

What is and What is not a Salient Object?

Learning Salient Object Detector by Ensembling Linear Exemplar Regressors

Changqun Xia¹, Jia Li^{1,2}, Xiaowu Chen^{1*}, Anlin Zheng¹, Yu Zhang¹

¹State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University

²International Research Institute for Multidisciplinary Science, Beihang University

Abstract

Finding what is and what is not a salient object can be helpful in developing better features and models in salient object detection (SOD). In this paper, we investigate the images that are selected and discarded in constructing a new SOD dataset and find that many similar candidates, complex shape and low objectness are three main attributes of many non-salient objects. Moreover, objects may have diversified attributes that make them salient. As a result, we propose a novel salient object detector by ensembling linear exemplar regressors. We first select reliable foreground and background seeds using the boundary prior and then adopt locally linear embedding (LLE) to conduct manifold-preserving foregroundness propagation. In this manner, a foregroundness map can be generated to roughly pop-out salient objects and suppress non-salient ones with many similar candidates. Moreover, we extract the shape, foregroundness and attention descriptors to characterize the extracted object proposals, and a linear exemplar regressor is trained to encode how to detect salient proposals in a specific image. Finally, various linear exemplar regressors are ensembled to form a single detector that adapts to various scenarios. Extensive experimental results on 5 dataset and the new SOD dataset show that our approach outperforms 9 state-of-art methods.

1. Introduction

Salient object detection (SOD) is a fundamental problem that attracts increasing research interests [16, 35, 36]. In SOD, a key step is to distinguish salient and non-salient objects using the visual attributes. However, in complex scenarios it is often unclear which attributes inherently make an object pop-out and how to separate salient and non-salient objects sharing some attributes (see Fig. 1). As a result, an investigation on what is and what is not a salient object is necessary before developing SOD models.

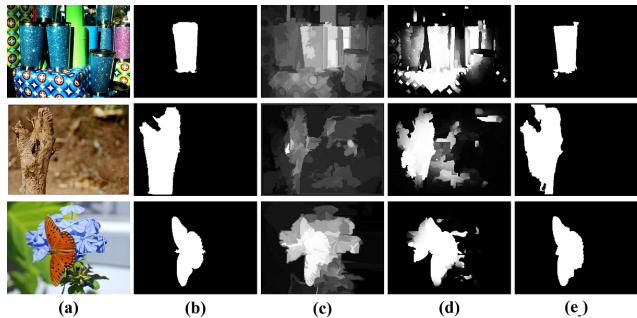


Figure 1. Salient objects can pop-out for having different visual attributes, which may be partially shared with non-salient objects. (a) Images, (b) ground-truth, (c) results of [16], (d) results of [36], (e) results of our approach. The main attributes of salient objects in the three images are location (1st row), shape (2nd row) and color (3rd row), while attributes shared with non-salient objects are shape (1st row), color (2nd row) and location (3rd row).

In the past decade, extensive efforts have been made to find a comprehensive and convincing definition of salient objects. For example, Jiang *et al.* [17] proposed that salient objects are characterized by the uniqueness, focusness, and objectness. In [10], salient objects were considered to be unique and have compact spatial distribution, or be distinctive with respect to both their local and global surroundings [13]. Based on these findings, heuristic features can be designed to identify whether a region [8, 24, 16, 44], a superpixel [19, 14] or a pixel [37, 34, 42] is salient or not. Typically, these models can achieve impressive performance when salient and non-salient objects are remarkably different. However, in complex scenarios that salient and non-salient objects may share some visual attributes, making them difficult to be separated (see Fig. 1). Although such scenarios can be partially addressed by training extremely complex models by using Deep Convolutional Neural Networks [15, 12, 21, 22, 25] or Recurrent Neural Networks [26, 20, 38], such deep models are often difficult to be trained or fine-tuned. Moreover, it is still unclear what visual attributes contribute the most in separating salient and non-salient objects due to the ‘black box’ characteristic of

* Corresponding author: Xiaowu Chen (email: chen@buaa.edu.cn).

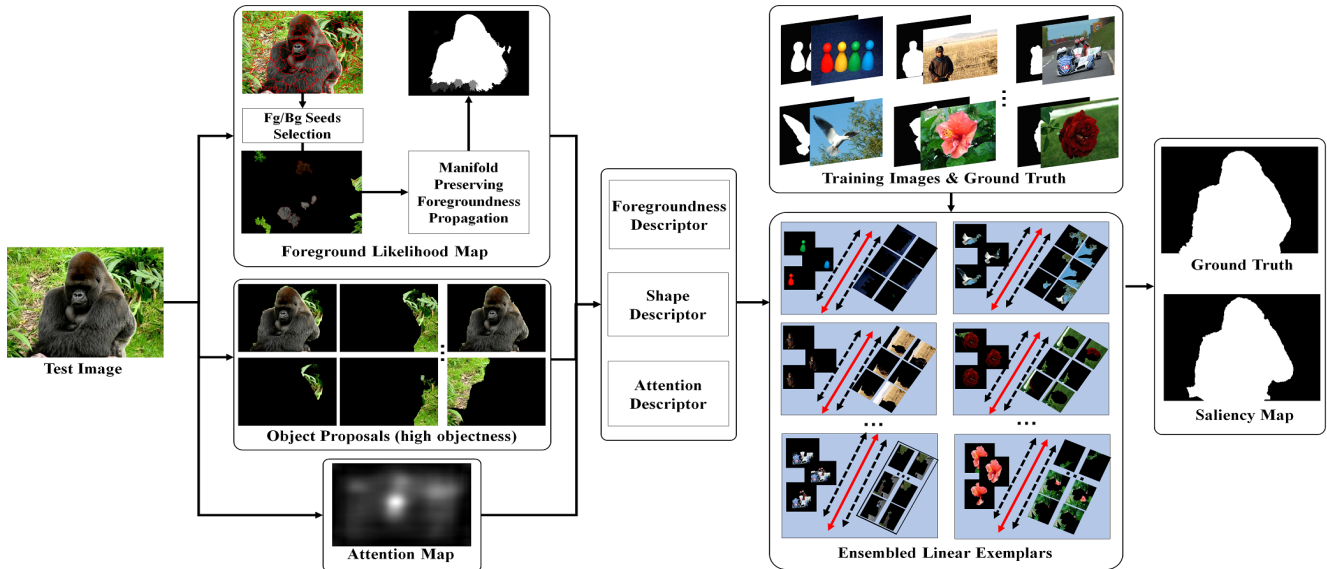


Figure 2. System framework of the proposed approach.

