Learning Noise-Aware Encoder-Decoder from Noisy Labels by Alternating Back-Propagation for Saliency Detection

Anonymous ECCV Supplementary Material

Paper ID 2760

We introduce the noise generator network, residual channel attention module details, more experimental results and model analysis in this document.

1 Network Details

We visualize details of the generator network and residual channel attention module in the encoder-decoder network.

1.1 Generator

We construct the noise generator by using six cascaded deconvolution layers, finishing with tanh activation to map the noise map Δ to the range of $[-1, 1]$, as shown in Fig. 1. "dec(11,11)" represents the deconvolutional layer that maps Z to a feature map size of 11×11. "dec(2)" indicates a deconvolutional layer of stride 2.

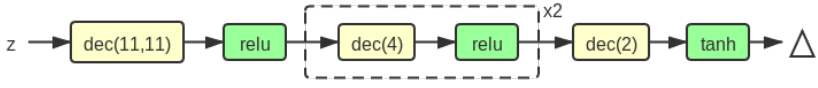

Fig. 1. Generator network, which takes low dimensional vector Z as input, and produce noise map Δ of the same size as our latent saliency map S.

1.2 Residual Channel Attention

We visualize the residual channel attention module in the paper, and show its details in Fig. 2, where "c1x1x64" represents 1×1 convolutional layer of channel size 64, "GAP" is the global average pooling layer, "Sig" is the Sigmoid operation.

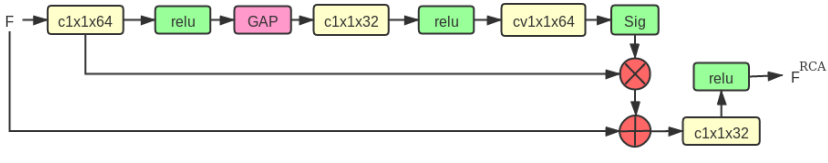

Fig. 2. Details of the residual channel attention module.

2 Visualization of Inferred Noise

We show more examples of the inferred noise Δ as well as the latent clean saliency map S in this section.

