

Deep Unsupervised Saliency Detection: A Multiple Noisy Labeling Perspective

Jing Zhang^{*1,2}, Tong Zhang^{*2,3}, Yuchao Dai^{†1}, Mehrtash Harandi^{2,3}, and Richard Hartley²

¹Northwestern Polytechnical University, Xi'an, China

²Australian National University, Canberra, Australia

³DATA61,CSIRO, Canberra, Australia

Abstract

The success of current deep saliency detection methods heavily depends on the availability of large-scale supervision in the form of per-pixel labeling. Such supervision, while labor-intensive and not always possible, tends to hinder the generalization ability of the learned models. By contrast, traditional handcrafted features based unsupervised saliency detection methods, even though have been surpassed by the deep supervised methods, are generally dataset-independent and could be applied in the wild. This raises a natural question that “Is it possible to learn saliency maps without using labeled data while improving the generalization ability?”. To this end, we present a novel perspective to unsupervised ¹ saliency detection through learning from multiple noisy labeling generated by “weak” and “noisy” unsupervised handcrafted saliency methods. Our end-to-end deep learning framework for unsupervised saliency detection consists of a latent saliency prediction module and a noise modeling module that work collaboratively and are optimized jointly. Explicit noise modeling enables us to deal with noisy saliency maps in a probabilistic way. Extensive experimental results on various benchmarking datasets show that our model not only outperforms all the unsupervised saliency methods with a large margin but also achieves comparable performance with the recent state-of-the-art supervised deep saliency methods.

1. Introduction

Saliency detection aims at identifying the visually interesting objects in images that are consistent with human perception, which is intrinsic to various vision tasks such as

^{*}These authors contributed equally in this work.

[†]Y. Dai (daiyuchao@nwpu.edu.cn) is the corresponding author.

¹There could be multiple definitions for unsupervised learning, in this paper, we refer unsupervised learning as learning without task-specific human annotations, e.g. dense saliency maps in our task.



Figure 1. Unsupervised saliency learning from weak “noisy” saliency maps. Given an input image x_i and its corresponding unsupervised saliency maps y_i^j , our framework learns the latent saliency map \bar{y}_i by jointly optimizing the saliency prediction module and the noise modeling module. Compared with SBF [35] which also learns from unsupervised saliency but with different strategy, our model achieves better performance.

context-aware image editing [36], image caption generation [31]. Depending on whether human annotations have been used, saliency detection methods can be roughly divided as: unsupervised methods and supervised methods. The former ones compute saliency directly based on various priors (e.g., center prior [9], global contrast prior [6], background connectivity prior [43] and etc.), which are summarized and described with human knowledge. The later ones learn direct mapping from color images to saliency maps by exploiting the availability of large-scale human annotated database.

Building upon the powerful learning capacity of convolutional neural network (CNN), deep supervised saliency detection methods [42, 11, 40] achieve state-of-the-art performances, outperforming the unsupervised methods by a wide margin. The success of these deep saliency methods strongly depend on the availability of large-scale training dataset with pixel-level human annotations, which is not only labor-intensive but also could hinder the generalization ability of the learned network models. By contrast, the unsupervised saliency methods, even though have been outperformed by the deep supervised methods, are generally dataset-independent and could be applied in the wild.

In this paper, we present a novel end-to-end deep learning framework for saliency detection that is free from human annotations, thus “unsupervised” (see Fig. 1 for a visualization). Our framework is built upon existing efficient and effective unsupervised saliency methods and the powerful capacity of deep neural network. The unsupervised saliency methods are formulated with human knowledge and different unsupervised saliency methods exploit different human designed priors for saliency detection. They are noisy (compared with ground truth human annotations) and could have method-specific bias in predicting saliency maps. By utilizing existing unsupervised saliency maps, we are able to remove the need of labor-intensive human annotations, also by jointly learn different priors from multiple unsupervised saliency methods, we are able to get complementary information of those unsupervised saliency.

To effectively leverage these noisy but informative saliency maps, we propose a novel perspective to the problem: *Instead of removing the noise in saliency labeling from unsupervised saliency methods with different fusion strategies [35], we explicitly model the noise in saliency maps.* As illustrated in Fig. 2, our framework consists of two consecutive modules, namely a saliency prediction module that learns the mapping from a color image to the “latent” saliency map based on current noise estimation and the noisy saliency maps, and a noise modeling module that fits the noise in noisy saliency maps and updates the noise estimation in different saliency maps based on updated saliency prediction and the noisy saliency maps. In this way, our method takes advantages of both probabilistic methods and deterministic methods, where the latent saliency prediction module works in a deterministic way while the noise modeling module fits the noise distribution in a probabilistic manner. Experiments suggest that our strategy is very effective and it only takes several rounds² till convergence.

To the best of our knowledge, the idea of considering unsupervised saliency maps as learning from multiple noisy labels is brand new and different from existing unsupervised deep saliency methods (e.g., [35]). Our main contributions can be summarized as:

- 1) We present a novel perspective to unsupervised deep saliency detection, and learn saliency maps from multiple noisy unsupervised saliency methods. We formulate the problem as joint optimization of a latent saliency prediction module and a noise modeling module.
- 2) Our deep saliency model is trained in an end-to-end manner without using any human annotations, leading to an extremely cheap solution.
- 3) Extensive performance evaluation on seven benchmarking datasets show that our framework outperforms ex-

²In our paper, an epoch means a complete pass through all the training data, an iteration means a complete pass through a batch, and a round means an update on noise module.

isting unsupervised methods with a wide margin while achieving comparable results with state-of-the-art deep supervised saliency detection methods [11, 40].

2. Related Work

Depending on whether human annotations are used or not, saliency detection techniques can be roughly grouped as unsupervised and supervised methods. Deep learning based methods are particular examples of the latter one. We will also discuss learning with multiple noisy labels.

