Salient Objects in Clutter: Bringing Salient Object Detection to the Foreground

Deng-Ping Fan[1]   Ming-Ming Cheng[1]*   Jiang-Jiang Liu[1]
Shang-Hua Gao[1]   Qibin Hou[1]   Ali Borji[2]

[1] CCCE, Nankai University   [2] CRCV, UCF
http://mmcheng.net/SOCBenchmark/

Abstract. In this paper, we provide a comprehensive evaluation of salient object detection (SOD) models. Our analysis identifies a serious design bias of existing SOD datasets which assumes that each image contains at least one clearly outstanding salient object in low clutter. This is an unrealistic assumption. The design bias has led to a saturated high performance for state-of-the-art SOD models when evaluated on existing datasets. The models, however, still perform far from being satisfactory when applied to real-world daily scenes. Based on our analyses, we first identify 7 crucial aspects that a comprehensive and balanced dataset should fulfill. Then, we propose a new high quality dataset and update the previous saliency benchmark. Specifically, our dataset called **SOC**, Salient Objects in Clutter, includes images with salient and non-salient objects from daily object categories. Beyond object category annotations, each salient image is accompanied by attributes (e.g., *appearance change, clutter*) that reflect common challenges in real-world scenes, and can help 1) gain a deeper insight into the SOD problem, 2) investigate the pros and cons of the SOD models, and 3) objectively assess models from different perspectives. Finally, we report attribute-based performance assessment on our SOC dataset. We believe that our dataset and results will open new directions for future research on salient object detection.

Keywords: Salient object detection · saliency benchmark · dataset · attribute

1   Introduction

This paper considers the task of salient object detection (SOD). Visual saliency mimics the ability of the human visual system to select a certain subset of the visual scene. SOD aims to detect the most attention-grabbing objects in a scene and then extract pixel-accurate silhouettes of the objects. The merit of SOD lies in it applications in many other computer vision tasks including: visual tracking [1], image retrieval [2], and weakly supervised semantic segmentation.

Our work is motivated by two observations. First, existing SOD datasets [3–11] are flawed either in the data collection procedure or quality of the data. Specifically, most datasets assume that an image contains at least one salient object, and thus discard

---

Fig. 1. Sample images from our new dataset including *non-salient* object images (first row) and salient object images (rows 2 to 4). For salient object images, instance-level ground truth map (different color), object attributes (Attr) and category labels are provided. Please refer to the supplemental material for more illustrations of our dataset.

images that do not contain salient objects. We call this *data selection bias*. Moreover, existing datasets mostly contain images with a single object or several objects (often a person) in low clutter. These datasets do not adequately reflect the complexity of images in the real world where scenes usually contain multiple objects amidst lots of clutter. As a result, all top performing models trained on the existing datasets have nearly saturated the performance (*e.g.*, > 0.9 *F-measure* over most current datasets) but unsatisfactory performance on realistic scenes (*e.g.*, < 0.45 *F-measure* in Table 3 and some examples in Fig. 7). Because current models may be biased towards ideal conditions, their effectiveness may be impaired once they are applied to real world scenes. To solve this problem, it is important to introduce a dataset that reaches closer to realistic conditions.

Second, only the overall performance of the models can be analyzed over existing datasets. None of the datasets contains various attributes that reflect challenges in real-world scenes. Having attributes helps 1) gain a deeper insight into the SOD problem, 2) investigate the pros and cons of the SOD models, and 3) objectively assess the model performances over different perspectives, which might be diverse for different applications.

Considering the above two issues, we make two contributions. Our main contribution is the collection of a new high quality SOD dataset, named the **SOC**, Salient Objects in Clutter. To date, SOC is the largest instance-level SOD dataset and contains 6,000 images from more than 80 common categories. It differs from existing datasets

in three aspects: 1) salient objects have category annotation which can be used for new research such as weakly supervised SOD tasks, 2) the inclusion of non-salient images which make this dataset closer to the real-world scenes and more challenging than the existing ones, and 3) salient objects have attributes reflecting specific situations faced in the real-wold such as *motion blur, occlusion and cluttered background*. As a consequence, our SOC dataset narrows the ***gap*** between existing datasets and the real-world scenes and provides a more realistic benchmark (see Fig. 1).

In addition, we provide a comprehensive evaluation of several state-of-the-art convolutional neural networks (CNNs) based models [11–23]. To evaluate the models, we introduce three metrics that measure the region similarity of the detection, the pixel-wise accuracy of the segmentation, and the structure similarity of the result. Furthermore, we give an attribute-based performance evaluation. These attributes allow a deeper understanding of the models and point out promising directions for further research.

We believe that our dataset and benchmark can be very influential for future SOD research in particular for application-oriented model development. The entire dataset and analyzing tools will be released freely to the public.

2   Related Works

In this section, we briefly discuss existing datasets designed for SOD tasks, especially in the aspects including annotation type, the number of salient objects per image, number of images, and image quality. We also review the CNNs based SOD models.

| (a) Image | (b) Pixel | (c) Instance | (d) Segment |
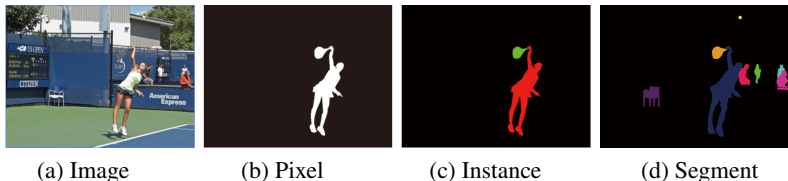
Fig. 2. Previous SOD datasets only annotate the image by drawing (b) pixel-accurate silhouettes of salient objects. Different from (d) MS COCO object segmentation dataset [24] (Objects are not necessarily being **salient**), our work focuses on (c) segmenting *salient object instances*.

