Supplementary Materials for Texture-guided Saliency Distilling for Unsupervised Salient Object Detection

Huajun Zhou[1], Bo Qiao[1], Lingxiao Yang[1], Jianhuang Lai[1,2,3], Xiaohua Xie[1,2,3*]
[1]School of Computer Science and Engineering, Sun Yat-sen University, China
[2]Guangdong Province Key Laboratory of Information Security Technology, China
[3]Key Laboratory of Machine Intelligence and Advanced Computing, Ministry of Education, China

Table 1. Setting comparison between USOD methods. "F" and "U" indicate fully-supervised and unsupervised pre-training. "IN" and "CS" are ImageNet [3] and CityScape [2] datasets, respectively.

Method	Training set	Input	Encoder	Pre-train	Saliency cues	Train time
EDNS [21]	DUTS-TR	352×352	VGG-16	F-IN	[15, 17, 27]	>8h
DCFD [9]	DUTS-TR	–	ResNet-50	F-IN	[8]	–
Ours	DUTS-TR	320×320	ResNet-50	U-IN	No	4.5h
SBF [18]	MSRA-B	224×224	VGG-16	F-IN	[16, 19, 20]	>3h
MNL [22]	MSRA-B	425×425	ResNet-101	F-IN	[6, 8, 16, 27]	>4h
USPS [11]	MSRA-B	432×432	ResNet-101	F-CS	[6, 8, 16, 27]	>30h
DCFD [9]	MSRA-B	–	ResNet-101	F-CS	[8]	–
A2S [25]	MSRA-B	320×320	ResNet-50	U-IN	No	1h
Ours	MSRA-B	320×320	ResNet-50	U-IN	No	1.3h

Setting comparison between USOD methods. As listed in Tab. 1, our method achieves better performance under disadvantage settings. Specifically, the 320^2 input of our method is small than most USOD methods, such as 425^2 for MNL [22] and 432^2 for USPS [11]. Moreover, we use ResNet-50 [4] as backbone, which is a weakened version of ResNet-101 used in many USOD methods [9,11,22]. As for pre-training, most existing methods employed the encoders pre-trained with manual annotations of some close-related datasets, such as ImageNet [3] for object recognition and Cityscape [2] for semantic segmentation. Such setting indicates that they benefit from the semantic knowledge of manual annotations, which violates the semantic-agnostic definition of the SOD task. On the contrary, the encoder of our method is pre-trained without using any human annotation. It means that no semantic knowledge is involved in the whole training process, which accords with the semantic-agnostic definition. Last, even excluding the additional time of existing methods [11, 18, 21, 22] to extract salience cues using traditional methods, the training time of our method is much less than that of most previous methods.

Qualitative comparison of different loss. In our manuscript, we exhibit the quantitative results of different losses in ablation study A. Here, we provide a qualitative comparison in Fig. 1. Our baseline A1 can accurately localize salient objects in images, but loses many details. Trained using the proposed losses, the network mines more detailed saliency knowledge progressively and thus precisely predicts the saliency boundaries.

Visualization of the learned saliency. In, Fig. 2, we visualize the learned saliency maps in our method during the training process. In the initial stage, our method is able to precisely localize the salient object based on the initial saliency cues, however, some small patches may still be misclassified. After subsequent tuning process, our method can learn more precise saliency knowledge and thus produce a high-quality pseudo label.

Effect of hyperparameters. The performance of our framework is affected by several hyperparameters, including α, λ_c, λ_b, λ_m, and k. We vary these hyperparameters and exhibit their results in Tab. 2. The results prove that the values $\alpha = 200$, $\lambda_c = 1$, $\lambda_b = 0.05$ and $\lambda_m = 1$ work best in practice. Our framework is robust to $\alpha \in [100, 300]$, and reports the best performance for $\alpha = 200$. Moreover, our framework achieves comparable performance for various λ_c and λ_m values within $[0.5, 1.5]$. Furthermore, our framework is sensitive to λ_b. Although we observe robust performance for $\lambda_b \in [0.03, 0.07]$, λ_b outside this range