2.1. Unsupervised Saliency Detection

Prior to the deep learning revolution, saliency methods mainly relied on different priors and handcrafted features [43, 7, 6, 9]. We refer interested readers to [2] and [3] for surveys and benchmark comparisons. Color contrast prior has been exploited at superpixel level in [6]. Shen and Wu [27] formulated saliency detection as a low-rank matrix decomposition problem by exploiting the sparsity prior for salient objects. Objectness, which highlights the object-like regions, has also been used in [15] to mark the regions that have higher possibilities of being an object. Zhu *et al.* [43] presented a robust background measure, namely “boundary connectivity” along with an optimization framework to measure backgroundness of each superpixel. Building upon the center prior, [9] detects the image regions that represent the scene, especially those that are near image center.

2.2. Supervised Saliency Detection

Conventional supervised techniques, such as [14, 17], formulate saliency detection as a regression problem, and a classifier is trained to assign saliency at pixel or superpixel level. Recently, deep neural networks have been adopted successfully for saliency detection [40, 26, 41, 29, 11, 22, 42, 19, 28, 20, 38, 39, 37]. Deep networks can encode high-level semantic features and hence capture saliency more effectively than both unsupervised saliency methods and non-deep supervised methods. Deep saliency detection methods generally train a deep neural network to assign saliency to each pixel or superpixel. Li and Yu [19] used learned features from an existing CNN model to replace the handcrafted features. Recently, Cheng *et al.* [11] proposed a deep supervised framework with multi-branch short connections embed both high- and low-level features for accurate saliency detection. With the same purpose, a multi-level deep feature aggregation framework is proposed in [40]. A top-down strategy and a loss function which penalizes errors on the edge is presented in [26].

2.3. Learning with Noisy Labels

Though deep techniques are methods of choice in saliency detection, very few studies have explicitly addressed the problem of saliency learning with unreliable

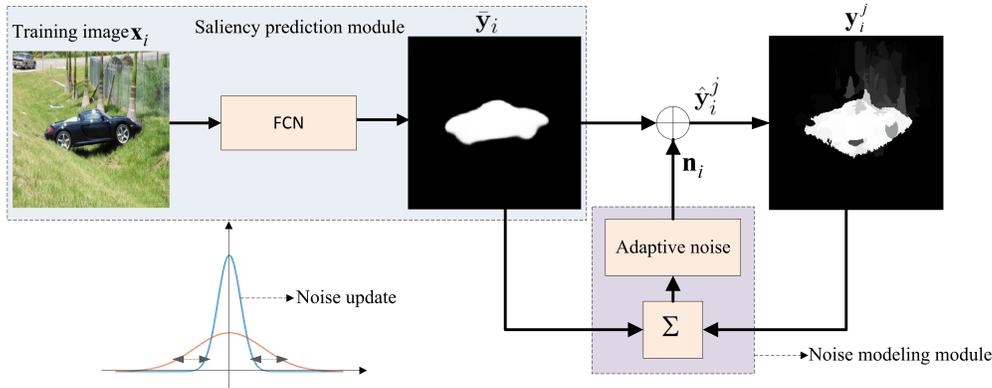


Figure 2. Conceptual illustration of our saliency detection framework, which consists of a “latent” saliency prediction module and a noise modeling module. Given an input image, noisy saliency maps are generated by handcrafted feature based unsupervised saliency detection methods. Our framework jointly optimizes both modules under a unified loss function. The saliency prediction module targets at learning latent saliency maps based on current noise estimation and the noisy saliency maps. The noise modeling module updates the noise estimation in different saliency maps based on updated saliency prediction and the noisy saliency maps. In our experiments, the overall optimization converges in several rounds.

and noisy labels [35]. Learning with noisy labels is mainly about learning classification models in the presence of inaccurate class labels. Whitehill *et al.* [30] solved the problem of picking the correct label based on the labels provided by many labelers with different expertise. Jindal *et al.* [16] proposed a dropout-regularized noise model by augmenting existing deep network with a noise model that accounts for label noise. Yao *et al.* [34] proposed a quality embedding model to infer the trustworthiness of noisy labels. Different from the above supervised learning with noisy labels methods, Lu *et al.* [25] proposed a weakly supervised semantic segmentation framework to deal with noisy labels.

To the best of our knowledge, [35] is the first and only deep method that learns saliency without human annotations, where saliency maps from unsupervised saliency methods are fused with manually designed rules in combining “intra-image” fusion stream and “inter-image” fusion stream to generate the learning curriculum. The method iteratively replaces inter-image saliency map of low reliability with its corresponding saliency map. Their recursive optimization depends on dedicated design and is computationally expensive. Different from [35], we formulate unsupervised saliency learning as the joint optimization of latent saliency and noise modeling. Our method is not only simpler and easier to implement, but also outperforms [35] and existing unsupervised saliency methods. Furthermore, our method produces competitive performances as compared to the most recent deep supervised saliency detection methods.

3. Our Framework

Targeting at achieving deep saliency detection without human annotations, we propose an end-to-end noise model

integrated deep framework, which builds upon existing efficient and effective unsupervised saliency detection methods and the powerful capacity of deep neural networks.

Given a color image x_i , we would like to learn a better saliency map from its M noisy saliency maps $y_i^j, j = 1, \dots, M$ using different unsupervised saliency methods [32, 13, 21, 43]. A trivial and direct solution would be using the noisy saliency maps as “proxy” human annotations and train a deep model with these noisy saliency maps as supervision. However, it is well-known that the network training is highly prone to the noise in supervision signals. A simple fusion of the multiple labels (training with averaging, treating as multiple labels) will also not work due to the strong inconsistency between labels. While there could be many other potentials in utilizing the noisy saliency maps, they are all based on human-designed pipelines, thus cannot effectively exploit the best manner. Instead, we propose a principled way to infer the saliency maps from using multiple noisy labels and simultaneously estimate the noise.