2.1   Datasets

Early datasets are either limited in the number of images or in their coarse annotation of salient objects. For example, the salient objects in datasets **MSRA-A** [3] and **MSRA-B** [3] are roughly annotated in the form of bounding boxes. **ASD** [25] and **MSRA10K** [6] mostly contain only one salient object in each image, while the **SED2** [4] dataset contains two objects in a single image but contains only 100 images. To improve the quality of datasets, researchers in recent years started to collect datasets with multiple objects in relatively complex and cluttered backgrounds. These datasets

include **DUT-OMRON** [9], **ECSSD** [8], **Judd-A** [7], and **PASCAL-S** [10]. These datasets have been improved in terms of annotation quality and the number of images, compared to their predecessors. Datasets **HKU-IS** [11], **XPIE** [26], and **DUTS** [27] resolved the shortcomings by collecting large amounts of pixel-wise labeled images (Fig. 2 (b)) with more than one salient object in images. However, they ignored the non-salient objects and did not offer instance-level (Fig. 2 (c)) salient objects annotation. Beyond these, researchers of [28] collected about 6k *simple background images* (most of them are pure texture images) to account for the non-salient scenes. This dataset is not sufficient to reflect real scenes as the real-world scenes are more complicated. The **ILSO** [29] dataset contains instance-level salient objects annotation but has boundaries roughly labeled as shown in Fig. 5 (a).

To sum up, as discussed above, existing datasets mostly focus on images with clear salient objects in simple backgrounds. Taking into account the aforementioned limitations of existing datasets, a more realistic dataset which contains realistic scenes with non-salient objects, textures "in the wild", and salient objects with attributes, is needed for future investigations in this field. Such a dataset can offer deep insights into weaknesses and strengths of SOD models.

2.2   Models

We divide the state-of-the-art deep models for SOD based on the number of tasks.

Single-task models have the single goal of detecting the salient objects in images. In **LEGS** [12], local information and global contrast were separately captured by two different deep CNNs, and were then fused to generate a saliency map. In [13], Zhao *et al.* presented a multi-context deep learning framework (**MC**) for SOD. Li *et al.* [11] (**MDF**) proposed to use multi-scale features extracted from a deep CNNs to derive a saliency map. Li *et al.* [14] presented a deep contrast network (**DCL**), which not only considered the pixel-wise information but also fused the segment-level guidance into the network. Lee *et al.* [15] (**ELD**) considered both high-level features extracted from CNNs and hand-crafted features. Liu *et al.* [16] (**DHS**) designed a two-stage network, in which a coarse downscaled prediction map was produced. It is then followed by another network to refine the details and upsample the prediction map hierarchically and progressively. Long *et al.* [30] proposed a fully convolutional network (**FCN**) to make dense pixel prediction problem feasible for end-to-end training. **RFCN** [17] used a recurrent FCN to incorporate the coarse predictions as saliency priors and refined the generated predictions in a stage-wise manner. The **DISC** [18] framework was proposed for fine-grained image saliency computing. Two stacked CNNs were utilized to obtain coarse-level and fine-grained saliency maps, respectively. **IMC** [19] integrated saliency cues at different levels through FCN. It could efficiently exploit both learned semantic cues and higher-order region statistics for edge-accurate SOD. Recently, a deep architecture [20] with short connections (**DSS**) was proposed. Hou *et al.* added connections from high-level features to low-level features based on the HED [31] architecture, achieving good performance. **NLDF** [21] integrated local and global features and added a boundary loss term into standard cross entropy loss to train an end-to-end network. **AMU** [22] was a generic aggregating multi-level convolutional feature framework. It integrated coarse semantics and fine detailed feature maps into multiple resolutions. Then

	No	Model	Year	Pub	#Training	*Training Set*	*Base Model*	*FCN*	*Sp*	*Proposal*	*Edge*
S-task	1	**LEGS** [12]	2015	CVPR	3,340	MB + P	—	✗	✗	✓	✗
	2	**MC** [13]	2015	CVPR	8,000	MK	GoogLeNet	✗	✓	✗	✗
	3	**MDF** [11]	2015	CVPR	2,500	MB	—	✗	✓	✗	✓
	4	**DCL** [14]	2016	CVPR	2,500	MB	VGGNet	✓	✓	✗	✗
	5	**ELD** [15]	2016	CVPR	9,000	MK	VGGNet	✗	✓	✗	✗
	6	**DHS** [16]	2016	CVPR	9,500	MK+D	VGGNet	✗	✗	✗	✗
	7	**RFCN** [17]	2016	ECCV	10,103	P2010	—	✓	✓	✗	✓
	8	**DISC** [18]	2016	TNNLS	9,000	MK	—	✗	✓	✗	✗
	9	**IMC** [19]	2017	WACV	6,000	MK	ResNet-101	✓	✓	✗	✗
	10	**DSS** [20]	2017	CVPR	2,500	MB	VGGNet	✓	✗	✗	✓
	11	**NLDF** [21]	2017	CVPR	2,500	MB	VGGNet	✓	✗	✗	✗
	12	**AMU** [22]	2017	ICCV	10,000	MK	VGGNet	✓	✗	✗	✓
	13	**UCF** [23]	2017	ICCV	10,000	MK	—	✓	✗	✗	✗
M-task	1	**DS** [32]	2016	TIP	10,000	MK	VGGNet	✓	✓	✗	✗
	2	**WSS** [27]	2017	CVPR	456K	ImageNet	VGGNet	✓	✓	✗	✗
	3	**MSR** [29]	2017	CVPR	5,000	MB + H	VGGNet	✓	✗	✓	✓

Table 1. CNNs based SOD models. We divided these models into single-task (S-task) and multi-task (M-task). ***Training Set:*** MB is the MSRA-B dataset [3]. MK is the MSRA-10K [6] dataset. ImageNet dataset refers to [33]. D is the DUT-OMRON [9] dataset. H is the HKU-IS [11] dataset. P is the PASCAL-S [10] dataset. P2010 is the PASCAL VOC 2010 semantic segmentation dataset [34]. ***Base Model:*** VGGNet, ResNet-101, AlexNet, GoogleNet are base models. ***FCN:*** whether model uses the fully convolutional network. ***Sp:*** whether model uses superpixels. ***Proposal:*** whether model uses the object proposal. ***Edge:*** whether model uses the edge or contour information.

it adaptively learned to combine these feature maps at each resolution and predicted saliency maps with the combined features. **UCF** [23] was proposed to improve the robustness and accuracy of saliency detection. They introduced a reformulated dropout after specific convolutional layers to construct an uncertain ensemble of internal feature units. Also, they proposed reformulated dropout after an effective hybrid up-sampling method to reduce the checkerboard artifacts of deconvolution operators in the decoder network.

