Weakly-Supervised Salient Object Detection via Scribble Annotations

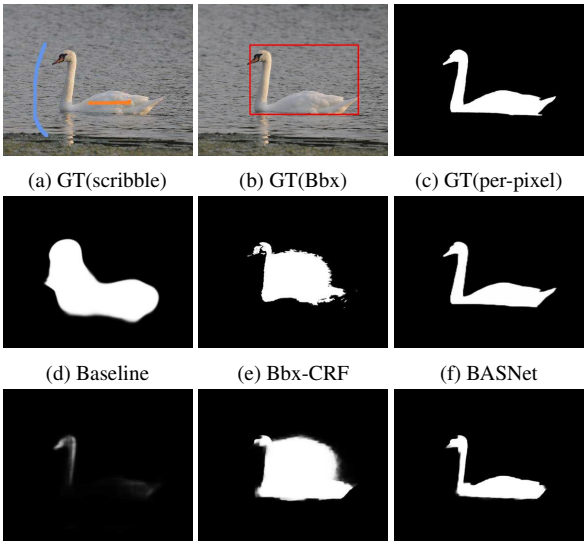
Jing Zhang[1,3,4] Xin Yu[1,3,5] Aixuan Li[2] Peipei Song[1,4] Bowen Liu[2] Yuchao Dai[2*]
[1] Australian National University, Australia [2] Northwestern Polytechnical University, China
[3] ACRV, Australia [4] Data61, Australia [5] ReLER, University of Technology Sydney, Australia

Abstract

Compared with laborious pixel-wise dense labeling, it is much easier to label data by scribbles, which only costs 1∼2 seconds to label one image. However, using scribble labels to learn salient object detection has not been explored. In this paper, we propose a weakly-supervised salient object detection model to learn saliency from such annotations. In doing so, we first relabel an existing large-scale salient object detection dataset with scribbles, namely S-DUTS dataset. Since object structure and detail information is not identified by scribbles, directly training with scribble labels will lead to saliency maps of poor boundary localization. To mitigate this problem, we propose an auxiliary edge detection task to localize object edges explicitly, and a gated structure-aware loss to place constraints on the scope of structure to be recovered. Moreover, we design a scribble boosting scheme to iteratively consolidate our scribble annotations, which are then employed as supervision to learn high-quality saliency maps. As existing saliency evaluation metrics neglect to measure structure alignment of the predictions, the saliency map ranking metric may not comply with human perception. We present a new metric, termed saliency structure measure, as a complementary metric to evaluate sharpness of the prediction. Extensive experiments on six benchmark datasets demonstrate that our method not only outperforms existing weakly-supervised/unsupervised methods, but also is on par with several fully-supervised state-of-the-art models[1].

1. Introduction

Visual salient object detection (SOD) aims at locating interesting regions that attract human attention most in an image. Conventional salient object detection methods [57, 14] based on hand-crafted features or human experience may fail to obtain high-quality saliency maps in complicated scenarios. The deep learning based salient object detection models [42, 50] have been widely studied, and significantly boost the saliency detection performance. However, these methods highly rely on a large amount of labeled data, which require time-consuming and laborious pixel-wise annotations. To achieve a trade-off between labeling efficiency and model performance, several weakly supervised or unsupervised methods [16, 47, 24, 52] have been proposed to learn saliency from sparse labeled data [16, 47] or infer the latent saliency from noisy annotations [24, 52].



(a) GT(scribble) (b) GT(Bbx) (c) GT(per-pixel)
(d) Baseline (e) Bbx-CRF (f) BASNet
(g) WSS (h) Bbx-Pred (i) Ours

Figure 1. (a) Our scribble annotations. (b) Ground-truth bounding box. (c) Ground-truth pixel-wise annotations. (d) Baseline model: trained directly on scribbles. (e) Refined bounding box annotation by DenseCRF [1]. (f) Result of a fully-supervised SOD method [26]. (g) Result of model trained on image-level annotations [34] (h) Model trained on the annotation (e). (i) Our result.

In this paper, we propose a new weakly-supervised salient object detection framework by learning from low-cost labeled data, (*i.e.*, scribbles, as seen in Fig. 1(a)). Here, we opt to scribble annotations because of their flexibility (although bounding box annotation is an option, it's not suitable for labeling winding objects, thus leading to inferior saliency maps, as seen in Fig. 1 (h)). Since scribble annotations are usually very sparse, object structure and details cannot be easily inferred. Directly training a deep

*Corresponding author: Yuchao Dai *(daiyuchao@gmail.com)*
[1]Our code and data is publicly available at: https://github.com/JingZhang617/Scribble_Saliency.

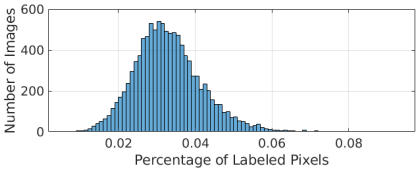


Figure 2. Percentage of labeled pixels in the S-DUTS dataset.

model with sparse scribbles by partial cross-entropy loss [30] may lead to saliency maps of poor boundary localization, as illustrated in Fig. 1 (d).

To achieve high-quality saliency maps, we present an auxiliary edge detection network and a gated structure-aware loss to enforce boundaries of our predicted saliency map to align with image edges in the salient region. The edge detection network forces the network to produce feature highlight object structure, and the gated structure-aware loss allows our network to focus on the salient region while ignoring the structure of the background. We further develop a scribble boosting manner to update our scribble annotations by propagating the labels to larger receptive fields of high confidence. In this way, we can obtain denser annotations as shown in Fig. 7 (g).

Due to the lack of scribble based saliency datasets, we relabel an existing saliency training dataset DUTS [34] with scribbles, namely S-DUTS dataset, to verify our method. DUTS is a widely used salient object detection dataset, which contains 10,553 training images. Annotators are asked to scribble the DUTS dataset according to their first impressions without showing them the ground-truth salient objects. Fig. 2 indicates the percentage of labeled pixels across the whole S-DUTS dataset. On average, around 3% of the pixels are labeled (either foreground or background) and the others are left as unknown pixels, demonstrating that the scribble annotations are very sparse. Note that, we only use scribble annotation as supervision signal during training, and we take RGB image as input to produce dense saliency map during testing.

Moreover, the rankings of saliency maps based on traditional mean absolute error (MAE) may not comply with human visual perception. For instance, in the 1^{st} row of Fig. 3, the last saliency map is visually better than the fourth one and the third one is better than the second one. We propose saliency structure measure (B_μ) as a complementary metric of existing evaluation metrics that takes the structure alignment of the saliency map into account. The measurements based on B_μ are more consistent with human perception, as shown in the 2^{nd} row of Fig. 3.

