CO-SALIENCY DETECTION VIA HIERARCHICAL CONSISTENCY MEASURE

Yonghua Zhang[1,3], *Liang Li*[1,3], *Runmin Cong*[2], *Xiaojie Guo*[1,3], *Hui Xu*[1,3], *Jiawan Zhang*[1,3,*]

[1]School of Computer Software, Tianjin University, Tianjin, China
[2]School of Electrical and Information Engineering, Tianjin University, Tianjin, China
[3]Key Research Center for Surface Monitoring and Analysis of Cultural Relics, SACH, China
{zhangyonghua, liangli, rmcong, huixu, jwzhang}@tju.edu.cn xj.max.guo@gmail.com

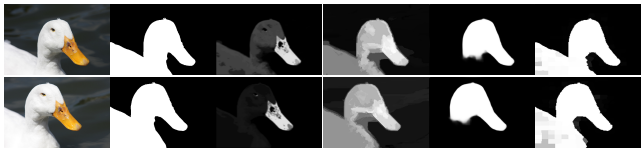
ABSTRACT

Co-saliency detection is a newly emerging research topic in multimedia and computer vision, the goal of which is to extract common salient objects from multiple images. Effectively seeking the global consistency among multiple images is critical to the performance. To achieve the goal, this paper designs a novel model with consideration of a hierarchical consistency measure. Different from most existing co-saliency methods that only exploit common features (such as color and texture), this paper further utilizes the shape of object as another cue to evaluate the consistency among common salient objects. More specifically, for each involved image, an intra-image saliency map is firstly generated via a single image saliency detection algorithm. Having the intra-image map constructed, the consistency metrics at object level and superpixel level are designed to measure the corresponding relationship among multiple images and obtain the inter saliency result by considering multiple visual attention features and multiple constrains. Finally, the intra-image and inter-image saliency maps are fused to produce the final map. Experiments on benchmark datasets are conducted to demonstrate the effectiveness of our method, and reveal its advances over other state-of-the-art alternatives.

Index Terms— Co-saliency detection, shape attribute, multi-feature similarity, hierarchical consistency measure

1. INTRODUCTION

Saliency detection is to discover the most visually salient objects from an image. It has been widely applied as the first step for a variety of multimedia and computer vision tasks, such as image segmentation [1], visual tracking [2], and image compression [3]. Over past years, the image co-saliency detection has been an emerging and rapidly growing research issue, which aims to capture common and salient objects or regions from a group of images [4]. As the extension of the traditional single image saliency detection, the co-saliency detection possesses an additional and significant property, *i.e.* all of co-salient objects should be similar in appearance. Intuitively and factually, due to this property, co-saliency detection is more useful in many tasks, for instance object co-segmentation and video foreground detection [5], object co-localization [6], to name just a few. Different from single saliency detection methods, which only rely on the contrast or uniqueness to compute the saliency map in an individual image, co-saliency detection also leverages the consistency of co-salient objects.



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

Fig. 1: Visual comparison of co-saliency maps. (a) and (b) are inputs and corresponding ground-truth in the iCoseg dataset. (c) and (d) show co-saliency maps by [5] and [7], respectively. (e) gives the single saliency maps generated by [8]. (f) is our final results.

To model the inter-image correspondence among images, many different methods have been designed. Fu *et al.* [5] formulated the inter-image relationship as a clustering process. Zhang *et al.* [9] captured the inter-image constraint through a Bayesian framework. In this paper, we formulate the inter-image constraint as a consistency matching problem. Two hierarchical consistency measures are proposed to evaluate the object-level similarity and superpixel-level similarity, respectively. The object-level consistency metric attempts to discover some reliable salient proposals based on multi-feature matching. The superpixel-level consistency metric is used to further refine the proposals and generate the inter saliency map.

Co-salient objects often show similar appearance, such as color and texture. However, most of existing co-saliency detection algorithms ignore an important appearance attribute, *i.e.* shape, when exploring the inter-image constraints. As shown in Fig.1(b), the co-salient objects have a similar shape. The shape feature owns powerful discrimination capability

* Corresponding author. This work was supported by National Social Science Foundation under Grant 15XMZ057, MSRA CCNP 2016, National Natural Science Foundation of China under Grants 61772512 and 61602338.

