EasyChair Preprint

№ 6770

Weakly-Supervised Salient Object Detection through Object Segmentation Guided by Scribble Annotations

Xiongying Wang, Zaid Al-Huda and Bo Peng

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

October 6, 2021

Weakly-Supervised Salient Object Detection through Object Segmentation Guided by Scribble Annotations

1st Xiongying Wang
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, China
wangxiongying@my.swjtu.edu.cn

2nd Zaid Al-Huda
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, China
eng.zaidalhuda@gmail.com

3rd Bo Peng
School of Computing and Artificial Intelligence
Southwest Jiaotong University
Chengdu, China
bpeng@swjtu.edu.cn

***Abstract*—With the advent of Neural Network, Fully-supervised salient object detection achieves great success. However, it takes plenty of efforts to obtain precise pixel-level annotations. In order to reduce human labeling efforts, some research adapt weak form annotations, but they still fall short of the fully-supervised. In this paper, we propose a novel weakly-supervised salient object detection framework, which can reduce labeling efforts by using scribble annotations. In the meantime, we also incorporate Deep Convolutional Network to achieve high performance. To this end, we utilize high-quality region hierarchies, which are generated by Convolutional Oriented Boundary (COB) network, to select optimal level for object segmentation. We build initial saliency maps and thoroughly annotate the images during the initialization phase by spreading labels information from scribbles to other regions. During the training phase, the salient object detection convolutional network is trained using the initial saliency maps. Then, we utilize Conditional Random Field (CRF) to refine saliency maps, which will then be used to retrain the network. To achieve quality saliency maps, we iteratively optimize the training process. Extensive experiments on six benchmarks demonstrate that our proposed method outperforms previous weakly-supervised algorithms.**

***Index Terms*—Salient object detection, Scribble annotations, Weakly-supervised, Hierarchical segmentation.**

I. Introduction

Salient object detection (SOD) is to identify the most appealing parts in an image based on human perception. To extract saliency, traditional techniques [1] [2] employ low-level characteristics such as colors and textures. Those hand-crafted features or human experience only work well in simple context, but fail in complex ones. Recently, the development of Deep Convolutional Neural Networks has boosted salient object detection [3] [4] [5]. However, the performance of these Convolutional Neural Networks based approaches come at the expense of large pixel-level annotations. It is tedious and often takes several minutes for an expert annotator to label one image. To reduce human efforts and keep high performance at the same time, several semi-supervised, weakly-supervised or unsupervised methods [6] [7] [8] [9] have been introduced. Those approaches have propensities for interpreting from sparse data [6] [7], or learning from noise data [8] [9].

On the other hand, recent research has looked into splitting an image into a multi-scale structure to capture objects at all scales. Hierarchical segmentation increases the likelihood of locating a whole or a portion of an object at a certain hierarchy level. Hierarchical algorithms suffer from instability as well. Since low-level features are used to build hierarchy algorithms, results are vulnerable to space and feature parameters selection (edges, colors, etc.). As a result, the object's scale is not enforced to be cohesive. In this paper, we select the best segmentation level based on the boundary maps predicted by the hierarchical image segmentation algorithm.

Even though some semi-supervised and unsupervised methods in literature can address the human efforts problems in some way, the performance is still far behind fully-supervised approaches. In order to achieve high performance, the optimal object segmentation level is selected from boundary maps predicted by Convolutional Oriented Boundary (COB) [10]. We also build a mapping model which use scribbles [5] over the object segmentation. In doing so, we can capture precise local structure while maintaining object contour. By mapping foreground scribbles on the contour map, the initial saliency annotations can be generated. The initial saliency maps are then used to train a convolutional network for SOD. The framework is updated by alternately training, predicting and upgrading the prediction maps. We apply Conditional Random Field [11] to refine prediction maps during the alternately iteration process to correct errors. After it converges, we choose the one with the minimum loss to predict benchmarks.

The following are the three main contributions of this paper: (1) We propose a novel weakly-supervised salient map generation framework; (2) We design an approach to apply

