Supervision by Fusion: Towards Unsupervised Learning of Deep Salient Object Detector

Dingwen Zhang, Junwei Han*, Yu Zhang
Northwestern Polytechnical University
{zhangdingwen2006yyy,junweihan2010}@gmail.com,zhangyuygss@mail.nwpu.edu.cn

Abstract

In light of the powerful learning capability of deep neural networks (DNNs), deep (convolutional) models have been built in recent years to address the task of salient object detection. Although training such deep saliency models can significantly improve the detection performance, it requires large-scale manual supervision in the form of pixel-level human annotation, which is highly labor-intensive and time-consuming. To address this problem, this paper makes the earliest effort to train a deep salient object detector without using any human annotation. The key insight is "supervision by fusion", i.e., generating useful supervisory signals from the fusion process of weak but fast unsupervised saliency models. Based on this insight, we combine an intra-image fusion stream and a inter-image fusion stream in the proposed framework to generate the learning curriculum and pseudo ground-truth for supervising the training of the deep salient object detector. Comprehensive experiments on four benchmark datasets demonstrate that our method can approach the same network trained with full supervision (within 2-5% performance gap) and, more encouragingly, even outperform a number of fully supervised state-of-the-art approaches.

1. Introduction

With the goal of discovering the object regions that can attract human visual attention in images, salient object detection has been gaining intensive research interest in recent years. Due to its capability in automatically revealing important and informative parts in each given image, it has been widely applied in image retrieval [7], object detection [43], event detection [3, 40], and so on.

1.1. Previous Works

The salient object detection approaches proposed in early ages mainly explored image saliency by evaluating the distinctiveness of each image region or image pixel with respect to the corresponding local context or global image scene [1, 13, 5, 8, 10]. For example, Achanta et al. [1] proposed to estimate center-surround contrast based on color and luminance features in the frequency domain. Klein et al. [13] proposed to calculate the Kullback-Leibler-Divergence (KLD) between the visual feature distribution at a center location and the surround context. In [5], Cheng et al. calculated the global contrast on the superpixel level where each superpixel's contrast is measured by a weighted integration of the differences between itself and all other superpixels in the image. These approaches can usually perform with little time cost. However, they generally suffer from the bottleneck in detection accuracy.

More recently, in light of the cutting edge learning capability of deep neural networks (DNNs), researchers have investigated several deep (convolutional) models for addressing the task of salient object detection [17, 32, 46, 25, 4, 16]. These approaches can usually obtain more promising performance due to the informative feature representation and hidden patterns learnt from the large-scale annotated training data. Specifically, Wang et al. [32] adopted a convolutional neural network (CNN) to predict saliency scores for each pixel in local context firstly and then refined the saliency score for each object proposal over the global view. Similarly, Zhao et al. [46] proposed a multi-context deep learning framework for salient object detection, which jointly modeled global context and local context in a unified framework. In [25], a coarse global prediction was generated by learning the global structured saliency cues firstly and then a hierarchical recurrent convolutional neural network was adopted to progressively integrate local context details.

1.2. Motivation and Contributions

Studies in this field have demonstrated that the DNN-based salient object detectors are highly effective and can achieve top results on modern benchmark datasets (see Fig. 1). However, all current DNN-based salient object detectors require the large-scale manual supervision in the form of pixel-level human annotation. Collecting such

*Corresponding author.

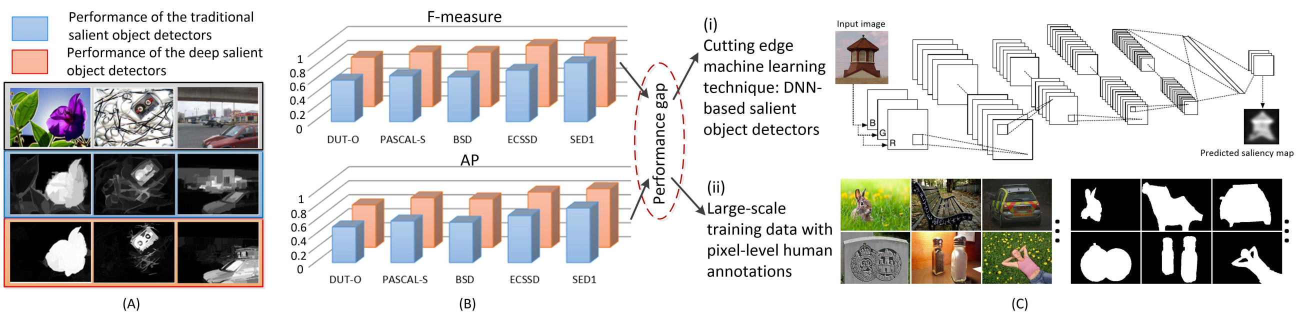

Figure 1. Illustration of our motivation. (A) shows some examples of the salient object detection results generated by the existing approaches, where the first row are the original images, the second row are the saliency maps obtained by the traditional salient object detector [12], and the third row are the saliency maps obtained by the deep salient object detector [46]. In (B), we show some statistics on five benchmark datasets, where the blue histograms indicate the average performance of the traditional salient object detectors [31, 27, 47, 35, 18, 45] and the red histograms indicate the average performance of the deep salient object detectors [17, 46, 25, 32, 16]. The performance gap between the traditional salient object detectors and the deep salient object detectors is mainly caused by (i) the powerful deep learning technique and (ii) the large-scale manually annotated training data, as shown in (C). Since collecting manual annotations is highly time-consuming and labor intensive, this paper makes the earliest attempt to train deep convolutional saliency model without using any human annotation.

annotated training data tends to be highly labor-intensive and time-consuming. In contrast, traditional unsupervised salient object detectors can be obtained much more cheaply. However, the shortcoming is that they cannot achieve satisfactory detection performance. In this paper, we make the earliest effort to explore: “Is pixel-level human annotation indispensable for building strong salient object detector?” and moreover, “Can deep salient object detectors be trained entirely without using human supervision?” Specifically, we propose a unsupervised learning framework[1] to train deep salient object detector by only using the raw image data, which can hopefully combine the advantages of the existing supervised DNN-based approaches (i.e., high performance) and the traditional unsupervised approaches (i.e., high convenience). The key is “supervision by fusion”.