3.1. Joint Saliency Prediction and Noise Modeling

By contrast to existing manually designed procedures and deep learning based pipeline [35], we propose a new perspective toward the problem of learning from unsupervised saliency. As illustrated in Fig. 2, our framework consists of two consecutive modules, namely a saliency prediction module that learns the mapping from a color image to the “latent” saliency map, and a noise modeling module that fits the noise. These two modules work collaboratively toward fitting the noisy saliency maps. By explicitly modeling noise, we are able to train a deep saliency prediction model without any human annotations and thus achieve unsupervised deep saliency detection.

3.2. Loss Function

We start with a set of training images, denoted as $\mathbf{X} = \{\mathbf{x}_i, i = 1, \dots, N\}$ and a set of M different saliency maps of these images, denoted as $\mathbf{Y} = \{\mathbf{y}_i^j, i = 1, \dots, N; j = 1, \dots, M\}$, where N is number of training images. These are precomputed by applying M different handcrafted ‘‘labellers’’. Throughout this discussion, i indexes the training image and j indexes the handcrafted labeller. We propose a neural network with parameter Θ for saliency detection, which computes a saliency map $\bar{\mathbf{y}}_i = f(\mathbf{x}_i, \Theta)$ of each image. Our idea is to model each of the handcrafted labellers as the sum of $\bar{\mathbf{y}}_i$ plus noise: $\mathbf{y}_i^j = \bar{\mathbf{y}}_i + \mathbf{n}_i^j$, where \mathbf{n}_i^j is a sample chosen from some probability (‘‘noise’’) distribution q_i , which is to be estimated. For simplicity in this work, it is assumed that the distribution q depends on \mathbf{x}_i , and not on the labeller j ³. We assume a simple model for the noise distributions q_i , namely that it is a zero-mean Gaussian, independent for each pixel of each image \mathbf{x}_i . Thus, the total distribution $\mathbf{q} = \{q_1, q_2, \dots, q_N\}$ is assumed independent for all i and pixel (m, n) , and is parametrized by a parameter set $\Sigma = \{\sigma_{mn}^i\}$, where i indexes the training image and (m, n) are pixel coordinates. Sometimes, distribution \mathbf{q} will be denoted as $\mathbf{q}(\Sigma)$ to emphasize the role of the parameters Σ . With this simple parameterization it is easy to generate noise samples \mathbf{n}_i^j for any i and j .

Given Θ , Σ , and an input image \mathbf{x}_i , one generates saliency map $\hat{\mathbf{y}}_i^j$ according to:

$$\hat{\mathbf{y}}_i^j = f(\mathbf{x}_i; \Theta) + \mathbf{n}_i^j = \bar{\mathbf{y}}_i + \mathbf{n}_i^j, \quad (1)$$

where each \mathbf{n}_i^j is a sample drawn from distribution $q_i(\Sigma)$. In the training process, the parameters Θ of the network and Σ of the noise model are updated to minimize an appropriate loss function. The loss function has two parts:

$$\mathcal{L}(\Theta, \Sigma) = \mathcal{L}_{\text{pred}}(\Theta, \Sigma) + \lambda \mathcal{L}_{\text{noise}}(\Theta, \Sigma), \quad (2)$$

where λ is the regularizer to balance these two terms. Under our optimization framework, increasing the variance in noise modeling will make the prediction loss $\mathcal{L}_{\text{pred}}$ large and decrease the $\mathcal{L}_{\text{noise}}$. Meanwhile, keeping the variance lower will decrease the cross-entropy loss $\mathcal{L}_{\text{pred}}$ but increase $\mathcal{L}_{\text{noise}}$. Thus our model balances between these two losses and converges to the state minimizing the overall loss. These two losses are described below:

Saliency Prediction: For the latent saliency prediction module, we use a fully convolutional neural network (FCN) due to its superior capability in feature learning and feature representation. We use the conventional cross-entropy loss and compute the loss function element-wisely across the whole training images.

³Assuming that distribution q is also dependent on the labeller j was observed not to improve results

The predictive loss $\mathcal{L}_{\text{Pred}}$ is designed to measure the agreement of the predicted labellings $\hat{\mathbf{y}}_i^j$ with handcrafted labellings \mathbf{y}_i^j . Cross-entropy loss is used for this purpose, and the cross-entropy loss between modeled value \hat{y} and ‘‘ground truth’’ value y (noisy label) is given by:

$$L_{\text{CE}} = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})). \quad (3)$$

This is applied to all pixel (m, n) , all labellers j and all the test images \mathbf{x}_i to give the total prediction loss.

$$\mathcal{L}_{\text{pred}}(\Theta, \Sigma) = \sum_{i=1}^N \sum_{j=1}^M \sum_{m,n} L_{\text{CE}}(\mathbf{y}_{i,mn}^j, \hat{\mathbf{y}}_{i,mn}^j), \quad (4)$$

where $\hat{\mathbf{y}}_{i,mn}^j$ is our noisy saliency map prediction at pixel (m, n) which can be easily computed by (1) element-wise, and $\hat{\mathbf{y}}_{i,mn}^j$ is truncated to lie in the range of $[0, 1]$.

Noise Modeling To effectively handle noisy saliency maps from different unsupervised saliency map labelers, we build a probabilistic model to approximate the noise, and connect it with our deterministic part (latent saliency prediction model as shown in Fig. 2). In this way, our entire model can be trained in an end-to-end manner to minimize the overall loss function Eq. (2).