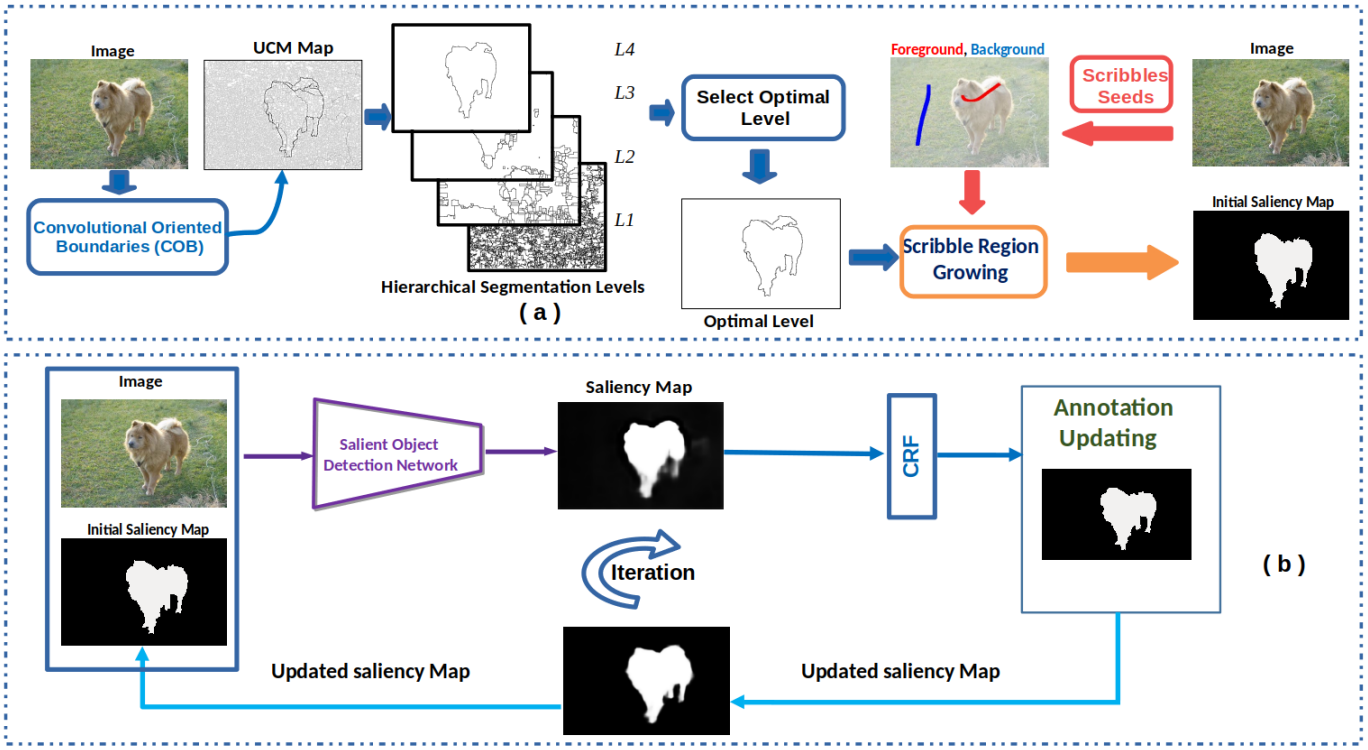

Fig. 1. Illustration of our framework. Two phases are included. Phase(a) uses contour maps to convey scribbles to initial saliency map. Phase(b) is an updating process to fine-tune the network.

hierarchical boundaries to propagate scribble to larger region; (3) In the experiments, our framework outperforms other state-of-the-art models on five benchmarks.

II. Related Work

A. Weak Annotations to Saliency

To achieve remarkable results, salient object detection (SOD) necessitates a large number of pixel-level annotations. Several recent researches seek to relieve the efforts of precise annotations and presented a variety of deep neural network based weakly supervised methods. Bounding box [12], and image-level labels [13] [6] are a few such weak annotation examples. Wang et al. [13] introduced a new way to extract salient objects with image-level annotation by combining Foreground Inference Network and an image classification network. Li et al. [6] leveraged coarse activation map generated from the unsupervised method in order to correct noise and generate better results. Fully connected CRF [14] used as a post-processing method in many studies [13] [6]. Zhang et al. [15] used "generating by fusing", which means to fuse the unsupervised outputs to guide the training. In this paper, we leveraged scribbles as weakly-guiding information for training process.

B. Hierarchical Image Segmentation

The bottom-up merging technique has been utilized by a huge number of algorithms to generate segmentation [16] [17] [18]. Arbelaze et al. [16] used spectral clustering to combine different local cues to generate global contour. They reduced the problem of image segmentation to contour detection. Pont-Tuset et al. [17] combined multi-scale information to generate proposals. They also proposed a one-scale approach, which ran faster since it only took one scale. To convey information from deep cues to unmarked regions, Al-Huda et al. [18] employed high-quality region hierarchies. In this paper, we use COB to generate contour maps, based on which initial maps were generated by growing information from scribbles to unknown areas.

III. Proposed Method

To take the advantages of both low efforts requirement from scribble and high performance from deep network, we propose a framework as is shown in Fig. 1. Two stages are included in this framework. The first stage focuses on getting initial saliency maps from scribbles. The second stage uses the initial saliency maps from the first stage to alternately update the network by correcting errors.

A. Generating high-quality initial saliency maps

In this section, we discuss three processes involved in the first phase of our framework. The steps included are shown in Fig. 1 (a): We generate contour hierarchy by applying COB on training dataset; then we choose the optimal contour level from generated hierarchy; finally, in order to generate high-quality initial saliency annotations as the supervision on network training, we grow scribble in the optimal contour

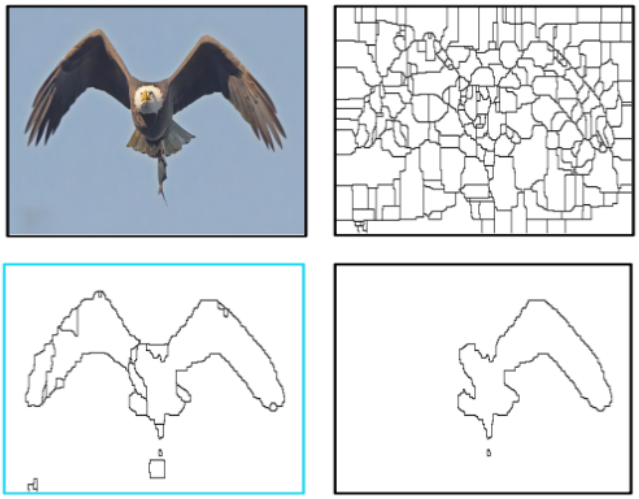

Fig. 2. Top left: Original Image. Top right: Over-Segmentation. Bottom left: Optimal Level. Bottom right: Under-Segmentation.

level.

