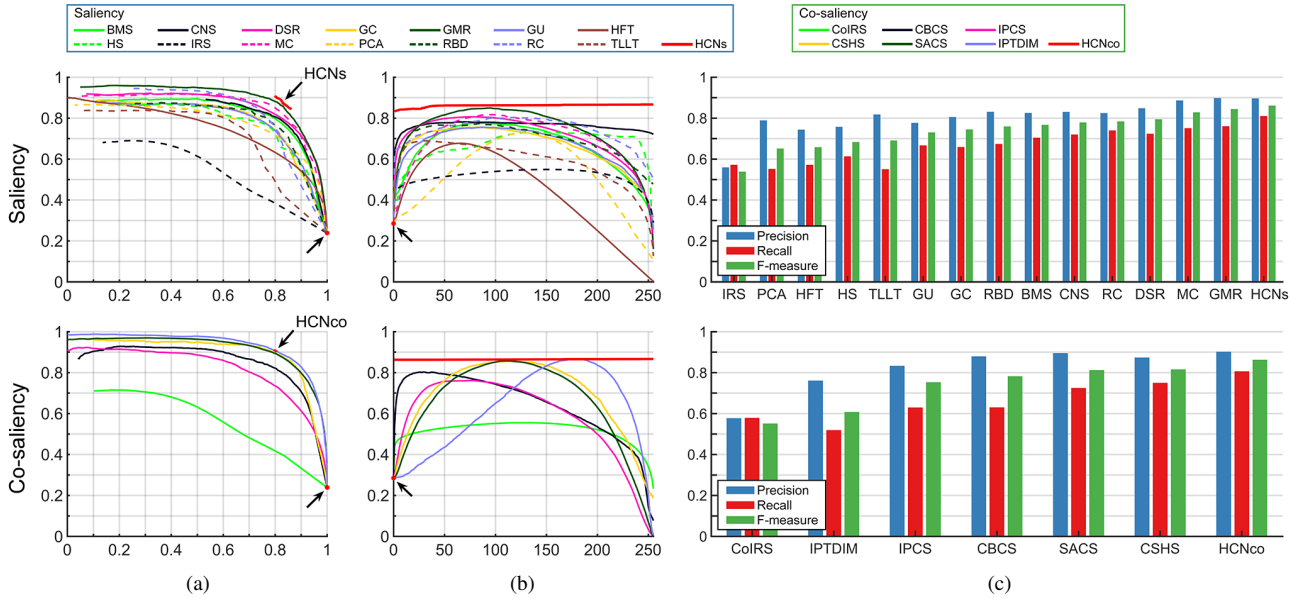


Figure 6: Performance of the proposed model compared with 14 saliency models (top) and 6 co-saliency models (bottom) on the Image Pair data set. (a) Precision (y-axis) and recall (x-axis) curves. (b) F-measure (y-axis) curves, where the x-axis denotes the fixed threshold $T_f \in [0, 255]$. (c) Precision-recall bars, sorted in ascending order of the F_β values obtained by adaptive thresholding.

simplicity and effectiveness of the proposed combination mechanism, which leverages both the surroundedness cue and the background measure that help in generating more accurate single-image saliency maps. Then a color names based cue is introduced to refine these maps and measure the color consistency of the common foreground regions. This paper is also a case study of the color attribute contrast based saliency/co-saliency detection. We show that the intra- and inter-saliency can benefit from the usage of color names. With regard to future work, we intend to incorporate more visual cues to improve performance, and extend the proposed co-saliency model to handle multiple images rather than an image pair.

Acknowledgment

The authors would like to thank Huan Wang, Andong Wang, Haiyang Zhang, and Wei Zhu for helpful discussions. They also thank Zun Li for providing some evaluation data. This work is supported by the National Natural Science Foundation of China (Nos. 61231014, 61403202, 61703209) and the China Postdoctoral Science Foundation (No. 2014M561654).

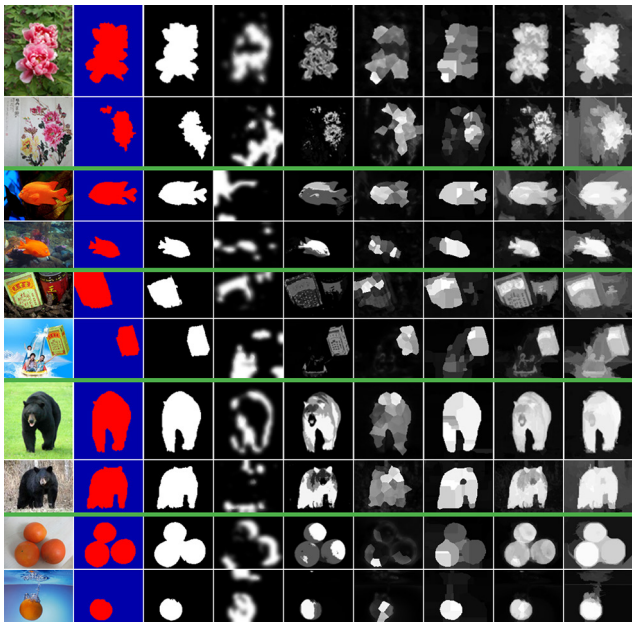


Figure 7: Visual comparison of co-saliency detection results. (a)-(b) Input images and ground truth masks [13]. Co-saliency maps produced using (c) the proposed model, (d) CoIRS [11], (e) CBCS [12], (f) IPCS [13], (g) CSHS [14], (h) SACS [15], and (i) IPTDIM [16], respectively.

