Supplementary Materials for Disentangled High Quality Salient Object Detection

Lv Tang Bo Li[1*] Yijie Zhong Shouhong Ding[1] Mofei Song[2,3]
[1]Youtu Lab, Tencent, Shanghai, China
[2]The School of Computer Science and Engineering,
[3]The Key Lab of Computer Network and Information Integration (Ministry of Education),
Southeast University, Nanjing, China
luckybird1994@gmail.com, libraboli@tencent.com, dun.haski@gmail.com,
ericshding@tencent.com, songmf@seu.edu.cn

1. Introduction

This supplemental material contains six parts:

- Section 2 gives more quantitative and qualitative experimental results to demonstrate the superiority of our novel disentangled framework.
- Section 3 provides more comprehensive analyses of the proposed disentangled framework to demonstrate its effectiveness.
- Section 4 shows some examples which have low quality annotations.
- Section 5 gives more details about the extra saliency supervision $L_{saliency}$ we used in LRSCN training.
- Section 6 gives more details about MECF module.
- Section 7 gives the formulas of two boundary evaluation metrics.

We hope this supplemental material can help you get a better understanding of our work.

2. More Quantitative and Qualitative Results

2.1. Quantitative Comparison on more datasets

We compare our method with other SOTA methods on another two conventional low-resolution datasets ECSSD [9] and PASCAL-S [6], which have 1000 and 850 images respectively. The results are reported in Table.1. It can be seen that our method consistently outperforms other methods across these two conventional datasets. We also show their PR curves in Fig.1. It should be noted that F_{max} represents F_{β}^{max}. We apologize for this writing error of Table.2 in the main text.

F-measure curves of different methods are displayed in Fig.2, for overall comparisons. One can observe that our approach noticeably outperforms all the other state-of-the-art methods. These observations demonstrate the efficiency and robustness of our proposed method across various challenging datasets.

SOC [1] is a new challenging dataset with nine attributes. In Table.2, we evaluate the mean F-measure score of our method as well as 11 state-of-the-art methods. We can see the proposed model achieves the competitive results among most of attributes and the overall score is best.

Model size and running time comparisons among different methods are also reported in Table.3. It can be seen that with the high-resolution input, our method is more efficient than HRNet. For fair, the running time analysis of our method is also conducted with the low-resolution input (352×352), and our method runs at a competitive efficiency.

2.2. Quantitative Comparison with different settings

Although the effectiveness of our method has been confirmed by existing quantitative comparison experiments, to further illustrate the superiority of our method in handling high-resolution SOD task, we modify the setting of existing methods to allow for a more comprehensive comparison.

First, we change the input for the current SOTA methods from low-resolution (e.g., typical size 320×320, 352×352) to high-resolution (1024×1024). The results are reported in Table.4. It can be found that all the compared SOTA methods perform better at low-resolution on most evaluation metrics. Therefore, we only compare our methods to

[*]Corresponding author and equal contribution to first author. This work was supported by National Natural Science Foundation of China 61906036 and the Fundamental Research Funds for the Central Universities (2242021k30056).

Table 1. Quantitative comparison with SOTA methods on another two conventional datasets.

Models	Training datasets	ECSSD				PASCAL-S			
		F_β^{max}	F_β	S_m	MAE	F_β^{max}	F_β	S_m	MAE
VGG-16 backbone									
Amulet(ICCV2017)	MK	0.915	0.868	0.894	0.059	0.828	0.757	0.818	0.100
DGRL(CVPR2018)	DUTS	0.922	0.903	0.906	0.043	0.849	0.807	0.834	0.074
DSS(TPAMI2019)	MB	0.921	0.904	0.882	0.052	0.831	0.802	0.798	0.094
CPD(CVPR2019)	DUTS	0.936	0.917	0.917	0.037	0.861	0.824	0.842	0.072
EGNET(ICCV2019)	DUTS	0.943	0.913	0.913	0.041	0.858	0.809	0.848	0.077
MINet(CVPR2020)	DUTS	0.943	0.922	0.917	0.036	0.865	0.829	0.854	0.064
ITSD(CVPR2020)	DUTS	0.939	0.875	0.914	0.040	0.869	0.773	0.853	0.068
GateNet(ECCV2020)	DUTS	0.941	0.896	0.917	0.041	0.870	0.797	0.853	0.068
HRNet(ICCV2019)	DUTS+HR	0.925	0.905	0.888	0.052	0.846	0.804	0.817	0.079
Ours	DUTS	0.948	0.931	0.918	0.034	0.874	0.845	0.854	0.063
Ours-DH	DUTS+HR-L	0.938	0.918	0.904	0.040	0.871	0.845	0.851	0.061
ResNet-50/ResNet-101/ResNeXt-101/Res2Net50 backbone									
R3Net(IJCAI2018)	MK	0.934	0.883	0.910	0.051	0.834	0.775	0.809	0.101
BasNet(CVPR2019)	DUTS	0.942	0.880	0.916	0.037	0.854	0.775	0.832	0.076
PFPN(AAAI2020)	DUTS	0.947	0.917	0.927	0.035	0.870	0.824	0.851	0.065
GCPA(AAAI2020)	DUTS	0.948	0.919	0.927	0.035	0.869	0.827	0.860	0.062
F3N(AAAI2020)	DUTS	0.945	0.925	0.924	0.036	0.872	0.840	0.855	0.062
LDF(CVPR2020)	DUTS	0.950	0.930	0.924	0.034	0.874	0.843	0.859	0.061
CSF(ECCV2020)	DUTS	0.950	0.925	0.927	0.033	0.874	0.823	0.858	0.069
Ours	DUTS	0.952	0.941	0.928	0.029	0.880	0.852	0.861	0.059
Ours-DH	DUTS+HR-L	0.953	0.941	0.926	0.030	0.878	0.852	0.859	0.060