978-1-5386-1737-3/18/$31.00 ©2018 IEEE

that is relatively stable to changes in lighting conditions. Therefore, we introduce the shape attribute into the object-level consistency metric to augment the evaluation of feature similarity.

In summary, the main contributions of the proposed method are as follows: (1) The hierarchical consistency measures are proposed to capture the inter-image correspondence at different scales; (2) The shape attribute is introduced into the consistency metric to augment the evaluation of feature similarity; (3) Extensive experimental evaluations are carried out to show that the proposed method achieves a superior performance, outperforming the current state-of-the-art approaches.

2. RELATED WORK

The goal of co-saliency detection is to discover the common and salient objects from a given image group. The existing co-saliency detection methods [4, 5, 9–12]can be roughly divided into three categories: bottom-up methods, fusion-based methods and learning-based methods.

Early attempts usually employ some low-level features to generate intra saliency and inter saliency. Li and Ngan [10] constructed a co-multilayer graph to compute the inter-image consistency and combined the single-image saliency map to generate the pair-wise co-saliency maps. Fu *et al.* [5] proposed a cluster-based approach, taking into account three visual attention cues to measure the cluster-level co-saliency. Liu *et al.* [13] provided a hierarchical segmentation based co-saliency model, which integrates the global similarity, intra-saliency and object prior to generate the co-saliency map.

Fusion-based methods aim to mine the common information from a set of single image saliency maps. Cao *et al.* [4] explored low-rank constraint to obtain the self-adaptive weights. Based on these weights, multiple saliency maps are fused to produce the co-saliency map. Huang *et al.* [14] constructed a multiscale superpixel pyramid and jointed low-rank analysis to obtain the fused saliency map, and then introduced a GMM-based co-saliency prior to generate the co-saliency maps. Tsai *et al.* [15] proposed a segmentation guided locally adaptive proposal fusion for co-saliency detection. They formulated co-saliency and co-segmentation as an energy optimization problem over a graph.

Recently, learning-based methods have become popular due to their promising performance. Zhang *et al.* [9] utilized the high-level semantic features extracted from deep layers of Convolutional Neural Network (CNN) to represent the region properties. Then, the co-saliency scores for object proposal windows are calculated via a Bayesian formulation. In [11], a self-paced multiple-instance learning framework is proposed for co-saliency detection. The designed self-learning scheme can iteratively estimate the appearance of co-salient objects and refine the generated co-saliency maps.

3. OUR METHOD

In this paper, we propose a simple yet effective co-saliency detection method. First, the intra-image saliency map for each image is generated by single image saliency detection method. Then, a hierarchical process is designed to capture the corresponding relationship among multiple images. Specially, the object level consistency score is computed to generate some reliable salient proposals by using multi-feature matching algorithms. Based on the obtained results and multiple constraints, the inter-image saliency map is computed on the superpixel level. Finally, the intra-image and inter-image saliency maps are fused by linear weights to generate the final co-saliency maps.

3.1. Intra-image saliency map

Given N RGB images $\{I^i\}_{i=1}^N$, we apply one of the existing state-of-the-art single image saliency detection algorithm to obtain the intra-image saliency map for each image, which is represented as $\{S_{intra}^i\}_{i=1}^N$. The more accurate the intra-image saliency map is, the better the final co-saliency map achieves. Here, we choose the DSS method [8] due to its superior performance for single image saliency detection.

3.2. Inter-image saliency map

In addition to the intra-image representation, the inter-image constraints should be introduced into co-saliency detection. In this paper, the inter-image correspondence is formulated as the consistency constraint, which includes the object-level consistency metric and superpixel-level consistency metric. The object-level consistency score is computed to generate some reliable salient proposals based on multi-feature matching. The superpixel-level consistency score is calculated by considering multiple constrains to generate the inter-image saliency map.

3.2.1. Object-level consistency metric

The common objects typically should have similar appearances, such as shape, color and texture. Under the circumstances, we can utilize multi-feature matching metrics to measure the object-level similarity and discover some positive, common and salient proposals from the intra-image saliency maps. Through the threshold segmentation and region filtering, some objects are determined as $O = \{o_j^i\}_{j=1}^{M_i}$, where M_i is the number of objects in an image. In order to evaluate the object similarity, three types of visual features, *i.e.*, shape, color and texture, are used to describe the appearance properties.