1) Generate object segmentation hierarchy: In order to generate accurate initial annotations, we apply hierarchical objection segmentation method for good performance. Considering higher accuracy and clearer boundaries as COB shows in [19] [20] [21], in this work we also use COB [10] to generate Ultrametric Contour Map (UCM). It applied deep neural network to guide the generation process, which means it has better performance in objection detection, which can in turn contribute to saliency map generation.

2) Optimal scale selection process: How to choose the candidate contour level will directly influence initial annotation results. As it is shown in Fig. 2 (top right), the low level contour map can capture more details, but with too much noise, not to mention the cost of computation. However, in high levels, such as (bottom right), even if it curbs the noise from the background, foreground parts merges too. So that vital information gets lost. Neither being too high nor too low a level is a good candidate, whereas middle levels can be a good balance. As it is shown in (bottom left), middle levels tend to curb noise whilst keeping enough information.

In this paper we choose the 30% level from the bottom. The equation used to calculate the number of regions is given as below:

$$N(p) = \frac{1}{n}\sum_{i=1}^{n} t(i), \tag{1}$$

where $t(i)$ is the numbers of regions in one contour map in a specific level p. n is the total image numbers in the training data set. Our goal is to find out the suitable level which satisfies:

$$l \leq N(p) \leq h. \tag{2}$$

Fig. 3. A illustration of extracting foreground scribble. The first row shows when both fore-scribble and back-scribble exist. The second row shows result after extracting fore-scribble.

Empirically, we set $l = 30$ and $h = 40$. By applying Eq. (1) and Eq. (2) in the boundary hierarchy, we find the 30% level from the bottom would be the best.

3) Scribble-guided object region growing process: Salience detection is to distinguish the foreground from the background - it is mutual-exclusive. A pixel could be either the foreground or the background. However, in an image, the most intriguing part (foreground) should be highlighted. In the proposed framework, we extract the foreground scribble from both fore-scribble and back-scribble. The result of choosing foreground is shown in Fig.3. The extracted the foreground scribble will be used to guide the region selection process.

Algorithm 1 Map foreground scribble to UCM

Input: UCM set $U = [u_1, u_2, ..., u_n]$and Foreground scribble set $F = [f_1, f_2, ..., f_n]$
Output: Initial saliency annotation

```
while i ≠ n do
    Pick up one pair (u_i, f_i)
    Find out all regions R = [r_1, r2, ..., r_m] in u_i.
    while j ≠ m do
        Check intersection In = r_j ∩ f_i
        if In ≥ θ then
            Mark r_j as salience
        else
            Mark r_j as background
        end if
    end while
end while
```

After obtaining the foreground scribble, we need to map fore-scribble to the UCM map for the initial saliency annotations. In this process, we need to find out all the closed regions in UCM. For regions which overlap with the fore-scribble, they will be assigned as the foreground. After scanning all

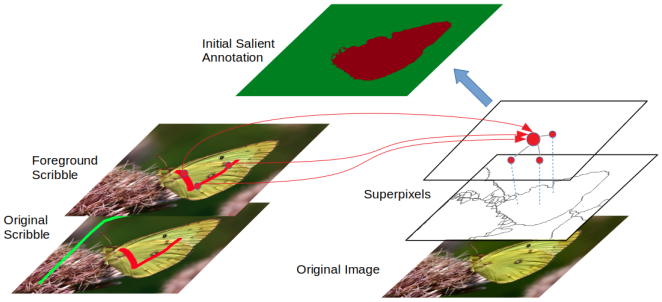

Fig. 4. An illustration of extracting foreground scribble. Then map the fore ground on superpixels to generate initial saliency annotation.

regions and marking all possible foreground regions, the rest regions will be assigned as the background. The process is illustrated in Algorithm 1. Empirically we set θ to be 50 pixels as a threshold to filter out errors in scribbles. The process of mapping the scribble to UCM is shown in Fig. 4.

B. Network Training And Updating

In this section, we discuss the second phase in our framework. This phase includes three steps, as is shown in Fig. 1(b). We use initial saliency annotations from first phase(Fig. 1(a)) to train a salient object detection network several epochs till it converges; then, we use the trained network to predict training saliency maps. We then apply CRF [11] to update the predicted saliency maps. Finally, we iterate the two steps mentioned before, until the network finally converges.