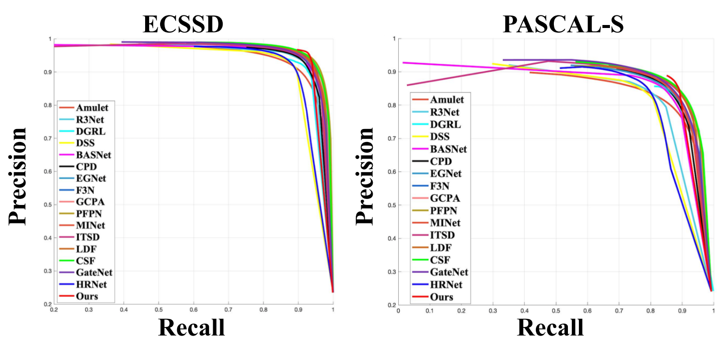

Figure 1. Comparison of PR curves across another two conventional low-resolution datasets.

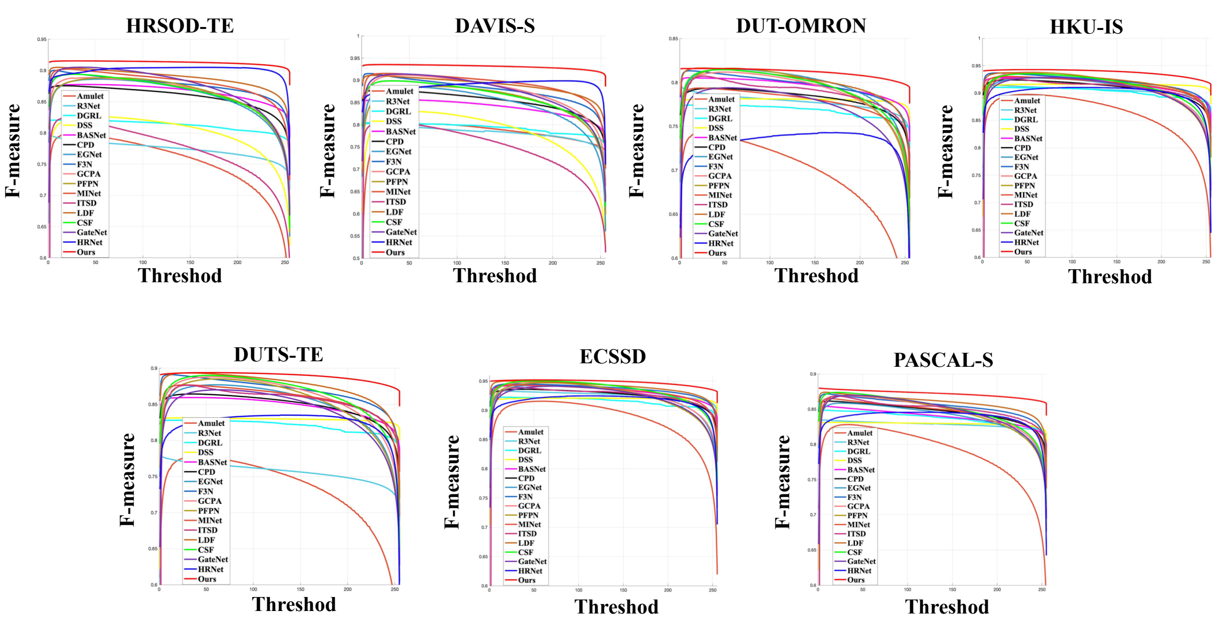

Figure 2. Comparison of the F-measure curves across on two high-resolution and five low-resolution datasets.

Table 2. Performance on SOC of different attributes. The last row shows the whole performance on the SOC dataset.