Multi-task models at present include three methods, **DS**, **WSS**, and **MSR**. The **DS** [32] model set up a multi-task learning scheme for exploring the intrinsic correlations between saliency detection and semantic image segmentation, which shared the information in FCN layers to generate effective features for object perception. Recently, Wang *et al.* [27] proposed a model named **WSS** which developed a weakly supervised learning method using image-level tags for saliency detection. First, they jointly trained Foreground Inference Net (FIN) and FCN for image categorization. Then, they used FIN fine-tuned with iterative CRF to enforce spatial label consistency to predict the saliency map. **MSR** [29] was designed for both salient region detection and salient object contour detection, integrated with multi-scale combinatorial grouping and a MAP-based [35] subset optimization framework. Using three refined VGG network streams with shared parameters and a learned attentional model for fusing results at different scales, the authors were able to achieve good results.

We benchmark a large set of the state-of-the-art CNNs based models (see Table 1) on our proposed dataset, highlighting the current issues and pointing out future research directions.

3   The Proposed Dataset

In this section, we present our new challenging SOC dataset designed to reflect the real-world scenes in detail. Sample images from SOC are shown in Fig. 1. Moreover, statistics regarding the categories and the attributes of SOC are shown in Fig. 4 (a) and Fig. 6, respectively. Based on the strengths and weaknesses of the existing datasets, we identify seven crucial aspects that a comprehensive and balanced dataset should fulfill.

1) Presence of Non-Salient Objects. Almost all of the existing SOD datasets make the assumption that an image contains at least one salient object and discard the images that do not contain salient objects. However, this assumption is an ideal setting which leads to *data selection bias*. In a realistic setting, images do not always contain salient objects. For example, some amorphous background images such as sky, grass and texture contain no salient objects at all [36]. The non-salient objects or background "stuff" may occupy the entire scene, and hence heavily constrain possible locations for a salient object. Xia *et al.* [26] proposed a state-of-the-art SOD model by judging what is or what is not a salient object, indicating that the non-salient object is crucial for reasoning about the salient object. This suggests that the non-salient objects deserve equal attention as the salient objects in SOD. Incorporating a number of images containing non-salient objects makes the dataset closer to real-world scenes, while becoming more challenging. Thus, we define the "*non-salient objects*" as images without salient objects or images with "stuff" categories. As suggested in [26, 36], the "stuff" categories including (a) densely distributed similar objects, (b) fuzzy shape, and (c) region without semantics, which are illustrated in Fig. 3 (a)-(c), respectively.
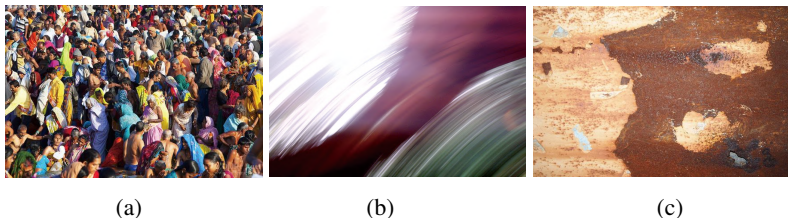

(a)   (b)   (c)

Fig. 3. Some examples of non-salient objects.

Based on the characteristics of non-salient objects, we collected 783 texture images from the DTD [37] dataset. To enrich the diversity, 2217 images including aurora, sky, crowds, store and many other kinds of realistic scenes were gathered from the Internet and other datasets [5, 10, 24]. We believe that incorporating enough non-salient objects would open up a promising direction for future works.

2) Number and Category of Images. A considerably large amount of images is essential to capture the diversity and abundance of real-world scenes. Moreover, with large amounts of data, SOD models can avoid over-fitting and enhance generalization. To this end, we gathered 6,000 images from more than 80 categories, containing 3,000 images with salient objects and 3,000 images without salient objects. We divide our dataset into training set, validation set and test set in the ratio of 6:2:2. To ensure fair-