The noise loss $\mathcal{L}_{\text{noise}}$ measures (for each training image \mathbf{x}_i) the agreement of the noise distribution $q_i(\Sigma)$ with the empirical variance of the measurements \mathbf{y}_i^j with respect to the output $\bar{\mathbf{y}}_i = f(\mathbf{x}_i; \Theta)$ of the network. More precisely, given an input \mathbf{x}_i , define $\hat{\mathbf{n}}_i^j = \mathbf{y}_i^j - \bar{\mathbf{y}}_i$, the empirical error of each \mathbf{y}_i^j with respect to the network prediction. For each pixel location (m, n) , this provides M samples from a zero-mean Gaussian probability distribution p_i , and its variance on every pixel can be written as $\hat{\sigma}_{i,mn}$. The complete set of parameters for p_i is denoted as $\hat{\Sigma} = \{\hat{\sigma}_{i,mn}\}$.

Since it is intractable to estimate the true posterior distribution of $\hat{\mathbf{n}}_i^j$, thus we propose to approximate it by sequentially optimizing the parameters of prior. We assume that the noise is generated by some random process, involving an unobserved continuous random variable set Σ . From an encoder perspective, the unobserved variable \mathbf{n} can be interpreted as a latent representation. Here, we model $\hat{\mathbf{y}}_i^j$ as a probabilistic encoder, since given an image \mathbf{x}_i and network parameters Θ it produces a distribution (e.g. a Gaussian) over possible values of the code \mathbf{n} . The process consists of two steps: (1) a noise map \mathbf{n}_i is generated from some prior distribution $q(\Sigma^*)$; (2) a noise map $\hat{\mathbf{n}}_i^j$ is produced and estimating the corresponding parameter $\hat{\sigma}_i$.

The corresponding noise loss is defined to be the KL divergence between distribution p_i and q_i .

$$\mathcal{L}_{\text{noise}}(\Theta, \Sigma) = \sum_i^N \text{KL}(q(\Sigma_i) \| p(\hat{\Sigma}_i)). \quad (5)$$

Since we employ the Gaussian distribution as the prior distribution for our noise model, the KL divergence has a closed-form solution as:

$$\text{KL}(q(\sigma) \| p(\hat{\sigma})) = \log(\hat{\sigma}/\sigma) + \frac{\sigma^2 + (\mu - \hat{\mu})^2}{2\hat{\sigma}^2} - \frac{1}{2}, \quad (6)$$

Based on this equation, we can update σ_i^2 for every coordinate (m, n) as

$$(\sigma_i^{t+1})^2 = (\sigma_i^t)^2 + \alpha((\hat{\sigma}_i^t)^2 - (\sigma_i^t)^2), \quad (7)$$

by differentiating Eq. (6) with respect to $\sigma_{i,mn}^2$, where α is the step size, and we set $\alpha = 0.01$ in this paper.

For different images we have the corresponding noise maps, which follows i.i.d. Gaussian distributions with different variance. Thus, it is hard to converge if simultaneously optimizing the FCN parameters Θ and noise parameters Σ . In order to train the whole network smoothly, we update the parameters of noise module after the prediction loss converges. Noise maps of a given image are sampled from the same distribution in a round, but they are updated in every round. At the first round, we initialize noise variance to be zero, and train the FCN until it converges. Based on the variance of the saliency prediction and noisy labels, we then update the noise variance for each image and retrain the network. Through minimizing the loss function Eq. (2) with this procedure, We can train the network and estimate the corresponding noise maps.

3.3. Deep Noise Model based Saliency Detector

Network Architecture We build our latent saliency prediction module upon the DeepLab network [4], where a deep CNN (ResNet-101 [10] in particular) originally designed for image classification is re-purposed by 1) transforming all fully connected layers to convolutional layers and 2) increasing feature resolution through dilated convolution [4]. Figure 2 shows the whole structure of our framework. Specifically, our model takes a rescaled image \mathbf{x}_i of 425×425 as input. For training, the noise model is used to iteratively update saliency prediction $\hat{\mathbf{y}}_i^j$, and it's excluded in testing stage, where the latent saliency prediction output $\bar{\mathbf{y}}_i$ in Fig. 2 is our predicted saliency map.

Implementation details: We trained our model using Caffe [12] with maximum epoch of 20. We initialized our model by using the Deep Residual Model trained for image classification [10]. We used the stochastic gradient descent method with momentum 0.9 and decreased learning rate 90% when the training loss did not decrease. Base learning rate is initialized as 1e-3 with the ‘‘poly’’ decay policy [12]. For validation, we set ‘‘test_iter’’ as 500 (test batch size 1) to cover the full 500 validation images. The training took 4 hours for one round with training batch size 1 and ‘‘iter_size’’ 20 on a PC with an NVIDIA Quadro M4000 GPU.

4. Experimental Results

In this section, we report experimental results on various saliency detection benchmarking datasets.

4.1. Setup

Dataset: We evaluated performance of our proposed model on 7 saliency benchmarking datasets. 3,000 images from the MSRA-B dataset[24] are used to get the noisy labels (where 2,500 images for training and 500 images for validation) and the remaining 2,000 images are kept for testing. Most of the images in MSRA-B dataset only have one salient object. The ECSSD dataset [32] contains 1,000 images of semantically meaningful but structurally complex images. The DUT dataset [33] contains 5,168 images. The SOD saliency dataset [14] contains 300 images, where many images contain multiple salient objects with low contrast. The SED2 [1] dataset contains 100 images with each image contains two salient objects. The PASCAL-S [23] dataset is generated from the PASCAL VOC [8] dataset and contains 850 images. The THUR dataset [5] contains 6,232 images of five classes, namely ‘‘butterfly’’, ‘‘coffee mug’’, ‘‘dog jump’’, ‘‘giraffe’’ and ‘‘plane’’.

Unsupervised Saliency Methods: In this paper, we learn unsupervised saliency from existing unsupervised saliency detection methods. In our experiment, we choose RBD [43], DSR [21], MC [13] and HS [32] due to their effectiveness and efficiency as illustrated in [3].