Attr	BASNet	CPD	EGNet	F3N	GCPA	PFPN	ITSD	LDF	MINet	CSF	GateNet	Ours	Ours-DH
AC	0.723	0.750	0.756	0.784	0.780	0.772	0.611	0.796	0.790	0.730	0.748	0.793	0.788
BO	0.511	0.794	0.702	0.791	0.882	0.837	0.499	0.807	0.814	0.825	0.737	0.858	0.848
CL	0.682	0.771	0.726	0.757	0.765	0.765	0.610	0.763	0.770	0.751	0.754	0.789	0.789
HO	0.772	0.777	0.756	0.790	0.780	0.777	0.685	0.797	0.792	0.779	0.788	0.817	0.817
MB	0.687	0.715	0.687	0.761	0.691	0.705	0.589	0.758	0.708	0.702	0.725	0.764	0.768
OC	0.686	0.719	0.702	0.724	0.720	0.729	0.629	0.739	0.729	0.703	0.728	0.771	0.771
OV	0.720	0.764	0.764	0.793	0.802	0.806	0.639	0.805	0.788	0.772	0.787	0.798	0.802
SC	0.708	0.723	0.683	0.747	0.707	0.697	0.592	0.746	0.726	0.690	0.715	0.785	0.782
SO	0.632	0.643	0.614	0.668	0.640	0.636	0.523	0.691	0.652	0.621	0.641	0.713	0.713
Avg	0.680	0.740	0.710	0.757	0.752	0.747	0.597	0.767	0.753	0.730	0.736	0.788	0.787

Table 3. Model size and running time comparisons between our approach and SOTA methods.

	Ours	Ours	DGRL	DSS	BASNet	EGNet	GCPA	PFPN	R3Net
Model Size(MB)	309.6	309.6	648	447.3	412.2	332.1	255.8	243.0	214.2
Time(s)	0.21	0.05	0.52	5.12	0.04	0.15	0.02	0.05	0.27
Size	1024×1024	352×352	384×384	224×224	256×256	400×300	320×320	256×256	256×256
	HRNet	MINet	CSF	Amulet	CPD	F3N	LDF	ITSD	GateNet
Model Size(MB)	129.6	181.4	139.3	132.6	111.5	97.4	95.9	63.7	-
Time(s)	0.39	0.01	0.01	0.05	0.02	0.03	0.02	0.02	0.03
Size	1024×1024	320×320	224×224	256×256	352×352	352×352	352×352	288×288	384×384

Table 4. Quantitative comparison with SOTA methods where the inputs are resized to high-resolution.

Models	HRSOD-TE						DAVIS-S					
	F_β^{max}	F_β	S_m	MAE	BDE	B_μ	F_β^{max}	F_β	S_m	MAE	BDE	B_μ
CPD(High-Resolution)	0.868	0.735	0.809	0.073	181.770	0.819	0.720	0.679	0.799	0.062	126.281	0.748
CPD(Low-Resolution)	0.876	0.829	0.887	0.039	72.686	0.824	0.878	0.822	0.903	0.025	36.649	0.703
EGNet(High-Resolution)	0.745	0.693	0.791	0.082	213.333	0.867	0.692	0.644	0.801	0.069	149.537	0.821
EGNet(Low-Resolution)	0.883	0.814	0.888	0.044	73.500	0.896	0.886	0.794	0.897	0.030	37.369	0.799
F3N(High-Resolution)	0.834	0.757	0.825	0.066	187.942	0.798	0.698	0.712	0.826	0.054	130.603	0.716
F3N(Low-Resolution)	0.900	0.853	0.897	0.035	65.901	0.817	0.915	0.845	0.913	0.020	45.106	0.719
GCPA(High-Resolution)	0.810	0.771	0.830	0.066	164.142	0.793	0.750	0.714	0.829	0.057	122.068	0.708
GCPA(Low-Resolution)	0.889	0.827	0.894	0.039	70.320	0.873	0.912	0.833	0.924	0.021	24.132	0.759
MINet(High-Resolution)	0.687	0.629	0.742	0.111	250.149	0.913	0.580	0.508	0.681	0.129	176.671	0.888
MINet(Low-Resolution)	0.902	0.851	0.903	0.032	76.291	0.849	0.915	0.864	0.926	0.019	32.304	0.742
LDF(High-Resolution)	0.650	0.586	0.673	0.133	208.545	0.898	0.590	0.553	0.696	0.101	150.540	0.844
LDF(Low-Resolution)	0.905	0.866	0.905	0.032	58.655	0.812	0.911	0.864	0.922	0.019	35.496	0.713
CSF(High-Resolution)	0.802	0.756	0.843	0.063	181.705	0.873	0.700	0.685	0.824	0.058	137.592	0.816
CSF(Low-Resolution)	0.894	0.832	0.900	0.038	71.293	0.922	0.899	0.822	0.912	0.025	30.488	0.848
Ours	0.918	0.902	0.912	0.027	48.468	0.711	0.933	0.919	0.933	0.015	15.676	0.536

Table 5. Quantitative comparison with SOTA methods which are finetuned on HRSOD-Training dataset.