We adapt VGG16 [22] as our back bone network to generate saliency map. We train this 2D prediction network based on initial annotations obtained from phase(a) in Fig.1. Here the loss function is defined as:

$$loss = -\frac{1}{n}\sum(y_i * \ln x_i + (1 - y_i) * \ln(1 - x_i)), \quad (3)$$

where y_n, x_n, i and n denote ground truth, prediction, current pixel and total pixel numbers in an image. The training stops when $loss \leq \gamma$, where $\gamma = 0.0001$, or over-fitting appears - loss starting to bounce back. We use VGG16 [22] pretrained on ImageNet [23] to initialize our network, and set base learning rate as 1e-4. It took for average 50 epochs to converge. After one iteration completed, we use the model with the lowest loss to predict on training dataset, and then use fully-connected CRF [11] to refine the new predicted result [24]. The whole process iterated several times until the network finally converges. After the whole training stops, we use pretrained network to predict on five benchmarks.

IV. Experimental Results

A. Setup

Competing methods: Six state-of-the-art weakly-supervised/unsupervised and eight fully-supervised methods are used for comparison.

Dataset: For training, the dataset we use is DUTS-TR [25]. The scribbles come from S-DUTS [5]. For testing, five common benchmarks are used:(1)ESSED [26]; (2)DUT [27]; (3)PASCAL-S [28]; (4)HKU-IS [29]; (5)DUTS testing dataset [25].

Metrics: Four commonly used metrics are utilized to evaluate the results: Mean Absolute Error(MAE), F-measure, PR-curve, and F-curve.

Mean Absolute Error(MAE) [30] - which is represented in our paper as $\mathcal{M}$ - calculates the absolute per-pixel difference between prediction map and ground truth map. It is defined as:

$$\mathcal{M} = \frac{1}{W \times H}\sum W_{x=1}\sum H_{y=1}|\overline{S}(x,y) - \overline{G}(x,y)|, \quad (4)$$

where $\overline{S}$ and $\overline{G}$ is prediction and ground truth, respectively. W and H mean the width and the height of an image. x and y denote the coordination of a pixel. For a dataset, MAE means the average of whole saliency maps over the whole dataset.

PR-curve, as a widely-accepted method to evaluate the saliency map, shows precision and call in a visual and intuitive way. Each pair of precision and recall is calculated by taking threshold in both prediction and ground truth. By varying threshold from the min value of the map to max of the map, we got a sequence of PR values, which then is used to plot PR curves.

To balance and get a comprehensive view of both precision and recall, F_β is defined in Eq. (5). It can be calculated based on each pair of precision and recall values.

$$F_\beta = \frac{(1 + \beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \quad (5)$$

where β^2 is set to be 0.3 according to [31]. The F-curve offers a direct visual way to show how F_β changes with different thresholds.

B. State-of-the-art Comparison

Quantitative Comparison: We compare our method with other state-of-the-art weakly-supervised/unsupervised methods in Table I and Fig. 6. As it is shown in Table I, our framework outperforms all the state-of-the-art weakly-supervised methods on F-measure. Compared with the state-of-the-art weakly-supervised method (WSA) about F-measure, we get 2.1% improvement on average on all test datasets. The highest F gain(2.7%) comes from dataset ECSSD. Images in ECSSD not only often contain multiple objects, but also have complex object structures, which in some way proved that our framework can better detect several objects with complex structures. The lowest F gain-1.7% is on dataset DUT-OMRON, which contains low image contrast and irregular object boundaries like a swimming pool, a road or a fence. These objects in the image often don't

TABLE I
RESULTS ON 5 BENCHMARKS DATASETS. THE BEST RESULTS ARE IN BOLD.

	Metric	Fully Sup. Models								weakly Sup./Unsup. Models						
		PiCANet [32]	NLDF [33]	MSNet [34]	CPD [35]	AFNet [36]	PFAN [37]	PAGRN [38]	BASNet [3]	SBF [15]	WSI [6]	WSS [13]	MNL [9]	MSW [7]	WSA [5]	Ours
ECSSD	$F_\beta \uparrow$	.8715	.8709	.8856	.9076	.9008	.8592	.8718	**.9128**	.7823	.7621	.7672	.8098	.7606	.8650	**.8929**
	$\mathcal{M} \downarrow$	.0543	.0656	.0479	.0434	.0450	.0467	.0644	**.0399**	.0955	.0681	.1081	.0902	.0980	.0610	**.0599**
DUT	$F_\beta \uparrow$	.7105	.6825	.7095	.7385	.7425	.7009	.6754	**.7668**	.6120	.6408	.5895	.5966	.5970	.7015	**.7210**
	$\mathcal{M} \downarrow$	.0722	.0796	.0636	.0567	.0574	.0615	.0709	**.0565**	.1076	.0999	.1102	.1028	.1087	.0684	**.0683**
PASCAL-S	$F_\beta \uparrow$	.7985	.7933	.8129	.8220	**.8241**	.7544	.7656	.8212	.7351	.6532	.6975	.7476	.6850	.7884	**.8054**
	$\mathcal{M} \downarrow$	.1284	.1454	.1193	.1215	**.1155**	.1372	.1516	.1217	.1669	.2055	.1843	.1576	.1780	.1399	**.0995**
HKU-IS	$F_\beta \uparrow$	.8543	.8711	.8780	.8948	.8877	.8717	.8638	**.9025**	.7825	.7625	.7734	.8196	.7337	.8576	**.8747**
	$\mathcal{M} \downarrow$	.0464	.0477	.0387	.0333	.0358	.0424	.0475	**.0322**	.0753	.0885	.0787	.0650	.0843	**.0470**	.0490
DUTS	$F_\beta \uparrow$	.7565	.7567	.7917	**.8246**	.8123	.7648	.7781	.8226	.6223	.5687	.6330	.7249	.6479	.7467	**.7731**
	$\mathcal{M} \downarrow$	.0621	.0652	.0490	**.0428**	.0457	.0609	.0555	.0476	.1069	.1156	.1000	.0749	.0912	**.0622**	.0642

