Multi-Source Uncertainty Mining for Deep Unsupervised Saliency Detection

Yifan Wang[1], Wenbo Zhang[1], Lijun Wang[1]*, Ting Liu[3], Huchuan Lu[1,2]
[1]Dalian University of Technology, [2]Peng Cheng Laboratory, [3]Alibaba Group

Abstract

Deep learning-based image salient object detection (SOD) heavily relies on large-scale training data with pixel-wise labeling. High-quality labels involve intensive labor and are expensive to acquire. In this paper, we propose a novel multi-source uncertainty mining method to facilitate unsupervised deep learning from multiple noisy labels generated by traditional handcrafted SOD methods. We design an Uncertainty Mining Network (UMNet) which consists of multiple Merge-and-Split (MS) modules to recursively analyze the commonality and difference among multiple noisy labels and infer pixel-wise uncertainty map for each label. Meanwhile, we model the noisy labels using Gibbs distribution and propose a weighted uncertainty loss to jointly train the UMNet with the SOD network. As a consequence, our UMNet can adaptively select reliable labels for SOD network learning. Extensive experiments on benchmark datasets demonstrate that our method not only outperforms existing unsupervised methods, but also is on par with fully-supervised state-of-the-art models.

1. Introduction

Image salient object detection (SOD) aims at identifying and segmenting the most prominent object in a scene. Existing SOD methods can be mainly divided into two categories, *i.e*., convolutional neural network (CNN) based and traditional handcrafted methods. Both of them have their unique pros and cons. On the one hand, driven by the strong model capacity of deep networks, CNN based SOD methods have achieved remarkable success. However, they heavily rely on large amounts of training data with pixel-wise annotations, which are labor-intensive and expensive to acquire. On the other hand, handcrafted SOD methods are more flexible to the data annotations, but they are fragile in practice due to the limitations of manually designed image features and priors.

With the above concern, one research topic termed deep unsupervised SOD [20,38,40,43] has been activated, which focuses on training the deep SOD networks using the noisy pseudo labels generated by traditional handcrafted SOD methods. Directly training networks using the noisy labels is not a wise choice since the deep network can easily fit to the corrupted labels [37]. One straightforward solution is first performing label refinement and then using the refined labels for network training [20]. Another popular line [38, 40, 43] devotes to modeling the noise of the pseudo labels. For instance, the work of [43] assumes that the label noise obeys a Gaussian distribution and builds a noise modeling module to fit such distribution. Zhang *et al.* [40] compute a dense confidence map based on the variance of network predictions among different training iterations. While promising results have been delivered, it is still an open problem to model the noisy labels and find the reliable ones in an unsupervised learning manner.

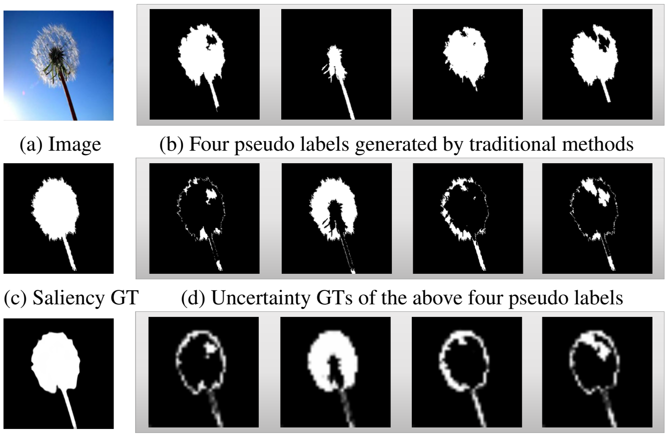

(a) Image (b) Four pseudo labels generated by traditional methods
(c) Saliency GT (d) Uncertainty GTs of the above four pseudo labels
(e) Saliency prediction (f) Our predicted uncertainty maps

Figure 1. Motivation. Given an input image (a) and its corresponding four pseudo labels generated by the traditional SOD methods (b), our UMNet predicts the uncertainty maps (f) of the pseudo labels, according to which our SODNet is learned under the supervision of the reliable labeling samples and generates promising saliency result (e). The ground truths of uncertainty maps (d) are obtained by the computing the difference between each pseudo label in (b) with the saliency ground truth (c), which are not available under the unsupervised learning setting.

In this paper, we establish a novel deep unsupervised SOD framework for effectively mining the reliable pixel-

[*]Corresponding author: Lijun Wang, ljwang@dlut.edu.cn

wise labels from multiple pseudo labels. As shown in Figure 1 (b), different handcrafted methods perform diversely for the same input image since they follow different manually designed principles. Nevertheless, each of the pseudo labels contains some reliable label samples (*cf*. the dark region of Figure 1 (d)). Accurately identifying these reliable/certain samples is troublesome when only observing single pseudo label without any other reference. Alternatively, it becomes much feasible if we simultaneously employ multiple labels of the same image for cross reference. Based on this insight, we design a novel Uncertainty Mining Network (UMNet) to densely capture the soft uncertainty from multi-source pseudo labels. It consists of multiple Merge-and-Split (MS) modules and infers the pixel-wise uncertainty map for each label by recursively analyzing the commonality and difference among multiple noisy labels. According to the predicted uncertainty by the UMNet, the Salient Object Detection Network (SODNet) can be learned using the reliable label samples.

For network training, another concern is encountered. Considering that the ground truth of uncertainty is not available under the unsupervised setting, it may lead to a trivial solution for the UMNet optimization, *e.g*., all the labels are uncertain. We attack this issue by modeling the noisy labels using Gibbs distribution under the Bayesian framework and developing an uncertainty weighted loss function for end-to-end training UMNet with SODNet. As a consequence, our UMNet is able to effectively identify the reliable pseudo labels while softly filtering out those of low qualities. The selected reliable pseudo labels are employed to provide supervision on SODNet, leading to more superior performance.