Models	HRSOD-TE						DAVIS-S					
	F_β^{max}	F_β	S_m	MAE	BDE	B_μ	F_β^{max}	F_β	S_m	MAE	BDE	B_μ
BASNet(finetune)	0.885	0.836	0.904	0.035	64.475	0.813	0.866	0.838	0.911	0.023	25.924	0.659
BASNet(original)	0.878	0.831	0.890	0.038	67.643	0.823	0.857	0.806	0.881	0.039	46.283	0.705
CPD(finetune)	0.890	0.846	0.899	0.035	80.857	0.783	0.890	0.871	0.925	0.020	29.376	0.671
CPD(original)	0.876	0.829	0.887	0.039	72.686	0.824	0.878	0.822	0.903	0.025	36.649	0.703
EGNet(finetune)	0.890	0.857	0.911	0.031	69.084	0.797	0.899	0.881	0.926	0.021	30.674	0.686
EGNet(original)	0.883	0.814	0.888	0.044	73.500	0.896	0.886	0.794	0.897	0.030	37.369	0.799
GCPA(finetune)	0.895	0.837	0.912	0.032	64.656	0.846	0.918	0.857	0.927	0.019	22.312	0.746
GCPA(original)	0.889	0.827	0.894	0.039	70.320	0.873	0.912	0.833	0.924	0.021	24.132	0.759
F3N(finetune)	0.905	0.865	0.909	0.033	60.803	0.787	0.920	0.860	0.921	0.019	29.106	0.661
F3N(original)	0.900	0.853	0.897	0.035	65.901	0.817	0.915	0.845	0.913	0.020	45.106	0.719
PFPN(finetune)	0.896	0.840	0.904	0.038	55.027	0.786	0.901	0.845	0.920	0.022	21.388	0.728
PFPN(original)	0.889	0.825	0.897	0.042	65.048	0.897	0.886	0.822	0.912	0.025	30.488	0.848
ITSD(finetune)	0.834	0.774	0.863	0.052	117.554	0.906	0.820	0.754	0.873	0.041	75.461	0.830
ITSD(original)	0.824	0.715	0.834	0.071	139.943	0.924	0.806	0.687	0.843	0.055	92.864	0.861
MINet(finetune)	0.908	0.871	0.908	0.029	66.089	0.749	0.923	0.879	0.928	0.017	25.408	0.692
MINet(original)	0.902	0.851	0.903	0.032	76.291	0.849	0.915	0.864	0.926	0.019	32.304	0.742
LDF(finetune)	0.910	0.862	0.910	0.031	77.098	0.812	0.920	0.867	0.922	0.018	42.226	0.727
LDF(original)	0.905	0.866	0.905	0.032	58.655	0.812	0.911	0.864	0.922	0.019	35.496	0.713
GateNet(finetune)	0.910	0.856	0.909	0.029	76.434	0.821	0.923	0.872	0.930	0.019	36.984	0.706
GateNet(original)	0.905	0.825	0.906	0.035	79.468	0.886	0.914	0.825	0.923	0.023	44.827	0.778
CSF(finetune)	0.902	0.859	0.909	0.029	56.425	0.884	0.910	0.870	0.931	0.017	24.669	0.791
CSF(original)	0.894	0.832	0.900	0.038	71.293	0.922	0.899	0.822	0.912	0.025	30.488	0.848
Ours	0.918	0.902	0.912	0.027	48.468	0.711	0.933	0.919	0.933	0.015	15.676	0.536

these SOTA methods' low-resolution results in our main paper. In particular, it is worth pointing out that due to GPU memory limitations, we cannot run BASNet, PFPN and ITSD at high-resolution. So we don't report their results in Table.4.

Then, we fine-tune 11 SOTA methods on high-resolution datasets (HRSOD-Training) which have high quality annotations, the results are reported on Table.5. As can be seen, high annotation quality can improve their original performance. However, even fine-tuned on HRSOD-Training datasets, our method (only trained on DUTS) still outperforms all of them by a large margin.

2.3. Qualitative Comparison

As shown in Fig.3, we provide a comprehensive qualitative comparison of our method with other 12 methods on challenging cases. These visual examples can further demonstrate that our method is able to restore accurate and complete boundaries of salient objects.