Competing methods: We compared our method against 10 state-of-the-art deep saliency detection methods (with clean labels): DSS [11], NLDF [26], Amulet [40], UCF [41], SRM [35], DMT [22], RFCN [28], DeepMC [42], MDF [19] and DC [20], 5 conventional handcrafted feature based saliency detection methods: DRFI [14], RBD [43], DSR [21], MC [13], and HS [32], which were proven in [3] as the state-of-the-art methods before the deep learning revolution, and the very recent unsupervised deep saliency detection method SBF [35].

Evaluation metrics: We use 3 evaluation metrics, including the mean absolute error (MAE), F-measure, as well as the Precision-Recall (PR) curve. MAE can provide a better estimate of the dissimilarity between the estimated and ground truth saliency map. It is the average per-pixel difference between the ground truth and the estimated saliency map, normalized to [0, 1], which is defined as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - GT(x, y)|, \quad (8)$$

where W and H are the width and height of the respective saliency map S , GT is the ground truth saliency map.

The F-measure (F_β) is defined as the weighted harmonic

Table 1. Performance of mean F-measure (F_β) and MAE for different methods including ours on seven benchmark datasets.

Methods	MSRA-B		ECSSD		DUT		SED2		PASCALS		THUR		SOD	
	F_β	MAE												
BL1	.7905	.0936	.7205	.1444	.5825	.1369	.7773	.1112	.6714	.2206	.5953	.1339	.6306	.1870
BL2	.6909	.1710	.6542	.2170	.4552	.2951	.7232	.1406	.6776	.2409	.5119	.2545	.5928	.2566
BL3	.8879	.0587	.8717	.0772	.7253	.0772	.8520	.0819	.8264	.1525	.7368	.0749	.7922	.1231
OURS	.8770	.0560	.8783	.0704	.7156	.0860	.8380	.0881	.8422	.1391	.7322	.0811	.7976	.1182

Table 2. Performance of mean F-measure (F_β) and MAE for different methods including ours on seven benchmark datasets (Best ones in bold). From DSS to DC are deep learning based supervised methods, from DRFI to HS are the handcrafted feature based unsupervised methods, SBF and OURS are deep learning based unsupervised saliency detection methods.

Methods	MSRA-B		ECSSD		DUT		SED2		PASCALS		THUR		SOD	
	F_β	MAE												
DSS [11]	.8941	.0474	.8796	.0699	.7290	.0760	.8236	.1014	.8243	.1546	.7081	.1142	.8048	.1118
NLDF [26]	.8970	.0478	.8908	.0655	.7360	.0796	-	-	.8391	.1454	-	-	.8235	.1030
Amulet [40]	-	-	.8825	.0607	.6932	.0976	.8745	.0629	.8371	.1292	.7115	.0937	.7729	.1248
UCF [41]	-	-	.8521	.0797	.6595	.1321	.8444	.0742	.8060	.1492	.6920	.1119	.7429	.1527
SRM [29]	.8506	.0665	.8260	.0922	.6722	.0846	.7447	.1164	.7766	.1696	.6894	.0871	.7246	.1369
DMT [22]	-	-	.7589	.1601	.6045	.0758	.7778	.1074	.6657	.2103	.6254	.0854	.6978	.1503
RFCN [28]	-	-	.8426	.0973	.6918	.0945	.7616	.1140	.8064	.1662	.7062	.1003	.7531	.1394
DeepMC [42]	.8966	.0491	.8061	.1019	.6715	.0885	.7660	.1162	.7327	.1928	.6549	.1025	.6862	.1557
MDF [19]	.7780	.1040	.8097	.1081	.6768	.0916	.7658	.1171	.7425	.2069	.6670	.1029	.6377	.1669
DC [20]	.8973	.0467	.8315	.0906	.6902	.0971	.7840	.1014	.7861	.1614	.6940	.0959	.7603	.1208
DRFI [14]	.7282	.1229	.6440	.1719	.5525	.1496	.7252	.1373	.5745	.2556	.5613	.1471	.5440	.2046
RBD [43]	.7508	.1171	.6518	.1832	.5100	.2011	.7939	.1096	.6581	.2418	.5221	.1936	.5927	.2181
DSR [21]	.7227	.1207	.6387	.1742	.5583	.1374	.7053	.1452	.5785	.2600	.5498	.1408	.5500	.2133
MC [13]	.7165	.1441	.6114	.2037	.5289	.1863	.6619	.1848	.5742	.2719	.5149	.1838	.5332	.2435
HS [44]	.7129	.1609	.6234	.2283	.5205	.2274	.7168	.1869	.5948	.2860	.5157	.2178	.5383	.2729
SBF [35]	-	-	.7870	.0850	.5830	.1350	-	-	.7780	.1669	-	-	.6760	.1400
OURS	.8770	.0560	.8783	.0704	.7156	.0860	.8380	.0881	.8422	.1391	.7322	.0811	.7976	.1182