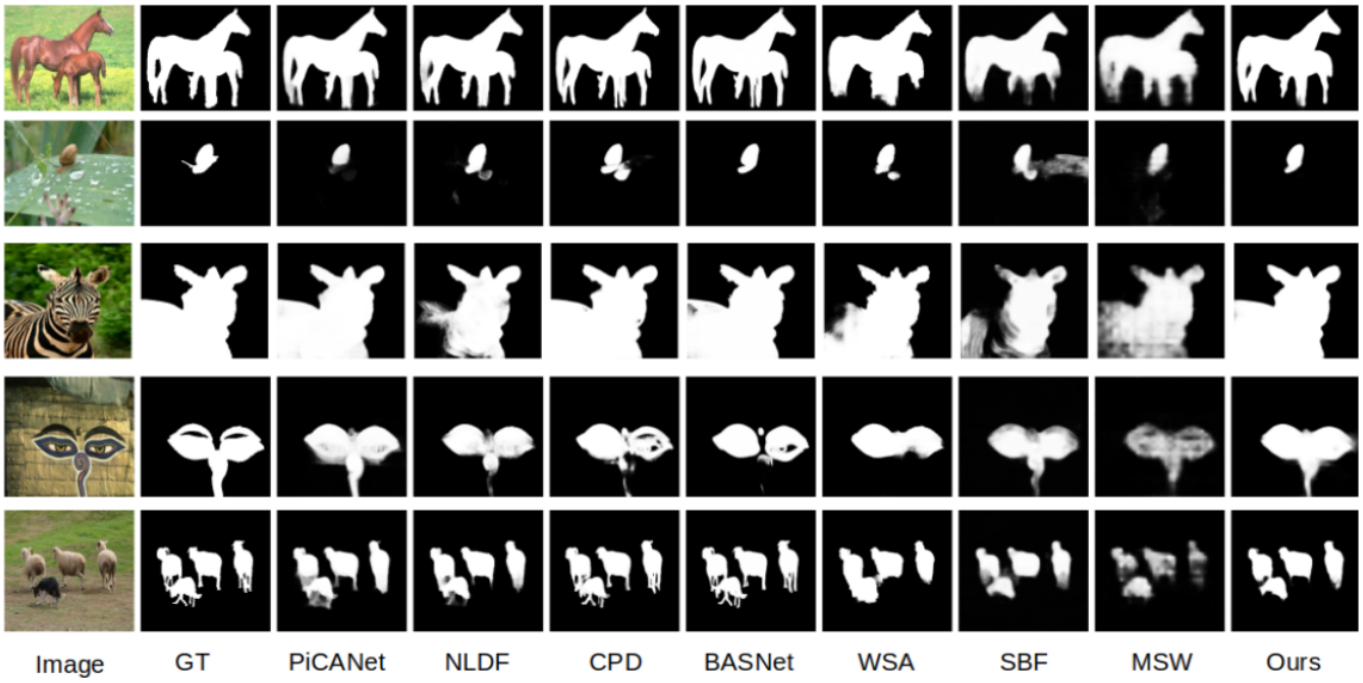

Fig. 5. Saliency maps comparison of the proposed framework with seven other weakly-supervised/unsupervised or fully-supervised methods.

have clear boundary or clear semantic meanings. In that case, our framework tends to predict the most salient part of those un-salient object(not complete), which in turn leads to relatively low F gains.

As for MAE($\mathcal{M}$), Table I shows that - compared to the second best method - our framework performs better on 3 out of the total 5 datasets. On ECSSD, DUT and PASCAL-S, we achieved 1.2% improvements on average, which means our framework can reduce disparities between the ground truth and the prediction. Even though, we do not gain the better results on HKU-IS and DUTS-TEST, the absolute difference(0.002) is tiny. In other words, we achieved comparable results on this two datasets, compared to the best method so far. In contrast to F-measure, MAE is harder to improve for several reasons: First, the calibrations are different. F-measure tends to focus on precision, so the more foreground the better. It makes sense, because saliency results are often used in the following image processing applications, so it is better not to miss salient information. In contrast, MAE mainly cares about the difference without considering the importance of precision or foreground, nor the semantic information like integrity, complexity and etc. Second, since our framework is based on scribbles – only 5% of the pixels on average are marked – the sparse information can definitely cause a loss in object structure and details. In such cases, minor thin things like hair, antenna, distal ends along side with low contrast like an animal with camouflage make the weakly-supervised method worse in MAE. So it is acceptable to have minor but comparable results with MAE.