We summarize our main contributions as: (1) we present a new weakly-supervised salient object detection method by learning saliency from scribbles, and introduce a new scribble based saliency dataset S-DUTS; (2) we propose a gated structure-aware loss to constrain a predicted saliency map to share similar structure with the input image in the salient region; (3) we design a scribble boosting scheme to expand our scribble annotations, thus facilitating high-quality saliency map acquisition; (4) we present a new evaluation metric to measure the structure alignment of predicted saliency maps, which is more consistent with human visual perception; (5) experimental results on six salient object detection benchmarks demonstrate that our method outperforms state-of-the-art weakly-supervised algorithms.

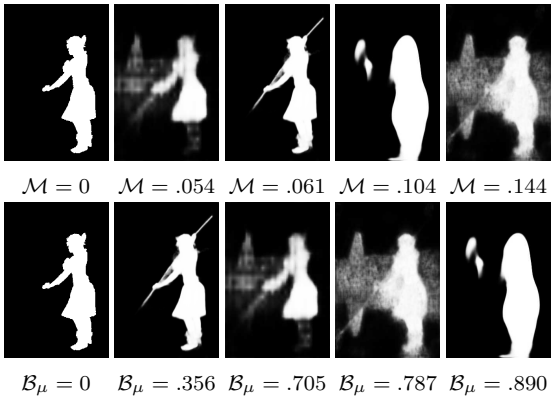


Figure 3. Saliency map ranking based on Mean Absolute Error (1^{st} row) and our proposed Saliency Structure Measure (2^{nd} row).

2. Related Work

Deep fully supervised saliency detection models [26, 55, 42, 50, 51, 36, 49] have been widely studied. As our method is weakly supervised, we mainly discuss related weakly-supervised dense prediction models and approaches to recover detail information from weak annotations.

2.1. Learning Saliency from Weak Annotations

To avoid requiring accurate pixel-wise labels, some SOD methods attempt to learn saliency from low-cost annotations, such as bounding boxes [29], image-level labels [34, 16], and noisy labels [52, 48, 24], *etc*. This motivates SOD to be formulated as a weakly-supervised or unsupervised task. Wang *et al*. [34] introduced a foreground inference network to produce saliency maps with image-level labels. With the same weak labels, Hsu *et al*. [10] presented a category-driven map generator to learn saliency from class activation map. Li *et al*. [16] adopted an iterative learning strategy to update an initial saliency map generated from unsupervised saliency methods by learning with image-level supervision. A fully connected CRF [1] was utilized in [34, 16] as post-processing to refine the produced saliency map. Zeng *et al*. [47] proposed to train saliency models with diverse weak supervision sources, including category labels, captions, and unlabeled data. Zhang *et al*. [48] fused saliency maps from unsupervised methods with heuristics within a deep learning framework. In a similar setting, Zhang *et al*. [52] proposed to collaboratively

update a saliency prediction module and a noise module to learn a saliency map from multiple noisy labels.

2.2. Weakly-Supervised Semantic Segmentation

Dai *et al*. [3] and Khoreva [13] proposed to learn semantic segmentation from bounding boxes in a weakly-supervised way. Hung *et al*. [12] randomly interleaved labeled and unlabeled data, and trained a network with an adversarial loss on the unlabeled data for semi-supervised semantic segmentation. Shi *et al*. [39] tackled the weakly-supervised semantic segmentation problem by using multiple dilated convolutional blocks of different dilation rates to encode dense object localization. Li *et al*. [37] presented an iterative bottom-up and top-down semantic segmentation framework to alternatingly expand object regions and optimize segmentation network with image tag supervision. Huang *et al*. [11] introduced a seeded region growing technique to learn semantic segmentation with image-level labels. Vernaza *et al*. [32] designed a random walk based label propagation method to learn semantic segmentation from sparse annotations.

2.3. Recovering Structure from Weak Labels

As weak annotations do not contain complete semantic region of the specific object, the predicted object structure is often incomplete. To preserve rich and fine-detailed semantic information, additional regularizations are often employed. Two main solutions are widely studied, including graph model based methods (*e.g*. CRF [1]) and boundary based losses [15]. Tang *et al*. [30] introduced a normalized cut loss as a regularizer with partial cross-entropy loss for weakly-supervised image segmentation. Tang *et al*. [31] modeled standard regularizers into a loss function over partial observation for semantic segmentation. Obukhov *et al*. [25] proposed a gated CRF loss for weakly-supervised semantic segmentation. Lampert *et al*. [15] introduced a constrain-to-boundary principle to recover detail information for weakly-supervised image segmentation.

2.4. Comparison with Existing Scribble Models

Although scribble annotations have been used in weakly-supervised semantic segmentation [19, 33], our proposed scribble based salient object detection method is different from them in the following aspects: (1) semantic segmentation methods target at class-specific objects. In this manner, class-specific similarity can be explored. On the contrary, salient object detection does not focus on class-specific objects, thus object category related information is not available. For instance, a leaf can be a salient object while the class category is not available in the widely used image-level label dataset [4, 20]. Therefore, we propose edge-guided gated structure-aware loss to obtain structure information from image instead of depending on image category. (2) although boundary information has been used in [33] to propagate labels, Wang *et al*. [33] regressed boundaries by an ℓ_2 loss. Thus, the structure of the segmentation may not be well aligned with the image edges. In contrast, our method minimizes the differences between first order derivatives of saliency maps and images, and leads to saliency map better aligned with image structure. (3) benefiting from our developed boosting method and the intrinsic property of salient objects, our method requires only scribble on any salient region as shown in Fig. 9, while scribbles are required to traverse all those semantic categories for scribble based semantic segmentation [19, 33].

3. Learning Saliency from Scribbles

Let's define our training dataset as: $D = \{x_i, y_i\}_{i=1}^{N}$, where x_i is an input image, y_i is its corresponding annotation, N is the size of the training dataset. For fully-supervised salient object detection, y_i is a pixel-wise label with 1 representing salient foreground and 0 denoting background. We define a new weakly-supervised saliency learning problem from scribble annotations, where y_i in our case is scribble annotation used during training, which includes three categories of supervision signal: 1 as foreground, 2 as background and 0 as unknown pixels. In Fig. 2, we show the percentage of annotated pixels of the training dataset, which indicates that around 3% of pixels are labeled as foreground or background in our scribble annotation.

There are three main components in our network, as illustrated in Fig. 4: (1) a saliency prediction network (SPN) to generate a coarse saliency map s^c, which is trained on scribble annotations by a partial cross-entropy loss [30]; (2) an edge detection network (EDN) is proposed to enhance structure of s^c, with a gated structure-aware loss employed to force the boundaries of saliency maps to comply with image edges; (3) an edge-enhanced saliency prediction module (ESPM) is designed to further refine the saliency maps generated from SPN.