deep models. As a result, finding what is not a salient object is as important as knowing what is a salient object, and the answer can be helpful for designing better features and developing better SOD models.

Toward this end, this paper first constructs a new SOD dataset and performs a comprehensive investigation on the images included or discarded in the construction process so as to find a more precise definition on what is and what is not a salient object. From the discarded images that are considered as ambiguous and confusing in determining the salient objects, we find that *similar candidates*, *complex shape and low objectness* are three major reasons that prevent an object being considered as unambiguously salient. Moreover, from the 10,000 images included in the dataset, we find that objects can become salient in diversified ways that may change remarkably in different scenes, which may imply that a SOD model should adaptively process various kinds of scenes for the perfect detection of salient objects and suppression of distractors.

Inspired by these two findings on what are salient and non-salient objects, we propose a simple yet effective approach for image-based SOD by ensembling plenty of linear exemplar regressors. The system framework of the proposed approach can be found in Fig. 2. Given a testing image, we first divide it into superpixels and extract a set of foreground and background seeds with respect to the boundary prior used in previous works [6, 16, 23, 39, 41]. Moreover, the locally linear embedding (LLE) algorithm is adopted to discover the relationship between each superpixel and its nearest neighbors in the feature subspace, and such relationships, together with the selected seeds, are used to guide the manifold preserving foregroundness propagation process so as to derive a foregroundness map that is capable

to roughly pop-out salient objects and suppress non-salient ones that have many similar candidates. Moreover, we generate an attention map by using a pre-trained deep fixation prediction model [29] and extract a set of object proposals with high objectness from the input image by using the Multiscale Combinatorial Grouping algorithm [4]. With the foregroundness and attention maps, the testing image can be characterized by its shape, attention and foregroundness descriptor, and such descriptor are then delivered into a salient object detector formed by ensembling various linear exemplar regressors so as to detect the salient proposals and suppress the non-salient ones. Note that each linear exemplar regressor is trained on a specific training image by using the same proposal descriptor, while each regressor encodes a specific way of separating salient objects from non-salient ones. As a result, their fusion can adaptively handle the SOD tasks in various scenarios, and the usage of shape descriptor and high-objectness proposals ensure the well suppression of non-salient objects. Extensive experimental results show that the proposed approach outperforms 9 state-of-the-art methods on 5 datasets and our new dataset.

The main contributions of this paper are summarized as follows: 1) We introduce a large SOD dataset with 10,000 images, which we promise to release so as to provide an additional source for training and testing SOD models; 2) We conduct an investigation on what is and what is not a salient object in constructing the dataset, based on which an effective salient object detector is proposed by ensembling linear exemplar regressors; and 3) we propose to compute the foregroundness map by using boundary prior and LLE, which is an effective cue in popping-out salient objects and suppress non-salient ones.



Figure 3. Images and annotations in XPIE. (a) Scene complexity: simple (top) to complex (bottom), (b) Number of salient objects: single (top) to multiple (bottom), (c) Object size: small (top) to large (bottom), (d) Object position: center (top) to border (bottom).

2. What is and What is Not a Salient Object

To obtain a comprehensive explanation on what is and what is not a salient object, a feasible solution is to investigate the whole process in constructing a new SOD dataset by observing the main characteristics of objects in images included into or discarded from the dataset. From these observations, we can infer the key attributes of salient and non-salient objects as well as the subjective bias that may inherently exist in image-based SOD datasets.

Toward this end, we construct a large SOD dataset (denoted as XPIE) and record all the details in the construction process. We first collect three kinds of images from three sources, including Panoramio [30], ImageNet [11], and two fixation datasets [7, 18]. The collection is fully automatic to avoid bringing in too much subjective bias. After that, we resize each image to have a maximum side length of 300 pixels and discard all gray images as well as the color images with minimum side length less than 128 pixels. Finally, we obtain 29,600 color images in three image subsets, denoted as **Set-P**, **Set-I** and **Set-E**, respectively. **Set-P** contains 8,800 images of places-of-interest with geographic information (e.g., GPS and tag), **Set-I** contains 19,600 images with object tags, and **Set-E** contains 1,200 images with human fixations.

Given these images, we ask two engineers to annotate them through two stages. In the first stage, an image is assigned a binary tag: ‘Yes’ for containing unambiguous salient objects, and ‘No’ otherwise. After the first stage, we have 21,002 images tagged with “Yes” and 8,598 images tagged with “No.” In the second stage, these two engineers are further asked to manually label the accurate boundaries of salient objects in 10,000 images tagged with “Yes.” Note that we have 10 volunteers involved in the whole process for cross-check the quality of annotations. Finally, we obtain the binary masks for 10,000 images. More statistics can be found in Table 1. As shown in Fig. 3, images in XPIE cover a variety of simple and complex scenes with different numbers, sizes and positions of salient objects. Thus XPIE can be used as an additional training/testing source in developing and benchmarking SOD models.