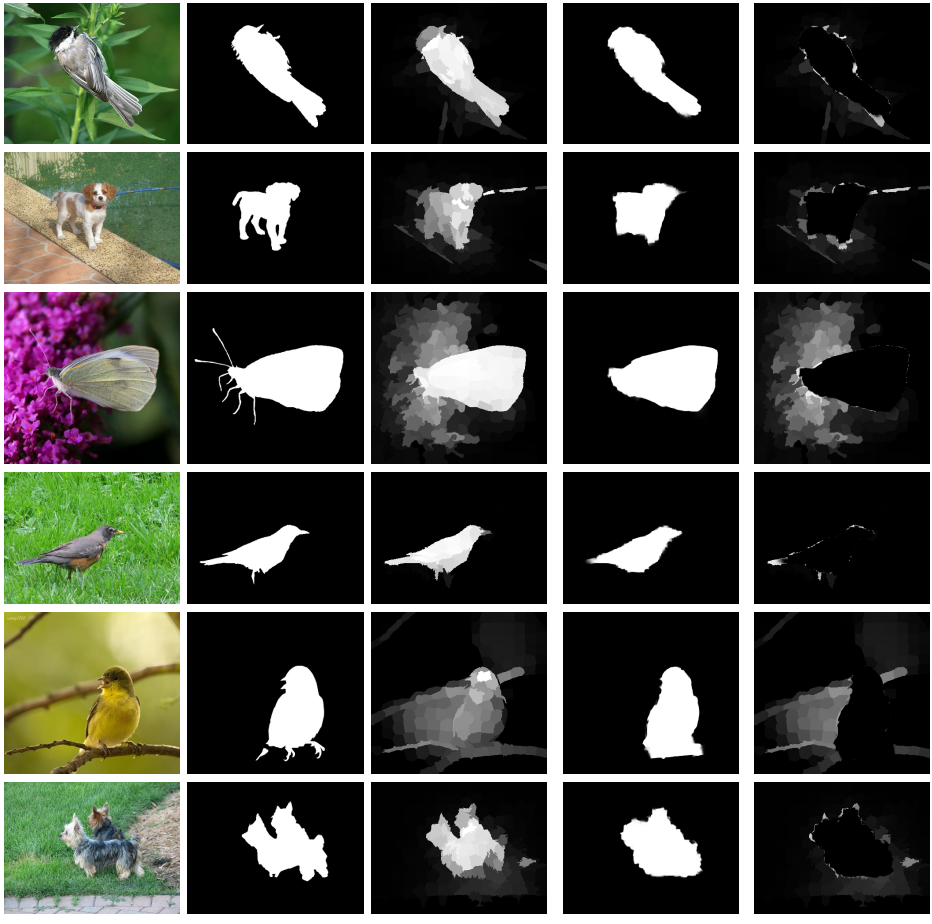

Fig. 3. Visualization of noisy label decomposition.

3 Evaluation on All Testing Datasets

In the paper, we report the 256-d F-measure and E-measure on four testing datasets due to page limitations. We show performance of our method and competing methods on all the five testing datasets in Fig. 4. On each dataset, the proposed method achieves consistently the best performance compared with the compared weakly supervised/unsupervised methods.

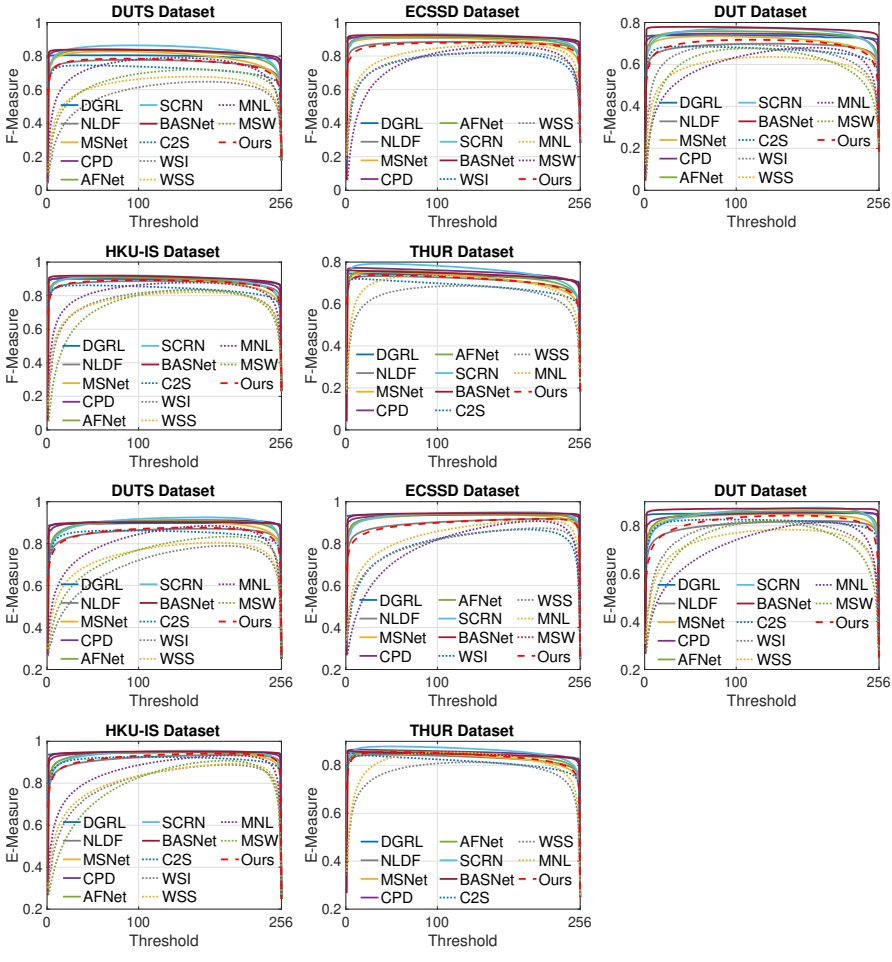

Fig. 4. F-measure and E-measure curves on five benchmark datasets (DUTS, ECSSD, DUT, HKU-IS, THUR). Best viewed on screen.