3.1. Weakly-Supervised Salient Object Detection

Saliency prediction network (SPN): We build our front-end saliency prediction network based on VGG16-Net [28] by removing layers after the fifth pooling layer. Similar to [43], we group the convolutional layers that generate feature maps of the same resolution as a stage of the network (as shown in Fig. 4). Thus, we denote the front-end model as $f_1(x, \theta) = \{s_1, ..., s_5\}$, where $s_m (m = 1, ..., 5)$ represents features from the last convolutional layer in the m-th stage ("relu1_2, relu2_2, relu3_3, relu4_3, relu5_3" in this paper), θ is the front-end network parameters.

As discussed in [39], enlarging receptive fields by different dilation rates can propagate the discriminative information to non-discriminative object regions. We employ a dense atrous spatial pyramid pooling (DenseASPP) module

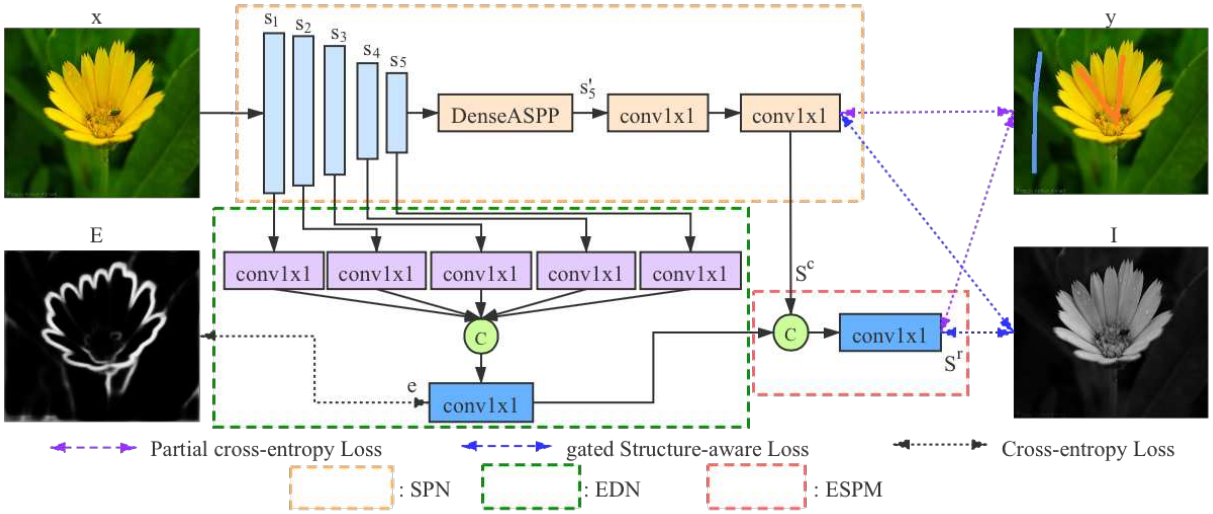


Figure 4. Illustration of our network. For simplicity, we do not show the scribble boosting mechanism here. "I" is the intensity image of input "x". "C": concatenation operation; "conv1x1": 1×1 convolutional layer.

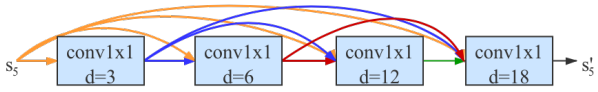


Figure 5. Our "DenseASPP" module. "conv1x1 d=3" represents a 1×1 convolutional layer with a dilation rate 3.

[46] on top of the front-end model to generate feature maps s'_5 with larger receptive fields from feature s_5. In particular, we use varying dilation rates in the convolutional layers of DenseASPP. Then, two extra 1×1 convolutional layers are used to map s'_5 to a one channel coarse saliency map s^c.

As we have unknown category pixels in the scribble annotations, partial cross-entropy loss [30] is adopted to train our SPN:

$$\mathcal{L}_s = \sum_{(u,v)\in J_l} \mathcal{L}_{u,v}, \tag{1}$$

where J_l represents the labeled pixel set, (u, v) is the pixel coordinates, and $\mathcal{L}_{u,v}$ is the cross-entropy loss at (u, v).

Edge detection network (EDN): Edge detection network encourages SPN to produce saliency features with rich structure information. We use features from the intermediate layers of SPN to produce one channel edge map e. Specifically, we map each $s_i (i = 1, ..., 5)$ to a feature map of channel size M with a 1×1 convolutional layer. Then we concatenate these five feature maps and feed them to a 1×1 convolutional layer to produce an edge map e. A cross-entropy loss $\mathcal{L}_e$ is used to train EDN:

$$\mathcal{L}_e = \sum_{u,v} (E \log e + (1 - E) \log(1 - e)), \tag{2}$$

where E is pre-computed by an existing edge detector [22].

Edge-enhanced saliency prediction module (ESPM): We introduce an edge-enhanced saliency prediction module to refine the coarse saliency map s^c from SPN and obtain an edge-preserving refined saliency map s^r. Specifically, we concatenate s^c and e and then feed them to a 1×1 convolutional layer to produce a saliency map s^r. Note that, we use the saliency map s^r as the final output of our network. Similar to training SPN, we employ a partial cross-entropy loss with scribble annotations to supervise s^r.

Gated structure-aware loss: Although ESPM encourages the network to produce saliency map with rich structure, there exists no constraints on scope of structure to be recovered. Following the "Constrain-to-boundary" principle [15], we propose a gated structure-aware loss, which encourages the structure of a predicted saliency map to be similar to the salient region of an image.

We expect the predicted saliency map having consistent intensities inside the salient region and distinct boundaries at the object edges. Inspired by the smoothness loss [9, 38], we also impose such constraint inside the salient regions. Recall that the smoothness loss is developed to enforce smoothness while preserving image structure across the whole image region. However, salient object detection intends to suppress the structure information outside the salient regions. Therefore, enforcing the smoothness loss across the entire image regions will make the saliency prediction ambiguous, as shown in Tabel 2 "M3".