Table 1. Statistics of the proposed XPIE dataset

	Set-P	Set-I	Set-E	XPIE
# Candidates	8,800	19,600	1,200	29,600
# Yes	2,433	17,875	694	21,002
# No	6,367	1,725	506	8,598
# Annotated	625	8,799	576	10,000

By observing the 10,000 images tagged with ‘Yes’ and 8,598 images tagged with “No” as well as the explanations from engineers and volunteers, we conclude three key reasons that an object is considered to be non-salient:

1) Many similar candidates. Some images contain a lot of candidate objects (*i.e.*, five or more in this study, see Fig. 4(a)) and it is very difficult to determine which ones are the most salient. In other words, several label ambiguity may inevitably arise when different subjects manually annotate the saliency objects in such images. Although such ambiguity can be alleviated by incorporating eye-tracking apparatus [24], the issue is still far from be addressed.

2) Complex shape. The more complex the shape of a object is, the less probably it is considered as salient. Note that an object may have complex shape for having fuzzy or complex boundaries or being occluded by some other objects (see Fig. 4 (b)). We also find these phenomena exist in other datasets like DUT-OMRON [41] and PASCAL-S [24].

3) Low objectness. Sometimes the most salient region is considered to be not an ‘object’ due to its semantic attributes (*e.g.*, the rock and road in Fig. 4 (c)). This may be caused by the fact that objects within these semantic categories often act as background in other images.

From these three reasons, we can also derive a definition of salient object that can be combined with previous definitions [5]. That is, a salient object should have a limited similar distractors, relatively clear and simple shape and high objectness. Moreover, we find that salient objects can pop-out for having specific visual attributes in different scenarios, which implies that a good SOD model should encode all probable ways that salient objects differ from non-salient ones and adaptively process all types of scenarios.

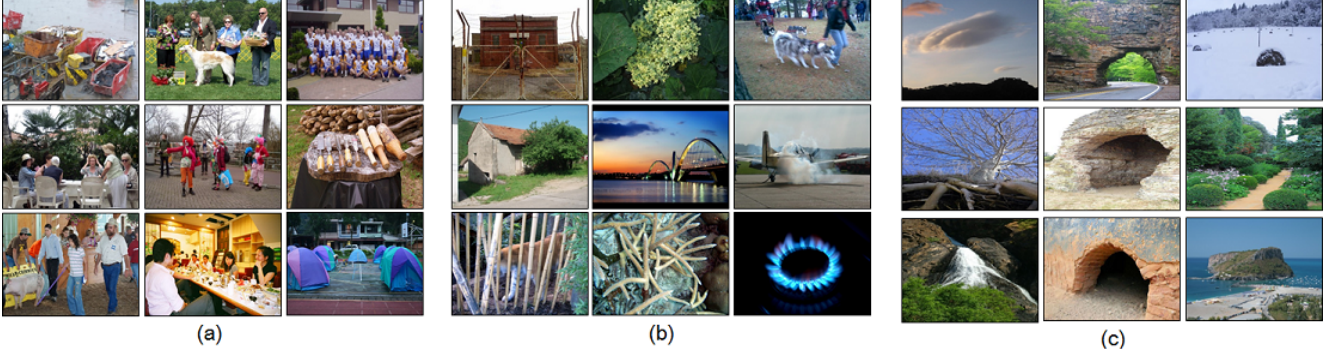


Figure 4. Three reasons that an object is considered as non-salient. (a) Many similar candidates, (b) Complex shape, (c) Low objectness.

3. Estimating Foregroundness Map via Manifold Preserving Propagation

As discussed in Sect. 2, objects with many similar candidates in the same image are likely to be non-salient. In other words, the inter-object similarities between objects are useful cues in separating salient and non-salient objects. Inspired by this fact, we propose to estimate a foregroundness map to depict where salient objects may reside by using such inter-object similarity. Toward this end, we first divide an image into a set of nearly regular superpixels using the SLIC method [2], and the number of superpixels, denoted as N , is empirically set to 200. For a superpixel \mathcal{S}_i , we represent it with its CIE LAB color vector \mathbf{c}_i and mean position vector \mathbf{p}_i .

Foreground and Background Seeds Selection. Given the features $\{\mathbf{c}_i, \mathbf{p}_i\}$, we aim to generate a set of highly reliable foreground and background seeds. For the sake of simplification, we adopt an indicator vector $\mathbf{y} = [y_1, \dots, y_N]^T$ whose component $y_i \in [0, 1]$ corresponds to the foregroundness of \mathcal{S}_i . To estimate \mathbf{y} , we first refer to the boundary prior widely, which assume that regions along the image boundary are more likely to be the background. Inspired by that, we initialize the indicator y_i to 0 if it falls at image boundary and 1 otherwise. After that, the refined foregroundness indicator vector $\hat{\mathbf{y}}$ can be updated by solving

$$\min_{\hat{\mathbf{y}}} \sum_{i=1}^N \|\hat{y}_i - y_i\|_2^2 + \lambda_\mu \sum_{i=1}^N \sum_{j \in \mathcal{N}_i^1} \alpha_{ij} (\hat{y}_i - \hat{y}_j)^2, \quad (1)$$

$$\text{s.t. } 0 \preceq \hat{\mathbf{y}} \preceq 1,$$

where \mathcal{N}_i^1 indicates the indices of superpixels that are adjacent to \mathcal{S}_i . λ_μ is a constant to incorporate the second smoothness term that enforce similar foregroundness scores at spatially adjacent superpixels. α_{ij} is a positive weight that measures the color similarity between \mathcal{S}_i and \mathcal{S}_j :

$$\alpha_{ij} = \exp\left(-\frac{\|\mathbf{c}_i - \mathbf{c}_j\|_2^2}{\sigma^2}\right). \quad (2)$$