Training deep salient object detector without using any human annotation is very challenging. Unlike[22], where the motion discontinuities among video sequences can be used to guide the unsupervised learning of the edge detectors, there is no such external information source that can be readily used to provide effective supervision in our task. Thus, in this paper, we propose to take advantage of the existing unsupervised salient object detector to provide the needed pseudo supervision. Along this direction, one naive strategy is to adopt the saliency maps generated by an existing unsupervised salient object detector to provide the initial pseudo ground-truth, and then train the DNN-based salient object detector in iterations by using the saliency prediction results of the current learning iteration as the supervision of the next learning iteration. Unfortunately, as shown in Fig. 2, this strategy cannot work well in practice. The underlying reasons are two folds: Firstly, only using one unsupervised salient object detector is not able to provide strong enough supervision. Training deep models under the generated pseudo ground-truth maps would inevitably lead the learner to build trivial feature representation and capture less informative saliency patterns. Secondly, the aforementioned learning strategy lacks a confidence weighting scheme, which plays an important role in guiding the learner gradually aggregating faithful knowledge from the confident training samples while refusing the noisy ones. As we know, due to the different ambiguities of the included contents, different images (or image regions) would have different difficulties for obtaining the truthful ground-truth. Treating all such training samples equally will introduce non-neglectable noise (the samples with totally wrong pseudo ground-truth label) to the learning procedure and thus further confuse the learner.

To address these problems, this paper proposes a novel unsupervised learning framework to train desirable deep salient object detector based on the “supervision by fusion” strategy: generating reliable supervisory signals from the fusion process of weak saliency models. As we know, the fusion process itself is a unsupervised inference procedure. Moreover, some modern fusion models can not only integrate the weak saliency models to obtain stronger saliency prediction but also automatically infer the reliabilities for each weak saliency model under the condition of the given image context at the same time. By involving such fusion process into the unsupervised learning framework, we can, on one hand, improve the overall strength of the pseudo ground-truth and, on the other hand, develop a dynamic learning curriculum to guide a robust learning procedure with the confidence weighting scheme. Specifically, the aforementioned problems can be addressed in the following ways:

Firstly, instead of directly adopting the saliency maps obtained from one weak saliency model to provide the pseu-

[1] In this paper, unsupervised learning refers to learning without using human annotation.

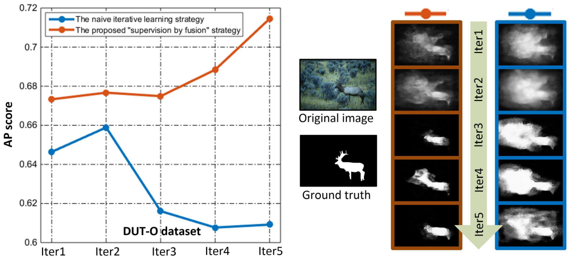

Figure 2. Examples to show that directly training the deep salient object detector in the native way cannot work well, while the proposed “supervision by fusion” strategy can guide a more robust unsupervised learning procedure.

do ground-truth for the training of the DNN-salient object detector, we propose to fuse the saliency maps of multiple weak but fast saliency models to obtain stronger pseudo supervision (both in the initialization stage and the subsequent learning iterations). As such saliency maps can be extracted in parallel, which is efficient, and the fusion process itself is not time-consuming[2], this strategy tends to be a cost-effective way to improve the unsupervised learning performance.

Secondly, during the fusion process, we can simultaneously infer the difficulty of each training data (both training images and more detailed superpixel regions), which provides a natural way to reflect their corresponding learning confidence. Basically, such difficulty is inferred based on the inconsistency of the fused weak salient object detectors. Thus, the training data inferred with less difficulty tends to obtain more consistent predictions and thus be more confident for the learning process and vice versa. Such information can help us to build an informative learning curriculum, which selects confident training samples against the noisy ones and assigns various important weights for the selected training samples (both in image level and superpixel level) during the learning procedure. In addition, after each learning iteration, the fusion process will again be used to update the difficulties of the training samples based on the current weak saliency predictions, which essentially forms the learning curriculum dynamically.

Based on the aforementioned strategies, we propose a novel unsupervised learning framework (see Fig. 3) to train deep salient object detector without using any human annotation. Specifically, we first use some fast unsupervised saliency models to extract the weak saliency maps of each training image (the blue blocks in Fig. 3). Then, intra-image fusion is performed within each individual image to obtain the superpixel-level confidence and superpixel-level fusion map (the pink blocks in Fig. 3) and inter-image fusion is performed on all the training images to obtain the image-level confidence and image-level fusion map (the green blocks in Fig. 3). Afterwards, the dynamic learning curriculum and pseudo ground-truth maps are generated to provide supervision for training the deep salient object detector (the yellow blocks in Fig. 3). Finally, the obtained deep salient object detector is used to update the weak saliency map collection for the next learning stage.

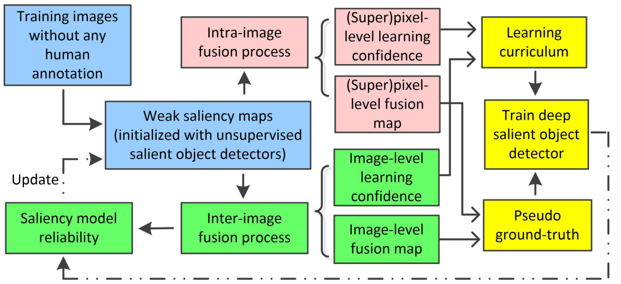
Training images without any human annotation
Intra-image fusion process
(Super)pixel-level learning confidence
Learning curriculum
(Super)pixel-level fusion map
Weak saliency maps (initialized with unsupervised salient object detectors)
Train deep salient object detector
Update
Image-level learning confidence
Saliency model reliability
Inter-image fusion process
Pseudo ground-truth
Image-level fusion map

Figure 3. The proposed unsupervised learning framework for training deep salient object detector under the supervision by fusion.