To mitigate this ambiguity, we employ a gate mechanism to let our network focus on salient regions only to reduce distraction caused by background structure. Specifically, we define the gated structure-aware loss as:

$$\mathcal{L}_b = \sum_{u,v} \sum_{d\in\overrightarrow{x},\overrightarrow{y}} \Psi(|\partial_d s_{u,v}| e^{-\alpha|\partial_d(G\cdot I_{u,v})|}), \tag{3}$$

where Ψ is defined as $\Psi(s) = \sqrt{s^2 + 1e^{-6}}$ to avoid cal-

(a) (b) (c) (d) (e)

Figure 6. Gated structure-aware constraint: (a) Initial predicted saliency map. (b) Image edge map. (c) Dilated version of (a). (d) Gated mask in Eq. 3. (e) Gated edge map.

culating the square root of zero, $I_{u,v}$ is the image intensity value at pixel (u, v), d indicates the partial derivatives on the $\overrightarrow{x}$ and $\overrightarrow{y}$ directions, and G is the gate for the structure-aware loss (see Fig .6 (d)). The gated structure-aware loss applies L1 penalty on gradients of saliency map s to encourages it to be locally smooth, with an edge-aware term ∂I as weight to maintain saliency distinction along image edges.

Specifically, as shown in Fig. 6, with predicted saliency map (a)) during training, we dilate it with a square kernel of size $k = 11$ to obtain an enlarged foreground region (c)). Then we define gate (d)) as binarized (c)) by adaptive thresholding. As seen in Fig. 6(e), our method is able to focus on the saliency region and predict sharp boundaries in a saliency map.

Objective Function: As shown in Fig. 4, we employ both partial cross-entropy loss $\mathcal{L}_s$ and gated structure-aware loss $\mathcal{L}_b$ to coarse saliency map s^c and refined map s^r, and use cross-entropy loss $\mathcal{L}_e$ for the edge detection network. Our final loss function is then defined as:

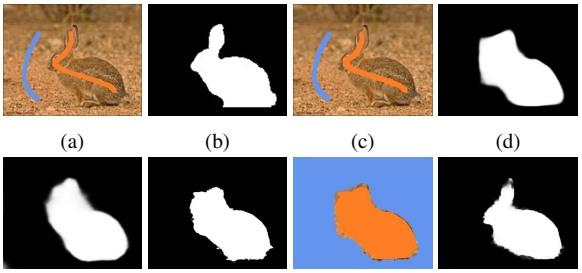
$$\begin{aligned}\mathcal{L} =& \mathcal{L}_s(s^c, y) + \mathcal{L}_s(s^r, y) \\ &+ \beta_1 \cdot \mathcal{L}_b(s^c, x) + \beta_2 \cdot \mathcal{L}_b(s^r, x) + \beta_3 \cdot \mathcal{L}_e,\end{aligned} \tag{4}$$

where y indicates scribble annotations. The partial cross-entropy loss $\mathcal{L}_s$ takes scribble annotation as supervision, while gated structure-aware loss $\mathcal{L}_b$ leverages image boundary information. These two losses do not contradict each other since $\mathcal{L}_s$ focuses on propagating the annotated scribble pixels to the foreground regions (relying on SPN), while $\mathcal{L}_b$ enforces s^r to be well aligned to edges extracted by EDN and prevents the foreground saliency pixels from being propagated to backgrounds.

3.2. Scribble Boosting

While we generate scribbles for a specific image, we simply annotate a very small portion of the foreground and background as shown in Fig. 1. Intra-class discontinuity, such as complex shapes and appearances of objects, may lead our model to be trapped in a local minima, with incomplete salient object segmented. Here, we attempt to propagate the scribble annotations to a denser annotation based on our initial estimation.

A straightforward solution to obtain denser annotations is to expand scribble labels by using DenseCRF [1], as shown in Fig. 7(c). However, as our scribble annotations are very sparse, DenseCRF fails to generate denser annotation from our scribbles (see Fig. 7(c)). As seen in Fig. 7(e), the predicted saliency map trained on (c) is still very similar to the one supervised by original scribbles (see Fig. 7(d)).



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 7. Illustration of using different strategies to enrich scribble annotations. (a) Input RGB image and scribble annotations. (b) Per-pixel wise ground-truth. (c) Result of applying DenseCRF to scribbles. (d) Saliency detection, trained on scribbles of (a). (e) Saliency detection, trained on scribbles of (c). (f) Applying DenseCRF to the result (d). (g) The confidence map between (d) and (f) for scribble boosting. Orange indicates consistent foreground, blue represents consistent background, and others are marked as unknown. (h) Our final result trained on new scribble (g).

Instead of expanding the scribble annotation directly, we apply DenseCRF to our initial saliency prediction s^{init}, and update s^{init} to s^{crf}. Directly training a network with s^{crf} will introduce noise to the network as s^{crf} is not the exact ground-truth. We compute difference of s^{init} and s^{crf}, and define pixels with $s^{\text{init}} = s^{\text{crf}} = 1$ as foreground pixels in the new scribble annotation, $s^{\text{init}} = s^{\text{crf}} = 0$ as background pixels, and others as unknown pixels. In Fig. 7 (g) and Fig. 7 (h), we illustrate the intermediate results of scribble boosting. Note that, our method achieves better saliency prediction results than the case of applying DenseCRF to the initial prediction (see Fig. 7 (f)). This demonstrates the effectiveness of our scribble boosting scheme. In our experiments, after conducting one iteration of our scribble boosting step, our performance is almost on par with fully-supervised methods.

3.3. Saliency Structure Measure

Existing saliency evaluation metrics (Mean Abosolute Error, Precision-recall curves, F-measure, E-measure [7] and S-measure [6]) focus on measuring accuracy of the prediction, while neglect whether a predicted saliency map complies with human perception or not. In other words, the estimated saliency map should be aligned with object structure of the input image. In [23], bIOU loss was proposed to penalize on saliency boundary length. We adapt the bIOU loss as an error metric B_μ to evaluate the structure alignment between saliency maps and their ground-truth.

Given a predicted saliency map s, and its pixel-wise ground truth y, their binarized edge maps are defined as g_s and g_y respectively. Then B_μ is expressed as: $B_\mu = 1 - \frac{2 \cdot \sum(g_s \cdot g_y)}{\sum(g_s^2 + g_y^2)}$, where $B_\mu \in [0, 1]$. $B_\mu = 0$ represents per-

Figure 8. The first two images show the original image edges. We dilate the original edges (last two images) to avoid misalignments due to the small scales of original edges.

fect prediction. As edges of prediction and ground-truth saliency maps may not be aligned well due to the small scales of edges, they will lead to unstable measurements (see Fig. 8). We dilate both edge maps with square kernel of size 3 before we compute the B_μ measure. As shown in Fig. 3, B_μ reflects the sharpness of predictions which is consistent with human perception.