Considering that (1) only consists of quadratic and linear terms with linear constraints, we can efficiently solving such a quadratic programming problem by using gradient descent algorithm. Note that only the color difference is considered in (2) since we aim to suppress all probable background regions that have similar color to the boundary regions. In practice, we initialize \mathbf{y} four times with the left, top, right and bottom boundaries, respectively. Let $\hat{\mathbf{y}}^l$, $\hat{\mathbf{y}}^t$, $\hat{\mathbf{y}}^r$ and $\hat{\mathbf{y}}^b$ be the refined foregroundness indicator, we can derive the i th component in the final indicator vector as

$$\hat{y}_i^* = \hat{\mathbf{y}}_i^l \cdot \hat{\mathbf{y}}_i^t \cdot \hat{\mathbf{y}}_i^r \cdot \hat{\mathbf{y}}_i^b. \quad (3)$$

Based on this $\hat{\mathbf{y}}_i^*$, we adopt two predefined thresholds, T_{high} and T_{low} , to get the most reliable foreground/background seeds. That is, foreground seeds are selected as $\hat{y}_i^* > T_{high}$ and background seeds with $\hat{y}_i^* < T_{low}$. In experiments, we empirically set T_{high} to be twice the mean foregroundness score of $\hat{\mathbf{y}}$ and T_{low} to be 0.05.

Manifold-preserving Foregroundness Propagation. In selecting foreground seeds, some non-salient objects will pop-out as well since we only use the simple color contrast. Recall that non-salient objects often have many similar candidates, we propose to further derive a foregroundness map via manifold preserving foreground propagation. Different from the seed selection step, we adopt the locally linear embedding (LLE) scheme to guide the propagation process. As shown in Fig. 5, we aim to maintain the same relationships (*i.e.*, color and location) between a superpixel and its nearest neighbors in the generated foregroundness map. In this manner, large salient objects can pop-out as a whole (see Fig. 5).

To model the spatial relationship between superpixels, we need to solve the problem

$$\min_{\{w_{ij}\}} \sum_{i=1}^N \|\mathbf{c}_i - \sum_{j \in \mathcal{N}_i} w_{ij} \mathbf{c}_j\|_2^2 + \|\mathbf{p}_i - \sum_{j \in \mathcal{N}_i^K} w_{ij} \mathbf{p}_j\|_2^2, \quad (4)$$

$$\text{s.t. } \sum_{j \in \mathcal{N}_i} w_{ij} = 1, \forall i = 1, 2, \dots, N,$$

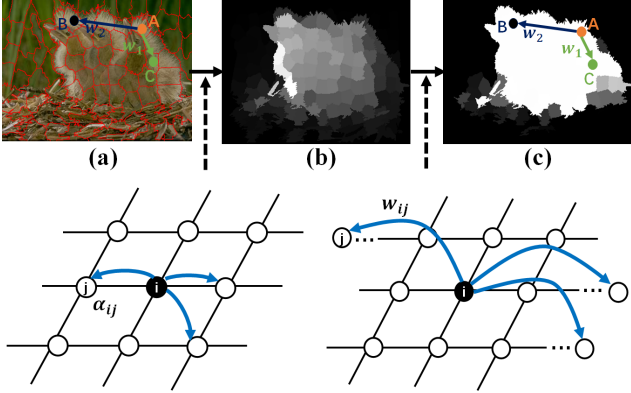


Figure 5. Foregroundness map estimation via manifold preserving foregroundness propagation. (a) Input image, (b) Initialized foregroundness map for foreground/background seed selection, (c) final foregroundness map.

where \mathcal{N}_i^K is the indices of the K nearest neighbors of \mathcal{S}_i , and we set $K = 5$ in experiments. In practice, we can optimize (4) using the method described in [32]. As a result, we can obtain a $N \times N$ relationship matrix $\mathbf{W} = [w_{ij}]$ that captures the manifold structure of all superpixels in the feature space. Based on this matrix, the foregroundness can be propagated as

$$\begin{aligned} \min_{\bar{\mathbf{y}}} \|\bar{\mathbf{y}} - \mathbf{W}\bar{\mathbf{y}}\|_2^2 + \lambda_{le} \sum_{i \in \mathbb{S}} (\bar{y}_i - g_i)^2 \\ \text{s.t. } 0 \preceq \bar{\mathbf{y}} \preceq 1, \end{aligned} \quad (5)$$

where \mathbb{S} is the indices of the selected foreground and background seeds. g_i is an indicator which equals to 1 or 0 if \mathcal{S}_i is the selected foreground seed and background seed, respectively. \bar{y}_i is the i th component of foregroundness vector $\bar{\mathbf{y}}$. In (4), the first term enforces the manifold-preserving foregroundness propagation, and the second term ensures the final foregroundness to be consistent with the seed selection results. λ_{le} controls the balance of the first term and the second term. The two terms are all squared errors. Thus, we can obtain the foregroundness vector by least-square algorithms. Finally, the foregroundness vector is converted to a foregroundness map by assigning the superpixel-based foregroundness to all pixels it contains.