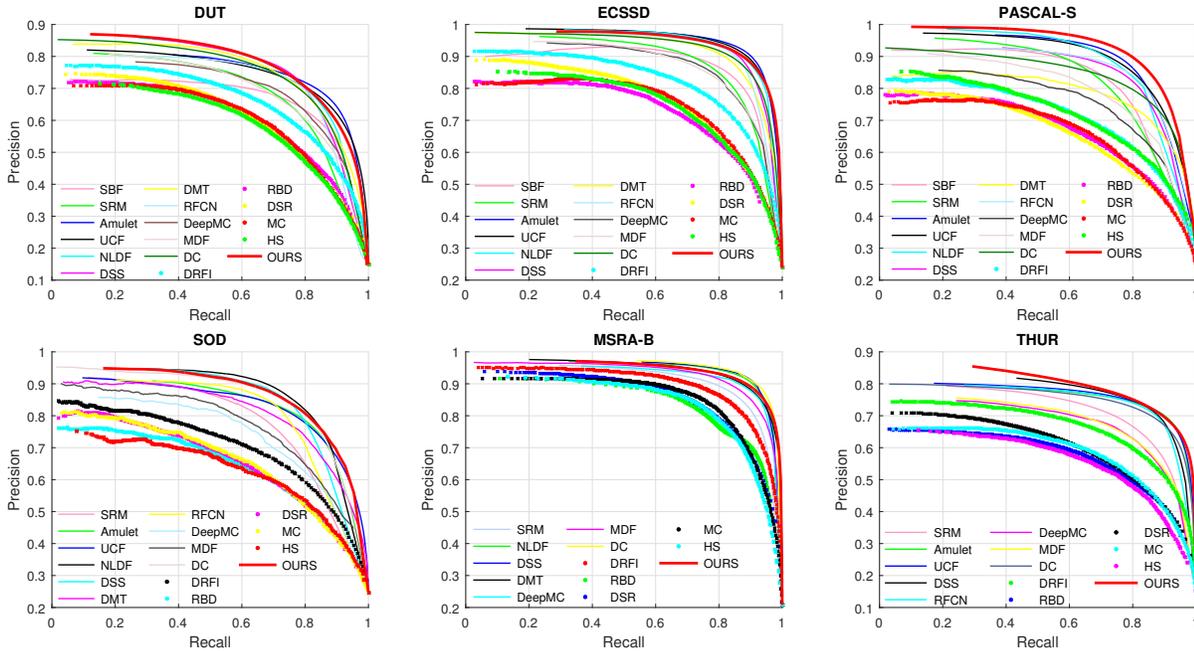


Figure 3. PR curves on six benchmark datasets (DUT, ECSSD, PASCAL-S, SOD, MSRA-B, THUR). Best Viewed on Screen.

mean of precision and recall:

$$F_\beta = (1 + \beta^2) \frac{Precision \times Recall}{\beta^2 Precision + Recall}, \quad (9)$$

where $\beta^2 = 0.3$, *Precision* corresponds to the percentage of salient pixels being correctly detected, *Recall* is the fraction of detected salient pixels in relation to the ground truth

number of salient pixels. The PR curves are obtained by thresholding the saliency map in the range of [0, 255].

4.2. Baseline Experiments

As there could be different ways to utilize the multiple noisy saliency maps, and for fair comparisons with straightforward solutions for our task, we run the following three baseline methods and the results are reported in Table 1.

Baseline 1 using noisy unsupervised saliency pseudo ground truth: For a given input image x_i and its M hand-crafted feature based saliency map $y_i^j, j = 1, \dots, M$, we get M image pairs with noisy label $\{x_i, y_i^j, j = 1, \dots, M\}$. Then we train a deep model [10] based on those noisy labels directly, and the results are shown as “BL1” in Table 1.

Baseline 2: using averaged unsupervised saliency as pseudo ground truth: Instead of using all the four unsupervised saliency as ground truth, we use the averaged saliency map of those unsupervised saliency as pseudo ground truth, and trained another baseline model “BL2” in Table 1.

Baseline 3: supervised learning with ground truth supervision: Our proposed framework consists of the saliency prediction module and the noise modeling module to effectively leverage the noisy saliency maps. To illustrate the best performance our model can achieve as well as to provide a baseline comparison for our framework, we train our latent saliency module directly with clean labels, which naturally gives an upper bound of the saliency detection performance. The results “BL3” are reported in Table 1.

Analysis: In Table 1, we compare our unsupervised saliency method with the above baseline configurations. Our method clearly outperforms both BL1 and BL2 with a wide margin, demonstrating the superiority of our end-to-end learning framework. As illustrated in Table 1, the performance of BL1 is better than the performance of BL2. This is because: 1) For BL1, we have 12,000 training image pairs (four unsupervised saliency methods), while for BL2, we have 3,000 averaged noisy labels; 2) as those unsupervised saliency methods tend to prefer different priors for saliency detection, and their saliency maps can be complementary or controversial to some extent. Simply averaging those saliency maps results in even worse proxy saliency map supervision. Compared with BL3, which is trained with ground truth clean labels and without noise, our unsupervised method achieves highly comparable results. This demonstrates that by jointly learning the latent saliency maps and modeling the noise in a unified framework, we are able to learn the desired reliable saliency maps even without any human annotations.

4.3. Comparison with the State-of-the-art

Quantitative Comparison We compared our method with eleven most recent deep saliency methods and five conventional methods. Results are reported in Table 2

and Fig. 3, where “OURS” represents the results of our model. Table 2 shows that on those seven benchmark datasets, deep supervised methods significantly outperform traditional methods with 2%-12% decrease in MAE, which further proves the superiority of deep saliency detection.

MSRA-B is a relatively simple dataset, where most salient objects dominate the whole image. The most recent deep supervised saliency methods [11] [26] [40] can achieve the highest mean F-measure of 0.8970, and our unsupervised method without human annotations can achieve a mean F-measure of 0.8770, which is only a slight worse. The DUT dataset has more than 25% of images with saliency occupation less than 4%. Small salient object detection is quite challenging which increase the difficulty of this dataset. We achieve the third highest mean F-measure compared with all the competing methods. The THUR dataset is the largest dataset we used in this paper, and most of the images have complex background. The state-of-the-art competing method achieves a mean F-measure/MAE as 0.7115/0.0854, while our method achieves the best mean F-measure and MAE as 0.7322/0.0811. SBF [35] uses inter- and intra-image confidence map as pseudo ground truth to train an unsupervised deep model based on unsupervised saliency, which is quite different from our formulation of predicting saliency from unsupervised saliency as learning from noisy labels. Table 2 shows that our framework leads to better performance, with 10% mean F-measure improvement and 3% decrease of MAE on average. Fig. 3 shows comparison between PR curves of our method and the competing methods on four benchmarking datasets. For the PASCAL-S and THUR dataset, our method ranks almost the 1st, and for the other three datasets, our method achieves competitive performance compared with the competing deep supervised methods. These experiments altogether proves the effectiveness our proposed unsupervised saliency detection framework.