ness, the test set is not published, but with the *on-line testing* provided on our website[1]. Fig. 4 (a) shows the number of salient objects for each category. It shows that the "person" category accounts for a large proportion, which is reasonable as people usually appear in daily scenes along with other objects.
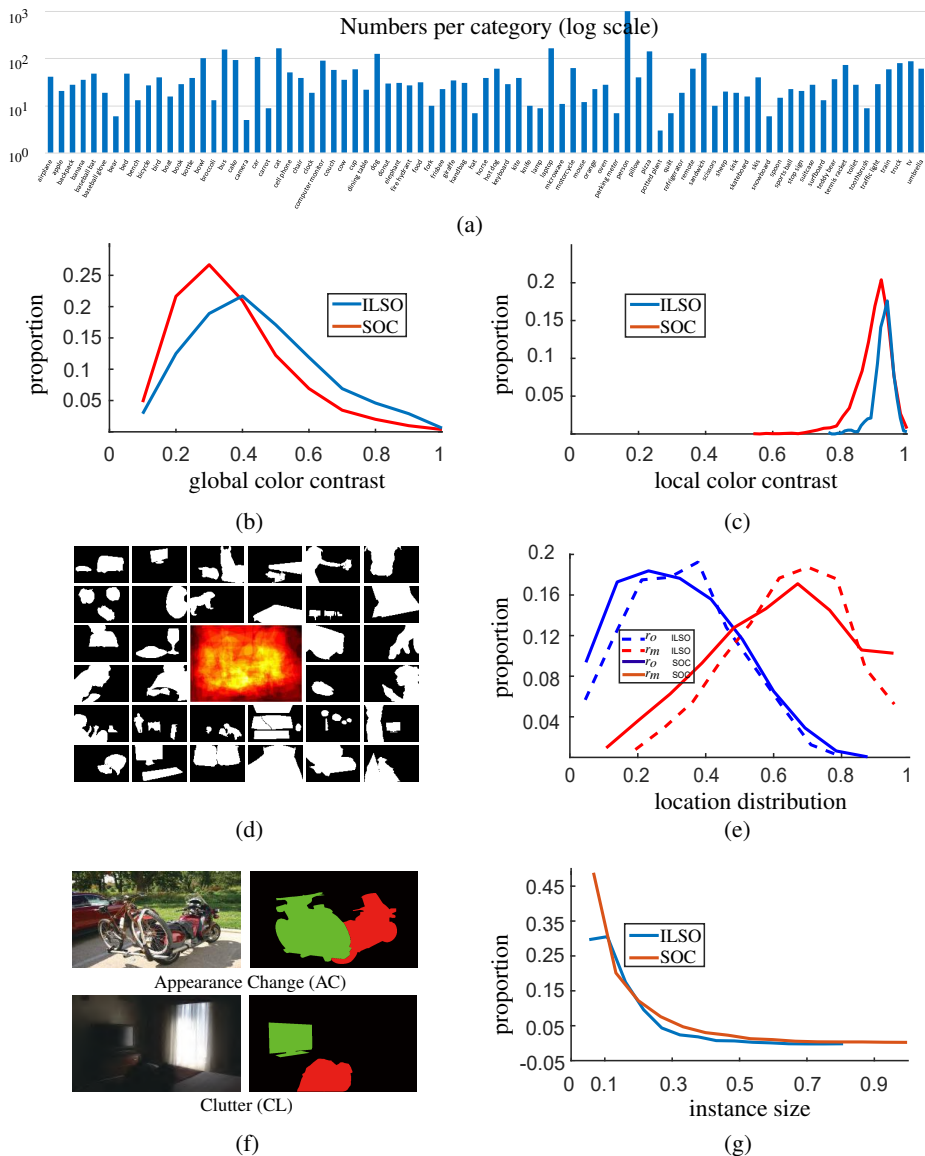
3) Global/Local Color Contrast of Salient Objects. As described in [10], the term "salient" is related to the global/local contrast of the foreground and background. It is essential to check whether the salient objects are easy to detect. For each object, we compute RGB color histograms for foreground and background separately. Then, χ^2 distance is utilized to measure the distance between the two histograms. The global and local color contrast distribution are shown in Fig. 4 (b) and (c), respectively. In comparison to ILSO, our SOC has more proportion of objects with low global color contrast and local color contrast.

4) Locations of Salient Objects. *Center bias* has been identified as one of the most significant biases of saliency detection datasets [10, 38, 39]. Fig. 4 (d) illustrates a set of images and their overlay map. As can be seen, although salient objects are located in different positions, the overlay map still shows that somehow this set of images is center biased. Previous benchmarks often adopt this incorrect way to analyze the location distribution of salient objects. To avoid this misleading phenomenon, we plot the statistics of two quantities r_o and r_m in Fig. 4 (e), where r_o and r_m denote how far an object center and the farthest (margin) point in an object are from the image center, respectively. Both r_o and r_m are divided by half image diagonal length for normalization so that $r_o, r_m \in [0, 1]$. From these statistics, we can observe that salient objects in our dataset do not suffer from center bias.

5) Size of Salient Objects. The size of an instance-level salient object is defined as the proportion of pixels in the image [10]. As shown in Fig. 4 (g), the size of salient objects in our SOC varies in a broader range, compared with the only existing instance-level ILSO [29] dataset. Also, medium-sized objects in SOC have a higher proportion.

6) High-Quality Salient Object Labeling. As also noticed in [40], training on the ECSSD dataset (1,000) allows to achieve better results than other datasets (*e.g.*, MSRA10K, with 10,000 images). Besides the scale, dataset quality is also an important factor. To obtain a large amount of high quality images, we randomly select images from the MSCOCO dataset [24], which is a large-scale real-world dataset whose objects are labeled with polygons (*i.e.*, coarse labeling). High-quality labels also play a critical role in improving the accuracy of SOD models [25]. Toward this end, we relabel the dataset with pixel-wise annotations. Similar to famous SOD task oriented benchmark datasets [3–6, 8, 11, 25–27, 29, 41], we did not use the eye tracker device. We have taken a number of steps to provide the high-quality of the annotations. These steps include two stages: **In the bounding boxes (bboxes) stage**, (i) we ask 5 viewers to annotate objects with bboxes that they think are salient in each image. (ii) keep the images which majority (≥ 3) viewers annotated the same (the IOU of the bbox > 0.8) object. After the first stage, we have 3,000 salient object images annotated with bboxes. **In the second stage**, we further manually label the accurate silhouettes of the salient objects according to the bboxes. Note that we have 10 volunteers involved in the whole steps for cross-check the quality of annotations. In the end, we keep 3,000 images with high-quality,

---

[1] The link will be made publicly available after the review.

(a)

(b)

(c)

(d)

(e)

(f)

(g)

Fig. 4. (a) Number of annotated instances per category in our SOC dataset. (b, c) The statistics of global color contrast and local color contrast, respectively. (d) A set of saliency maps from our dataset and their overlay map. (e) Location distribution of the salient objects in SOC. (f) Attribute visual examples. (g) The distribution of instance sizes for the SOC and ILSO [29].

instance-level labeled salient objects. As shown in Fig. 5 (b,d), the boundaries of our object labels are precise, sharp and smooth. During the annotation process, we also add some new categories (*e.g.*, *computer monitor, hat, pillow*) that are not labeled in the MSCOCO dataset [24].