Shape similarity. The canny edge detection algorithm is utilized to obtain object boundaries from the binary saliency masks. Then, Shape Context method [16] is introduced to measure the shape similarity among multiple objects. For

each object, Shape Context method first samples the contour with roughly uniform spacing and obtains hundreds of points as representative of its shape. Then, the shape context of each point is constructed by the histogram of the relative coordinates of the remaining points. The cost for matching points is computed by the chi-square distance test. Regularized thin-plate splines provide the transformation maps for aligning the shapes. Finally, a sum of the matching errors between corresponding points and the magnitude of the aligning transform is introduced to evaluate the dissimilarity between two shapes. The shape similarity s^{mn} between two objects is computed as:

$$s^{mn} = \exp\left(-\frac{C^{mn} + A^{mn}}{\sigma_s^2}\right), \tag{1}$$

where $m, n = \{1, 2, ..., M\}$, $M = \sum_{i=1}^{N} M_i$ is the total number of objects. Further, C^{mn} is the shape context matching cost, and A^{mn} is the affine transformation cost. Besides, σ_s is a non-negative parameter controlling the transition bandwidth. More details can be found in [16].

Color similarity. To evaluate the color similarity between two objects, we extract two types of color features for each object, *e.g.*, the average color value in three color spaces (*i.e.* RGB, CIELAB, and HSV), and the color histogram in RGB and HSV color spaces. The color similarity between two objects is defined as:

$$c^{mn} = \exp\left(-\frac{\| \boldsymbol{\mu}_m - \boldsymbol{\mu}_n \|_2 + \sum_{k=1}^{K} \frac{[\boldsymbol{h}_m(k) - \boldsymbol{h}_n(k)]^2}{\boldsymbol{h}_m(k) + \boldsymbol{h}_n(k)}}{\sigma_c^2}\right), \tag{2}$$

where $\| \cdot \|_2$ denotes the ℓ_2-norm, $\boldsymbol{\mu}_m$ and $\boldsymbol{\mu}_n$ are the m^{th} and n^{th} object's normalized average color value, respectively. $\boldsymbol{h}(k)$ is the normalized histogram valve, K is the number of histogram bins and $K = 512$ in our experiments.

Texture similarity. The Gabor filter responses with 8 orientations and three scales are exploited to represent the texture attribute of one object. The magnitude vector of Gabor filter for each object is computed by combing the 24 filters output. Then the texture similarity is generated as:

$$t^{mn} = \exp\left(-\frac{\| \boldsymbol{t}_m - \boldsymbol{t}_n \|_2}{\sigma_t^2}\right), \tag{3}$$

where $\boldsymbol{t}_m$ and $\boldsymbol{t}_n$ are the m^{th} and n^{th} object's normalized Gabor filter magnitude map, respectively. We notice that the parameters σ_c and σ_t play a similar role as in (1). In our experiments, we set $\sigma_s^2 = \sigma_c^2 = \sigma_t^2 = 0.1$.

Finally, three types of feature similarity are combined to evaluate the consistency relationship between each pair of objects. The object-level similarity is defined as:

$$\phi^{mn} = \alpha_s \cdot s^{mn} + \alpha_c \cdot c^{mn} + \alpha_t \cdot t^{mn}, \tag{4}$$

where $\alpha_s, \alpha_c, \alpha_t$ are coefficients for shape, color and texture similarity, respectively. In our experiments, $\alpha_s = 1/3$, $\alpha_c = 1/3$, and $\alpha_t = 1/3$ can perform reasonably well. The larger ϕ^{mn} is, the higher the similarity between two objects achieves. The final similarity value for each object is defined as $\Phi(o^m) = \arg\max(\phi^{mn})$, where $n \neq m$. All objects with $\Phi(o^m)$ greater than a threshold T_1 are regarded as positive common salient objects and denoted as O_p.