The contributions of this work can be summarized into three folds as follows.

(1) We develop a novel deep unsupervised SOD paradigm which automatically learns to mine the reliable labels from noisy pseudo ones of multiple sources, leading to more effective unsupervised learning.

(2) We present a Merge-and-Split module that helps the uncertainty mining network to effectively capture the per-pixel reliability of the pseudo labels by simultaneously analyzing the commonality and difference of multi-source noisy labels.

(3) We propose an uncertainty weighted loss function that models the noisy labels as Gibbs distribution in a principled way, allowing the whole networks to be jointly trained in an elegant manner without uncertainty annotations.

Experiments on popular SOD benchmark datasets show that the proposed method can effectively facilitate the SOD network learning with noisy labels and achieves the state-of-the-art performance.

2. Related Works

2.1. Fully-supervised SOD

With the development of deep learning technique and SOD benchmark datasets [16, 27, 35], it has achieved great evolution in the SOD research community [23,30,33]. Early works [11], [26] utilize multi-layer perception (MLP) classifiers to detect salient regions patch by patch, which fail to effectively capture spatial information of images and are also time-consuming. Later on, FCN-based methods such as [17, 21, 24, 28, 45] are dominant in this field, achieving more competitive performance in terms of both accuracy and speed. However, the fully-supervised SOD methods mainly depend on large-scale pixel-level labeled training data, which are expensive to obtain in practice.

2.2. Semi-/Weakly- supervised SOD

To relieve the burden of handcrafted labeling, researchers explore to learn the SOD networks by using some weak supervisions, such as image-level category labels [27], scribbles [41], image captions [36], *etc*. For instance, Wang *et al.* [27] design a foreground inference network to capture the potential salient regions by learning the image-level category prediction task. The work of [41] introduces an auxiliary edge detection task to learn salient object detection using scribble annotations, which only costs no more than 2 seconds to label an image. Li *et al.* [15] propose to integrate a new branch with a well-trained contour detection network to estimate saliency score for each pixel. In [36], a unified framework is developed to train saliency detection models with both category labels and image captions. Besides, inspired by the self-paced learning technique, [39] designs an adversarial-paced learning based framework to learn SOD task using only a few pixel-level training labels, which can be seen as semi-supervised SOD task. While promising performance has been achieved, these works are still reliant to annotations.

2.3. Unsupervised SOD

Traditional SOD works [8], [22], [31], [35] can be classified as unsupervised learning methods relying on manually designed image features and data priors. While being free from data annotations, the generalization ability of these methods is limited in especially complex situations. Recently, unsupervised SOD has been promoted by employing deep neural networks, where the SOD network is supervised with the noisy labels that are generated by using the handcrafted methods. The earliest effort is made by [38], which proposes a "supervision by fusion" (SBF) strategy that generates reliable supervisory labels by the fusion process of handcrafted SOD models in iterative learning stages. The work of [43] introduces a noise modeling module with a strong assumption that the label noise obeys a Gaussian

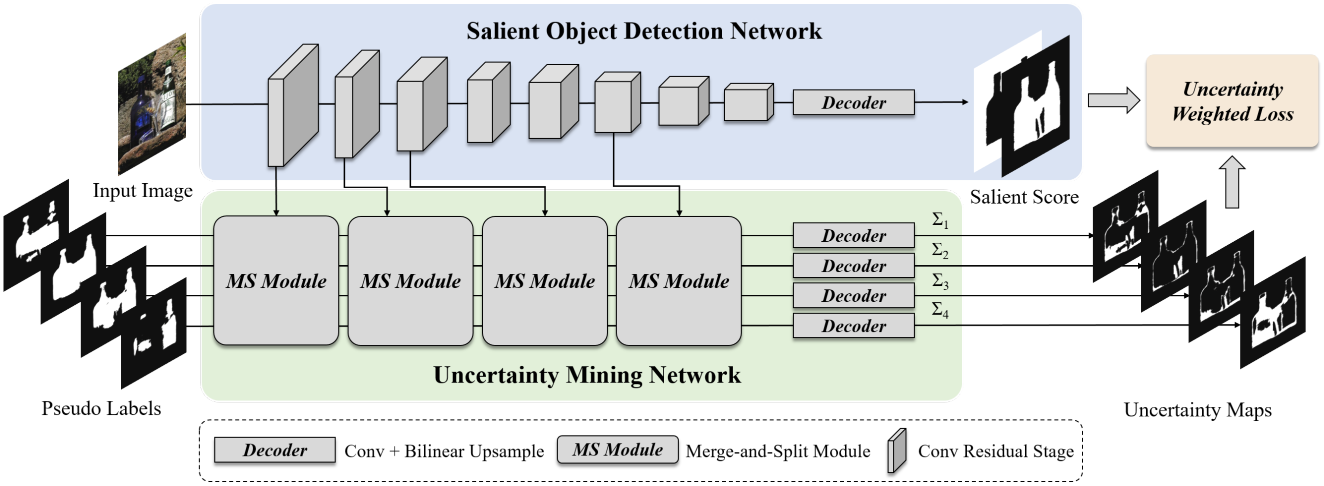


Figure 2. Overview of the proposed deep unsupervised SOD framework. Given the input image and M noisy pseudo labels, the salient object detection network (top) predicts the two-channel saliency score map including the salient foreground and background. Meanwhile, the uncertainty mining network (bottom) containing multiple merge-and-split modules takes M pseudo labels and multi-stage image features as input, and produces M dense uncertainty maps. The whole networks are jointly trained under the supervision of the uncertainty weighted loss. In practice, we adopt four handcrafted SOD methods to generate the pseudo labels, *i.e.* $M = 4$. At inference, only the trained SOD network is employed for saliency prediction.

distribution. Zhang *et al.* [40] propose to select the reliable labels by computing a dense confidence map, which is used to reweight the cost function per-pixel. The current leading method [20] first refines the noisy labels in multiple iterations by using moving average and fully-connected CRF, and then trains the SOD network directly using all the refined labels. While delivering favorable SOD results, it is not an optimal solution to use all the refined labels without any selection since there still exists much of noise. Although the prior works [38, 40, 43] have devoted to modeling the label noise or computing the label confidence, it involves dedicated assumptions and manual designs. In contrast, in this work, we mine the reliable labels via the uncertainty mining network (UMNet) integrating our merge-and-split modules, and train the whole networks in the data-driven manner, which help UMNet to produce more robust uncertainty predictions and further facilitate the SOD task.