To sum up, we mainly have three-fold contributions:

1) We make the earliest effort to train powerful deep salient object detector without using any pixel-level human annotation. It is of great significance as it can integrate the advantages of the supervised DNN-based approaches (i.e., high performance) and the traditional unsupervised approaches (i.e., high convenience).

2) We reveal the insight of “supervision by fusion”, i.e., generating reliable supervisory signals from the fusion process of weak saliency models in iterative learning stages. Through fusion, we can use the obtained fusion map to provide more reliable supervision and the sample confidence weights to generate the dynamic learning curriculum.

3) Through comprehensive experiments on four benchmark datasets, we demonstrate that our insight can be successfully implemented via a novel unsupervised learning framework based on the two-stream fusion.

2. The Proposed Approach

Given N training images $\{\mathbf{I}_n\}, n \in [1, N]$, we use three unsupervised salient object detectors [29, 45, 44][3] to generate the initial weak saliency maps $\{\mathbf{WSM}_n^m\}, m \in [1, M], M = 3$ due to their efficiency. Then, the superpixel-level confidence maps $\{\mathbf{b}_n\}$ and superpixel-level fusion maps $\{\mathbf{SFM}_n\}$ are obtained by intra-image fusion (see Sec. 2.1), while the image-level confidence weights $\{\beta_n\}$ and image-level fusion maps $\{\mathbf{IFM}_n\}$ are obtained by inter-image fusion (see Sec. 2.2). Afterwards, $\{\mathbf{b}_n\}$, $\{\mathbf{SFM}_n\}$, $\{\beta_n\}$, $\{\mathbf{IFM}_n\}$ are used to guide the training of deep salient object detector (see Sec. 2.3). The saliency maps generated by the method that inferred as having the lowest prediction reliability (in inter-image fusion) are then replaced by

[2] The GLAD fusion model [36] takes 10 minutes for 1 million images using a single core of a Xeon 2.8 GHz processor.

[3] Here we choose to use three unsupervised salient object detectors by considering the tradeoff of the effectiveness and efficiency in fusing different number of weak models.

the saliency maps generated by the learnt deep salient object detector, which forms the new weak saliency maps for guiding the learning in the next iteration. The above process is iterated until convergence (typically 4 to 5 iterations suffice). Based on our observation, after 4 to 5 learning iterations, the weak saliency maps tend to converge to a single map, and the complementarity among them becomes weak.

2.1. Intra-Image Fusion

Given each training image $\mathbf{I}_n$ and the corresponding weak saliency maps $\{\mathbf{WSM}_n^m\}_{m=1}^M$, the goal of intra-image fusion is to infer the superpixel-level reliability of each weak salient object detector to integrate the weak saliency maps with considering the different difficulties of various local image regions. Specifically, we first extract N_r superpixel regions $sp_{n,i}$, where $i \in [1, N_r]$, from the n-th image using [6][4]. For each weak saliency map, the weak saliency value of each superpixel region $s_{n,i}^m$ is the mean value of the pixels residing in the superpixel. Denoting the average saliency value for each superpixel region as $asv_{n,i} = \frac{1}{M}\sum_m s_{n,i}^m$, we use $\tau_{n,i}^m = |s_{n,i}^m - asv_{n,i}|/asv_{n,i}$ to reflect the agreement between each weak saliency prediction and the average saliency value. Then, the weak saliency label of each superpixel $l_{n,i}^m$ is obtained by:

$$l_n^m = \begin{cases} 1, & \tau_{n,i}^m \le t_n \\ 0, & others \end{cases}, \quad t_n = \frac{1}{MR_n}\sum_{i=1}^{N_r}\sum_{m=1}^{M}\tau_{n,i}^m. \tag{1}$$

Afterwards, for inferring the superpixel-level reliabilities of the weak salient object detectors $\{a_n^m\}$ and the difficulties of various superpixel regions $\{b_{n,i}\}$, we adopt the GLAD fusion model [36], which is a probabilistic model providing a principled way to approach the fusion problem. Basically, it models the superpixel-level reliability and difficulty as latent variables. The probability of correct labeling is defined as:

$$p(l_{n,i}^m = z_{n,i} | a_n^m, b_{n,i}) = \frac{1}{1 + e^{-a_n^m b_{n,i}}}, \tag{2}$$

where $z_{n,i}$ indicates the underlying true saliency label of the i-th superpixel, $b_{n,i} \in [0, \infty)$ and $a_n^m \in (-\infty, +\infty)$. In this model, $b_{n,i} = 0$ means the superpixel region is very ambiguous and hence even the best weak salient object detector has a 50% chance of predicting it correctly, while $b_{n,i} = \infty$ means the superpixel region is so easy that even the most obtuse salient object detector can always predict it correctly. Thus, a larger $b_{n,i}$ indicates a higher learning confidence. For a_n^m, a very large value that closes to $+\infty$ means the weak salient object detector always predicts correctly, while a very small value that closes to $-\infty$ means the salient object detector always predicts incorrectly. Given the observed weak saliency labels $\{l_{n,i}^m\}$, GLAD infers the values of the involved variables $\{z_{n,i}\}$, $\{a_n^m\}$, and $\{b_{n,i}\}$ by maximizing the likelihood estimates using Expectation-Maximization approach. Finally, the superpixel-level learning confidence maps $\{\mathbf{b}_n\}$ can be formed by the inferred $\{b_{n,i}\}$ and the superpixel-level fusion maps can be obtained by[5]:

$$\mathbf{SFM}_n = \sum_{m=1}^{M} a_n^m \cdot \mathbf{WSM}_n^m, \tag{3}$$

2.2. Inter-Image Fusion

Different from intra-image fusion, inter-image fusion integrates the weak saliency maps from the entire training image collection instead of the content of each single image, and the fusion process considers the global image scenes rather than the local superpixel regions. During the fusion process, it infers the image-level reliability of each weak salient object detector and the difficulties of various global image scenes. Specifically, given the training image collection $\{\mathbf{I}_n\}$ and the weak saliency map collection $\{\mathbf{WSM}_n^m\}$, we first calculate the average saliency map $\{\mathbf{ASM}_n\}$ of each image and the distance $\{\Gamma_n^m\}$ between each weak saliency map and the average saliency map as:

$$\begin{aligned} \mathbf{ASM}_n &= \frac{1}{M}\sum_{m=1}^{M}\mathbf{WSM}_n^m, \\ \Gamma_n^m &= |\mathbf{WSM}_n^m - \mathbf{ASM}_n|_1 / |\mathbf{ASM}_n|_1. \end{aligned} \tag{4}$$

Then, the binary label of the m-th weak salient object detector on the n-th training images $\{L_n^m\}$ is obtained by:

$$L_n^m = \begin{cases} 1, & \Gamma_n^m \le T \\ 0, & others \end{cases}, \quad T = \frac{1}{MN}\sum_{n=1}^{N}\sum_{m=1}^{M}\Gamma_n^m, \tag{5}$$

which reflects the agreement between the weak saliency map and the corresponding average saliency map. Afterwards, we adopt the GLAD fusion model to infer the image-level reliabilities of the weak salient object detectors $\{\alpha^m\}$ and the difficulties of various training images $\{\beta_n\}$. Finally, the image-level fusion maps can be obtained by[6]:

$$\mathbf{IFM}_n = \sum_{m=1}^{M} \alpha^m \cdot \mathbf{WSM}_n^m, \tag{6}$$

2.3. Training Deep Salient Object Detector under Supervision by Fusion

2.3.1 The Network Architecture

We build our deep salient object detector based on the DHSNet [25] due to its effectiveness and efficiency. The

[4] The intra-image fusion can also perform on pixel-level, while this paper uses superpixels for pursuing lower computational cost.

[5] Here we name the fusion results obtained from the intra-image fusion process as the superpixel-level fusion maps because the basic computational unit in the intra-image fusion is each superpixel region.

[6] Here we name the fusion results obtained from the inter-image fusion process as the image-level fusion maps because the basic computational unit in the inter-image fusion is each image.

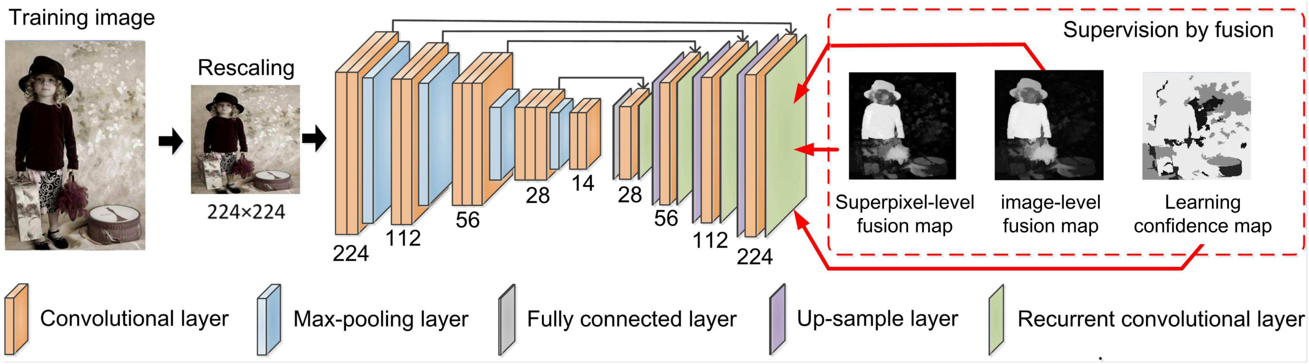

Figure 4. The architecture of the network for the adopted deep salient object detector.

network architecture is shown in Fig. 4. Specifically, it takes an rescaled image of size 224 × 224 as input, then the 13 convolutional layers of VGG 16-layer network [30] is adopted to extract deep feature maps. Afterwards, on top of the last convolutional feature map (with the size of $14 \times 14 \times 512$), a fully connected layer with sigmoid activation function and $28 \times 28 \times 1$ nodes is deployed, which can generate the coarse saliency prediction. To progressively render image details to improve the spatial resolution, 4 recurrent convolutional layers [25] are used subsequently. For each recurrent convolutional layer, the input is the 65-channel feature maps which concatenate a fully connected layer/up-sampling layer and a corresponding convolutional layer squashed by a 64 convolutional kernels and a sigmoid activation function (shown as the black arrows in Fig. 4). The output is the 1-channel refined saliency map with the same size of the input feature maps. The final predicted saliency map is obtained by the last recurrent convolutional layer, with the size $D = 224 \times 224$. For training such deep salient object detector without using any human annotation, we introduce three channels of supervisory signals, including the superpixel-level fusion maps $\{\mathbf{SFM}_n\}$, the image-level fusion maps $\{\mathbf{IFM}_n\}$, and the learning confidence maps $\{\mathbf{LCM}_n\}$, where $\mathbf{LCM}_n = \mathbf{b}_n \cdot \beta_n$ [7].