3.2.2. Superpixel-level consistency metric

In this subsection, superpixel-level consistency scores for all images are computed based on the obtained positive salient objects. First, the superpixels $R^i = \{r_d^i\}_{d=1}^{D_i}$ for each RGB image I^i are extracted by using SLIC algorithm [17], where D_i is the number of superpixels in image I^i. The set of superpixels $\{r_p^u\}_{u=1}^{U}$ inside of the positive objects O_p is denoted as R_p, where U is the total number of superpixels in R_p. Meanwhile, the set of superpixels outside of O_p is denoted as R_c. Here, one superpixel is inside of one object if $Area(r_d \bigcap o_m)/TotalArea(r_d) \geqslant 0.2$, where $Area(z)$ is the number of pixels inside of z. In this paper, the similarity between each pair of superpixels is calculated based on color and texture features by using the same computation process as the object-level's. Two superpixels are similar if their similarity is greater than a threshold T_2.

Initial consistency score. The initial consistency score for each superpixel r_d^i is calculated according to the similarity among the r_p^u in R_p. We first define the initial consistency score for r_p^u in R_p. Specifically, for each superpixel r_p^u, the more the number of its similar superpixels in R_p, the higher its consistency score is. It is formulated as:

$$\bar{S}(r_p^u) = \frac{2}{1 + \exp(-\frac{n(u)}{U \cdot \delta_s})} - 1, \tag{5}$$

where $n(u)$ is the number of similar superpixels in R_p with r_p^u, and δ_s is a parameter to control the consistency score. In our experiments, we set $\delta_s = 0.05$.

For each superpixel $\{r_c^v\}_{v=1}^{V}$ in R_c, its most similar superpixel $r_p^{\hat{u}}$ in R_p is searched based on the similarity measure, where V is the total number of superpixels in R_c. Then based on intra-image saliency value and the distance constrain, the consistency score for each superpixel r_c^v is calculated by:

$$\bar{S}(r_c^v) = 0.3 \cdot \exp(-\frac{\bar{d}(v)}{10}) \cdot \phi^{\hat{u}v} \cdot \bar{S}(r_p^{\hat{u}}) \cdot \bar{S}_{intra}(r_c^v), \tag{6}$$

where $\phi^{\hat{u}v}$ is the similarity between r_c^v and $r_p^{\hat{u}}$, $\bar{d}(v)$ is the minimum Euclidean distance between the centroid of superpixel r_c^v and the boundary points of objects which located in the same image, $\bar{S}_{intra}(r_c^v)$ is the mean intra-image saliency value of pixels inside of r_c^v.

Refined consistency score. The intra-image saliency maps generated by DSS method often exist some misses due to the lack of consistency constrain, especially when the salient objects connect with the border of image. In order to

Algorithm 1: Object Growing Algorithm

Input: O, T_2, R_c.
for m *from* 1 *to* M **do**
  while o^m *has no change* **do**
    Search the outmost superpixels $\{r_m^q\}_{q=1}^{Q_m}$ of o_m;
    for q *from* 1 *to* Q_m **do**
      Search r_m^q's nearest neighbour superpixels $\{r^b\}_{b=1}^{B}$ and compute the similarity ϕ^{qb};
      if $(\phi^{qb} > T_2)\&\&(r^b \in R_c)$ **then**
        $S_r(r^b) = \bar{S}(r_m^q)$;
        o_m=$o_m + r^b$;
      end
    end
  end
end
Output: Merged superpixel set R_r and the refined consistency score $S_r(r^b)$.

improve the detection performance, an object growing algorithm is designed for consistency score refinement of R_c, as shown in Algorithm 1. Similar to region growing algorithm, if a superpixel r_c^v in R_c is similar with the nearest neighbour superpixel which is inside of an object, the consistency score of r_c^v will be changed and the r_c^v is merged into the object. The merged superpixels set after this process is denoted as R_r and their consistency scores are S_r.