3. Method

This work focuses on learning deep SOD network without the annotated ground truth. One potential way is directly using the pseudo labels generated by the traditional handcrafted SOD methods as supervision. However, as shown in Figure 1 (b), these pseudo labels contain large amounts of noise with strong inconsistency, which could inevitably hinder the network learning and degrade the final SOD performance. To circumvent this issue, we resort to evaluate the reliability of each label pixel by learning a dense uncertainty map for each given label, upon which the SOD network is supervised.

The proposed framework is depicted in Figure 2, which consists of a Salient Object Detection Network (SODNet) and a Multi-source Uncertainty Mining Network (UMNet). Given an input image $\mathbf{I}$, we first obtain M pseudo labels $\{\mathbf{Y}_m\}_{m=1}^{M}$ using M different handcrafted SOD methods. Inspired by [20], all the pseudo labels are separately refined using the first stage of method [20] to improve their quality. SODNet aims to detect the salient object in $\mathbf{I}$ and predicts the saliency mask. Meanwhile, all the refined pseudo labels as well as the multi-scale image features extracted by SODNet are fed into UMNet, and the uncertainty maps $\{\mathbf{\Sigma}_m\}_{m=1}^{M}$ for the pseudo labels are generated. The whole networks including SODNet and UMNet are jointly trained under the supervision of the proposed uncertainty weighted loss, which helps UMNet to accurately estimate the dense uncertainty maps of all the pseudo labels and further facilitates the SODNet learning under the unsupervised setting.

As shown on the top row of Figure 2, given the input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, SODNet outputs the saliency score map $\mathbf{S} \in \mathbb{R}^{H\times W\times N_c}$ of N_c channels which can be further normalized into the saliency probability map using either Sigmoid ($N_c = 1$) or Softmax ($N_c = 2$) function. The architecture design for SODNet is not the focus of this paper, and many existing SOD models can be used in our framework. We adopt the same network architectures used in [20, 43] for fair comparison. Specifically, it builds upon the dilated residual network (DRN) [2], which modifies the original ResNet101 by replacing all the fully connected layers with convolutional layers and using the atrous convolutional layers to preserve feature resolutions. The output resolution of the last convolutional layer is 1/8 of the input image, which is finally upsampled to the original input resolution using the nearest neighbour interpolation. As shown in Figure 2, SODNet contains eight convolutional residual stages [2]. We combine the features extracted from the first three and the sixth stages of SODNet as hierarchical guidance for label uncertainty prediction in UMNet. In the following, we will describe the proposed UMNet (Section 3.1) and the uncertainty weighted loss (Section 3.2) in detail.

3.1. Uncertainty Mining Network with Merge-and-Split Modules

The uncertainty mining network (UMNet) is designed to identify the pixel-wise reliability for each given pseudo label by observing all the pseudo labels and the input image. Our key insight is that it is hard for the network to predict the uncertainty based on only single pseudo label since there is no additional reference. In comparison, by simultaneously considering the diversity of multiple pseudo labels produced by different SOD methods, it can help the network to more effectively capture the reliable label samples through analyzing the commonality and difference among different pseudo labels.

As shown in Figure 2, the proposed UMNet consists of M branches. Each branch extracts features from one pseudo label and produces the corresponding uncertainty map. Features among pseudo labels hierarchically interact with each other via the proposed merge-and-split (MS) modules, which enhances the features to generate more robust uncertainty results by gathering and analyzing information from all the pseudo labels.

The architecture details of the proposed MS module are depicted in Figure 3. Given M label features and the image feature generated by a specific stage of SODNet, they are firstly fed into different residual blocks [7] respectively as a pre-processing step. Then an information aggregation operation including the channel-wise concatenation and an additional residual block is applied to merge the information from all the pseudo labels and the input image. Each label feature is then combined with the merged features via channel-wise concatenation, followed by a residual block to produce the output feature. The merge-and-split mechanism provides the opportunity for each network branch dedicated to one pseudo label to see more comprehensive information, and thus empowers the network to predict more robust and accurate uncertainty maps.

In this work, four MS modules are adopted as the core components in the UMNet, which receive the image features from the 1st, 2nd, 3rd, and 6th residual stages of SODNet, respectively. Besides, the output resolution of the first three MS modules are downsampled to 1/2 of the input resolution through strided convolutional layers. The output M label features from the last MS module are finally passed through M decoders comprising a convolutional and a nearest neighbor upsampling layer, which generate the uncertainty maps for the M pseudo labels, respectively.