3.4. Network Details

We use VGG16-Net [28] as our backbone network. In the edge detection network, we encode s_m to feature maps of channel size 32 through 1×1 convolutional layers. In the "DenseASPP" module (Fig. 5), the first three convolutional layers produce saliency features of channel size 32, and the last convolutional layer map the feature maps to s_5' of same size as s_5. Then we use two sequential convolutional layers to map s_5' to one channel coarse saliency map s^c. The hyper-parameters in Eq. 3 and Eq. (4) are set as: $\alpha = 10$, $\beta_1 = \beta_2 = 0.3$, $\beta_3 = 1$.

We train our model for 50 epochs using Pytorch, with the SPN initialized with parameters from VGG16-Net [28] pre-trained on ImageNet [4]. The other newly added convolutional layers are randomly initialized with $\mathcal{N}(0, 0.01)$. The base learning rate is initialized as 1e-4. The whole training takes 6 hours with a training batch size 15 on a PC with a NVIDIA GeForce RTX 2080 GPU.

4. Experimental Results

4.1. Scribble Dataset

In order to train our weakly-supervised salient object detection method, we relabel an existing saliency dataset with scribble annotations by three annotators (S-DUTS dataset). In Fig. 9, we show two examples of scribble annotations by different labelers. Due to the sparsity of scribbles, the annotated scribbles do not have large overlaps. Thus, majority voting is not conducted. As aforementioned, labeling one image with scribbles is very fast, which only takes 1∼2 seconds on average.

4.2. Setup

Datasets: We train our network on our newly labeled scribble saliency dataset: S-DUTS. Then, we evaluate our method on six widely-used benchmarks: (1) DUTS testing dataset [34]; (2) ECSSD [44]; (3) DUT [45]; (4) PASCAL-S [18]; (5) HKU-IS [17] and (6) THUR [2].

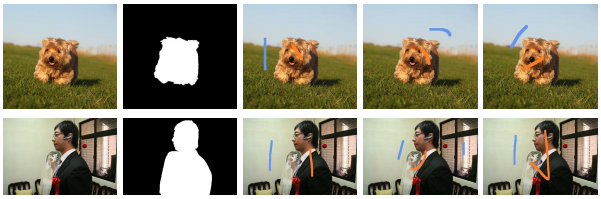


Figure 9. Illustration of scribble annotations by different labelers. From left to right: input RGB images, pixel-wise ground-truth labels, scribble annotations by three different labelers.

Competing methods: We compare our method with five state-of-the-art weakly-supervised/unsupervised methods and eleven fully-supervised saliency detection methods.
Evaluation Metrics: Four evaluation metrics are used, including Mean Absolute Error (MAE $\mathcal{M}$), Mean F-measure (F_β), mean E-measure (E_ξ [7]) and our proposed saliency structure measure (B_μ).

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

Quantitative Comparison: In Table 1 and Fig. 11, we compare our results with other competing methods. As indicated in Table 1, we achieves consistently the best performance compared with other weakly-supervised or unsupervised methods under these four saliency evaluation metrics. Since state-of-the-art weakly-supervised or unsupervised models do not impose any constraints on the boundaries of predicted saliency maps, these methods cannot preserve the structure in the prediction and produce high values on B_μ measure. In contrast, our method explicitly enforces a gated structure-aware loss to the edges of the prediction, and achieves lower B_μ. Moreover, our performance is also comparable or superior to some fully-supervised saliency models, such as DGRL and PiCANet. Fig. 11 shows the E-measure and F-measure curves of our method as well as the other competing methods on HKU-IS and THUR datasets. Due to limits of space, E-measure and F-measure curves on the other four testing datasets are provided in the supplementary material. As illustrated in Fig. 11, our method significantly outperforms the other weakly-supervised and unsupervised models with different thresholds, demonstrating the robustness of our method. Furthermore, the performance of our method is also on par with some fully-supervised methods as seen in Fig. 11.
Qualitative Comparison: We sample four images from the ECSSD dataset [44] and the saliency maps predicted by six competing methods and our method are illustrated in Fig. 10. Our method, while achieving performance on par with some fully-supervised methods, significantly outperforms other weakly-supervised and unsupervised models. In Fig. 10, we further show that directly training with scribbles produces saliency maps with poor localization ("M1"). Benefiting from our EDN as well as gated structure-aware loss, our network is able to produce sharper saliency maps

Table 1. Evaluation results on six benchmark datasets. ↑ & ↓ denote larger and smaller is better, respectively.