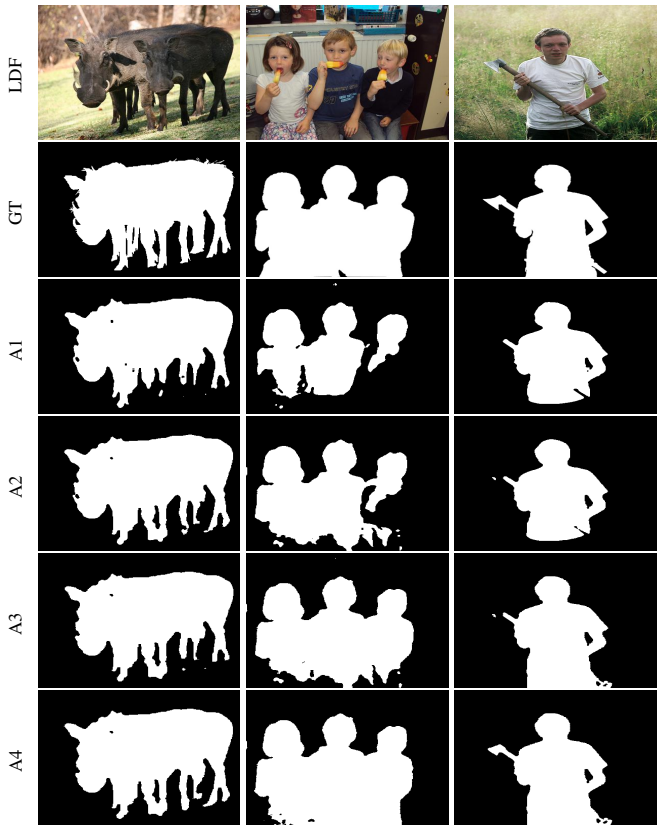

Figure 1. Saliency predictions of our method with different losses.

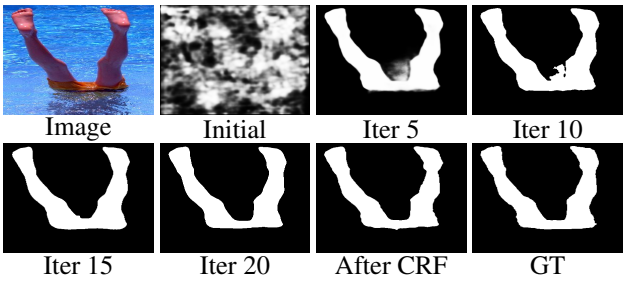

Figure 2. The learned saliency maps during training.

Table 2. Effect of different hyperparameters.

Parameter	Value	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
α	100	.911	.932	.046
	150	.914	.937	.043
	200	.917	.945	.038
	250	.915	.942	.039
	300	.914	.943	.039
λ_c	0.5	.911	.937	.042
	0.7	.915	.942	.039
	1	.917	.945	.038
	1.2	.914	.943	.038
	1.5	.913	.941	.039
λ_b	0.01	.869	.915	.054
	0.03	.910	.938	.042
	0.05	.917	.945	.038
	0.07	.914	.945	.039
	0.09	.908	.942	.040
λ_m	0.5	.915	.943	.038
	0.75	.915	.945	.038
	1	.917	.945	.038
	1.25	.915	.943	.039
	1.5	.914	.941	.039
k	3	.913	.937	.042
	5	.917	.945	.038
	7	.914	.943	.039

Table 3. Different losses for the second stage.

Loss	DUT-OMRON			ECSSD		
	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
BCE	.708	.834	.066	.891	.924	.047
BCE+IOU	.726	.846	.065	.894	.923	.048
BIS	.716	.838	.067	.886	.919	.049
CTLoss	.743	.862	.061	.907	.914	.057
IOU	.745	.863	.061	.916	.938	.044

(*e.g.*, $\lambda_b = 0.01$) seems to induce significant performance drops. We attribute this performance drop to the weakened effect of pixels with similar appearances.