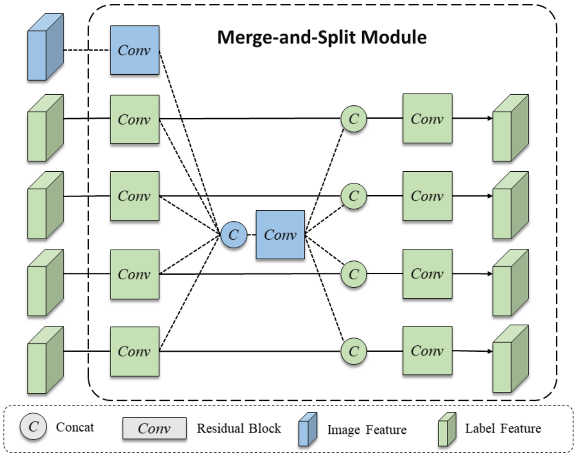


Figure 3. Illustration of the proposed merge-and-split module.

3.2. Network Learning with Multi-source Uncertainty Mining

Since the ground truth of uncertainty maps are not available, training UMNet to generate the desired results is not trivial issue. Let us denote the pseudo label of pixel i generated by the m-th handcrafted SOD method as $\mathbf{Y}_m^i = c$, with $c = 1$ indicating salient foreground and $c = 0$ otherwise. The predicted uncertainty and saliency score are represented by $\mathbf{\Sigma}_m^i$ and $\mathbf{S}^i$, respectively. Inspired by [9, 10], we model the noisy pseudo label as a random variable y subject to Gibbs distribution under the Bayesian theory. When using Softmax function to normalize the saliency score, the probability distribution of y can be computed as:

$$p(y|\mathbf{S}^i, \mathbf{\Sigma}_m^i) = \text{Softmax}\left(\frac{1}{(\mathbf{\Sigma}_m^i)^2}\mathbf{S}^i\right). \tag{1}$$

The magnitude of the learned uncertainty $\mathbf{\Sigma}_m^i$ determines the uniform/flat degree of the distribution. Given the observed pseudo label $\mathbf{Y}_m^i = c$, the negative log likelihood can then be derived as follows:

$$\begin{aligned} &-\log p(y = \mathbf{Y}_m^i|\mathbf{S}^i, \mathbf{\Sigma}_m^i) \\ =& -\frac{1}{(\mathbf{\Sigma}_m^i)^2}s_c^i + \log\sum_{c'}\exp\left(\frac{1}{(\mathbf{\Sigma}_m^i)^2}s_{c'}^i\right) \\ \approx& \frac{1}{(\mathbf{\Sigma}_m^i)^2}L_{CE}\left(\mathbf{Y}_m^i, \mathbf{S}^i\right) + \log\mathbf{\Sigma}_m^i, \end{aligned} \tag{2}$$

where $L_{CE}(\mathbf{Y}_m^i, \mathbf{S}^i)$ denotes the cross-entropy loss computed using the pseudo label $\mathbf{Y}_m^i$ and unnormlaized score $\mathbf{S}^i$ (*i.e.*, the logit). Detailed derivation can be found in [10].

Since Sigmoid function is a special case for Softmax function, the above derivation using Softmax also holds when using Sigmoid function to normalize the saliency scores. The first term on the right hand side of (2) shows that the uncertainty $\mathbf{\Sigma}_m^i$ with large value will decrease the contribution of $L_{CE}(\mathbf{Y}_m^i, \mathbf{S}^i)$, whereas with small value will increase its contribution. Meanwhile, the last term can be seen as a regularization imposed on $\mathbf{\Sigma}_m^i$ predicted by UMNet, which will be penalized if the value of $\mathbf{\Sigma}_m^i$ is too large.

We extend the above formulation to the whole image with M pseudo labels, and obtain the final loss $\mathcal{L}(\Theta_{sod}, \Theta_\sigma)$ as follows:

$$\begin{aligned}\mathcal{L}(\Theta_S, \Theta_\Sigma) &= -\sum_{m=1}^{M}\sum_{i}\log p(y = \mathbf{Y}_m^i | \mathbf{S}^i, \mathbf{\Sigma}_m^i) \\ &= \sum_{m=1}^{M}\sum_{i}^{H\times W}\frac{1}{(\mathbf{\Sigma}_m^i)^2}L_{CE}(\mathbf{Y}_m^i, \mathbf{S}^i) + \log \mathbf{\Sigma}_m^i\end{aligned} \tag{3}$$

where Θ_S and Θ_Σ are the trainable parameters of SODNet and UPNet, respectively. H and W are the height and width of the input image, respectively. We denote (3) as the uncertainty weighted loss, which allows the whole networks to be jointly learned in a principled way.

In practice, we made two modifications on (3) to improve training stability. Instead of directly generating the uncertainty map $\mathbf{\Sigma}_m$, our UMNet predicts its logarithmic form, *i.e.* $\mathbf{E}_m^i = \log(\mathbf{\Sigma}_m^i)^2$. Besides, we adopt the Sigmoid unit as the output layer of our UMNet and further normalize the predicted logarithmic uncertainty values to the interval of $[-\tau, \tau]$. As a consequence, (3) can be rewritten as:

$$\mathcal{L}(\Theta_S, \Theta_\Sigma) = \sum_{m=1}^{M}\sum_{i}^{H\times W}\exp(-\mathbf{E}_m^i)L_{CE}(\mathbf{Y}_m^i, \mathbf{S}^i) + \frac{1}{2}\mathbf{E}_m^i. \tag{4}$$

3.3. Implement Details

Our training settings mainly follow the recent leading work [20]. The training data and validation data consist of 2500 and 500 images, respectively, from MSRA-B dataset. Four traditional SOD methods (*i.e.* $M = 4$) are employed including RBD [46], DSR [14], MC [8], and HS [47] to produce the pseudo labels for the training images. The pseudo labels from each handcrafted method are separately refined by using the first stage of [20] to improve their quality. The threshhold to binary the pseudo labels is empirically set as 0.5, which works well in practice. All the input images are resized to the spatial size of 320 × 320. Data augmentation including random flipping and rotation is adopted for the training data. Following [20], the parameters of SODNet are initialized by using the pretrained model of [2], and the learning rate of these parameters is set as $2e-5$. The parameters of MSNet are randomly initialized by using the method of [6] with a learning rate of $2e-4$. The whole networks are jointly trained end-to-end using the ADAM optimizer with the bath size of 16. The whole training process takes about 200 epoches on the platform with one Geforce 3090 GPUs. At inference, only the learned SODNet is employed to produce the saliency masks.