For the superpixels in R_p, we define their consistency score is equal to 1 since they are reliable. Finally, the consistency score of each superpixel is regarded as its saliency value and the inter-image saliency map for each image is generated as:

$$S_{inter}^i(r_d^i) = \begin{cases} 1, & \text{if } r_d^i \in R_p; \\ S_r(r_d^i), & \text{if } r_d^i \in R_r; \\ \bar{S}(r_d^i), & \text{others .} \end{cases} \tag{7}$$

3.3. Co-saliency map

The intra-image saliency map and inter-image saliency map for each image are fused by using a weighted way to obtain the final co-saliency maps:

$$S_{co}^i(r_d^i) = \begin{cases} S_{inter}^i(r_d^i), & \text{if } r_d^i \in R_r; \\ \lambda_1 \cdot S_{intra}^i(r_d^i) + \lambda_2 \cdot S_{inter}^i(r_d^i), & \text{otherwise,} \end{cases} \tag{8}$$

where λ_1 and λ_2 is the weighted coefficient. In our experiments, simply adopting $\lambda_1 = \lambda_2 = 0.5$ works well.

4. EXPERIMENTS

We evaluate the performance of our method, and compare it with the state-of-the-art methods on two public datasets: the iCoseg dataset and the MSRC dataset. The former contains 38 image sets of totally 643 images with manually labeled ground truth. The latter contains 7 image sets of totally 240 images with manually labeled ground truth.

4.1. Parameter settings and evaluation metrics

We first introduce the parameter settings in our experiments. The number of superpixels for each image is 300, *i.e.* $D_i = 300$. We set the thresholds $T_1 = 0.8$ and $T_2 = 0.7$.

To evaluate the performance of our method, an object comparison is performed based on generated co-saliency map and the ground truth mask. Four evaluation metrics including Precision-Recall (PR) curve, F-measure (F), AUC, and Mean Absolute Error (MAE) are calculated. The precision and recall scores are produced by thresholding the saliency map into binary salient object masks with a series of fixed integers from 0 to 255. AUC is the under area of ROC curve, and the larger, the better.

4.2. Comparison with state-of-the-art methods

We compare our result with three single image saliency detection methods (*i.e.* BSCA [18], SMD [19], and DSS [8]) and four co-saliency detection methods (*i.e.* DW [9], MSPL [11], CCS [5], and IPDM [7]).

The quantitative comparison results in terms of the PR curves, F-measure, AUC scores, and MAE scores are reported in Fig.2. As can be seen, the proposed method reaches the highest level in all curves on two datasets. Moreover, the proposed method achieves the best performance on two datasets with the highest F-measure and AUC scores and the smallest MAE value compared with other methods. For the F-measure, the performance gain of the proposed method over the best competing approach MSPL [11] reaches 17.13% on the MSRC dataset. The percentage gain of MAE score also achieves 35.38%. On the iCoseg dataset, the minimum percentage gains of F-measure and MAE score of the proposed method achieve 15.65% and 54.72%, respectively. Although the DSS method also has approximate F-measure score compared with the proposed method, its PR curve, AUC and MAE are much inferior to ours. In addition, when the test image dataset contains several unrelated images, its performance will decline sharply.

Some visual comparisons of different methods on two datasets are illustrated in Fig.3 and Fig.4. Obviously, the existing co-saliency detection methods generate many false alarms and misses, while our method obtains more accurate and homogeneous results. Take Fig.3 as an example, single image saliency detection methods tend to capture the salient objects which located in the center of image. Thus their results are often incomplete when applied to co-saliency detection due to the lack of consistency constrain. Compared with the existing co-saliency detection methods, the proposed method highlights the common salient object more. The proposed object growing algorithm refines the intra-image saliency results and obtains more complete salient objects.