3. More analyses of the proposed disentangled framework

As described, high-resolution salient object detection task should be disentangled into two tasks. One can be viewed as a classic classification task, while the other one is a typical regression task. To further illustrate the validity of our theory, we conduct additional experiments. Specifically, we consider these two tasks as regression or classification tasks simultaneously. The results are reported in Table.6. Compared with our proposed method, if we take the disentangled framework as the combination of the two regression or classification tasks, the performance will be degraded. Because the purpose of the proposed disentangled framework is to capture sufficient semantics at low-resolution (LRSCN Stage) and refine accurate boundary at high-resolution (HRRN Stage), which should be viewed as a classic classification task and a typical regression task. Fig.4 shows some examples that our proposed HRRN can further refine accurate boundary, guided by trimaps. Specifically, column.3 and column.4 show the saliency maps and trimaps generated by LRSCN, and column.5 shows the results refined by HRRN. From Fig.4, guided by trimaps, our proposed HRRN can further refine the pixels value in uncertain regions to get more clear saliency results.

Aforementioned work LDF [12] has also introduced concepts related to decoupling. However, they still try to address the SOD task under a single regression framework. Their approach is essentially an expansion of additional boundary supervision, which barely touches the very nature of the SOD. As illustrated in our experiments, it is more natural to disentangle the SOD into two different tasks.

4. Annotation Problems

As described in [13], widely used saliency datasets have some problems in annotation quality. So, to quantify the annotation quality problem, we randomly select 100 images from DUT-TR, and 10 of them have easily spotted annotation errors. We manually relabel the 10 images. The B_μ between the two different annotations is 0.49 and 42% of the boundary pixel annotations are inaccurate. Fig.5 shows some examples which have annotation problems, including wrong semantic annotation (row 1 and row 2), boundary annotation shifting (row 3) and low contour accuracy (row4, row5 and row 6). In conclusion, the DUTS-TR training dataset does have annotation problems [13], and we relabeled some examples to demonstrate these problems in the supplemental material. Since correcting annotations for the whole DUT-TR is a time-consuming task, we will provide an accurate GT of DUT-TR in the future for statistical analysis

5. Details of $L_{saliency}$

As described, to guarantee the arruracy of trimap, we add extra saliency supervision $L_{saliency}$ as the supplement of trimap supervision. Here we give more details about $L_{saliency}$.

After LRSCN, the prediction saliency map is S, and the binary groundtruth is G. In SOD, binary cross entropy (BCE) is the most widely used loss function, and it is a pixel-wise loss which is defined as:

$$L_{Pixel} = -(Glog(S) + (1-G)log(1-S)). \quad (1)$$

To learn the structural information of the salient objects, following the setting of [10, 2], we use the sliding window fashion to model region-level similarity between groundtruth and saliency map. The corresponding regions are denoted as $S_i = \{S_i : i = 1, ...M\}$ and $G_i = \{G_i : i = 1, ...M\}$, where M is the total number of region. Then we use SSIM to evaluate the similarity between S_i and G_i, which is defined as:

$$SSD_i = \frac{(2\mu_s\mu_g + C_1)(2\sigma_{sg} + C_2)}{(\mu_s^2 + \mu_g^2 + C_1)(\sigma_s^2 + \sigma_g^2 + C_2)} \quad (2)$$

where local statistics μ_s, σ_s is mean and std vector of S_i, μ_g, σ_g is mean and std vector of G_i. The overall loss function is defined as:

$$L_{Region} = 1 - \frac{1}{M}\sum_{i=1}^{M} SSD_i. \quad (3)$$

Finally, inspired by [15], we directly optimize the F-measure to learn the global information from groundtruth.