2.3.2 The Learning Strategy

Denote the trainable parameters of the network as $\mathbf{W}$, which can be optimized using the back propagation algorithm [28] via minimizing the cost function $L(\mathbf{W})$:

$$L(\mathbf{W}) = -\frac{1}{B}\sum_{n=1}^{B}\sum_{d=1}^{D}[\mathbf{LCM}_{n,d} \cdot H(\mathbf{SFM}_{n,d}, \Psi(\mathbf{I}_n|\mathbf{W})_d) + \mathbf{LCM}_{n,d} \cdot H(\mathbf{IFM}_{n,d}, \Psi(\mathbf{I}_n|\mathbf{W})_d)] + \lambda r(\mathbf{W}), \quad (7)$$

where $H(p,q) = p\log(q) + (1-p)\log(1-q)$ is the cross-entropy loss, B indicates the size of the minibatch, $LCM_{n,d} \in [0,1]$ indicates the d-th element of the learning confidence map $\mathbf{LCM}_n$, $SFM_{n,d} \in \{0,1\}$ and $IFM_{n,d} \in \{0,1\}$ indicate the d-th element of the binarized superpixel-level fusion map $\mathbf{SFM}_n$ and image-level fusion map $\mathbf{IFM}_n$, respectively, $\Psi(\mathbf{I}_n|\mathbf{W})_d \in [0,1]$ indicates the d-th element of the predicted saliency map $\Psi(\mathbf{I}_n|\mathbf{W})$, $r(\cdot)$ is the squared ℓ_2-norm, and λ is the weight decay factor, d indicates the pixel index for the fusion maps and the confidence maps. To facilitate learning, we also adopt the deep supervision [15] scheme, where the supervisory signals are applied to supervise each of the recurrent convolutional layer.

In Eq. (7), the first term penalizes the predictions which are inconsistent with the superpixel-level fusion maps, while the second term penalizes the predictions which are inconsistent with the image-level fusion maps. As either the superpixel-level fusion maps or the image-level fusion maps are not perfect, collaborating these two kinds of fusion maps could provide supervision that is complementary to each other, leading to more effective learning performance as demonstrated in Sec. 3.3. The learning confidence maps are used to provide the superpixel-wise learning weight for the learning process, which naturally compiles to the modern self-paced learning (SPL) and curriculum learning (CL) regimes [14, 42, 9, 2]. Specifically, SPL and CL are the weighting-based robust learning regime. The core of SPL is to alternately infer the learning confidence of each training sample and learn the model in iterations while the core of CL is to use a pre-defined learning curriculum to guide the learning from easy examples to more complex ones. Besides, the proposed learning scheme also has some interesting differences compared with SPL and CL: 1) Instead of inferring the self-paced learning weights via the "internal force" (the learnt classifier itself), the learning weights of the proposed learning scheme are inferred by the "external force" (the fusion process). 2) Rather than pre-defining a learning curriculum and fixing it during the entire learning procedure, the learning curriculum formed by the learning weights are updated along the learning iterations. Moreover, SPL and CL are usually used under the supervision

[7] Before calculating $\{\mathbf{LCM}_n\}$, values in each $\mathbf{b}_n$ are normalized to [0,1] and $\{\beta_n\}$ are normalized to [0,1].