4. Experiments

We first compare our model against several related methods, followed by extensive ablation studies to explore the contributions of different components.

4.1. Dataset and Evaluation Metrics

We evaluate the proposed model on five public SOD benchmark datasets, including DUTS test dataset [27], DUT-OMRON [35], ECSSD [34] and HKU-IS [12]. For quantitative evaluation, four popular evaluation metrics are utilized, *i.e.*, Mean Absolute Error (MAE), max F-measure (F_{max}) [1], S-measure (S) [3], and E-measure (E) [4].

4.2. Overall Comparison

We compare the proposed method against four groups of state-of-the-art SOD works: (1) seven fully-supervised methods that training deep networks using clean pixel-level annotations, including DCL [13], Amulet [44], SRM [29], NLDF [19], PiCANet [18], AFNet [5], and MSNet [32]; (2) four weakly-supervised methods that using some weak labels, including WSS [27], MWS [36], SODSA [42], and C2S [15]. (3) seven handcrafted unsupervised SOD methods including HS [47], RBD [46], SF [22], GS [31], MFR [35], MC [8], DSR [14]; (4) four deep unsupervised SOD methods that are most related to ours, including SBF [38], USD [43], E-BigBiGAN [25], and DeepUSPS [20]. All the results of the compared methods are provided by the authors or obtained from public data.

Table 1 reports the evaluation results of all the compared methods. It shows that the proposed method achieves superior performance among the unsupervised approaches. Among them, the recent leading method DeepUSPS [20] can be seen as a baseline of our method, which uses the same training settings including training data, SOD network, the parameter initialization, *etc*. The difference between DeepUSPS and ours is that DeepUSPS directly trains the final SOD network using all the refined pseudo labels without any label selection. While DeepUSPS employs DenseCRF as post-processing at inference, it is still inferior to ours, which verifies the effectiveness of our the uncertainty mining mechanism for label selection. Different from USD [43] that learns the noise of multiple pseudo labels by minimizing the KL divergence between the noise model prediction and the assumed Gaussian distribution, ours implicitly trains the noisy label model (*i.e.* Uncertainty Mining Network) jointly with SODNet using the designed

Method	DUTS				DUT-OMRON				ECSSD				HKU-IS			
	MAE↓	F_{max} ↑	S↑	E↑	MAE↓	F_{max} ↑	S↑	E↑	MAE↓	F_{max} ↑	S↑	E↑	MAE↓	F_{max} ↑	S↑	E↑
Fully supervised SOD																
DCL† [13]	8.78	78.20	79.55	81.71	7.97	75.67	77.04	82.84	6.79	90.07	86.84	90.97	4.8	90.7	87.7	-
Amulet [44]	8.46	77.78	80.39	80.30	9.76	74.29	78.04	78.41	5.89	91.53	89.39	91.21	5.07	89.74	88.56	91.38
NLDF [19]	6.52	81.24	81.55	85.45	7.96	75.32	76.98	81.66	6.26	90.50	87.47	91.19	4.77	90.20	87.83	92.95
SRM [29]	5.87	82.63	83.48	86.73	6.94	76.90	79.70	84.28	5.44	91.73	89.49	92.73	4.59	90.58	88.66	93.79
MSNet [32]	4.55	84.27	84.97	89.83	5.57	79.00	81.75	86.44	3.80	93.24	90.97	**94.79**	3.37	91.95	90.25	94.95
AFNet [5]	4.58	86.29	**86.60**	89.49	5.74	79.72	**82.53**	85.95	4.18	93.50	91.33	94.14	3.58	92.26	**90.49**	94.75
PiCANet† [18]	**4.04**	**86.66**	86.15	**91.37**	**5.43**	**80.38**	82.46	**87.35**	**3.45**	**94.04**	**91.57**	94.26	**3.08**	**92.68**	90.41	**95.14**
Weakly-supervised SOD																
WSS [27]	9.94	73.92	74.90	79.66	10.97	69.08	73.11	76.79	10.35	85.56	81.12	86.89	7.92	85.87	82.17	89.56
MWS [36]	9.12	76.74	75.84	81.57	10.87	71.76	75.58	76.42	9.64	87.77	82.75	88.47	8.43	85.60	81.81	89.57
SODSA [42]	**6.22**	78.87	80.29	**86.89**	**6.84**	**75.32**	**78.43**	**84.48**	**5.90**	88.80	86.54	**91.72**	4.70	88.05	86.45	**93.22**
C2S [15]	6.64	**79.02**	**81.68**	84.23	7.90	73.34	77.93	81.76	5.93	**89.57**	**88.17**	90.73	**4.60**	**89.87**	**88.85**	92.90
Handcrafted Unsupervised SOD																
HS [47]	24.32	56.99	60.05	69.50	22.74	61.61	63.26	71.29	22.75	72.66	68.52	72.70	21.50	70.76	67.42	76.27
SF [22]	15.23	45.78	50.48	69.47	14.68	49.53	54.13	70.17	21.88	54.79	47.91	67.66	17.44	58.38	51.17	72.36
GS [31]	18.12	53.27	62.15	67.15	17.32	55.59	63.84	67.83	20.58	66.08	66.03	75.11	16.81	67.71	69.08	78.43
MFR [35]	19.36	59.28	62.54	71.21	18.74	61.02	64.51	72.55	18.93	73.57	68.92	77.49	17.82	70.63	66.87	78.47
MC [8]	19.88	61.01	62.46	**72.20**	18.63	62.73	64.91	**72.77**	20.24	**73.93**	**69.24**	**78.74**	18.40	72.34	68.38	80.40
RBD [46]	15.31	59.14	64.64	71.09	14.38	**63.04**	**68.14**	72.04	17.14	71.62	68.83	78.68	14.28	72.30	**70.61**	**81.18**
DSR [14]	**14.78**	**61.59**	**65.22**	71.59	**13.87**	62.70	67.27	72.16	**17.13**	73.48	68.51	78.65	**14.21**	**73.51**	69.95	80.79
Deep Unsupervised SOD																
E-BigBiGAN [25]	19.53	66.86	68.61	71.54	23.20	60.73	64.27	70.29	16.26	82.59	78.96	81.07	15.50	80.41	77.60	83.28
SBF [38]	10.69	69.83	74.27	78.17	10.76	68.49	74.72	76.95	8.80	85.32	83.23	87.61	7.53	83.93	82.91	89.33
USD [43]	7.49	-	**81.28**	85.25	10.28	-	73.32	71.24	9.02	-	84.56	83.57	6.50	-	86.02	85.79
DeepUSPS† [20]	6.78	78.42	78.67	84.87	**6.25**	77.31	79.46	84.82	**6.32**	90.07	86.11	90.39	**4.12**	90.20	87.60	93.06
Ours	**6.67**	**79.87**	80.21	**86.28**	6.31	**78.71**	**80.40**	**85.96**	6.36	**90.29**	**86.77**	**90.42**	**4.12**	**90.76**	**88.62**	**93.94**