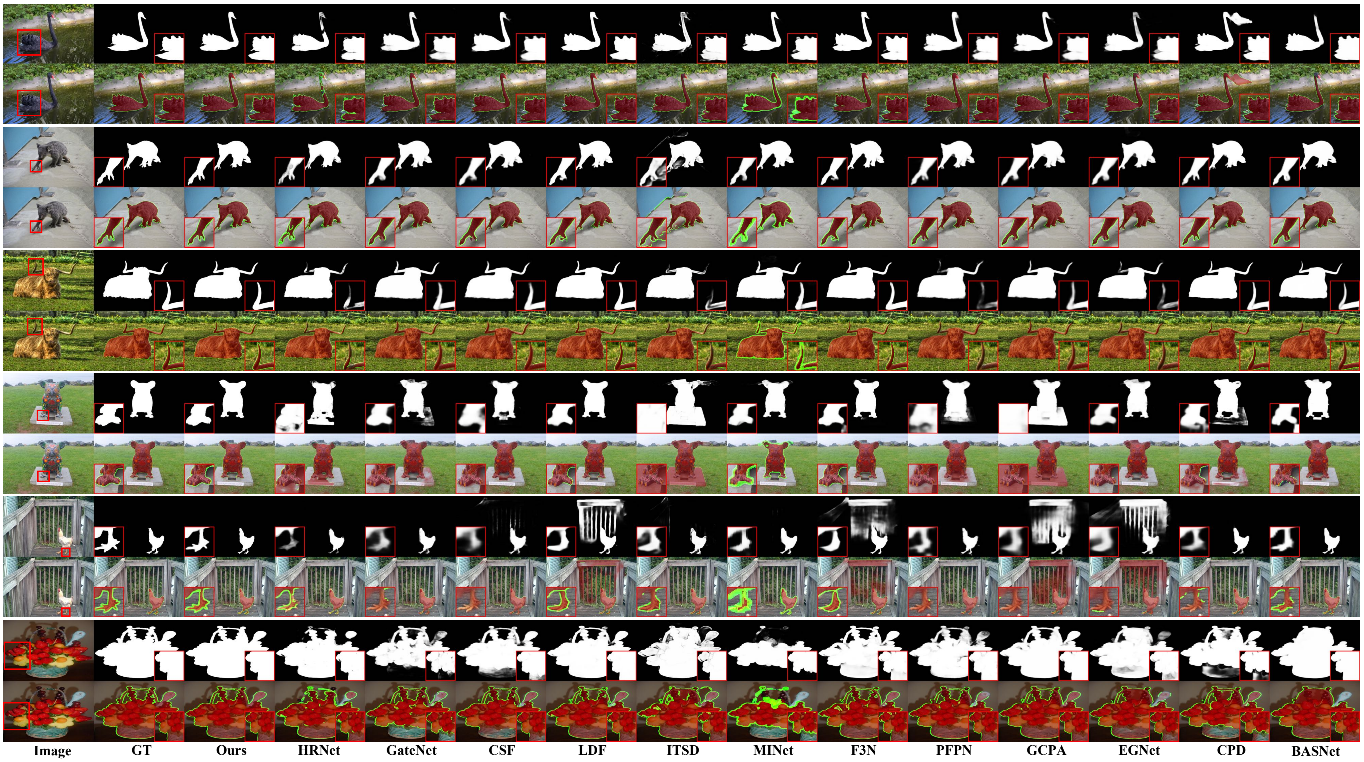

Figure 3. Visual comparison between our method and other SOTA methods. Each sample occupies two rows. Best viewed by zooming in. It can be clearly observed that our method achieves impressive performance in all these cases.

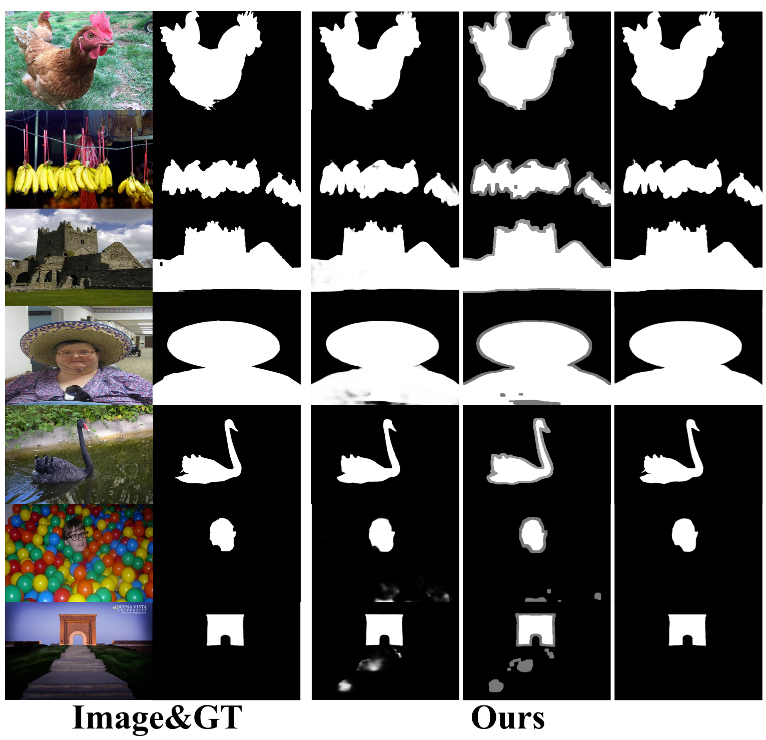

Figure 4. Examples of coarse saliency maps, trimaps and refined saliency map.

Table 6. Ablation Studies of disentangled framework.