Table 1. Performance of ablation study models.

	Metric	DUTS	ECSSD	DUT	HKU-IS	SOD	PASCAL-S	THUR
GS	$S_\alpha \uparrow$	.6192	.6566	.6372	.6903	.6088	.6031	.6356
	$F_\beta \uparrow$	.4139	.5071	.4373	.5342	.4675	.5115	.4274
	$E_\xi \uparrow$	.6228	.6221	.6333	.6600	.5954	.5738	.6338
	$\mathcal{M} \downarrow$	.1835	.2078	.1751	.1690	.2319	.2634	.1755
*GS**	$S_\alpha \uparrow$	.7870	.8261	.7373	.8374	.7144	.7066	.8037
	$F_\beta \uparrow$	.7400	.8360	.6520	.8429	.6998	.7343	.7233
	$E_\xi \uparrow$	.8114	.8434	.7525	.8654	.7230	.7045	.8400
	$\mathcal{M} \downarrow$	.0711	.0873	.0832	.0620	.1414	.1729	.0713

4 More Ablation Study Results

We show more ablation studies in this section to further illustrate effectiveness of the proposed method, and performance is shown in Table 1.

1) Train the encoder-decoder network with other noisy label set. We report our results by using three sets of noisy labels. In the ablation study section, we trained the encoder-decoder model with a single noise label (noisy labels from RBD [8]) as shown in "Noi_ED" and "MR". To test how the encoder-decoder performs on the other noisy label set ("GS" [5]), we show performance by using a noisy label from "GS", performance is shown in Table 1 "GS", representing training the encoder-decoder by using noisy labels generated by GS. The performance of "GS" is not good enough, which is consistent with the conclusion that deep neural networks are not robust to noise [7].

2) Train the proposed noise-aware encoder-decoder network with other noisy label set. We report our performance in using single noisy labels in the ablation study section as "Sin_G" and "MR*", representing training the proposed model with noisy label generated by conventional handcrafted feature based method RBD and MR respectively. We then trained with noise label set from "GS" [5], and performance is shown in Table 1 "GS*", representing training the noise-aware encoder-decoder by using noisy labels from GS. Compared "GS"' with "GS*", we can observe significant performance improvement, which further prove effectiveness of our solution.

5 More Model Analysis Results

We reported in the paper that the proposed solution can be treated as boosting technique for existing fully supervised saliency detection methods. We further show our performance of boosting other three state-of-the-art fully supervised models in Table 2, where bold numbers indicate better performance compared with performance of the raw methods.

1) Boost performance of other fully supervised models. We claim in the paper that our method can also be treated as a boosting technique, which can boost performance of those fully supervised models (raw methods). We reported results on BASNet [4], and we show the result of boosting other three fully supervised

Table 2. Performance of model analysis models.

	Metric	DUTS	ECSSD	DUT	HKU-IS	SOD	PASCAL-S	THUR
*NLDF**	$S_\alpha \uparrow$	**.8353**	**.8766**	**.7931**	**.8913**	.7742	**.7782**	**.8164**
	$F_\beta \uparrow$	**.7902**	**.8807**	**.7095**	**.8884**	.7635	**.8150**	**.7329**
	$E_\xi \uparrow$	**.8727**	**.9047**	**.8244**	**.9292**	.7960	**.7974**	**.8465**
	$\mathcal{M} \downarrow$	**.0540**	**.0605**	**.0660**	**.0400**	.1108	**.1322**	**.0701**
*AFNet**	$S_\alpha \uparrow$	.8603	.8914	.8215	.9010	**.7878**	.7937	**.8289**
	$F_\beta \uparrow$	**.8168**	.8944	.7394	**.8960**	**.7782**	**.8300**	**.7490**
	$E_\xi \uparrow$	**.8966**	.9171	.8322	**.9354**	**.8097**	.8218	**.8557**
	$\mathcal{M} \downarrow$	**.0456**	.0492	.0601	**.0336**	**.0966**	**.1151**	**.0654**
*PiCANet**	$S_\alpha \uparrow$	.8377	.8901	.7967	**.8953**	.7951	**.7983**	.8131
	$F_\beta \uparrow$	**.7793**	**.8866**	.7068	**.8845**	**.7801**	**.8300**	**.7206**
	$E_\xi \uparrow$	**.8784**	**.9218**	**.8315**	**.9349**	**.8282**	**.8260**	**.8399**
	$\mathcal{M} \downarrow$	**.0534**	**.0509**	**.0671**	**.0385**	.0976	**.1162**	**.0747**