Table 1. Evaluation results on the popular SOD benchmark datasets measured in % of MAE, F_{max}, S, and E metrics. ↑ and ↓ indicate that the larger and smaller scores are better, respectively. † denotes that dense CRF is adopted for post-processing at inference. **Bold** numbers indicate the best performance in each group.

uncertainty weighted loss, and delivers promising performance. In addition, our method without any clean training annotations achieves on par with the supervised methods [13, 19, 29, 44] and is better than the weakly-supervised ones [27, 36].

Figure 4 show some visual comparisons. Our method yields high quality saliency results in various challenging scenarios and outperforms the compared methods.

4.3. Ablation

To further verify our main contributions, we evaluate different variants of our method on DUTS test dataset and DUT-OMRON dataset. For fair comparison, all the compared models are trained and tested using the same training protocols as ours. Results are summarized in Figure 5 measured in % of F_{max} and S metrics.

4.3.1 Oracle and Baselines

We first conduct the "Oracle" experiments, where the SODNet is trained using the ground truth annotations with the cross-entropy loss for supervision. Besides, two other baselines are designed to verify our main idea of uncertainty learning as well as the proposed uncertainty weighted loss. The first one is directly mixing all the multi-source pseudo labels and training SODNet with the cross-entropy loss, meaning that all the pseudo labels are treated equally without considering their uncertainties. We denote this variant as "Baseline1", which is equivalent to the method of DeepUSPS [20]. Nevertheless, for fair comparison, we reimplement this variant and remove the DenseCRF at inference. In addition, instead of learning the uncertainty via deep network and the proposed loss, we are also interested to see the performance of using the handcrafted manner to select the pseudo labels for network training. Therefore, "Baseline2" is the one that selecting the reliable labels by using