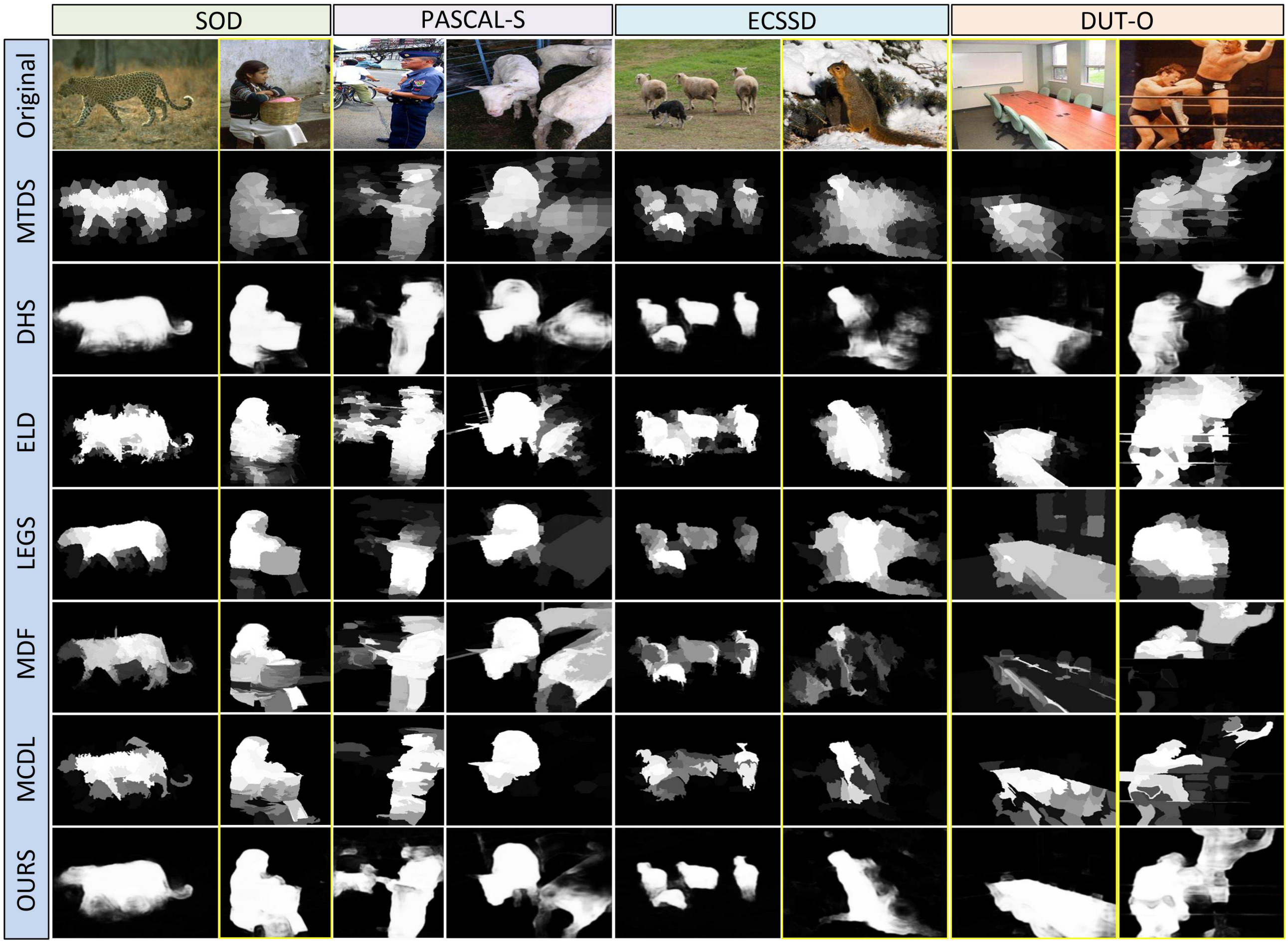

Figure 5. Some visualization examples of the saliency maps obtained by the proposed unsupervised learning framework as well as the other supervised state-of-the-art methods. Yellow columns indicate the examples when our method outperforms the supervised state-of-the-arts.

in some forms of the human annotation, while the proposed learning scheme works in unsupervised learning scenario.

3. Experiments

3.1. Experimental Settings

We implemented comprehensive experiments by using five widely used salient object benchmark datasets, which are the MSRA10K [5], ECSSD [37], SOD [26], DUT-O [38], and PASCAL-S [21], respectively. Specifically, M-SRA10K is the largest dataset which contains 10,000 images with various object categories. SOD, ECSSD, DUT-O, and PASCAL-S consist of 300, 1000, 5168, and 850 images, respectively. We used the MSRA10K dataset for training our network as well as all the compared baseline networks and used the rest four for testing.

To quantitatively evaluate the experimental results, we utilize five evaluation metrics, which are the Precision-Recall (PR) curve, AP score, F-measure, mean absolute error (MAE), and SOV, respectively. Specifically, the PR curve reflects the mean precision and recall of saliency maps at different thresholds, while AP score is obtained by accumulating the area of the PR curve. The F-measure is calculated by $F_\iota = \frac{(1+\iota^2)Precision \times Recall}{\iota^2 \times Precision + Recall}$, where $Precision$ and $Recall$ are obtained using twice the mean saliency value of saliency maps as the threshold and ι^2 is set to 0.3 according to [25, 38]. The MAE is the average pixel-wise difference between the predicted saliency map and the corresponding ground truth. The SOV [19] is the intersection-over-union overlap between the ground truth mask and the predicted saliency map binarized by using the same adaptive threshold as during the calculation of F-measure.

Before training, we increased the training set through image augmentation implemented by horizontal-flipping and image cropping as suggested by [25]. During the training process, the parameters in the 13 convolutional layers were initialized by the VGG net while other parameters were initialized randomly. As suggested by [25], we set the mini-batch size to 12 and the iteration step to 50,000. The learning rate was set to 0.03 in the 4 recurrent convolutional layers while 1/1000 smaller in the 13 convolutional layers. It was halved every 5,000 iterations. The momentum was set

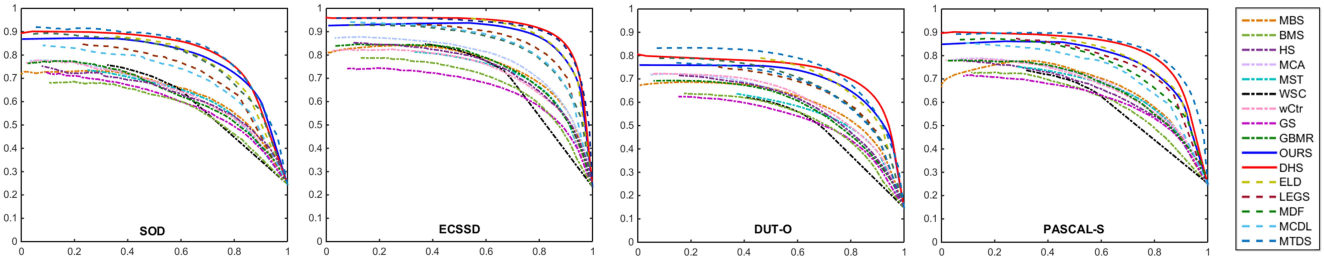

Figure 6. Comparison with other state-of-the-art methods on four saliency detection datasets in terms of the PR curve.

Table 1. Comparison with other state-of-the-art methods on four saliency detection datasets in terms of AP, F_ι, SOV (higher values indicate better results), and MAE (lower values indicate better results). The results listed in the top block are obtained from the state-of-the-art unsupervised methods. The results listed in the bottom block are obtained from the state-of-the-art supervised methods (using deep models). The bold numbers indicate the best unsupervised performance.

	SOD				ECSSD				DUT-O				PASCAL-S			
	AP	F_ι	MAE	SOV	AP	F_ι	MAE	SOV	AP	F_ι	MAE	SOV	AP	F_ι	MAE	SOV
MBS	.636	.580	.222	.419	.750	.674	.168	.534	.598	.519	.199	.426	.665	.608	.202	.443
BMS	.594	.530	.255	.375	.695	.637	.204	.480	.564	.487	.228	.392	.626	.564	.235	.400
HS	.638	.498	.252	.333	.750	.627	.189	.454	.614	.513	.245	.397	.660	.528	.237	.353
MCA	.645	.570	.240	.413	.780	.702	.173	.556	.602	.509	.235	.422	.677	.600	.218	.434
MST	.662	.583	.205	.423	.758	.677	.155	.539	.624	.517	.172	.417	.690	.609	.194	.453
WSC	.661	.590	.191	.431	.743	.680	.149	.529	.593	.496	.146	.395	.633	.568	.192	.424
wCtr	.641	.584	.202	.429	.727	.676	.157	.531	.616	.528	.183	.433	.667	.600	.197	.439
GS	.612	.546	.247	.405	.664	.608	.205	.477	.553	.465	.227	.383	.626	.562	.226	.416
GBMR	.638	.569	.243	.403	.751	.689	.163	.524	.589	.527	.228	.424	.670	.607	.217	.438
OURS	**.789**	**.676**	**.140**	**.545**	**.889**	**.787**	**.085**	**.677**	**.715**	**.583**	**.135**	**.505**	**.791**	**.680**	**.141**	**.549**
DHS	.860	.724	.121	.578	.916	.832	.070	.720	.728	.637	.104	.552	.823	.729	.115	.597
MTDS	.822	.708	.138	.570	.915	.817	.083	.707	.751	.606	.132	.529	.842	.726	.120	.591
ELD	.798	.706	.134	.555	.913	.806	.081	.698	.754	.607	.110	.522	.818	.716	.125	.578
LEGS	.747	.674	.169	.506	.868	.783	.100	.632	.694	.586	.152	.483	.791	.694	.147	.530
MDF	.779	.687	.139	.523	.846	.749	.108	.612	.686	.596	.132	.487	.767	.721	.146	.500
MCDL	.721	.667	.151	.513	.849	.777	.097	.649	.678	.605	.110	.511	.742	.674	.144	.524