(a) ILSO   (b) SOC
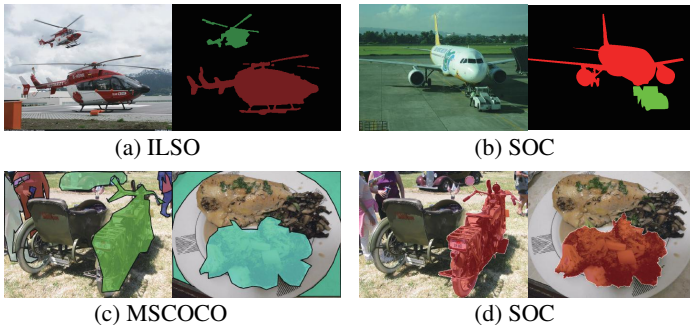
(c) MSCOCO   (d) SOC

Fig. 5. Compared with the recent new (a) instance-level ILSO dataset [29] which is labeled with discontinue coarse boundaries, (c) MSCOCO dataset [24] which is labeled with polygons, our (b, d) SOC dataset is labeled with smooth fine boundaries.

| Attr	Description
AC	*Appearance Change.* The obvious illumination change in the object region.
BO	*Big Object.* The ratio between the object area and the image area is larger than 0.5.
CL	*Clutter.* The foreground and background regions around the object have similar color. We labeled images that their global color contrast value is larger than 0.2, local color contrast value is smaller than 0.9 with clutter images (see Sec. 3).
HO	*Heterogeneous Object.* Objects composed of visually distinctive/dissimilar parts.
MB	*Motion Blur.* Objects have fuzzy boundaries due to shake of the camera or motion.
OC	*Occlusion.* Objects are partially or fully occluded.
OV	*Out-of-View.* Part of object is clipped by image boundaries.
SC	*Shape Complexity.* Objects have complex boundaries such as thin parts (*e.g.*, the foot of animal) and holes.
SO	*Small Object.* The ratio between the object area and the image area is smaller than 0.1.

Table 2. The list of salient object image attributes and the corresponding description. By observing the characteristics of the existing datasets, we summarize these attributes. Some visual examples can be found in Fig. 1 and Fig. 4 (f). For more examples, please refer to the supplementary materials.

7) Salient Objects with Attributes. Having attributes information regarding the images in a dataset helps objectively assess the performance of models over different types of parameters and variations. It also allows the inspection of model failures. To this end, we define a set of attributes to represent specific situations faced in the real-wold scenes such as *motion blur*, *occlusion* and *cluttered background* (summarized in Table 2). Note that one image can be annotated with multiple attributes as these attributes are not exclusive.

Inspired by [42], we present the distribution of attributes over the dataset as shown in Fig. 6 Left. Type *SO* has the largest proportion due to accurate instance-level (*e.g.*, tennis racket in Fig. 2) annotation. Type *HO* accounts for a large proportion, because the real-world scenes are composed of different constituent materials. *Motion blur* is more common in video frames than still images, but it also occurs in still images sometimes. Thus, type *MB* takes a relatively small proportion in our dataset. Since a realistic image usually contains multiple attributes, we show the dominant dependencies among
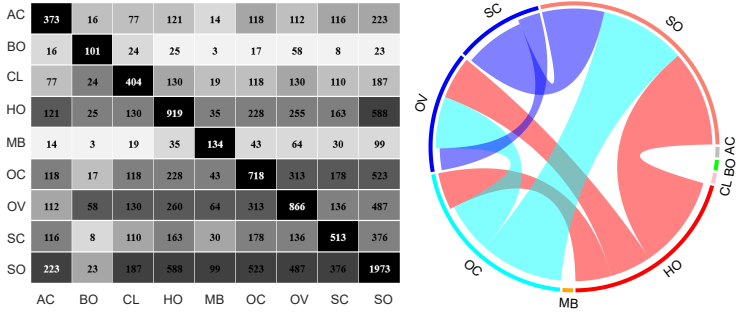
Fig. 6. Left: Attributes distribution over the salient object images in our SOC dataset. Each number in the grids indicates the image number of occurrences. Right: The **dominant dependencies** among attributes base on the frequency of occurrences. Larger width of a link indicates higher probability of an attribute to other ones.

attributes based on the frequency of occurrences in the Fig. 6 Right. For example, a scene containing lots of heterogeneous objects is likely to have a large number of objects blocking each other and forming complex spatial structures. Thus, type *HO* has a strong dependency with type *OC*, *OV*, and *SO*.

4   Benchmarking Models

In this section, we present the evaluation results of the sixteen SOD models on our SOC dataset. Nearly all representative CNNs based SOD models are evaluated. However, since the codes of some models are not publicly available, we do not consider them here. In addition, most models are not optimized for non-salient objects detection. Thus, to be fair, we only use the test set of our SOC dataset to evaluate SOD models. We describe the evaluation metrics in Sec. 4.1. Overall model performance on SOC dataset is presented in Sec. 4.2 and summarized in Table 3, while the attribute level performance (*e.g.*, performance of the appearance changes) is discussed in Sec. 4.3 and summarized in Table 4. The evaluation scripts are publicly available, and on-line evaluation test is provided on our website [2].

4.1   Evaluation Metrics

In a supervised evaluation framework, given a predicted map *M* generated by a SOD model and a ground truth mask *G*, the evaluation metrics are expected to tell which model generates the best result. Here, we use three different evaluation metrics to evaluate SOD models on our SOC dataset.

Pixel-wise Accuracy ε. The region similarity evaluation measure does not consider the true negative saliency assignments. As a remedy, we also compute the normalized ([0,1]) mean absolute error (MAE) between *M* and *G*, defined as:

---

[2] The link will be made publicly available.