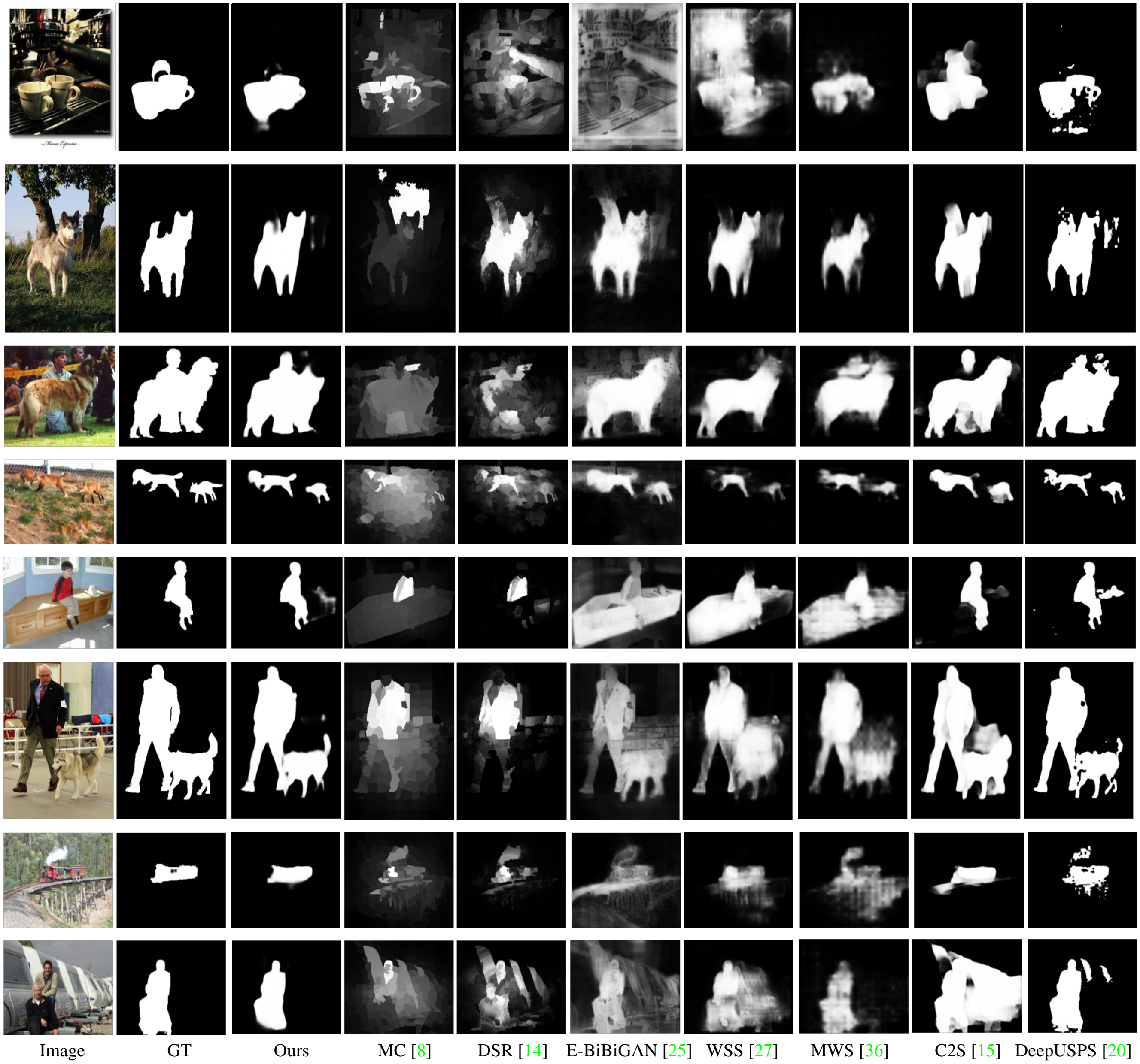


Figure 4. Some visual examples of the saliency detection results obtained by our approach and other state-of-the-art methods.

the majority voting strategy. It means that given the multi-source pseudo labels, each label value is deemed to be real if its value is equivalent to the labels from at least two other sources. Compared with the "Oracle" model, the performance drop of our method is acceptable under the unsupervised learning setting (see Figure 5). From the comparison with "Baseline1", it can be observed that under the supervision of the proposed loss, ours can effectively provide the reliable label samples for SODNet and thus facilitate the final SOD performance. In addition, the superiority of the our learned uncertainty method is non-negligible compared to the handcrafted selection strategy (Ours *vs*. Baseline2).

4.3.2 Merge-and-Split Module

Merge-and-Split (MS) module is the core component of the proposed UMNet. To demonstrate its effectiveness, we design two following variants of UMNet: (1) "UMNet_v1", where the feature concatenation among multi-source labels are removed so that each branch in UMNet is independently trained without any interaction. (2) "UMNet_v2", which contains single branch with the comparable parameters of

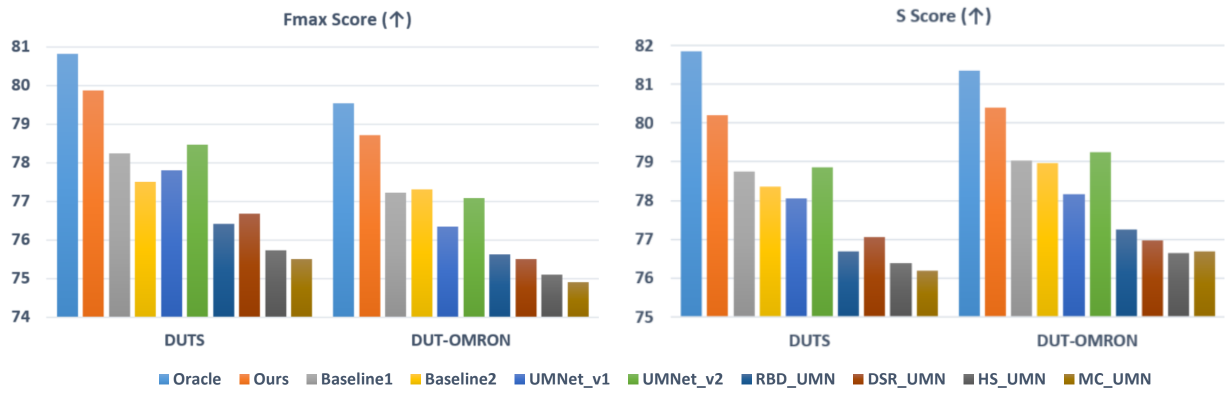


Figure 5. Results of ablation studies investigating the effectiveness of our main ingredients.

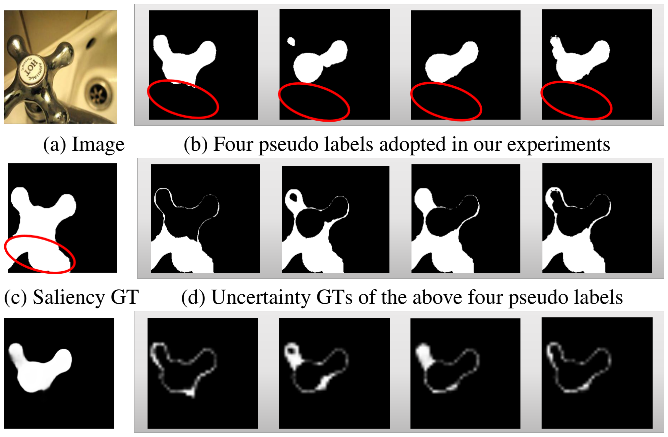
the proposed UMNet. The multi-source pseudo labels are first concatenated along channels and then fed into UMNet_v2. The superiority of our method compared to the two variants demonstrate the effectiveness of our merge and split strategy for uncertainty prediction.