to 0.9 and weight decay factor was set to 0.0005. The whole network was implemented by using the caffe [11] toolbox.

3.2. Comparison with the State-of-the-arts

In this section, we compare the proposed approach with other 15 state-of-the-art approaches, which include the unsupervised approaches such as MBS [45], BMS [44], HS [29], MCA [27], MST [31], WSC [18], wCtr [47], GS [35], GBMR [38], and the supervised deep saliency models such as DHS [25], MTDS [20] MCDL [46], MDF [17], LEGS [32], ELD [16]. Some of the reported results might differ from those in other papers due to the usage of different training data. Some results are visualized in Fig. 5.

For quantitative evaluation, we report comparison results with PR curves in Fig. 6 and those with AP, F-measure, MAE, and SOV scores in Table 1. In Table 1, the methods listed in the top block are the state-of-the-art unsupervised methods. "OURS" and DHS listed in the middle block are the proposed approach as well as its supervised baseline, respectively. More specifically, "OURS" and DHS utilized the same network architecture. The only difference is that the supervision of DHS comes from the pixel-wise human annotation whereas the supervision of our approach is automatically generated by the "supervision by fusion" framework. Thus, by training with the human labelled ground truth, DHS actually provides the upper boundary performance of the proposed approach. The methods listed in the bottom block are a number of state-of-the-art supervised deep saliency models, which were trained on the same image data with our approach but with additional human annotation. From the comparison with the unsupervised methods, we can observe that the proposed approach could significantly improve the performance in terms of all the evaluation metrics on all datasets. Notably, by only using additional unlabelled images, our approach can outperform the best unsupervised approach by more than 10% in terms of the AP score. From the comparison with DHS, we can observe that the proposed method can approach the supervised baseline within 2% to 5% performance gap, which indicates that our method can tremendously alleviate the human labor in annotating the training data[8] while only obtaining relatively insignificant performance drop. Thus, it demonstrates the cost-effectiveness of the proposed approach. More encouragingly, when compared with a number of other state-of-the-art supervised deep salient object detectors, our un-

[8] Based on our statistics, manual annotation needs around 31 seconds per-image, while our approach only needs around 0.7 second to obtain the weak saliency maps for each training image.

Table 2. Evaluation of the supervisory signals on four saliency detection datasets in terms of AP, F_ι, SOV (higher values indicate better results), and MAE (lower values indicate better results).

	SOD				ECSSD				DUT-O				PASCAL-S			
	AP	F_ι	MAE	SOV	AP	F_ι	MAE	SOV	AP	F_ι	MAE	SOV	AP	F_ι	MAE	SOV
-SFM	.750	.647	.154	.520	.852	.760	.099	.657	.640	.541	.162	.471	.754	.649	.151	.528
-b	.754	.658	.150	.526	.864	.768	.093	.662	.667	.561	.148	.487	.759	.664	.148	.541
-IFM	.749	.649	.150	.525	.844	.757	.096	.657	.649	.552	.152	.481	.750	.653	.149	.534
-β	.754	.674	.148	.534	.876	.781	.095	.660	.677	.579	.142	.498	.777	.679	.144	.549
OURS	.789	.676	.140	.545	.889	.787	.085	.677	.715	.583	.135	.505	.791	.680	.141	.549

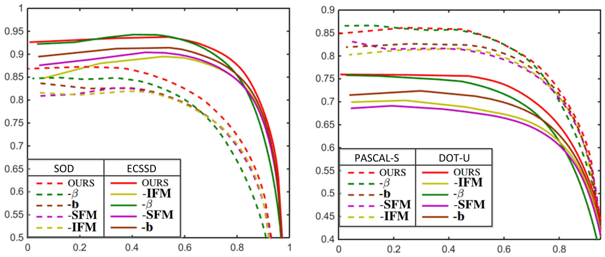

Figure 7. Evaluation of the supervisory signals on four saliency detection datasets in terms of PR curve.

supervised learning framework can still obtain competitive performance. Notably, in terms of AP, our approach can even outperform LEGS, MDF, and MCDL.