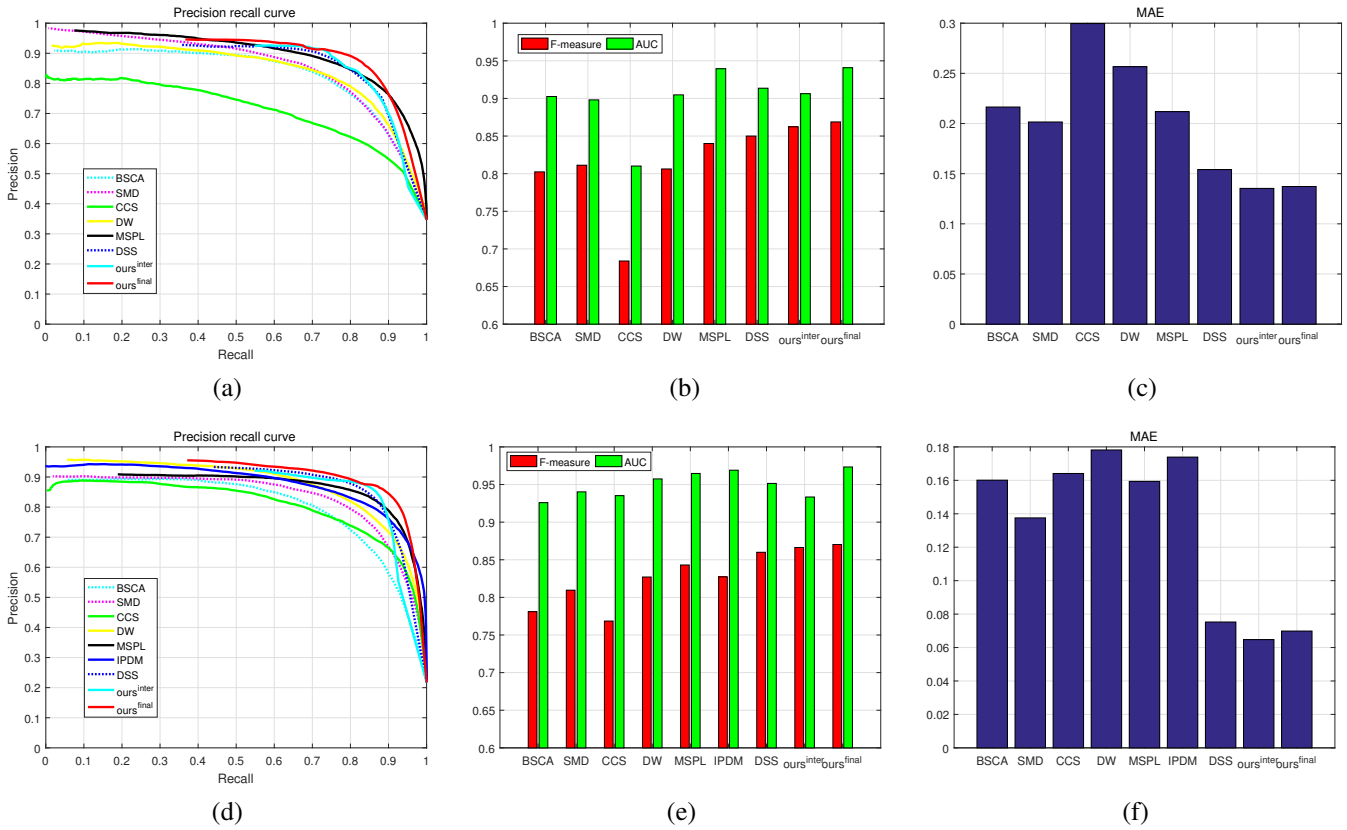


Fig. 2: The quantitative performances of different methods on two datasets. (a)-(c) are PR curves, F-measure and AUC scores, and MAE on the MSRC dataset, respectively. (d)-(f) are PR curves, F-measure and AUC scores, and MAE on the iCoseg dataset, respectively.

5. CONCLUSION

In this paper, we present a co-saliency detection method based on hierarchical consistency measure by exploring the multi-feature similarity and inter-image constrains among multiple images. Since the co-salient objects often have similar appearance, the shape attribute is introduced to constrain the object-level similarity and evaluate the consistency among different objects. In addition, an object growing algorithm is designed to refine the superpixel-level consistency measure and generate the inter saliency map. Experiment results on two public datasets have demonstrated that the proposed method outperforms other state-of-the-art methods.

References

[1] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *ICCV*, 2009.

[2] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *CVPR*, 2012.

[3] S. Han and N. Vasconcelos, "Image compression using object-based regions of interest," in *ICIP*, 2006.

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

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

[6] K. Tang, A. Joulin, L. Li, and F. Li, "Co-localization in real-world images," in *CVPR*, 2014.

[7] D. Zhang, J. Han, J. Han, and L. Shao, "Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining," *IEEE TNNLS*, vol. 27, no. 6, pp. 1163, 2016.

[8] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017.

[9] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *IJCV*, vol. 120, no. 2, pp. 215–232, 2016.