Qualitative Comparison Figure 4 demonstrates several visual comparisons, where our method consistently outperforms the competing methods, especially those four unsupervised saliency we used to train our model. The first image is a simple scenario, and most of the competing methods can achieve good results, while our method achieves the best result with most of the background region suppressed. Background of the third image is very complex, and all the competing methods fail to detect salient object. With proper noisy labels, we achieve the best results compared with both unsupervised saliency methods and deep saliency methods. The fourth image is in very low-contrast, where most of the competing methods failed to capture the whole salient objects with the last penguin mis-detected, especially for those unsupervised saliency methods. Our method captures all the three penguins properly. The salient objects in the last row are quite small, and the competing methods failed to

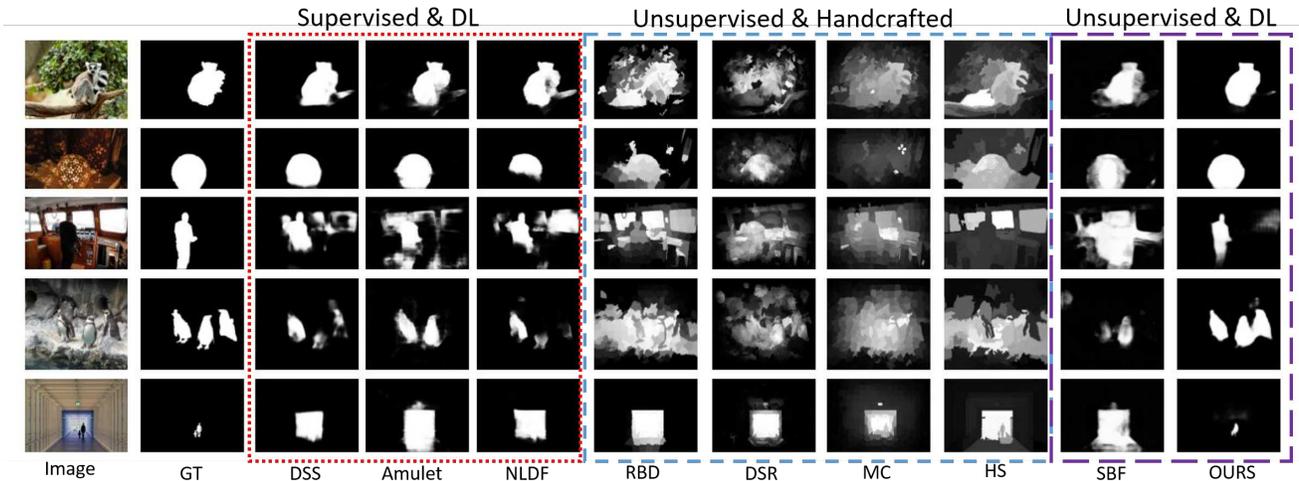


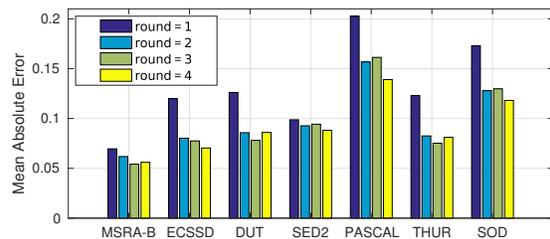
Figure 4. Visual comparison between our method and other competing methods.

capture salient regions, while our method capture the whole salient region with high precision.

Ablation Studies: In this paper, we propose to iteratively update the noise modeling module and the latent saliency prediction model to achieve accurate saliency detection. As the two modules work collaboratively to optimize the overall loss function, it is interesting to see how the saliency prediction results evolves with respect to the increase of updating round. In Fig. 5, we illustrate both the performance metric (MAE) with respect to updating round and an example saliency detection results. Starting with the zero noise initialization, our method consistently improves the performance of saliency detection with the updating of noise modeling. Also, only after several updating rounds, our method convergences to desired state as shown in Fig. 5.

5. Conclusions

In this paper, we propose an end-to-end saliency learning framework without the need of human annotated saliency maps in network training. We represent unsupervised saliency learning as learning from multiple noisy saliency maps generated by various efficient and effective conventional unsupervised saliency detection methods. Our framework consists of a latent saliency prediction module and an explicit noise modeling models, which work collaboratively. Extensive experimental results on various benchmarking datasets prove the superiority of our method, which not only outperforms traditional unsupervised methods with a wide margin but also achieves highly comparable performance with current state-of-the-art deep supervised saliency detection methods. In the future, we plan to investigate the challenging scenarios of multiple saliency object detection and small salient object detection under our



(a) MAE of each round on 7 datasets

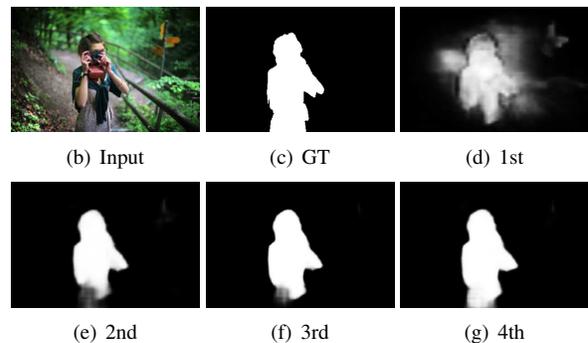


Figure 5. Performance of each round. Top: MAE of each dataset. Bottom: an example image, ground-truth and intermedia results generated by each updating round.

framework. Extending our framework to dense prediction tasks such as semantic segmentation [25] and monocular depth estimation [18] could be interesting directions.