In Fig. 6, we also compared with other competing methods in 3 benchmarks: ECSSD, DUT, HKU-IS. We will compare from perspectives of methods, metrics and then benchmarks.

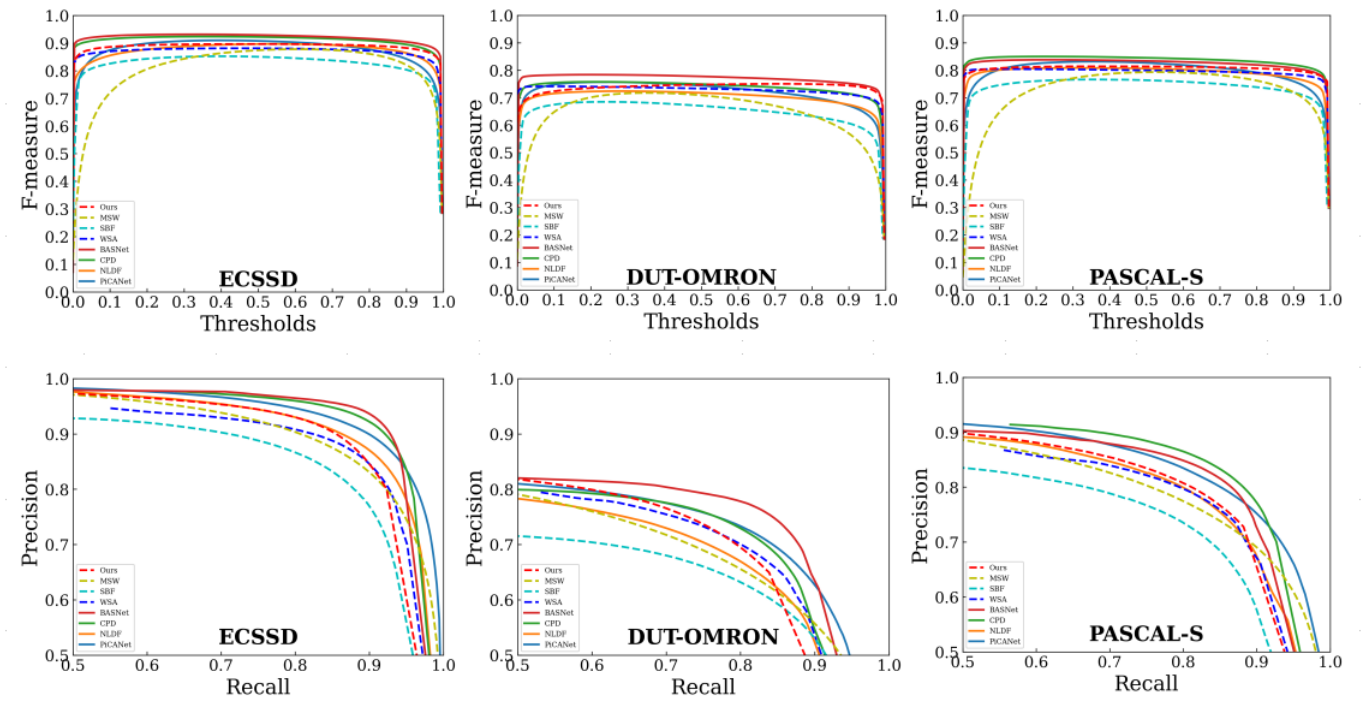

Fig. 6. F-measure curves(1st row) and Precision-Recall Curves(2nd row) on three benchmarks.

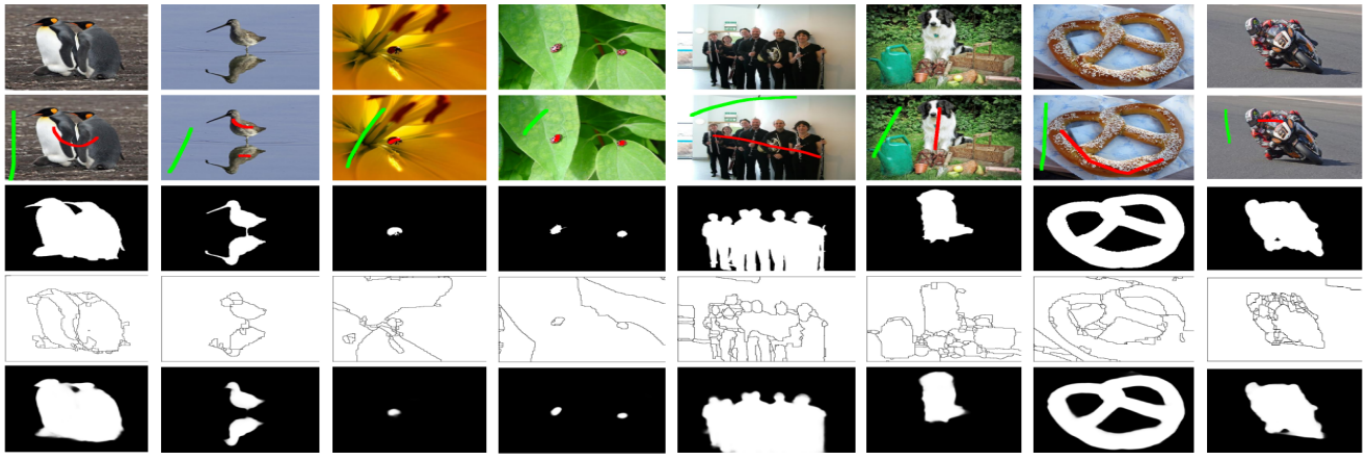

Fig. 7. An illustration of how UCM and scribbles used to generate initial input contributes to the final prediction. Five rows in the figure are: Input image (1st. row); Scribbles on input images; GT (3nd. row); (3rd row); UCM (4th. row) and Our prediction (5th. row).