	Single-task													Multi-task		
Type	LEGS	MC	MDF	DCL	AMU	RFCN	DHS	ELD	DISC	IMC	UCF	DSS	NLDF	DS	WSS	MSR
	[12]	[13]	[11]	[14]	[14]	[17]	[16]	[15]	[18]	[19]	[23]	[20]	[21]	[32]	[27]	[29]
F_{all} ↑	.276	.291	.307	.339	.341	**.435**	.360	.317	.288	.352	.333	.341	.352	.347	.327	**.380**
S_{all} ↑	.677	.757	.736	.771	.737	.814	.804	.776	.737	.664	.657	.807	**.818**	.779	.785	**.819**
ε_{all} ↓	.230	.138	.150	.157	.185	.113	.118	.135	.173	.269	.282	.111	**.104**	.155	.133	.113

Table 3. The performance of SOD models under three metrics. F stands for region similarity, ε is the mean absolute error, and S is the structure similarity. ↑ stand for the higher the number the better, and vice versa for ↓. The evaluation results are calculated according to Eqn. (3) over our SOC dataset. $S_{all}, F_{all}, \varepsilon_{all}$ indicate the overall performance using the metric of S, F, ε, respectively. The best result is highlighted in **bold**.

$$\varepsilon = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} ||M(x,y) - G(x,y)||, \tag{1}$$

where W and H are the width and height of images, respectively.

Region Similarity F. To measure how well the regions of the two maps match, we use the F-measure, defined as:

$$F = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \tag{2}$$

where $\beta^2 = 0.3$ is suggested by [25] to trade-off the *recall* and *precision*. However, the black (all-zero matrix) ground truth is not well defined in F-measure when calculating *recall* and *precision*. Under this circumstances, different foreground maps get the same result 0, which is apparently unreasonable. Thus, F-measure is not suitable for measuring the results of non-salient object detection.

However, both metrics of ε and F are based on pixel-wise errors and often ignore the structural similarities. Behavioral vision studies have shown that the human visual system is highly sensitive to structures in scenes [43]. In many applications, it is desired that the results of the SOD model retain the structure of objects.

Structure Similarity S. *S-measure* proposed by Fan *et al.* [43] evaluates the structural similarity, by considering both regions and objects. Therefore, we additionally use *S-measure* to evaluate the structural similarity between M and G. Note that the next overall performance we evaluated and analyzed are based on the *S-measure*.

4.2 Metric Statistics

To obtain an overall result, we average the scores of the evaluation metrics η ($\eta \in \{F, \varepsilon, S\}$), denoted by:

$$M_\eta(D) = \frac{1}{|D|} \sum_{I \in D} \bar{\eta}(I_i), \tag{3}$$

where $\bar{\eta}(I_i)$ is the evaluation score of the image I_i within the image dataset D.

Single-task: For the single-task models, the best performing model on the entire SOC dataset (S_{all} in Table 3) is **NLDF** [21] ($M_S = 0.818$), followed by **RFCN** [17]

($M_S = 0.814$). MDF [11] and AMU [22] use edge cues to promote the saliency map but fail to achieve the ideal goal. Aiming at using the local region information of images, MC [13], MDF [11], ELD [15], and DISC [18] try to use superpixel methods to segment images into regions and then extract features from these regions, which is complex and time-consuming. To further improve the performance, UCF [23], DSS [20], NLDF [21], and AMU [22] utilize the FCN to improve the performance of SOD (S_{sal} in Table 4). Some other methods such as DCL [14] and IMC [19] try to combine superpixels with FCN to build a powerful model. Furthermore, RFCN [17] combines two related cues including edges and superpixels into FCN to obtain the good performance ($M_F = 0.435$, $M_S = 0.814$) over the overall dataset.

Multi-task: Different from models mentioned above, MSR [29] detects the instance-level salient objects using three closely related steps: estimating saliency maps, detecting salient object contours, and identifying salient object instances. It creates a multi-scale saliency refinement network that results in the highest performance (S_{all}). Other two multi-task models DS [32] and WSS [27] utilize the segmentation and classification results simultaneously to generate the saliency maps, obtaining a moderate performance. It is worth mentioning that although WSS is a weakly supervised multi-task model, it still achieves comparable performance to other single-task, fully supervised models. So, the weakly-supervised and multi-task based models can be promising future directions.

4.3   Attributes-based Evaluation

We assign the salient images with attributes as discussed in Sec. 3 and Table 2. Each attribute stands for a challenging problem faced in the real-world scenes. The attributes allow us to identify groups of images with a dominant feature (*e.g.*, presence of clutter), which is crucial to illustrate the performance of SOD models and to relate SOD to application-oriented tasks. For example, sketch2photo application [44] prefers models with good performance on big objects, which can be identified by attributes-based performance evaluation methods.

Results. In Table 4, we show the performance on subsets of our dataset characterized by a particular attribute. Due to space limitation, in the following parts, we only select some representative attributes for further analysis. More details can be found in the supplementary material.

Big Object (BO) scenes often occur when objects are in a close distance with the camera, in which circumstances the tiny text or patterns would always be seen clearly. In this case, the models which prefer to focus on local information will be mislead seriously, leading to a considerable (*e.g.*, 28.9% loss for DSS [20], 20.8% loss for MC [13] and 23.8% loss for RFCN [17]) loss of performance.

However, the performance of IMC [19] model goes up for a slight margin of 3.2% instead. After taking a deeper look of the pipeline of this model, we came up a reasonable explanation. IMC uses a coarse predicted map to express semantics and utilizes over-segmented images to supplement the structural information, achieving a satisfying result on type *BO*. However, over-segmented images cannot make up the missing details, causing 4.6% degradation of performance on the type of *SO*.