Acknowledgement. J. Zhang would like to thank Prof. Mingyi He for his immeasurable support and encouragement. T. Zhang was supported by the Australian Research Council (ARC) Discovery Projects funding scheme (project DP150104645). Y. Dai was supported in part by National 1000 Young Talents Plan of China, Natural Science Foundation of China (61420106007, 61671387), and ARC grant (DE140100180).

References

- [1] S. Alpert, M. Galun, A. Brandt, and R. Basri. Image segmentation by probabilistic bottom-up aggregation and cue integration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(2):315–327, Feb 2012. **5**
- [2] A. Borji, M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. *CoRR*, abs/1411.5878, 2014. **2**
- [3] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Trans. Image Proc.*, 24(12):5706–5722, 2015. **2, 5**
- [4] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP(99):1–1, 2017. **5**
- [5] M. Cheng, N. J. Mitra, X. Huang, and S. Hu. Salienshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014. **5**
- [6] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 409–416, 2011. **1, 2**
- [7] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook. Efficient salient region detection with soft image abstraction. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1529–1536, 2013. **2**
- [8] M. Everingham, S. M. A. Eslami, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *Int. J. Comp. Vis.*, 111(1):98–136, 2015. **5**
- [9] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(10):1915–1926, Oct 2012. **1, 2**
- [10] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 770–778, June 2016. **5, 7**
- [11] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr. Deeply supervised salient object detection with short connections. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3203–3212, July 2017. **1, 2, 5, 6, 7**
- [12] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. ACM Int. Conf. Multimedia*, pages 675–678, 2014. **5**
- [13] B. Jiang, L. Zhang, H. Lu, C. Yang, and M. Yang. Saliency detection via absorbing markov chain. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1665–1672, 2013. **3, 5, 6**
- [14] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li. Salient object detection: A discriminative regional feature integration approach. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2083–2090, 2013. **2, 5, 6**
- [15] P. Jiang, H. Ling, J. Yu, and J. Peng. Salient region detection by UFO: Uniqueness, focusness and objectness. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1976–1983, 2013. **2**
- [16] I. Jindal, M. Nokleby, and X. Chen. Learning deep networks from noisy labels with dropout regularization. In *Proc. IEEE Int. Conf. Data Mining.*, pages 967–972, Dec 2016. **3**
- [17] J. Kim, D. Han, Y.-W. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 883–890, 2014. **2**
- [18] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1119–1127, June 2015. **8**
- [19] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5455–5463, June 2015. **2, 5, 6**
- [20] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 478–487, June 2016. **2, 5, 6**
- [21] X. Li, H. Lu, L. Zhang, X. Ruan, and M. Yang. Saliency detection via dense and sparse reconstruction. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2976–2983, Dec 2013. **3, 5, 6**
- [22] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang. Deepsaliency: Multi-task deep neural network model for salient object detection. *IEEE Trans. Image Proc.*, 25(8):3919–3930, Aug 2016. **2, 5, 6**
- [23] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 280–287, 2014. **5**
- [24] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum. Learning to detect a salient object. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1–8, 2007. **5**
- [25] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(3):486–500, Mar 2017. **3, 8**
- [26] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin. Non-local deep features for salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, July 2017. **2, 5, 6, 7**
- [27] X. Shen and Y. Wu. A unified approach to salient object detection via low rank matrix recovery. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 853–860, 2012. **2**
- [28] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *Proc. Eur. Conf. Comp. Vis.*, pages 825–841, 2016. **2, 5, 6**
- [29] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stage-wise refinement model for detecting salient objects in images. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2017. **2, 6**
- [30] J. Whitehill, T. fan Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Proc. Adv. Neural Inf. Process. Syst.*, pages 2035–2043. 2009. **3**
- [31] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proc. Int. Conf. Mach. Learn.*, volume 37, pages 2048–2057, 2015. **1**
- [32] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1155–1162, 2013. **3, 5**

- [33] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang. Saliency detection via graph-based manifold ranking. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3166–3173, 2013. [5](#)
- [34] J. Yao, J. Wang, I. Tsang, Y. Zhang, J. Sun, C. Zhang, and R. Zhang. Deep Learning from Noisy Image Labels with Quality Embedding. *ArXiv e-prints*, Nov. 2017. [3](#)
- [35] D. Zhang, J. Han, and Y. Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proc. IEEE Int. Conf. Comp. Vis.*, Oct 2017. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#)
- [36] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, and R. R. Martin. A shape-preserving approach to image resizing. *Computer Graphics Forum*, 28(7):1897–1906, 2009. [1](#)
- [37] J. Zhang, Y. Dai, B. Li, and M. He. Attention to the scale: Deep multi-scale salient object detection. In *Proc. Int. Conf. on Digital Image Computing: Techniques and Applications*, pages 1–7, Nov 2017. [2](#)
- [38] J. Zhang, Y. Dai, and F. Porikli. Deep salient object detection by integrating multi-level cues. In *Proc. IEEE Winter Conference on Applications of Computer Vision*, pages 1–10, March 2017. [2](#)
- [39] J. Zhang, B. Li, Y. Dai, F. Porikli, and M. He. Integrated deep and shallow networks for salient object detection. In *Proc. IEEE Int. Conf. Image Process.*, pages 1537–1541, Sept 2017. [2](#)
- [40] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, Oct 2017. [1](#), [2](#), [5](#), [6](#), [7](#)
- [41] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, Oct 2017. [2](#), [5](#), [6](#)
- [42] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1265–1274, 2015. [1](#), [2](#), [5](#), [6](#)
- [43] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2814–2821, 2014. [1](#), [2](#), [3](#), [5](#), [6](#)
- [44] W. Zou and N. Komodakis. Harf: Hierarchy-associated rich features for salient object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 406–414, Dec 2015. [6](#)