Fig. 6 shows that our method surpasses other weakly-supervised/unsupervised approaches in both F-measure and PR-curves, which demonstrates robustness of our framework. In particular, our method gains by a large margin when it compares to weakly-supervised methods like MSW and SBF. Moreover, our method is even comparable to some fully supervised methods. Particularly in dataset DUT-OMRON, it is better than fully supervised NLDF. However, if we check metrics separately, for F-measure, our framework tends to obtain high scores. Since it focuses on extracting foreground so that pixel with low scores are more likely to be marked as background by the framework. Precision tends to get the best point when the threshold is 0.5, which also suggests pixels with scores less than threshold are prone to be background. When we compare different benchmarks, it shows that methods are more likely to achieve better performance on ECSSD and PASCAL-S, but lower on DUT-OMRON. It also suggests that different benchmarks could have different data composition, and using various datasets can better evaluate methods' performance.

Qualitative Comparison: To further illustrate superior performance of our framework, we visualize three aspects:

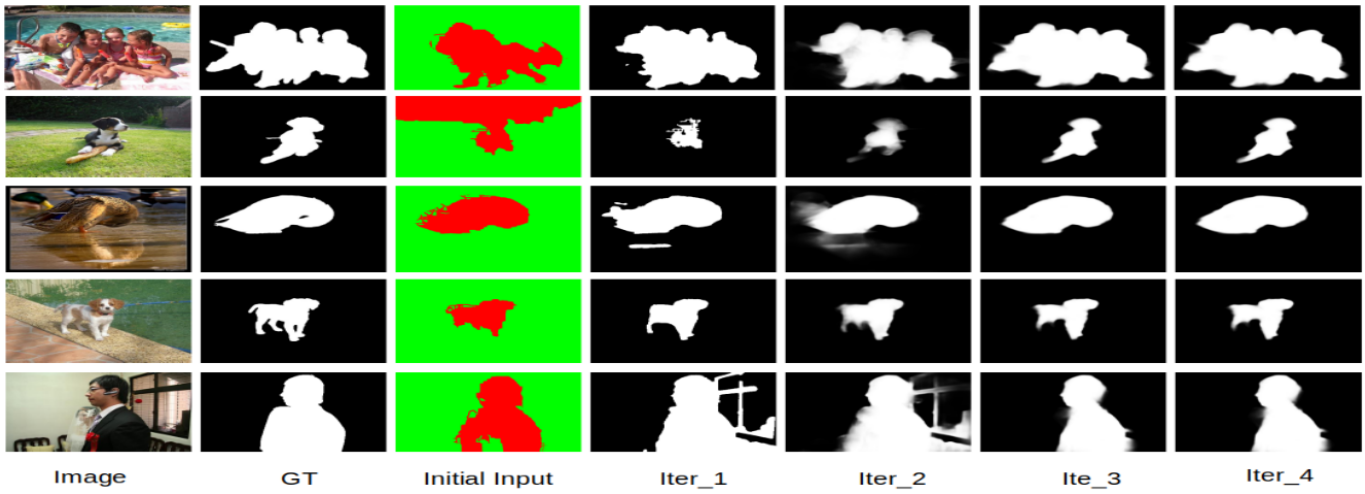

Fig. 8. A demonstration of how iterations help with fine-tuning the network, Iter in the figure means Iteration.

how different methods works, the role of UCM in our framework, and the influence of iteration.

In Fig. 5, we randomly pick out five images from test dataset to show our strengths compared to other weakly-supervised/unsupervised or fully supervised methods. As is shown in Fig.5 (1st row), thanks to boundary correcting post-processing iterations, our method tends to have clearer and more complete object limbs. It still holds when it is small target (2nd row). The alternate post-processing can suppress noise during iteration. Those weakly-supervised/unsupervised methods used for comparison are either miss parts or predict noise when it comes to small objects. Fully supervised methods tend to predict more facing small objects, since over-fitting often comes after training too much. Our framework can do a better job when targets take up most of the image and is composed of high contrast parts(complex structure) like zebra in 3rd row. Even facing human-made highly-semantically-meaningful targets – which other methods don't interpret complete, our method can still pull it off (4th row). Our framework still holds its ground when it's about small objects and multiple ones like it is in the 5th row.