Loss for training extra saliency detectors. For fully-supervised SOD methods, there are many choices for the loss functions, such as BCE loss [5, 23], BCE+IOU loss [12, 24], CTLoss [1, 26], BIS(BCE+IOU+SSIM) loss [13]. We employ these losses to train our saliency detector with the generated pseudo labels, as exhibited in Tab. 3. In summary, training our detector with IOU loss achieves the best results compared to other losses. BCE and CTLoss provide pixel-wise supervised signals, which means that training with these losses is easy to overfit the noises and thus degrade the generalization ability of our detector. Similarly, SSIM is based on regional statistics and thus is sensitive to noisy regions in pseudo labels. Unlike the above losses, IOU is robust to pixel-level or region-level noises because it is based on global statistics of saliency predictions.

Necessity of unsupervised SOD. Ideally, a well supervisedly trained class-agnostic SOD model can handle all scenarios, whereas is hard to obtain in practical. First, SOD methods trained on datasets with limited classes may not perform well on unseen classes, even if class labels are not used during training. Second, SOD methods trained on a certain style of images (e.g., natural images) do not perform well on other styles of images (e.g., medical images). To prove this point, we show the results of supervised LDF [14]

Table 4. Performance of SOD methods on X-ray images.

	Training set	$F_\beta \uparrow$	$E_\xi \uparrow$	$\mathcal{M} \downarrow$
LDF [14]	DUTS-TR	.296	.508	.315
Ours	DUTS-TR	.530	.664	.309
Ours*	DUTS-TR+X-ray	.924	.943	.056

Figure 3. Examples of the predicted saliency maps.

and our unsupervised method on chest X-ray images [1] in Tab. 4 and Fig. 3. It proves that SOD methods trained on existing SOD datasets perform poorly on X-ray images, while our unsupervised method can achieve significant performance without using extra human annotations for specific scenarios. Third, many weakly-supervised methods leverage USOD methods as pre-processing or auxiliary loss, such as [7, 10]. In addition, our method can be considered as a novel self-supervised learning paradigm.

References

[1] Zixuan Chen, Huajun Zhou, Jianhuang Lai, Lingxiao Yang, and Xiaohua Xie. Contour-aware loss: Boundary-aware learning for salient object segmentation. *IEEE Transactions on Image Processing*, 30:431–443, 2020. 2

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Computer Vision and Pattern Recognition*, 2016. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 2

[6] Bowen Jiang, Lihe Zhang, Huchuan Lu, Chuan Yang, and Ming-Hsuan Yang. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE international conference on computer vision*, pages 1665–1672, 2013. 1

[7] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5495–5505, 2021. 3

[8] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE international conference on computer vision*, pages 2976–2983, 2013. 1

[9] Xiangru Lin, Ziyi Wu, Guanqi Chen, Guanbin Li, and Yizhou Yu. A causal debiasing framework for unsupervised salient object detection. In *Thirty-sixth AAAI conference on artificial intelligence*, 2022. 1

[10] Wenfeng Luo, Meng Yang, and Weishi Zheng. Weakly-supervised semantic segmentation with saliency and incremental supervision updating. *Pattern Recognition*, 115:107858, 2021. 3

[11] 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. *arXiv preprint arXiv:1909.13055*, 2019. 1

[12] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020. 2

[13] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 2

[14] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13025–13034, 2020. 2, 3

[15] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *European conference on computer vision*, pages 29–42. Springer, 2012. 1

[16] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013. 1

[17] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 1

[18] Dingwen Zhang, Junwei Han, and Yu Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017. 1

[19] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013. 1

[20] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Minimum barrier salient

[1] https://lhncbc.nlm.nih.gov/LHC-downloads/downloads.html

object detection at 80 fps. In *Proceedings of the IEEE international conference on computer vision*, pages 1404–1412, 2015. 1

[21] Jing Zhang, Jianwen Xie, and Nick Barnes. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. In *European conference on computer vision*, pages 349–366. Springer, 2020. 1

[22] Jing Zhang, Tong Zhang, Yuchao Dai, Mehrtash Harandi, and Richard Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9029–9038, 2018. 1

[23] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017. 2

[24] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 8779–8788, 2019. 2

[25] Huajun Zhou, Peijia Chen, Lingxiao Yang, Xiaohua Xie, and Jianhuang Lai. Activation to saliency: Forming high-quality labels for unsupervised salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 1

[26] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9141–9150, 2020. 2

[27] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2814–2821, 2014. 1