Configurations	HRSOD-TE						DAIVS-S					
	F_β^{max}	F_β	S_m	MAE	BDE	B_μ	F_β^{max}	F_β	S_m	MAE	BDE	B_μ
Regression-Regression	0.912	0.894	0.899	0.031	56.251	0.814	0.923	0.909	0.918	0.019	22.737	0.649
Classification-Classification	0.913	0.895	0.898	0.030	54.143	0.809	0.921	0.907	0.921	0.020	23.892	0.662
Ours	0.918	0.902	0.912	0.027	48.468	0.711	0.933	0.919	0.933	0.015	15.676	0.536

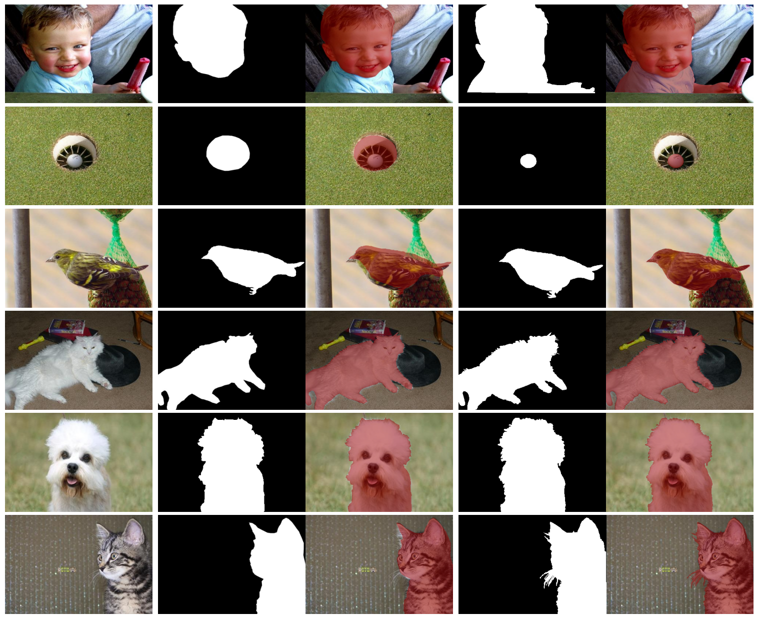

Image Original Annotation Relabeled Annotation

Figure 5. Examples that have annotation quality problem. Best viewed by zooming in.

Table 7. Ablation Studies of $L_{saliency}$.

Configurations	HRSOD-TE						DAIVS-S					
	F_β^{max}	F_β	S_m	MAE	BDE	B_μ	F_β^{max}	F_β	S_m	MAE	BDE	B_μ
LRSCN(L_{trimap})+HRRN	0.895	0.870	0.883	0.035	75.732	0.879	0.900	0.880	0.890	0.026	41.221	0.733
LRSCN($L_P + L_{trimap}$)+HRRN	0.912	0.898	0.908	0.029	53.040	0.764	0.925	0.910	0.926	0.018	19.022	0.569
LRSCN($L_P + L_R + L_{trimap}$)+HRRN	0.917	0.900	0.910	0.029	52.048	0.743	0.932	0.914	0.930	0.017	17.688	0.552
LRSCN($L_P + L_R + L_O + L_{trimap}$)+HRRN	0.918	0.902	0.912	0.027	48.468	0.711	0.933	0.919	0.933	0.015	15.676	0.536

For easy remembering, we denote F-measure as F_β in the following. F_β is defined as:

$$precision = \frac{\sum S \cdot G}{\sum S + \epsilon}, \quad recall = \frac{\sum S \cdot G}{\sum G + \epsilon}, \tag{4}$$

$$F_\beta = \frac{(1+\beta^2) \cdot precision \cdot recall}{\beta^2 \cdot precision + recall}, \tag{5}$$

where $\cdot$ means pixel-wise multiplication, $\epsilon = 1e^{-7}$ is a regularization constant to avoid division of zero. L_{Object} loss function is defined as:

$$L_{Object} = 1 - F_\beta. \tag{6}$$

The whole loss is defined as:

$$L = L_{Object} + L_{Region} + L_{Pixel}. \tag{7}$$

Besides, following [8, 11], we used multi-levels saliency supervision to facilitate sufficient training, so the whole saliency loss is defined as:

$$L_{saliency} = \sum_{i=1}^{4} \frac{1}{2^{i-1}} L_i, \tag{8}$$

where i means the i-th level.

To further validate the role of $L_{saliency}$, we train the LRSCN with different loss functions and the results are reported on Table.7. As can be can, without $L_{saliency}$, the performance is dropped lot. Because the trimap groundtruth is randomly generated from binary groundtruth, so only using L_{trimap} cannot maintain consistency between trimap and saliency map. When we only add L_P on multi-levels, the model can already achieve the largest performance boost. A better performance has been achieved through the combination of L_P, L_R and L_O.