		Fully Sup. Models											Weakly Sup./Unsup. Models					
	Metric	DGRL [35]	UCF [53]	PiCANet [21]	R3Net [5]	NLDF [23]	MSNet [40]	CPD [41]	AFNet [8]	PFAN [56]	PAGRN [54]	BASNet [26]	SBF [48]	WSI [16]	WSS [34]	MNL [52]	MSW [47]	**Ours**
ECSSD	$B_\mu \downarrow$	.4997	.6990	.5917	.4718	.5942	.5421	.4338	.5100	.6601	.5742	**.3642**	.7587	.8007	.8079	.6806	.8510	**.5500**
	$F_\beta \uparrow$	.9027	.8446	.8715	**.9144**	.8709	.8856	.9076	.9008	.8592	.8718	.9128	.7823	.7621	.7672	.8098	.7606	**.8650**
	$E_\xi \uparrow$	.9371	.8870	.9085	**.9396**	.8952	.9218	.9321	.9294	.8636	.8869	.9378	.8354	.7921	.7963	.8357	.7876	**.9077**
	$\mathcal{M} \downarrow$	.0430	.0705	.0543	.0421	.0656	.0479	.0434	.0450	.0467	.0644	**.0399**	.0955	.0681	.1081	.0902	.0980	**.0610**
DUT	$B_\mu \downarrow$	.6188	.8115	.6846	.6061	.7148	.6415	.5491	.6027	.6443	.6447	**.4803**	.8119	.8392	.8298	.7759	.8903	**.6551**
	$F_\beta \uparrow$	.7264	.6318	.7105	.7471	.6825	.7095	.7385	.7425	.7009	.6754	**.7668**	.6120	.6408	.5895	.5966	.5970	**.7015**
	$E_\xi \uparrow$	.8446	.7597	.8231	.8527	.7983	.8306	.8450	.8456	.7990	.7717	**.8649**	.7633	.7605	.7292	.7124	.7283	**.8345**
	$\mathcal{M} \downarrow$	.0632	.1204	.0722	.0625	.0796	.0636	.0567	.0574	.0615	.0709	**.0565**	.1076	.0999	.1102	.1028	.1087	**.0684**
PASCAL-S	$B_\mu \downarrow$	.6479	.7832	.7037	.6623	.7313	.6708	.6162	.6586	.7097	.6915	**.5819**	.8146	.8550	.8309	.7762	.8703	**.6648**
	$F_\beta \uparrow$	**.8289**	.7873	.7985	.7974	.7933	.8129	.8220	.8241	.7544	.7656	.8212	.7351	.6532	.6975	.7476	.6850	**.7884**
	$E_\xi \uparrow$	**.8353**	.7953	.8045	.7806	.7828	.8219	.8197	.8269	.7464	.7545	.8214	.7459	.6474	.6904	.7408	.6932	**.7975**
	$\mathcal{M} \downarrow$	**.1150**	.1402	.1284	.1452	.1454	.1193	.1215	.1155	.1372	.1516	.1217	.1669	.2055	.1843	.1576	.1780	**.1399**
HKU-IS	$B_\mu \downarrow$	.4962	.6788	.5608	.4765	.5525	.4979	.4211	.4828	.5302	.5329	**.3593**	.7336	.7824	.7517	.6265	.8295	**.5369**
	$F_\beta \uparrow$	.8844	.8189	.8543	.8923	.8711	.8780	.8948	.8877	.8717	.8638	**.9025**	.7825	.7625	.7734	.8196	.7337	**.8576**
	$E_\xi \uparrow$	.9388	.8860	.9097	.9393	.9139	.9304	.9402	.9344	.8982	.8979	**.9432**	.8549	.7995	.8185	.8579	.7862	**.9232**
	$\mathcal{M} \downarrow$	.0374	.0620	.0464	.0357	.0477	.0387	.0333	.0358	.0424	.0475	**.0322**	.0753	.0885	.0787	.0650	.0843	**.0470**
THUR	$B_\mu \downarrow$	.5781	-	.6589	-	.6517	.6196	.5244	.5740	.7426	.6312	**.4891**	.7852	-	.7880	.7173	-	**.5964**
	$F_\beta \uparrow$	.7271	.	.7098	-	.7111	.7177	**.7498**	.7327	.6833	.7395	.7366	.6269	-	.6526	.6911	-	**.7181**
	$E_\xi \uparrow$	.8378	.	.8211	-	.8266	.8288	**.8514**	.8398	.8038	.8417	.8408	.7699	-	.7747	.8073	-	**.8367**
	$\mathcal{M} \downarrow$	.0774	.	.0836	-	.0805	.0794	.0935	.0724	.0939	**.0704**	.0734	.1071	-	.0966	.0860	-	**.0772**
DUTS	$B_\mu \downarrow$	.5644	.7956	.6348	-	.6494	.5823	.4618	.5395	.6173	.5870	**.4000**	.8082	.8785	.7802	.7117	.8293	**.6026**
	$F_\beta \uparrow$	.7898	.6631	.7565	-	.7567	.7917	**.8246**	.8123	.7648	.7781	.8226	.6223	.5687	.6330	.7249	.6479	**.7467**
	$E_\xi \uparrow$	.8873	.7750	.8529	-	.8511	.8829	**.9021**	.8928	.8301	.8422	.8955	.7629	.6900	.8061	.8525	.7419	**.8649**
	$\mathcal{M} \downarrow$	.0512	.1122	.0621	-	.0652	.0490	**.0428**	.0457	.0609	.0555	.0476	.1069	.1156	.1000	.0749	.0912	**.0622**

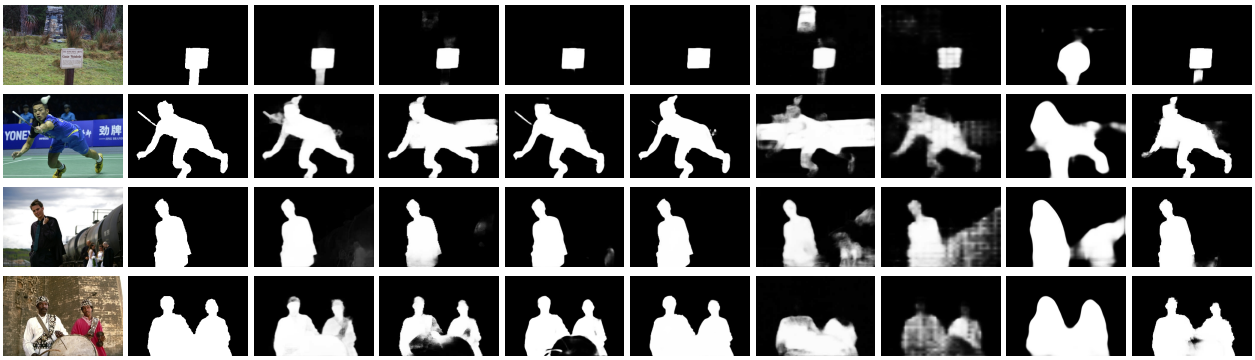


Figure 10. Comparisons of saliency maps. "M1" represents the results of a baseline model marked as "M1" in Section 4.4.

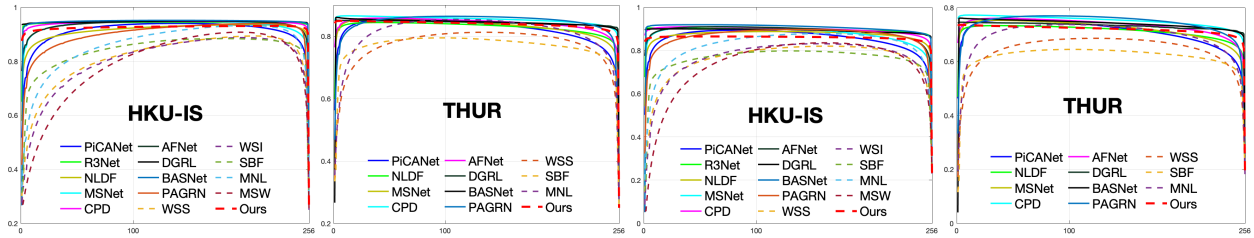


Figure 11. E-measure (1st two figures) and F-measure (last two figures) curves on two benchmark datasets. Best Viewed on screen.

than other weakly-supervised and unsupervised ones.

4.4. Ablation Study

We carry out nine experiments (as shown in Table 2) to analyze our method, including our loss functions ("M1", "M2" and "M3"), network structure ("M4"), Dense-CRF post-processing ("M5"), scribble boosting strategy ("M6"), scribble enlargement ("M7") and robustness analysis ("M8", "M9"). Our final result is denoted as "M0".

Direct training with scribble annotations: We employ the partial cross-entropy loss to train our SPN in Fig. 4 with scribble labels. The performance is marked as "M1". As expected, "M1" is much worse than our result "M0" and the high B_μ measure also indicates that object structure is not well preserved if only using the partial cross-entropy loss.