4. Learning a Salient Object Detector by Ensembling Linear Exemplar Regressors

4.1. Training Linear Exemplar Regressors

Given the foregroundness map, we can train a salient object detector by ensembling linear exemplar regressors. Let \mathbb{I} be the set of training images, G be the ground-truth mask of an image $\mathcal{I} \in \mathbb{I}$. Inspired by the annotation process of salient objects in [24], we extract a set of object proposals from \mathcal{I} by using the Multiscale Combinatorial Grouping

algorithm [4], which are denoted as \mathbb{O} . Moreover, we use the fixation prediction model proposed in [29] to predict an attention map (*i.e.*, a fixation density map) that reveals the most attractive regions. Furthermore, we derive a ground-truth saliency of a proposal $\mathcal{O} \in \mathbb{O}_{\mathcal{I}}$ as

$$G(\mathcal{O}) = \frac{1}{|\mathcal{O}|} \sum_{p \in \mathcal{O}} G(p), \quad (6)$$

where p is a pixel in \mathcal{O} . In training the detectors, we only select positives from \mathcal{I} with $G(\mathcal{O}) > 0.7$ and negatives with $G(\mathcal{O}) < 0.3$, which are denoted as $\mathbb{O}_{\mathcal{I}}^+$ and $\mathbb{O}_{\mathcal{I}}^-$, respectively. After that, we represent each object proposal \mathcal{O} in \mathbb{O}^+ and \mathbb{O}^- with a heuristic feature vector $\mathbf{v}_{\mathcal{O}}$, including 14d shape descriptor proposed in [4] and additional 11d shape descriptor like center of gravity, width-height-ratio, orientation, eccentricity, etc., 27d foregroundness descriptor based on the foregroundness map and 27d attention descriptor based on the attention map (as in [24]). Finally, each object proposal \mathcal{O} is represented by a 79d descriptor $\mathbf{v}_{\mathcal{O}}$.

With these feature vectors, a linear exemplar regressor $\phi(\mathbf{v})$ can be trained to separate salient and non-salient objects on a specific image by minimizing

$$\begin{aligned} \min_{\phi} \frac{1}{2} \|\mathbf{w}\|_2^2 + C^+ \sum_{\mathcal{O} \in \mathbb{O}^+} \zeta_{\mathcal{O}} + C^- \sum_{\mathcal{O} \in \mathbb{O}^-} \zeta_{\mathcal{O}}, \\ \text{s.t. } \forall \mathcal{O} \in \mathbb{O}^+, \mathbf{w}^T \mathbf{v}_{\mathcal{O}} + b \geq 1 - \zeta_{\mathcal{O}}, \zeta_{\mathcal{O}} \geq 0, \\ \forall \mathcal{O} \in \mathbb{O}^-, \mathbf{w}^T \mathbf{v}_{\mathcal{O}} + b \leq \zeta_{\mathcal{O}} - 1, \zeta_{\mathcal{O}} \geq 0, \end{aligned} \quad (7)$$

where C^+ and C^- are empirically set to $1/|\mathbb{O}^+|$ and $1/|\mathbb{O}^-|$ to balance the influence of probable salient objects and distractors. \mathbf{w} and b are the parameters of $\phi(\mathbf{v})$.

4.2. Ensembling for Salient Object Detection

Given all linear exemplar regressors, a proposal \mathcal{O} in a testing image gains $|\mathbb{I}|$ saliency scores, denoted as $\{\phi_{\mathcal{I}}(\mathbf{v}_{\mathcal{O}}) | \mathcal{I} \in \mathbb{I}\}$. However, the scores of each linear exemplar regressor may fall in different dynamic ranges so that their direct fusion will lead to inaccurate saliency maps (see Fig. 6(c)-(d)).

To fuse these predictions, we propose an enhancement operation on $\{\phi_{\mathcal{I}}(\mathcal{O})\}$, which increases the probability to adopt the most relevant linear exemplar regressors and decreases the influence of irrelevant ones. The enhancement adopts a sigmoid function as

$$f(x) = \frac{1}{1 + e^{-a(x-b)}} \quad (8)$$

where $x \in \{\phi_{\mathcal{I}}(\mathcal{O})\}$. a, b are predefined parameters to control the degree of enhancement so that regressors that output intermediate scores will be inhibited (*e.g.*, $x = 0.5$), while regressors with high and low scores will be maintained (*e.g.*,

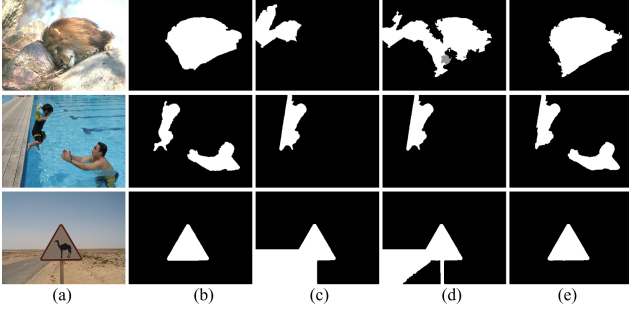


Figure 6. Different fusion strategy for SOD. (a) Image, (b) ground-truth, (c) direct fusion by computing the maximum saliency value, (d) direct fusion by computing the mean saliency value, (e) enhanced fusion by computing the mean saliency value after an enhancement operation using a sigmoid function.

$x = 0.0$ or $x = 1.0$). In this manner, we can calibrate all regressors and emphasize more on the ones that output confident predictions on an object proposal. In this manner, regressors with the capability of separating salient and non-salient objects in this scene can be emphasized, making the ensemble model scene adaptive. Finally, we sum up the enhanced saliency scores to derive the saliency at a pixel p :

$$Sal(p) = \frac{1}{|\mathcal{O}|} \sum_{\mathcal{O} \in \mathcal{O}} \xi(p \in \mathcal{O}) \cdot \sum_{\mathcal{I} \in \mathbb{I}} f(\phi_{\mathcal{I}}(\mathbf{v}_{\mathcal{O}})), \quad (9)$$

where $\xi(p \in \mathcal{O})$ is an indicator function which equals to 1 if $p \in \mathcal{O}$ and 0 otherwise. After that, we normalize the pixel-wise saliency into the dynamic range of $[0, 1]$ and adopt the exponential operations proposed in [42] to enhance the contrast of saliency maps, followed by a morphological operation to obtain a smooth salient object (see Fig. 6).