To illustrate how UCM contribute to our superior performance, we randomly pick up a few images from training data set to show how the result becomes after the network converges, as is shown in Fig. 7. Thanks to COB [10] and UCM level choosing process in Algorithm 1, we can achieve a lot in the following complex scenarios: object composed with high contrast parts – complex structure; highly semantic-related mirror; small targets in similar environment; multiple small objects; many objects; saliency in several un-salient noise objects; object with irregular boundaries; and strong semantic-related targets, as shown in Fig. 7, columns 1st, 2nd, 3rd, 4th, 5th, 6th, 7th, and 8th, respectively.

We also illustrate how iterations contribute to our superior performance in Fig. 8. In general, as is shown, with iterations(Here the iteration is not epoch; One iteration contains many epochs) alternately going, the network converges towards a finer result. After each iteration, CRF is applied to retouch prediction, and it takes about four iterations to converge. The benefits of our alternately iteration can be summarized as follows: When more than one salient object exists and initial input do not catch them completely in Fig. 8 (1st and 2nd row), iteration can fix the missing parts gradually; as is shown in 3rd row, if the salience exists in a environment with similar color and texture patterns(unclear boundary), even though CRF can usher in errors, the iteration can still filter it out; In the 4th row, alternate iterations also partially solved one of the biggest problems in weakly supervised saliency – missing details; Moreover, in 5th row, the iteration process also shows its ability to catch high level semantic meanings and suppress errors.

V. Conclusion

In this paper, we proposed a weakly-supervised salient object detection (SOD) framework with supervision from information-sparse scribble annotations. To overcome the low information density with scribble, we introduced contour generation method COB as guidance to generate saliency map as initial input. With more information-dense initial saliency map as input, we alternately trained a network. After applying CRF on predicted saliency maps at each iteration, our network corrected errors and optimized gradually on training dataset. Extensive experiments show that on metrics F-measure, our framework outperforms all the other weakly-supervised/unsupervised methods. Our method also achieve the best result on most benchmarks about MAE. The PR-curves also show robustness of our approach. Furthermore, our framework is even on par with some fully-supervised methods.

ACKNOWLEDGMENT

This work was supported by the National Science Foundation of China (Nos. 61772435, 61961038), Sichuan highway science and technology project (No. 2019-01) and the Fundamental Research Funds for the Central Universities (No. 2682021ZTPY069).

REFERENCES

[1] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 883–890, 2014.

[2] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2814–2821, 2014.

[3] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7479–7489, 2019.

[4] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 8779–8788, 2019.

[5] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12546–12555, 2020.

[6] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, 2018.

[7] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6074–6083, 2019.

[8] D. T. Nguyen, M. Dax, C. K. Mummadi, T. P. N. Ngo, T. H. P. Nguyen, Z. Lou, and T. Brox, "Deepusps: Deep robust unsupervised saliency prediction with self-supervision," *arXiv preprint arXiv:1909.13055*, 2019.

[9] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9029–9038, 2018.

[10] K.-K. Maninis, J. Pont-Tuset, P. Arbeláez, and L. Van Gool, "Convolutional oriented boundaries," in *European conference on computer vision*, pp. 580–596, Springer, 2016.

[11] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, pp. 109–117, 2011.

[12] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3238–3245, 2013.

[13] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 136–145, 2017.

[14] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[15] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4048–4056, 2017.

[16] P. Arbelaez, M. Maire, C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 5, pp. 898–916, 2010.

[17] J. Pont-Tuset, P. Arbelaez, J. T. Barron, F. Marques, and J. Malik, "Multiscale combinatorial grouping for image segmentation and object proposal generation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 1, pp. 128–140, 2016.

[18] Z. Al-Huda, B. Peng, Y. Yang, R. N. A. Algburi, M. Ahmad, F. Khurshid, and K. Moghalles, "Weakly supervised semantic segmentation by iteratively refining optimal segmentation with deep cues guidance," *Neural Computing and Applications*, pp. 1–26, 2021.

[19] B. Peng, Z. Al-Huda, Z. Xie, and X. Wu, "Multi-scale region composition of hierarchical image segmentation," *Multimedia Tools and Applications*, vol. 79, no. 43, pp. 32833–32855, 2020.

[20] Z. Al-Huda, B. Peng, Y. Yang, and R. N. A. Algburi, "Object scale selection of hierarchical image segmentation with deep seeds," *IET Image Processing*, vol. 15, no. 1, pp. 191–205, 2021.

[21] Z. Al-Huda, D. Zhai, Y. Yang, and R. N. A. Algburi, "Optimal scale of hierarchical image segmentation with scribbles guidance for weakly supervised semantic segmentation," *International Journal of Pattern Recognition and Artificial Intelligence*, p. 2154026, 2021.

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

[23] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," *arXiv preprint arXiv:1412.7062*, 2014.

[25] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017.

[26] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1155–1162, 2013.

[27] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3166–3173, 2013.

[28] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 280–287, 2014.

[29] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5455–5463, 2015.

[30] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE conference on computer vision and pattern recognition*, pp. 733–740, IEEE, 2012.

[31] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 678–686, 2016.

[32] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3089–3098, 2018.

[33] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp. 6609–6617, 2017.

[34] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8150–8159, 2019.

[35] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3907–3916, 2019.

[36] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1623–1632, 2019.

[37] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3085–3094, 2019.

[38] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 714–722, 2018.