3.3. Ablation Studies

In this section, we conducted ablation studies on four saliency detection datasets to comprehensively evaluate the importance of different supervisory signals used in the proposed unsupervised learning framework. Specifically, we in turn removed a certain supervisory signal and used the performance drop to estimate the importance of the corresponding supervisory signal to the full system. The experimental results are shown in Table. 2 and Fig. 7, where "**-SFM**", "**-b**","**-IFM**","-β" indicate the learning framework without using the supervisory signal of the superpixel-level fusion map, the superpixel-level confidence map, the image-level fusion map, and the image-level confidence weight, respectively. From the experimental results, we can observe that:

1) All the four types of supervisory signals are beneficial to the proposed learning framework as without using each of them can cause obvious performance drop, especially in terms of AP and F_ι; 2) Basically, the average performance drop (i.e., the importance) of the used supervisory signals follows: **SFM** (0.483[9]) > **IFM** (0.431) > **b** (0.312) > β (0.186), which demonstrates that the superpixel-level fusion map and the image-level fusion map provide the major supervision for the network while the superpixel-level confidence map and the image-level confidence weight provide additionally helpful supervision; 3) Consistent and obvious performance gain can be obtained by the full learning framework as compared with other baselines, which demonstrates the rationality of the proposed "supervision by fusion" strategy as well as the established unsupervised learning framework. 4) The performance drop on the relatively more challenging salient object detection datasets, e.g., DUT-O, tends to be more significant than it on the relatively less challenging ones, e.g., ECSSD, which indicates that the explored supervisory signals are more valuable in dealing with more challenging scenarios.

4. Conclusion

This paper has proposed a novel unsupervised learning framework to train the DNN-based salient object detector. It revealed the insight of "supervision by fusion" and established a novel two-stream framework to generate useful supervisory signals through the intra-image fusion and inter-image fusion processes. Comprehensive experiments on four benchmark datasets have demonstrated the effectiveness of the proposed approach as well as each of the used supervisory signal. Notably, our method can approach the the same network trained with full supervision (within 2-5% performance gap) and, more encouragingly, outperformed a number of fully supervised state-of-the-art approaches. For future work, we will apply the proposed unsupervised deep learning technique into solving a wide range of computer vision tasks, such as co-saliency detection [39, 41], semantic segmentation [33, 34], and object parsing [23, 24].

Acknowledgement: This work was supported in part by the National Science Foundation of China under Grant 61473231 and Grant 61522207.

References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[3] X. Chang, Y. Yang, E. P. Xing, and Y.-L. Yu. Complex event detection using semantic saliency and nearly-isotonic svm. In *ICML*, 2015.

[4] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li. Disc: Deep image saliency computing via progressive representation learning. *TNNLS*, 27(6):1135–1149, 2016.

[5] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu. Global contrast based salient region detection. In *CVPR*, 2011.

[9] The reported number is the average performance drop across all the evaluation metrics and datasets.

[6] P. F. Felzenszwalb and D. P. Huttenlocher. Efficient graph-based image segmentation. *IJCV*, 59(2):167–181, 2004.

[7] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu. Visual-textual joint relevance learning for tag-based social image search. *TIP*, 22(1):363–376, 2013.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. *TPAMI*, 34(10):1915–1926, 2012.

[9] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, and J. Yang. Multi-modal curriculum learning for semi-supervised image classification. *TIP*, 25(7):3249–3260, 2016.

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[11] 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 *ACM-MM*, 2014.

[12] 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.

[13] D. A. Klein and S. Frintrop. Center-surround divergence of feature statistics for salient object detection. In *ICCV*, 2011.

[14] M. P. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[15] C.-Y. Lee, S. Xie, P. W. Gallagher, Z. Zhang, and Z. Tu. Deeply-supervised nets. In *AISTATS*, 2015.

[16] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *ICCV*, 2016.

[17] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *CVPR*, 2015.

[18] N. Li, B. Sun, and J. Yu. A weighted sparse coding framework for saliency detection. In *CVPR*, 2015.

[19] X. Li, Y. Li, C. Shen, A. Dick, and A. Van Den Hengel. Contextual hypergraph modeling for salient object detection. In *ICCV*, 2013.

[20] 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. *TIP*, 25(8):3919–3930, 2016.

[21] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *CVPR*, 2014.

[22] Y. Li, M. Paluri, J. M. Rehg, and P. Dollár. Unsupervised learning of edges. *arXiv preprint arXiv:1511.04166*, 2015.

[23] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *TPAMI*, 37(12):2402–2414, 2015.

[24] X. Liang, X. Shen, D. Xiang, J. Feng, L. Lin, and S. Yan. Semantic object parsing with local-global long short-term memory. In *CVPR*, 2016.

[25] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, 2016.

[26] V. Movahedi and J. H. Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPRW*, 2010.

[27] Y. Qin, H. Lu, Y. Xu, and H. Wang. Saliency detection via cellular automata. In *CVPR*, 2015.

[28] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. Technical report, DTIC Document, 1985.

[29] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 38(4):717–729, 2016.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[31] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien. Real-time salient object detection with a minimum spanning tree. In *CVPR*, 2016.

[32] L. Wang, H. Lu, X. Ruan, and M.-H. Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015.

[33] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. *CVPR*, 2017.

[34] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan. Stc: A simple to complex framework for weakly-supervised semantic segmentation. *TPAMI*, 2016.

[35] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *ECCV*, 2012.

[36] J. Whitehill, T.-f. 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 *NIPS*, 2009.

[37] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *CVPR*, 2013.

[38] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013.

[39] D. Zhang, J. Han, J. Han, and L. Shao. Cosaliency detection based on intrasaliency prior transfer and deep intersaliency mining. *TNNLS*, 27(6):1163–1176, 2016.

[40] D. Zhang, J. Han, L. Jiang, S. Ye, and X. Chang. Revealing event saliency in unconstrained video collection. *TIP*, 26(4):1746–1758, 2017.

[41] D. Zhang, J. Han, C. Li, J. Wang, and X. Li. Detection of co-salient objects by looking deep and wide. *IJCV*, 120(2):215–232, 2016.

[42] D. Zhang, D. Meng, and J. Han. Co-saliency detection via a self-paced multiple-instance learning framework. *TPAMI*, 39(5):865–878, 2017.

[43] D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. *IJCAI*, 2016.

[44] J. Zhang and S. Sclaroff. Exploiting surroundedness for saliency detection: a boolean map approach. *TPAMI*, 38(5):889–902, 2016.

[45] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech. Minimum barrier salient object detection at 80 fps. In *ICCV*, 2015.

[46] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *CVPR*, 2015.

[47] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *CVPR*, 2014.