5. Experiments

In experiments, we compare our approach, denoted as ELE (*i.e.*, Ensembling Linear Exemplars), with 9 state-of-the-art methods, including MST [36], BL [35], BSCA [31], MBS [42], RBD [43], DRFI [16], MR [41], DSR [23] and GS [39]. Moreover, we extend ELE by replacing the negative instance used in training each linear exemplar regressor with a huge bag of negatives collected from all training images, and the extend model is denoted as ELE+.

For the 11 models, comparisons are conducted over 5 public datasets and our new dataset **XPIE**, including:

- 1) **SED1** [3] contains 100 images with a dominant salient object per image.
- 2) **PASCAL-S** [24] contains 850 natural images annotated with pre-segmented objects and eye-tracking data.
- 3) **ECSSD** [40] contains 1,000 structurally complex images manually annotated by 5 subjects.

4) **MSRA-B** [27] contains 5,000 images and we select 2,500 images for training and 500 images for validation in training linear exemplar regressors. This is also the same setting used in [16]. Only the rest 2000 images are used for testing all the models.

5) **THUR15K** [9] contains 6,233 images about butterfly, coffee mug, dog jump, giraffe, and plane.

6) **XPIE** contains 10,000 images with pixel-wise masks of salient objects. It covers many complex scenes with different numbers, sizes and positions of salient objects.

In the comparisons, we adopt the F-measure curves, adaptive F-measure, weighted F-measure and mean absolute error (MAE) as the evaluation metrics. In computing F-measure curves, the precision and recall are first computed by binarizing the saliency maps with a threshold sliding from 0 to 255 and compare the binary maps with ground-truth maps. At each threshold, F-measure is computed as

$$F\text{-measure} = \frac{(1 + \beta^2) \cdot \text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}, \quad (10)$$

where β is set to 0.3 as in [1]. Besides, we report adaptive F-measure (FM) using an adaptive threshold for generating a binary saliency map. The adaptive threshold is computed as twice the mean of a saliency map. Meanwhile, the unified weighted F-measure (wFM, see [28]) is computed to reflect the overall performance. In addition, MAE is calculated as the average absolute per-pixel difference between the gray-scale saliency maps and the ground-truth saliency maps.

5.1. Comparison with State-of-the-art Methods

The performance scores of our approaches and the other 9 methods are shown in Table 2. The curves of F-measure are shown in Fig. 8, and we also show some representative results in Fig. 7. From these results, we find that ELE and ELE+ outperform the other 9 approaches on all datasets and achieve the lowest MAE and the highest wFM.

The success of our approach can be explained from three aspects. First, the investigation about what is and what is not a salient object provides useful cues in designing effective features for separating salient and non-salient objects. In particular, the foregroundness map computed via manifold preserving propagation can inhibit non-salient objects with many similar candidates. Second, with the finding that non-salient objects have low objectness, we extract object proposals containing high objectness that are more tightly correlated with the semantic attributes of objects. Actually, the proposal-based framework is similar to the way that human perceives salient objects. Third, various ways of separating salient and non-salient objects in diversified scenes are isomorphically represented with the exemplar-based linear regressors. The enhancement-based fusion strategy for combining exemplar scores makes the learned salient detector emphasize more on the most relevant linear exemplar

Table 2. Comparison of quantitative results including adaptive F-measure (FM, larger is better), weighted F-measure (wFM, larger is better) and MAE(smaller is better). The best three results are shown in red, blue, and green, respectively.

Dataset	Metric	GS	DSR	MR	DRFI	RBD	MBS	BSCA	BL	MST	ELE	ELE+
SED1	MAE	0.176	0.160	0.153	0.164	0.143	0.172	0.154	0.185	0.134	0.105	0.108
	FM	0.751	0.803	0.820	0.817	0.803	0.805	0.819	0.808	0.804	0.849	0.854
	wFM	0.571	0.616	0.626	0.598	0.652	0.579	0.611	0.555	0.697	0.773	0.767
PASCAL-S	MAE	0.221	0.205	0.221	0.219	0.199	0.220	0.222	0.247	0.184	0.161	0.159
	FM	0.601	0.628	0.649	0.666	0.639	0.642	0.640	0.632	0.660	0.692	0.710
	wFM	0.414	0.420	0.416	0.418	0.449	0.366	0.432	0.400	0.527	0.576	0.581
ECSSD	MAE	0.255	0.226	0.237	0.239	0.225	0.245	0.233	0.262	0.208	0.183	0.183
	FM	0.624	0.687	0.697	0.717	0.679	0.668	0.697	0.691	0.688	0.740	0.749
	wFM	0.435	0.490	0.478	0.472	0.490	0.432	0.495	0.449	0.578	0.649	0.650
MSRA-B	MAE	0.144	0.119	0.127	0.139	0.110	0.140	0.130	0.171	0.094	0.069	0.071
	FM	0.754	0.795	0.826	0.828	0.811	0.801	0.809	0.807	0.812	0.853	0.861
	wFM	0.563	0.629	0.613	0.570	0.650	0.548	0.601	0.520	0.725	0.797	0.795
THUR15K	MAE	0.176	0.142	0.178	0.167	0.150	0.159	0.182	0.220	0.141	0.121	0.111
	FM	0.518	0.579	0.586	0.615	0.566	0.595	0.574	0.575	0.607	0.630	0.654
	wFM	0.370	0.422	0.378	0.399	0.421	0.366	0.387	0.341	0.499	0.549	0.567
XPIE	MAE	0.181	0.155	0.177	0.160	0.149	0.162	0.181	0.213	0.146	0.128	0.121
	FM	0.612	0.655	0.672	0.698	0.665	0.670	0.658	0.653	0.675	0.701	0.722
	wFM	0.435	0.476	0.451	0.478	0.503	0.515	0.455	0.407	0.560	0.603	0.611