6. Details of MECF Module

As described, we develop a multi-scale feature extraction module (ME) and cross-level feature fusion module (CF) to

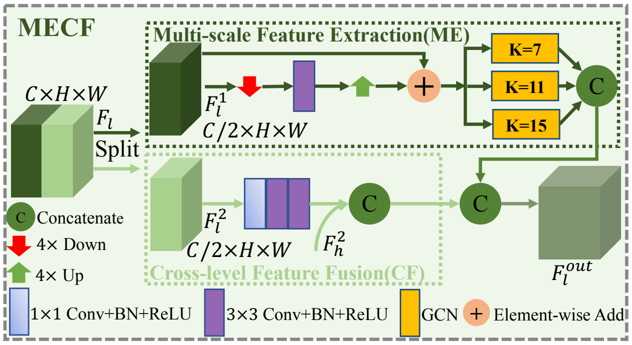

Figure 6. Architecture of MECF Module.

help LRSCN capture sufficient semantics at low-resolution. Here we give more details about MECF module. The architecture of MECF Module is shown in Fig.6.

Multi-scale feature extraction module can allow each spatial location to view the local context at small scale spaces and capture multi-scale contextual information, which can enlarge the feature F_l^1 receptive field. Specifically, we first use an average pooling and a 3×3 convolutional layer to downsample F_l^1. Then upsampled feature from small scale is added with F_l^1. Finally, Global Convolutional Network (GCN) [7] is used to further enlarge the feature receptive field. Because F_3^1 and F_4^1 are close to the input and receptive field is relatively small, we use GCNs with $k = 7, 11, 15$ to fully enlarge receptive field. Receptive fields of F_5^1 and F_6^1 are relatively bigger, we only use GCNs with $k = 7, 11$ and $k = 7$.

Low-level features have rich details but full of background noises, so we design cross-level feature fusion module, which can leverage the rich semantics of high-level feature F_h^2 and help restrain the non-salient regions in low-level features. Specifically, we first use a 1×1 convolutional layer to compress the channels of F_l^2, then use two 3×3 convolutional layer to transfer the feature for SOD task. Finally, the transferred feature is fused with high-level feature F_h^2 as the output of this module. Each of these convolution layers is followed by a batch normalization [5] and a ReLU activation [4].

7. Formulas of Evaluation Metrics

Following [13] and [14], we use Boundary Displacement Error(BDE) [3] and B_μ metrics to evaluate the boundary quality.

BDE measures the average displacement error of boundary pixels between two predictions, which can be formulated as:

$$BDE(X,Y) = \frac{\sum_x inf_{y\in Y} d(x,y)}{2N_X} + \frac{\sum_y inf_{x\in X} d(x,y)}{2N_Y}, \tag{9}$$

where X and Y are two boundary pixel sets which represent saliency prediction and their corresponding groundtruth, and x, y are pixels in them. N_x and N_y denote the number pixels in X and Y. inf represents for the infimum and $d(\cdot)$ denotes Euclidean distance.

B_μ evaluates the structure alignment between saliency map and their groundtruth, it can be expressed as:

$$B_\mu = 1 - \frac{2\sum(g_s g_y)}{\sum(g_s^2 + g_y^2)}, \tag{10}$$

where g_s and g_y represent the binarized edge maps of predicted saliency map and groundtruth. Following [14], we use Canny edge detector to compute edge maps. B_μ reflect the sharpness of predictions which is consistent with human perception. Both two evaluation codes are provided in the Github link in our main paper.

References

[1] Deng-Ping Fan, Ming-Ming Cheng, Jiangjiang Liu, Shanghua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV (15)*, volume 11219 of *Lecture Notes in Computer Science*, pages 196–212. Springer, 2018.

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

[3] Jordi Freixenet, Xavier Muñoz, David Raba, Joan Martí, and Xavier Cufí. Yet another survey on image segmentation: Region and boundary information integration. In *ECCV (3)*, volume 2352 of *Lecture Notes in Computer Science*, pages 408–422. Springer, 2002.

[4] Richard H. R. Hahnloser and H. Sebastian Seung. Permitted and forbidden sets in symmetric threshold-linear networks. In *NIPS*, pages 217–223. MIT Press, 2000.

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML(2015)*, volume 37, pages 448–456. JMLR.org, 2015.

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

[7] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters - improve semantic segmentation by global convolutional network. In *CVPR*, 2017.

[8] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jägersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489. Computer Vision Foundation / IEEE, 2019.

[9] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended CSSD. *Trans. Pattern Anal. Mach. Intell.*, 38(4):717–729, 2016.

[10] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[11] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. *CoRR*, abs/1911.11445, 2019.

[12] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13022–13031. IEEE, 2020.

[13] Yi Zeng, Pingping Zhang, Zhe L. Lin, Jianming Zhang, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7233–7242. IEEE, 2019.

[14] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, pages 12543–12552. IEEE, 2020.

[15] Kai Zhao, Shanghua Gao, Wenguan Wang, and Ming-Ming Cheng. Optimizing the f-measure for threshold-free salient object detection. In *ICCV*, pages 8848–8856. IEEE, 2019.