4.3.3 Impact of Multi-source Labels

To further investigate the impact of using multi-source pseudo labels under the proposed framework, we conduct experiments using only single-source labels. Therefore, the UMNet only contains single branch with comparable model capacities with our multiple branch counterpart, and the MS module is replaced with residual blocks. We denote these variants as “RBD_UMN”,“DSR_UMN”,”MC_UMN”, and “HS_UMN” to distinguish them from the original ones. All the variants are trained using the proposed uncertainty weighted loss. We observe a significant performance drop when using single-source labels, which may attribute to the inaccurate uncertainty prediction. This also verifies our intuitive idea that evaluating the uncertainty of one label sample by simultaneously analyzing multi-source labels is more robust under the unsupervised setting.

4.4. Limitations

As shown in Figure 6, when all the pseudo labels reach an agreement but with wrong label values, our UMNet fails to capture such noise and makes wrong estimation, hindering the SODNet learning. Such issue is challenging to circumvent without any other reference under the unsupervised setting. Exploring more sophisticated prior knowledge or learning mechanisms may be a promising solution, which we would like to leave as our future work.



(a) Image (b) Four pseudo labels adopted in our experiments

(c) Saliency GT (d) Uncertainty GTs of the above four pseudo labels

(e) Saliency prediction (f) Our predicted uncertainty maps

Figure 6. Limitations of the proposed method. For the foreground region that denoted by the red ellipses, all the pseudo labels (b) believe it belongs to the background, which misleads UPNet to trust the noisy label samples and hinders the final saliency prediction.

5. Conclusion

This paper presents a novel multi-source uncertainty mining approach for deep unsupervised SOD, which aims to learn to select the reliable label samples from noisy pseudo labels generated by handcrafted SOD methods. To obtain the robust dense uncertainty maps of the noisy pseudo labels, a Merge-and-Split (MS) module is designed to simultaneously analyse the commonalities and diversities among multi-source pseudo labels, which enables the uncertainty mining network (UMNet) to effectively capture the reliable label samples to supervise the SODNet. In addition, an uncertainty weighted loss is further developed by modeling the noise labels using Gibbs distribution and effectively facilitates network learning without any human annotations. Extensive experiments demonstrate the effectiveness of the proposed method as well as the main contributions.

Acknowledgement. This paper is supported in part by the National Key R&D Program of China (2018AAA0102001), National Natural Science Foundation of China (62006036, 61906031, 61725202, U1903215, 61829102), and Dalian Innovation leader’s support Plan (2018RD07).

References

[1] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 5

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2018. 3, 4, 5

[3] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017. 5

[4] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In Jérôme Lang, editor, *IJCAI*, pages 698–704, 2018. 5

[5] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, 2019. 5, 6

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, pages 1026–1034, 2015. 5

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4

[8] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *ICCV*, 2013. 2, 5, 6, 7

[9] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *NeurIPS*, pages 5574–5584, 2017. 4

[10] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *CVPR*, pages 7482–7491, 2018. 4

[11] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016. 2

[12] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015. 5

[13] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, 2016. 5, 6

[14] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, 2013. 5, 6, 7

[15] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, 2018. 2, 5, 6, 7

[16] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 2

[17] Jiangjiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019. 2

[18] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018. 5, 6

[19] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A. Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. 5, 6

[20] Duc Tam Nguyen, Maximilian Dax, Chaithanya Kumar Mummadi, Thi-Phuong-Nhung Ngo, Thi Hoai Phuong Nguyen, Zhongyu Lou, and Thomas Brox. Deepusps: Deep robust unsupervised saliency prediction with self-supervision. *NeurIPS*, 2019. 1, 3, 5, 6, 7

[21] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 2

[22] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 2, 5, 6

[23] Xuebin Qin, Zichen Vincent Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jägersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 2

[24] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *ICCV*, 2019. 2

[25] Andrey Voynov, Stanislav Morozov, and Artem Babenko. Object segmentation without labels with large-scale generative models. In *Int. Conf. Mach. Learn.*, volume 139, pages 10596–10606, 2021. 5, 6, 7

[26] Lijun Wang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Deep networks for saliency detection via local estimation and global search. In *CVPR*, 2015. 2

[27] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017. 2, 5, 6, 7

[28] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Salient object detection with recurrent fully convolutional networks. *IEEE TPAMI*, 41(7):1734–1746, 2019. 2

[29] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, 2017. 5, 6

[30] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, 2018. 2

[31] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, 2012. 2, 5, 6

[32] Runmin Wu, Mengyang Feng, Wenlong Guan, Dong Wang, Huchuan Lu, and Errui Ding. A mutual learning method for salient object detection with intertwined multi-supervision. In *CVPR*, 2019. 5, 6

[33] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019. 2

[34] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013. 5

[35] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2, 5, 6

[36] Yu Zeng, Yun-Zhi Zhuge, Huchuan Lu, Lihe Zhang, Mingyang Qian, and Yizhou Yu. Multi-source weak supervision for saliency detection. In *CVPR*, 2019. 2, 5, 6, 7

[37] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017. 1

[38] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *ICCV*, pages 4068–4076, 2017. 1, 2, 3, 5, 6

[39] Dingwen Zhang, Haibin Tian, and Jungong Han. Few-cost salient object detection with adversarial-paced learning. In *NeurIPS*, 2020. 2

[40] Jing Zhang, Yuchao Dai, Tong Zhang, Mehrtash Harandi, Nick Barnes, and Richard Hartley. Learning saliency from single noisy labelling: A robust model fitting perspective. *IEEE TPAMI*, 43(8):2866–2873, 2021. 1, 3

[41] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020. 2

[42] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020. 5, 6

[43] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard I. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *CVPR*, 2018. 1, 2, 3, 5, 6

[44] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017. 5, 6

[45] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. In *ECCV*, pages 35–51, 2020. 2

[46] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *CVPR*, 2014. 5, 6

[47] Wenbin Zou and Nikos Komodakis. HARF: hierarchy-associated rich features for salient object detection. In *ICCV*, pages 406–414, 2015. 5, 6