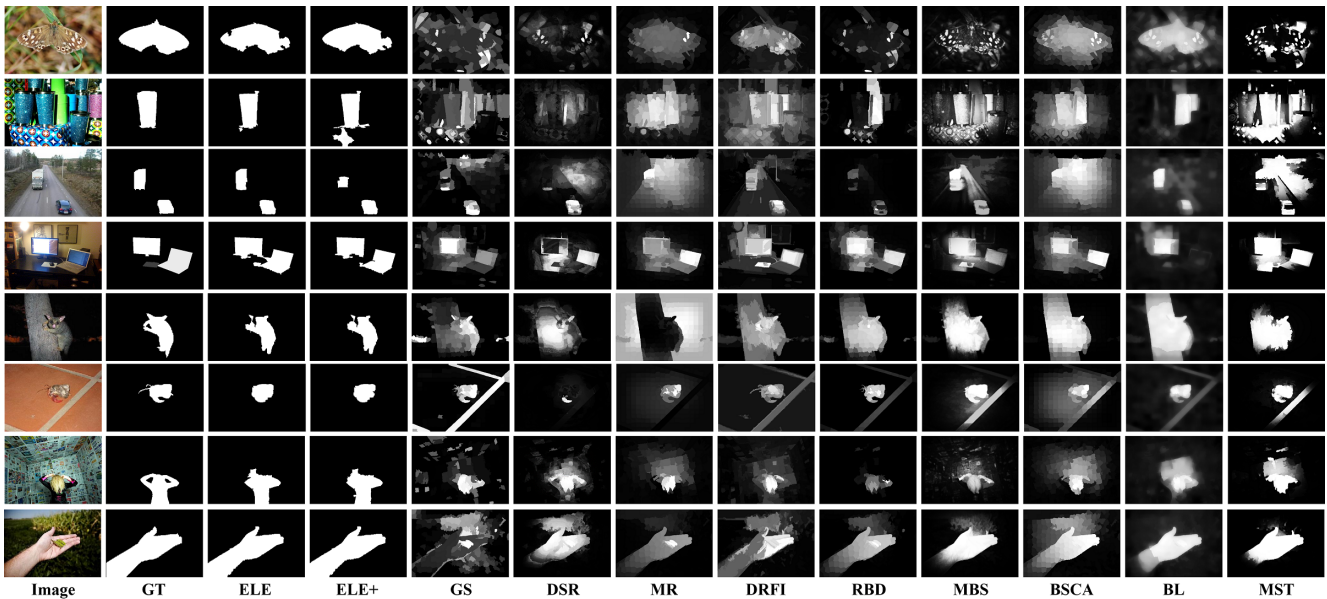


Figure 7. Representative results of our approach (ELE and ELE+) and 9 state-of-the-art methods.

regressors, while the regressors making ambiguous predictions will be somehow ignored. In this manner, the saliency maps become less noisy.

5.2. Performance Analysis

We conduct four small experiments on the 1000 testing images of ECSSD to further validate the effectiveness of the proposed approach.

In the first experiment, we check the effectiveness of the generated foregroundness maps by treating them as saliency maps. In this experiment, we compare the foregroundness maps with results from similar models like MST [36], RBD [43] and MR [41]. We find that such foregroundness maps have a best weighted F-measure of 0.602 among them, while the weighted F-measure scores of MST, RBD and MR reach only 0.578, 0.490 and 0.478, respectively. This val-

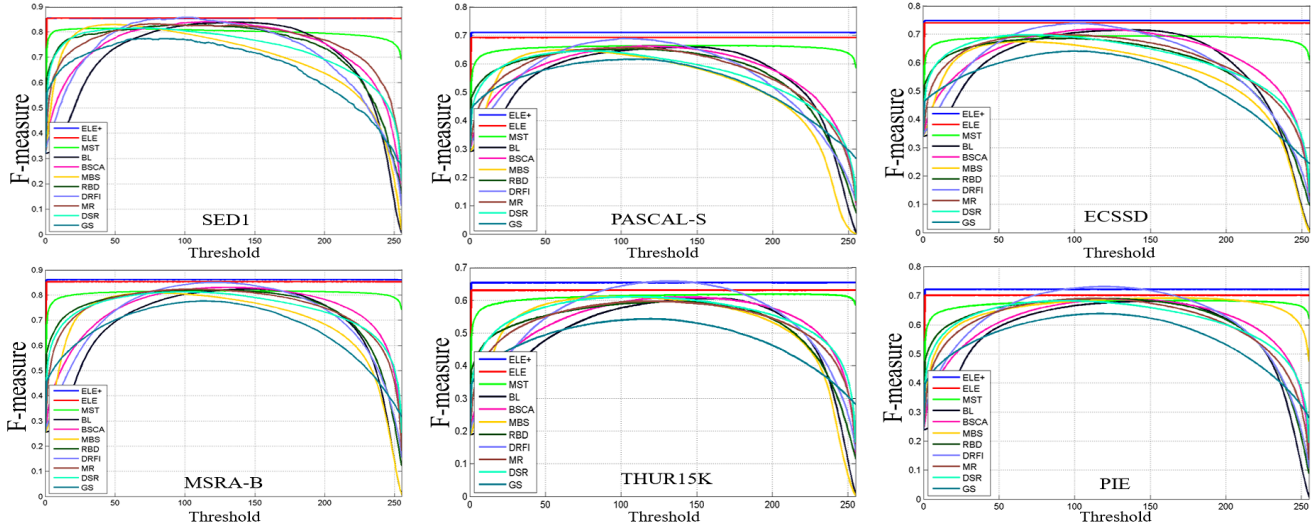


Figure 8. The F-measure curves of our approaches (ELE and ELE+) and 9 state-of-the-art methods over 6 datasets (larger is better).

indicates that the proposed foregroundness maps are very effective and contribute a lot to the impressive performance of ELE and ELE+.