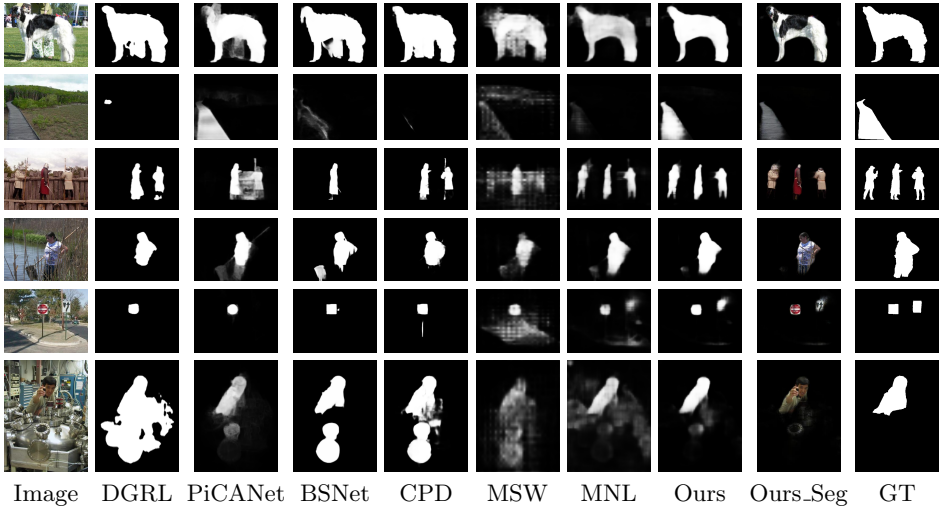

Fig. 5. Comparisons of saliency maps, where x-axis represents image, it's predicted saliency map by different competing methods (DGRL, PiCANet, BASNet, CPD, MSW, MNL) and ours, the segmented foreground and ground truth saliency map.

methods, including NLDF [3], AFNet [1], PiCANet [2], and performance is shown in Table 1 "NLDF*", "AFNet*" and "PiCANet*" respectively. We compared the boosted performance with their original performance, and we can conclude that our method can help those deep models achieve better performance or at least comparable, which further indicates the effectiveness of the proposed method.

6 More Visual Comparison

We show more visual comparison between our method and competing methods in Fig. 5 to further illustrate superior performance of our method.

7 Learning Saliency from Noisy Labels via VAE

We introduced an alternative generative model (VAE in particular) to learn saliency from a noisy label, which uses an encoder (including four cascade convolutional layers) to approximate the posterior distribution $p_\phi(Z|Y,X)$. The encoder maps the input image X and noisy label Y to latent noise vector Z. As we illustrated in the main paper (ablation study section), our final loss function includes a reconstruction loss $\|Y_i - f(X_i, Z_i, \theta)\|^2$, a KL-divergence loss $\mathrm{KL}(p_\phi(Z|Y,X)\|p_\theta(Z|Y,X))$ and a edge-aware smoothness loss. Both the reconstruction loss and edge-aware smoothness loss is well-explained in the paper, the KL-divergence loss can be learned as:

- Given image X and noisy label Y, the encoder maps them to a $2d$ (d is the size of latent space) vector, with the first d elements represent the mean vector μ, and last d elements form the standard deviation σ.
- We sample from $\mathcal{N}(\mu, \sigma^2)$ to obtain noise vector $Z = \sigma * \epsilon + \mu$, where $\epsilon \in \mathcal{N}(0,1)$ is standard Gaussian white noise.
- Then the KL-divergence loss $\mathrm{KL}(p_\phi(Z|Y,X)\|p_\theta(Z|Y,X))$ is defined as: $\mathrm{KL}(p_\phi(Z|Y,X)\|p_\theta(Z|Y,X)) = -0.5 * \sum(1 + \log(\sigma^2) - \mu^2 - \sigma^2)$
- The final objective to train the VAE based model is defined as: $\|Y_i - f(X_i, Z_i, \theta)\|^2 + \mathrm{KL}(p_\phi(Z|Y,X)\|p_\theta(Z|Y,X)) + l_s$, where l_s is the edge-aware smoothness loss.

8 Limitation of Our Solution

We also investigate the limitations of our solution in this section, including scales of salient objects and extreme noise level.

8.1 Scale of Salient Objects

We show some failed samples in Fig. 6, and compute their foreground ratio. We find that the existing deep models (including ours) can predict well on images with medium size (around 25% of the image) salient objects, and perform worse on those images with very big or small salient objects. The main reason comes from the effective receptive field. Although much work has been done to enlarge the theoretical receptive field, the scale of the salient object still remains an open problem for the whole salient object detection community. We will investigate scale-robustness in the future work.

8.2 Extreme Noise Level

To test how our model performs with extreme noise levels, we trained our noise-aware encoder-decoder with random noise (Y is random noise in this scenario), and obtain really bad results. The main reason is that the large amount of noise in the training set dominates the network to result in rather noisy prediction. Although the edge-aware smoothness loss intend to guide to latent saliency map S to share similar structure as input image X, the log-likelihood objective $\mathcal{L}(\theta)$ pushes S to be similar to random noise.

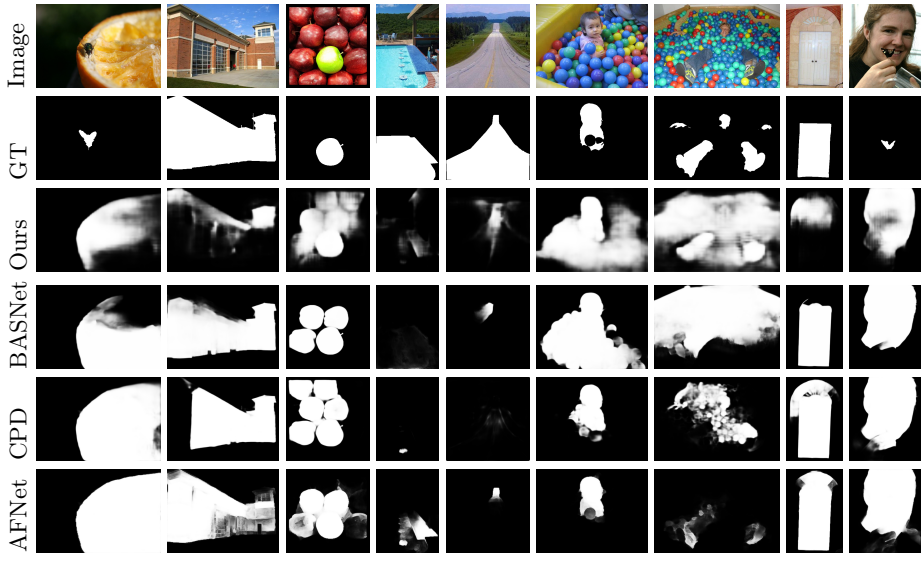

Fig. 6. Failed samples of our method compared with compething methods of BASNet [4], CPD [6] and AFNet [1].

References

1. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
2. Liu, N., Han, J., Yang, M.H.: Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2018)
3. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 6609–6617 (2017)
4. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
5. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: Proc. Eur. Conf. Comp. Vis. pp. 29–42 (2012)
6. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. (2019)
7. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning requires rethinking generalization. In: Proc. Int. Conf. Learning Representations (2017)
8. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: Proc. IEEE Conf. Comp. Vis. Patt. Recogn. pp. 2814–2821 (2014)