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

[11] D. Zhang, D. Meng, and J. Han, "Co-saliency detection via a self-paced multiple-instance learning framework," *IEEE PAMI*, vol. 39, no. 5, pp. 865–878, 2017.

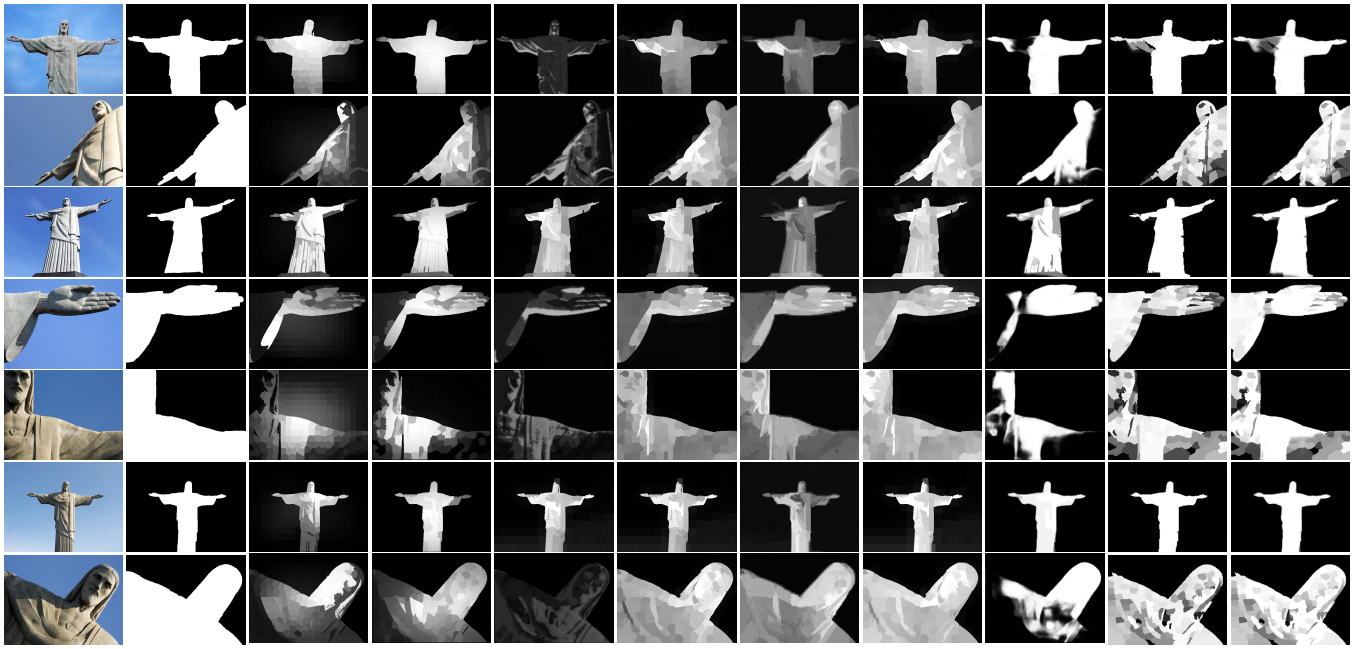


Fig. 3: Visual comparison of competing methods on the iCoseg dataset.

Fig. 4: Visual comparison of competing methods on the MSRC dataset.

[12] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation," *IEEE TIP*, vol. 27, no. 2, pp. 568–579, 2018.

[13] Z. Liu, W. Zou, L. Li, L. Shen, and O. Meur, "Co-saliency detection based on hierarchical segmentation," *IEEE SPL*, vol. 21, no. 1, pp. 88–92, 2013.

[14] R. Huang, W. Feng, and J. Sun, "Saliency and co-saliency detection by low-rank multiscale fusion," in *ICME*, 2015.

[15] C. Tsai, X. Qian, and Y. Lin, "Segmentation guided local proposal fusion for co-saliency detection," in *ICME*, 2017.

[16] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE PAMI*, vol. 24, no. 24, pp. 509–522, 2002.

[17] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE PAMI*, vol. 34, no. 11, pp. 2274–2282, 2012.

[18] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015.

[19] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. Maybank, "Salient object detection via structured matrix decomposition," *IEEE PAMI*, vol. 39, no. 4, pp. 818–832, 2017.