In the second experiment, we compare the heuristic features and deep features extracted by deep models. In this experiment, we use GoogleNet in [33] to compute an 1024 feature descriptor for each proposal. We find that with the same training processing, the weighted F-measure of ELE drops sharply from 0.649 to 0.578. This may be caused by the over-fitting risk in using the high-dimensional CNN features (e.g., 1024d, 2048d or 4096d), especially when the amount of training data is very small.

In the third experiment, we test various fusion strategies of exemplar scores. We compare 3 different fusion ways as shown in Fig. 6 (b),(c) and (d) respectively. We find that the weighted F-measure of using the max and mean value of raw exemplar scores as the final saliency value of a proposal is 0.540 and 0.588, while the weighted F-measure of using enhancement-based fusion is 0.649. This indicates that the enhancement operation on raw exemplar scores is useful in selecting the most relevant exemplar linear regressors and suppressing irrelevant regressors that tend to output medium responses to all candidates.

In the four experiment, we compare ELE and ELE+. The quantitative results in Table 2 show that using huge negative bags has positive impact to the performance of each linear exemplar regressor. Actually, saliency is a relative concept and with the training data from only one image we can simply infer how to separate specific salient objects from specific non-salient ones. On the contrary, the extended bag of negatives can tell us more about how to separate specific salient objects from massive non-salient ones. In this manner, the generalization ability of the detector can be improved, leading to better performance.

6. Conclusions

Knowing what is and what is not a salient object is important for designing better features and developing better models for image-based SOD. Toward this end, this paper first constructs a new SOD dataset and explores how salient objects in the selected images behave and how the non-salient objects in the discard images look like. By investigating the visual attributes of salient and non-salient objects, we find that non-salient objects often have many similar candidates, complex shape and low objectness, while salient objects can pop-out for having diversified attributes.

Inspired by the two findings derived from the construction process of a new dataset, we compute a novel foregroundness map by manifold-preserving foregroundness propagation, which can be used to extract effective features for separating salient and non-salient objects. Moreover, we train an effective salient object detector by ensembling plenty of linear exemplar regressors. Experimental results show that the proposed detector outperforms 9 state-of-the-art methods.

In the future work, we will extend the proposed approach in two ways. Considering the highly flexible architecture of the ensembling-based framework, we will try to design a scheme with which the detector can gradually evolve by learning from incremental training data. Moreover, more features will be incorporated into the framework, especially the high-level features such as semantic descriptor extracted by pre-trained models.

Acknowledgement. This work was partially supported by the National Natural Science Foundation of China (61532003 & 61672072 & 61325011 & 61421003) and the Fundamental Research Funds for the Central Universities.

References

- [1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.
- [2] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *TPAMI*, 34(11):2274–2282, 2012.
- [3] S. Alpert, M. Galun, A. Brandt, and R. Basri. Image segmentation by probabilistic bottom-up aggregation and cue integration. *TPAMI*, 2012.
- [4] P. Arbeláez, J. Pont-Tuset, J. Barron, F. Marques, and J. Malik. Multiscale combinatorial grouping. In *CVPR*, 2014.
- [5] A. Borji. What is a salient object? a dataset and a baseline model for salient object detection. *TIP*, 2015.
- [6] A. Borji, D. N. Sihite, and L. Itti. Salient object detection: A benchmark. In *ECCV*, 2012.
- [7] N. D. Bruce and J. K. Tsotsos. Saliency based on information maximization. In *NIPS*, pages 155–162, Vancouver, BC, Canada, 2005.
- [8] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai. Fusing generic objectness and visual saliency for salient object detection. In *ICCV*, 2011.
- [9] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu. Salienshape: Group saliency in image collections. *The Visual Computer*, 2014.
- [10] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *CVPR*, 2013.
- [11] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.
- [12] L. Gayoung, T. Yu-Wing, and K. Junmo. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016.
- [13] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *TPAMI*, 2012.
- [14] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang. Saliency propagation from simple to difficult. In *CVPR*, 2015.
- [15] S. He, R. Lau, W. Liu, Z. Huang, and Q. Yang. SuperCNN: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 2015.
- [16] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *CVPR*, 2013.
- [17] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by ufo: Uniqueness, focusness and objectness. In *CVPR*, 2013.
- [18] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.
- [19] J. Kim, D. Han, Y.-W. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *CVPR*, 2014.
- [20] J. Kuen, Z. Wang, and G. Wang. Recurrent attentional networks for saliency detection. In *CVPR*, 2016.
- [21] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015.
- [22] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016.
- [23] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013.
- [24] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.
- [25] X. R. Lijun Wang, Huchuan Lu and M.-H. Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015.
- [26] N. Liu and J. Han. DHSNet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.
- [27] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. *TPAMI*, 2011.
- [28] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014.
- [29] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O’Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, June 2016.
- [30] Panoramio. <http://www.panoramio.com>.
- [31] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *CVPR*, 2015.
- [32] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 2000.
- [33] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. 2015.
- [34] Y. Tian, J. Li, S. Yu, and T. Huang. Learning complementary saliency priors for foreground object segmentation in complex scenes. *IJCV*, 2015.
- [35] N. Tong, H. Lu, X. Ruan, and M.-H. Yang. Salient object detection via bootstrap learning. In *CVPR*, 2015.
- [36] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien. Real-time salient object detection with a minimum spanning tree. 2016.
- [37] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009.
- [38] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, 2016.
- [39] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*. 2012.
- [40] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, 2013.
- [41] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.
- [42] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *ICCV*, 2015.
- [43] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.
- [44] W. Zou and N. Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *ICCV*, 2015.