Impact of gated structure-aware loss: We add our gated structure-aware loss to "M1", and the performance is denoted by "M2". The gated structure-aware loss improves the performance in comparison with "M1". However, without using our EDN, "M2" is still inferior to "M0".

Table 2. Ablation study on six benchmark datasets.

	Metric	M0	M1	M2	M3	M4	M5	M6	M7	M8	M9
ECSSD	$B_\mu \downarrow$	.550	.896	.592	.616	.714	.582	.554	.771	.543	.592
	$F_\beta \uparrow$	.865	.699	.823	.804	.778	.845	.835	.696	.868	.839
	$E_\xi \uparrow$	.908	.814	.874	.859	.865	.898	.890	.730	.908	.907
	$\mathcal{M} \downarrow$	.061	.117	.083	.094	.091	.068	.074	.136	.059	. 070
DUT	$B_\mu \downarrow$	.655	.925	.696	.711	.777	.685	.665	.786	.656	.708
	$F_\beta \uparrow$	.702	.518	.656	.626	.580	.679	.658	.556	.691	.671
	$E_\xi \uparrow$	.835	.699	.807	.774	.743	.823	.805	.711	.823	.816
	$\mathcal{M} \downarrow$	.068	.134	.083	.102	.116	.074	.081	.108	.069	.080
PASCAL-S	$B_\mu \downarrow$	.665	.921	.732	.760	.787	.693	.676	.792	.664	.722
	$F_\beta \uparrow$	.788	.693	.748	.727	.741	.772	.768	.657	.792	.771
	$E_\xi \uparrow$	.798	.761	.757	.731	.795	.791	.782	.664	.800	.804
	$\mathcal{M} \downarrow$	.140	.171	.160	.173	.152	.145	.152	.204	.136	.143
HKU-IS	$B_\mu \downarrow$	.537	.892	.567	.609	.670	.574	.559	.747	.535	.564
	$F_\beta \uparrow$	.858	.651	.813	.789	.747	.835	.812	.646	.857	.821
	$E_\xi \uparrow$	.923	.799	.904	.878	.867	.911	.900	.761	.920	.907
	$\mathcal{M} \downarrow$	.047	.113	.060	.083	.080	.055	.062	.123	.047	.058
THUR	$B_\mu \downarrow$	.596	.927	.637	.677	.751	.635	.606	.780	.592	.650
	$F_\beta \uparrow$	.718	.520	.660	.641	.596	.696	.683	.586	.718	.690
	$E_\xi \uparrow$	.837	.687	.803	.773	.750	.824	.814	.718	.834	.804
	$\mathcal{M} \downarrow$	.077	.150	.099	.118	.123	.085	.087	.125	.078	.086
DUTS	$B_\mu \downarrow$	.603	.923	.681	.708	.763	.639	.634	.745	.604	.687
	$F_\beta \uparrow$	.747	.517	.688	.652	.607	.728	.685	.578	.743	.728
	$E_\xi \uparrow$	.865	.699	.833	.805	.776	.857	.828	.719	.856	.855
	$\mathcal{M} \downarrow$	.062	.135	.079	.101	.106	.068	.080	.106	.061	.080

Impact of gate: We propose gated structure-aware loss to let the network focus on salient regions of images instead of the entire image as in the traditional smoothness loss [38]. To verify the importance of the gate, we compare our loss with the smoothness loss, marked as "M3". As indicated, "M2" achieves better performance than "M3", demonstrating the gate reduces the ambiguity of structure recovery.

Impact of the edge detection task: We add edge detection task to "M1", and use cross-entropy loss to train the EDN. Performance is indicated by "M4". We observe that the B_μ measure is significantly decreased compared to "M1". This indicates that our auxiliary edge-detection network provides rich structure guidance for saliency prediction. Note that, our gated structure-aware loss is not used in "M4".

Impact of scribble boosting: We employ all the branches as well as our proposed losses to train our network and the performance is denoted by "M5". The predicted saliency map is also called our initial estimated saliency map. We observe decreased performance compared with "M0", where one iteration of scribble boosting is employed, which indicates effectiveness of the proposed boosting scheme.

Employing DenseCRF as post-processing: After obtaining our initial predicted saliency map, we can also use post-processing techniques to enhance the boundaries of the saliency maps. Therefore, we refine "M5" with DenseCRF, and results are shown in "M6", which is inferior to "M5". The reason lies in two parts: 1) the hyperparameters for DenseCRF is not the best; 2) DenseCRF recover structure information without considering saliency of the structure, causing extra false positive region. Using our scribble boosting mechanism, we can always achieve boosted or at least comparable performance as indicated by "M0".

Using Grabcut to generate pseudo label: Given scribble annotation, one can enlarge the annotation by using Grabcut [27]. We carried out experiment with pseudo label y' obtained by applying Grabcut to our scribble annotations y, and show performance in "M7". During training, we employ the same loss function as in Eq. 4, except that we use cross-entropy loss for $\mathcal{L}_s$. Performance of "M7" is worse than ours. The main reason is that pseudo label y' contains noise due to limited accuracy of Grabcut. Training directly with y' will overwhelm the network remembering the noisy label instead of learning useful saliency information.

Robustness to different scribble annotations: We report our performance "M0" by training the network with one set of scribble dataset. We then train with another set of the scribble dataset ("M8") to test robustness of our model. We observe staple performance compared with "M0". This implies that our method is robust to the scribble annotations despite their sparsity and few overlaps annotated by different labelers. We also conduct experiments with merged scribbles of different labelers as supervision signal and show performance of this experiment in the supplementary material.

Different edge detection methods: We obtain the edge maps E in Eq. 2 from RCF edge detection network [22] to train EDN. We also employ a hand-crafted edge map detection method, "Sobel", to train EDN, denoted by "M9". Since Sobel operator is more sensitive to image noise compared to RCF, "M9" is a little inferior to "M0". However, "M9" still achieves better performance than the results without using EDN, such as "M1", "M2" and "M3", which further indicates effectiveness of the edge detection module.

5. Conclusions

In this paper, we proposed a weakly-supervised salient object detection (SOD) network trained on our newly labeled scribble dataset (S-DUTS). Our method significantly relaxes the requirement of labeled data for training a SOD network. By introducing an auxiliary edge detection task and a gated structure-aware loss, our method produces saliency maps with rich structure, which is more consistent with human perception measured by our proposed saliency structure measure. Moreover, we develop a scribble boosting mechanism to further enrich scribble labels. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art weakly-supervised or unsupervised methods and is on par with fully-supervised methods.

Acknowledgment. This research was supported in part by Natural Science Foundation of China grants (61871325, 61420106007, 61671387), the Australia Research Council Centre of Excellence for Robotics Vision (CE140100016), and the National Key Research and Development Program of China under Grant 2018AAA0102803. We thank all reviewers and Area Chairs for their constructive comments.