|						Single-task								Multi-task		
Attr	LEGS [12]	MC [13]	MDF [11]	DCL [14]	AMU [14]	RFCN [17]	DHS [16]	ELD [15]	DISC [18]	IMC [19]	UCF [23]	DSS [20]	NLDF [21]	DS [32]	WSS [27]	MSR [29]
S_{sal}	.607	.619	.610	.705	.705	.709	**.728**	.664	.629	.679	.678	.698	.714	.719	.676	**.748**
AC	.625	.631	.614	.734	.736	.744	**.745**	.673	.644	.702	.714	.726	.737	.764	.691	**.789**
BO	.509	.490	.461⁻	.610	.569	.540	.590	.576	.517	**.701**⁺	.636	.496⁻	.568	.685	.566	.667
CL	.620	.635	.566	.699	.708	.714	**.743**	.658	.635	.696	.704	.677⁻	.713	.729	.678	**.756**
HO	.666	.666	.648	.745	.755	.759	**.766**	.706	.681	.715	.744	.748	.755	.756	.707	**.777**
MB	.543⁻	.603	.615	.693	.706	.715	**.722**	.639	.600	.689	.682	.695	.685	.711	.641	**.757**
OC	.609	.617	.608	.708⁺	**.725**⁺	.711	.716	.658	.630	.672	**.701**⁺	.689	.709	.725⁺	.672	**.740**
OV	.548	.584	.568	.699	**.708**⁺	.687	.706	.637	.573	**.693**⁺	**.685**⁺	.665	.688	.722⁺	.624	**.743**
SC	.608	.620	.669⁺	.738	.731	.735	**.763**	.688	.653	.690	**.722**⁺	**.746**⁺	.745	.724	.677	**.773**
SO	.573⁻	.601	.621	.691	.685	.698	**.713**	.644	.614	.648⁻	.650	.696⁻	.703	.696	.659	**.730**

Table 4. Attributes-based performance on our SOC salient objects sub-dataset. For each model, the score corresponds to the average structure similarity M_S (in Sec. 4.1) over all datasets with that specific attribute (*e.g.*, CL). The higher score the better the performance. The best result is highlighted in **bold**. The average salient-object performance S_{sal} is presented in the first row using the structure similarity S. The symbol of $^+$ and $^-$ indicates *increase* and *decrease* compared to the average (S_{sal}) result, respectively.

Small Object (SO) is tricky for all SOD models. All models encounter performance degradation (*e.g.*, from DSS [20] -0.3% to LEGS [12] -5.6%), because *SOs* are easily ignored during down-sampling of CNNs. DSS [20] is the only model that has a slight decrease of performance on type *SO*, while it has the biggest (28.9%) loss of performance on type *BO*. MDF [11] uses multi-scale superpixels as the input of network, so it retains the details of small objects well. However, due to the limited size of superpixels, MDF can not efficiently sense the global semantics, causing a big failure on type *BO*.

Occlusions (OC) scenes in which objects are partly obscured. Thus, it requires SOD models to capture global semantics to make up for the incomplete information of objects. To do so, DS [32] & AMU [22] made use of the multi-scale features in the down-sample progress to generate a fused saliency map; UCF [23] proposed an uncertain learning mechanism to learn uncertain convolutional features. All these methods try to get saliency maps containing both global and local features. Unsurprisingly, these methods have achieved pretty good results on type *OC*. Based on the above analyses, we also find that these three models perform very well on the scenes requiring more semantic information like type *AC*, *OV* and *CL*.

Heterogeneous Object (HO) is a common attribute in nature scenes. The performance of different models on type *HO* gets some improvement to their average performances respectively, all fluctuating from 3.9% to 9.7%. We suspect this is because type *HO* accounts for a significant proportion of all datasets, objectively making models more fitting to this attribute. This result in some degree confirms our statistics in Fig. 6.

Motion Blur (MB) often occurs when capturing a moving object, *e.g.*, running dog and sporting person. Note that this attribute just accounts for a relatively small proportion (4.46%) in our dataset, since these circumstances only happen occasionally. But it still poses a challenge to several models. According to the performance of models on type *MB*, this attribute influences the results with a maximum up to 10.5% (*e.g.*, LEGS [12]) compared with S_{sal}. Due to the utilizing of proposal technique, MSR model
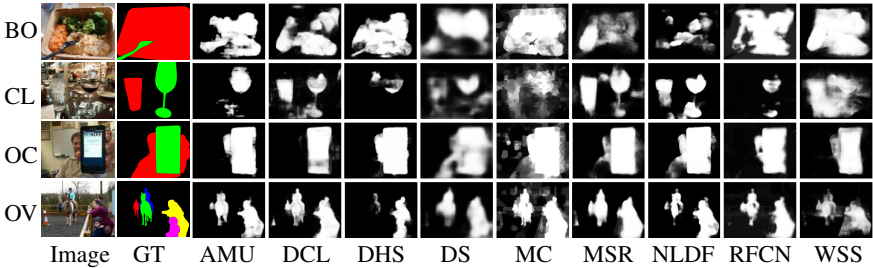
Fig. 7. Visual comparisons between deep models. Each row represents different attributes.

achieves a negligible increase of performance. Some specific applications like video processing can be a more suitable choice.

Out of view (OV) shares some similarities with *OC*. For example, they all lack complete information about the object. But the difference between them is also obvious. *OV* needs a rough estimate of the shape and size of the object, while *OC* needs to exclude the interference information brought by the object in front. With proper global semantic information produced by the high-level layers of CNNs, models perform well on type *OC* (*e.g.*, UCF [23], AMU [22], DS [32], and MSR [29]) also achieve a good performance on this attribute. On the contrary, models such as MDF [11] based on superpixels still suffer from a large margin of decrease, since they can not fix the missing part.

5   Discussion and Conclusion

To our best knowledge, this work presents the currently largest scale performance evaluation of CNNs based salient object detection models. Our analysis points out a serious *data selection bias* in existing SOD datasets. This design bias has lead to state-of-the-art SOD algorithms almost achieve saturated high performance when evaluated on existing datasets, but are still far from being satisfactory when applied to real-world daily scenes. Based on our analysis, we first identify 7 important aspects that a comprehensive and balanced dataset should fulfill. We firstly introduces a high quality SOD dataset, **SOC**. It contains salient objects from daily life in their natural environments which reaches closer to realistic settings. The SOC dataset will evolve and grow over time and will enable research possibilities in multiple directions, *e.g.*, salient object subitizing [45], instance level salient object detection [29], weakly supervised based salient object detection [27], *etc*. Then, a set of attributes (*e.g.*, *Appearance Change*) is proposed in the attempt to obtain a deeper insight into the SOD problem, investigate the pros and cons of the SOD algorithms, and objectively assess the model performances over different perspectives/requirements. Finally, we report attribute-based performance assessment on our SOC dataset. The results open up promising future directions for model development and comparison.