References

[1] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Proc. Int. Conf. Comput. Vis. Syst.*, 2008, pp. 66–75.

[2] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[3] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1155–1162.

[4] C. Gong, D. Tao, W. Liu, S. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2531–2539.

[5] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[6] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 32: 1–20, 2008.

[7] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013,

Table III: F_β statistics of saliency models

#	Model	AvgF	MaxF	AdaptF	Average
1	BMS [7]	.6592	.7763	.7666	.7340
2	CNS [17]	.7612	.7817	.7787	.7738
3	DSR [30]	.7098	.8063	.7945	.7702
4	GC [31]	.6634	.7553	.7446	.7211
5	GMR [23]	.7391	.8493	.8442	.8109
6	GU [31]	.6642	.7553	.7303	.7166
7	HFT [8]	.4421	.6772	.6575	.5923
8	HS [3]	.6688	.7345	.6826	.6953
9	IRS [11]	.5149	.5491	.5380	.5340
10	MC [32]	.6933	.8171	.8280	.7795
11	PCA [33]	.5251	.7277	.6506	.6345
12	RBD [26]	.6950	.7727	.7587	.7422
13	RC [19]	.7383	.8031	.7840	.7751
14	TLLT [4]	.5885	.6892	.6908	.6561
15	HCN_s	.8587	.8663	.8611	.8621
Average		.6614	.7574	.7407	.7198

Table IV: F_β statistics of co-saliency models

#	Model	AvgF	MaxF	AdaptF	Average
1	CoIRS [11]	.5150	.5548	.5512	.5403
2	CBCS [12]	.6433	.8028	.7816	.7425
3	IPCS [13]	.5855	.7612	.7526	.6998
4	CSHS [14]	.6894	.8559	.8157	.7870
5	SACS [15]	.6499	.8571	.8114	.7728
6	IPTDIM [16]	.6161	.8671	.6070	.6968
7	HCN_{co}	.8620	.8665	.8625	.8637
Average		.6516	.7951	.7403	.7290

pp. 153–160.

[8] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, 2013.

[9] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[10] D. Zhang, H. Fu, J. Han, and F. Wu, "A review of co-saliency detection technique: Fundamentals, applications, and challenges," *arXiv:1604.07090v3 [cs.CV]*, pp. 1–18, 2017.

[11] Y.-L. Chen and C.-T. Hsu, "Implicit rank-sparsity decomposition: Applications to saliency/co-saliency detection," in *Proc. Int. Conf. Pattern Recognit.*, 2014, pp. 2305–2310.

[12] H. Fu, X. Cao, and Z. Tu, "Cluster-based co-saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 10, pp. 3766–3778, 2013.

[13] H. Li and K. N. Ngan, "A co-saliency model of image pairs," *IEEE Trans. Image Process.*, vol. 20, no. 12, pp. 3365–3375, 2011.

[14] Z. Liu, W. Zou, L. Li, L. Shen, and O. Le Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE Signal Process. Lett.*, vol. 21, no. 1, pp. 88–92, 2014.

[15] X. Cao, Z. Tao, B. Zhang, H. Fu, and W. Feng, "Self-adaptively weighted co-saliency detection via rank constraint," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 4175–4186, 2014.

[16] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 27, no. 6, pp. 1163–1176, 2016.

[17] J. Lou, H. Wang, L. Chen, Q. Xia, W. Zhu, and M. Ren, "Exploiting color name space for salient object detection," *arXiv:1703.08912 [cs.CV]*, pp. 1–13, 2017.

[18] J. van de Weijer, C. Schmid, and J. Verbeek, "Learning color names from real-world images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2007, pp. 1–8.

[19] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 409–416.

[20] E. Rubin, "Figure and ground," in *Readings in Perception*, 1958, pp. 194–203.

[21] P. Soille, *Morphological Image Analysis: Principles and Applications*. Springer-Verlag, 1999.

[22] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.

[23] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 3166–3173.

[24] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, 1998.

[25] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 2376–2383.

[26] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2814–2821.

[27] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–698, 1986.

[28] L. Vincent, "Morphological grayscale reconstruction in image analysis: Applications and efficient algorithms," *IEEE Trans. Image Process.*, vol. 2, no. 2, pp. 176–201, 1993.

[29] K.-Y. Chang, T.-L. Liu, and S.-H. Lai, "From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 2129–2136.

[30] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2976–2983.

[31] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1529–1536.

[32] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 1665–1672.

[33] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1139–1146.