References

[1] 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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2017. 1, 2, 3, 5

[2] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014. 6

[3] Jifeng Dai, Kaiming He, and Jian Sun. Boxsup: Exploiting bounding boxes to supervise convolutional networks for semantic segmentation. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1635–1643, 2015. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 248–255, 2009. 3, 6

[5] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R^3Net: recurrent residual refinement network for saliency detection. In *Proc. IEEE Int. Joint Conf. Artificial Intell.*, pages 684–690, 2018. 7

[6] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 4548–4557, 2017. 5

[7] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Proc. IEEE Int. Joint Conf. Artificial Intell.*, pages 698–704, 2018. 5, 6

[8] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1623–1632, 2019. 7

[9] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 270–279, 2017. 4

[10] Kuang-Jui Hsu, Yen-Yu Lin, and Yung-Yu Chuang. Weakly supervised saliency detection with a category-driven map generator. In *Proc. Brit. Mach. Vis. Conf.*, 2017. 2

[11] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7014–7023, 2018. 3

[12] Wei-Chih Hung, Yi-Hsuan Tsai, Yan-Ting Liou, Yen-Yu Lin, and Ming-Hsuan Yang. Adversarial learning for semi-supervised semantic segmentation. In *Proc. Brit. Mach. Vis. Conf.*, 2018. 3

[13] Anna Khoreva, Rodrigo Benenson, Jan Hosang, Matthias Hein, and Bernt Schiele. Simple does it: Weakly supervised instance and semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 876–885, 2017. 3

[14] Jiwhan Kim, Dongyoon Han, Yu-Wing Tai, and Junmo Kim. Salient region detection via high-dimensional color transform. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 883–890, 2014. 1

[15] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 695–711, 2016. 3, 4

[16] Guanbin Li, Yuan Xie, and Liang Lin. Weakly supervised salient object detection using image labels. In *Proc. AAAI Conf. Artificial Intelligence*, 2018. 1, 2, 7

[17] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 5455–5463, 2015. 6

[18] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 280–287, 2014. 6

[19] Di Lin, Jifeng Dai, Jiaya Jia, Kaiming He, and Jian Sun. Scribblesup: Scribble-supervised convolutional networks for semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3159–3167, 2016. 3

[20] Maire Michael Belongie Serge Hays James Perona Pietro Ramanan Deva Dollár Piotr Zitnick C. Lawrence Lin, Tsung-Yi. Microsoft coco: Common objects in context. In *Proc. Eur. Conf. Comp. Vis.*, pages 740–755, 2014. 3

[21] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3089–3098, 2018. 7

[22] Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3000–3009, 2017. 4, 8

[23] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6609–6617, 2017. 5, 7

[24] 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. *Proc. Adv. Neural Inf. Process. Syst.*, 2019. 1, 2

[25] Anton Obukhov, Stamatios Georgoulis, Dengxin Dai, and Luc Van Gool. Gated crf loss for weakly supervised semantic image segmentation. In *Proc. Adv. Neural Inf. Process. Syst.*, 2019. 3

[26] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7479–7489, 2019. 1, 2, 7

[27] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. Grabcut -interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (SIGGRAPH)*, 2004. 8

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. Int. Conf. Learning Representations*, 2014. 3, 6

[29] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learn-

ing for object saliency and detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3238–3245, 2013. 2

[30] Meng Tang, Abdelaziz Djelouah, Federico Perazzi, Yuri Boykov, and Christopher Schroers. Normalized cut loss for weakly-supervised cnn segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1818–1827, 2018. 2, 3, 4

[31] Meng Tang, Federico Perazzi, Abdelaziz Djelouah, Ismail Ben Ayed, Christopher Schroers, and Yuri Boykov. On regularized losses for weakly-supervised cnn segmentation. In *Proc. Eur. Conf. Comp. Vis.*, pages 524–540, 2018. 3

[32] Paul Vernaza and Manmohan Chandraker. Learning random-walk label propagation for weakly-supervised semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2953–2961, 2017. 3

[33] Bin Wang, Guojun Qi, Sheng Tang, Tianzhu Zhang, Yunchao Wei, Linghui Li, and Yongdong Zhang. Boundary perception guidance: A scribble-supervised semantic segmentation approach. In *Proc. IEEE Int. Joint Conf. Artificial Intell.*, pages 3663–3669, 2019. 3

[34] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 136–145, 2017. 1, 2, 6, 7

[35] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3127–3135, 2018. 7

[36] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2019. 2

[37] Xiang Wang, Shaodi You, Xi Li, and Huimin Ma. Weakly-supervised semantic segmentation by iteratively mining common object features. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1354–1362, 2018. 3

[38] Yang Wang, Yi Yang, Zhenheng Yang, Liang Zhao, Peng Wang, and Wei Xu. Occlusion aware unsupervised learning of optical flow. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4884–4893, 2018. 4, 8

[39] Yunchao Wei, Huaxin Xiao, Honghui Shi, Zequn Jie, Jiashi Feng, and Thomas S Huang. Revisiting dilated convolution: A simple approach for weakly- and semi- supervised semantic segmentation. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 7268–7277, 2018. 3

[40] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8150–8159, 2019. 7

[41] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3907–3916, 2019. 7

[42] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. 1, 2

[43] Saining Xie and Zhuowen Tu. Holistically-nested edge detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 1395–1403, 2015. 3

[44] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 1155–1162, 2013. 6

[45] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3166–3173, 2013. 6

[46] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3684–3692, 2018. 4

[47] Yu Zeng, Yunzi Zhuge, Huchuan Lu, and Lihe Zhang. Multi-source weak supervision for saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6074–6083, 2019. 1, 2, 7

[48] 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.*, pages 4068–4076, 2017. 2, 7

[49] Jing Zhang, Yuchao Dai, and Fatih Porikli. Deep salient object detection by integrating multi-level cues. In *Proc. IEEE Winter Conf. on App. of Comp. Vis.*, pages 1–10, 2017. 2

[50] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2020. 1, 2

[51] Jing Zhang, Bo Li, Yuchao Dai, Fatih Porikli, and Mingyi He. Integrated deep and shallow networks for salient object detection. In *Proc. IEEE Int. Conf. Image Process.*, pages 1537–1541, 2017. 2

[52] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 9029–9038, 2018. 1, 2, 7

[53] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 212–221, 2017. 7

[54] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 714–722, 2018. 7

[55] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. 2

[56] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3085–3094, 2019. 7

[57] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 2814–2821, 2014. 1