We hope that the public availability of this dataset and the identified areas for potential future works can attract more interests in such an active and fundamentally important field of SOD.

Acknowledgments

References

1. Borji, A., Frintrop, S., Sihite, D.N., Itti, L.: Adaptive object tracking by learning background context. In: IEEE CVPRW, IEEE (2012) 23–30 1
2. He, J., Feng, J., Liu, X., Cheng, T., Lin, T.H., Chung, H., Chang, S.F.: Mobile product search with bag of hash bits and boundary reranking. In: CVPR, IEEE (2012) 3005–3012 1
3. Liu, T., Sun, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: CVPR, IEEE (2007) 1–8 1, 3, 5, 7
4. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: CVPR, IEEE (2007) 1–8 1, 3, 7
5. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: ICCV. Volume 2., IEEE (2001) 416–423 1, 6, 7
6. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H.S., Hu, S.M.: Global contrast based salient region detection. IEEE TPAMI **37**(3) (2015) 569–582 1, 3, 5, 7
7. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: a benchmark. In: ECCV, Springer-Verlag (2012) 414–429 1, 4
8. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 1155–1162 1, 4, 7
9. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR, IEEE (2013) 1, 4, 5
10. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR, IEEE (2014) 280–287 1, 4, 5, 6, 7
11. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR, IEEE (2015) 5455–5463 1, 3, 4, 5, 7, 11, 12, 13, 14
12. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: CVPR, IEEE (2015) 3183–3192 3, 4, 5, 11, 13
13. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: CVPR, IEEE (2015) 1265–1274 3, 4, 5, 11, 12, 13
14. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: CVPR, IEEE (2016) 3, 4, 5, 11, 12, 13
15. Gayoung, L., Yu-Wing, T., Junmo, K.: Deep saliency with encoded low level distance map and high level features. In: CVPR, IEEE (2016) 3, 4, 5, 11, 12, 13
16. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: CVPR, IEEE (2016) 678–686 3, 4, 5, 11, 13
17. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: European Conference on Computer Vision, Springer (2016) 825–841 3, 4, 5, 11, 12, 13
18. Chen, T., Lin, L., Liu, L., Luo, X., Li, X.: Disc: Deep image saliency computing via progressive representation learning. IEEE transactions on neural networks and learning systems **27**(6) (2016) 1135–1149 3, 4, 5, 11, 12, 13
19. Zhang, J., Dai, Y., Porikli, F.: Deep salient object detection by integrating multi-level cues. In: Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on, IEEE (2017) 1–10 3, 4, 5, 11, 12, 13

20. Hou, Q., Cheng, M.M., Hu, X.W., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: CVPR, IEEE (2017) 3, 4, 5, 11, 12, 13

21. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: CVPR, IEEE (July 2017) 3, 4, 5, 11, 12, 13

22. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: ICCV, IEEE (2017) 3, 4, 5, 12, 13, 14

23. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: ICCV. (2017) 3, 5, 11, 12, 13, 14

24. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV, Springer (2014) 740–755 3, 6, 7, 8, 9

25. Achanta, R., Hemami, S., Estrada, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: CVPR, IEEE (2009) 3, 7, 11

26. Xia, C., Li, J., Chen, X., Zheng, A., Zhang, Y.: What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In: CVPR, IEEE (2017) 4, 6, 7

27. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR, IEEE (2017) 136–145 4, 5, 7, 11, 12, 13, 14

28. Jiang, H., Cheng, M.M., Li, S.J., Borji, A., Wang, J.: Joint salient object detection and existence prediction. Front. Comput. Sci (2017) 4

29. Li, G., Xie, Y., Lin, L., Yu, Y.: Instance-level salient object segmentation. In: CVPR, IEEE (2017) 4, 5, 7, 8, 9, 11, 12, 13, 14

30. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR, IEEE (2015) 3431–3440 4

31. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV, IEEE (2015) 1395–1403 4

32. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: Multi-task deep neural network model for salient object detection. IEEE TIP **25**(8) (2016) 3919–3930 5, 11, 12, 13, 14

33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision **115**(3) (2015) 211–252 5

34. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results 5

35. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Unconstrained salient object detection via proposal subset optimization. In: CVPR, IEEE (2016) 5733–5742 5

36. Caesar, H., Uijlings, J., Ferrari, V.: COCO-Stuff: Thing and stuff classes in context. arXiv preprint arXiv:1612.03716 (2016) 6

37. Lazebnik, S., Schmid, C., Ponce, J.: A sparse texture representation using local affine regions. IEEE TPAMI **27**(8) (2005) 1265–1278 6

38. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Transactions on Image Processing **24**(12) (2015) 5706–5722 7

39. Judd, T., Durand, F., Torralba, A.: A benchmark of computational models of saliency to predict human fixations. In: MIT Technical Report. (2012) 7

40. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. IEEE TPAMI (2018) 7

41. Jiang, H., Cheng, M.M., Li, S.J., Borji, A., Wang, J.: Joint Salient Object Detection and Existence Prediction. Front. Comput. Sci. (2017) 7

42. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR, IEEE (2016) 724–732 9

43. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV, IEEE (2017) 11
44. Chen, T., Cheng, M.M., Tan, P., Shamir, A., Hu, S.M.: Sketch2photo: Internet image montage. ACM Transactions on Graphics (TOG) **28**(5) (2009) 124 12
45. Zhang, J., Ma, S., Sameki, M., Sclaroff, S., Betke, M., Lin, Z., Shen, X., Price, B., Mech, R.: Salient object subitizing. In: CVPR, IEEE (2015) 4045–4054 14