BBS-Net: RGB-D Salient Object Detection with a Bifurcated Backbone Strategy Network (Supplementary Material)

Anonymous ECCV submission

Paper ID 1547

Abstract. In this supplementary material, we introduce the evaluation metrics in detail, and provide more exprimental results to demonstrate the effectiveness of our model and for better comparison with state-of-the-art methods.

1 Evaluation Metric

In this section, we introduce the detailed definations of the metrics for the binary foreground map evaluation used in our experiments, including F-measure (F_β), S-measure (S_α), E-measure (E_ξ), MAE and PR curve.

F-measure [1] is a common evaluation metric based on region similarity. It is defined as:

$$F_\beta^i = \frac{(1+\beta^2)P^i \times R^i}{\beta^2 \times P^i + R^i}, \tag{1}$$

where P^i and R^i are the corresponding precision and recall for the threshold i ($i \in \{1, 2, \cdots, 255\}$), respectively. β controls the trade-off between P^i and R^i. In our experiments, we set β^2 to 0.3, as proposed by [1]. In the manuscript, we utilize the maximum F-measure (for all thresholds) to evaluate different methods. Besides, we also provide results of the adaptive F-measure (*i.e.*, the threshold is defined as the double mean value of the saliency map) and mean F-measure (*i.e.*, the average F-measure for all thresholds) in this material.

S-measure [16] is a structure measure which combines the region-aware structural similarity (S_r) and object-aware structural similarity (S_o). It is on the basis of behavioral vision studies that huaman vision pays more attention to the structures and is represented by:

$$S_\alpha = \alpha \times S_o + (1-\alpha) \times S_r, \tag{2}$$

where $\alpha \in [0, 1]$ is a hyper-parameter to balance the (S_r) and (S_o). We set it to 0.5 as the default setting.

E-measure is recently proposed by [17], which is based on the cognitive vision studies and utilizes both image-level and local pixel-level statistics for evaluating the binary saliency map. It is defined as:

$$E_\xi = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} \phi(x, y), \tag{3}$$

where w and h are the width and height of the saliency map. ϕ is the enhanced alignment matrix. Similar to the aforementioned F-measure, we also provide the results of max E-measure, adaptive E-measure and mean E-measure.

The **MAE** represents the average absolute error between the predicted saliency map and the ground truth. It is denoted as:

$$M = \frac{1}{N}|S - G|, \tag{4}$$

where S and G are the prediced saliency map and ground-truth binary map, respectively. N represents the total number of pixels.

PR curve is generated by a series of Precision-Recall (PR) pairs, which is calculated by the binarized saliency map with its thresholds varying from $0-255$. In detail, the Precision (P) and Recall (R) are calculated by:

$$P = \frac{|S' \cap G|}{|S'|}, R = \frac{|S' \cap G|}{|G|}, \tag{5}$$

where S' is the binary mask for the predicted map S according to the threshold.

2 Additional Experimental Results

Tab. 1. We show additional results on the recent proposed DUT [45] dataset.

Tab. 2. To further demonstrate the effectiveness of GCM module in the cascaded decoder, we conduct experiments without using GCM (*i.e.*, replace GCM with a $1\times$ 1 convolution).

Tab. 3. We make an analysis of different data augumentation strategies, *i.e.*, random flipping, border clipping and random rotating.

Tab. 4. We test the speed of the proposed model using different settings.

Tab. 5. We compare the runtime of different models. The timings are borrowed from original paper or provided by authors.

Tab. 6. We make quantitative comparisons of different models using 4 more metrics (*i.e.*, adaptive F-measure, adaptive E-measure, mean F-measure and mean E-measure) on 7 public datasets.

Fig. 1. We show 6 representative failure cases selected from several datasets.

Fig. 2. We give a rank of 19 state-of-the-art models using the metric of max F-measure on 7 datasets (Tab. 6).

Fig. 3. We shows the complete version of PR curves for various methods on 7 datasets.

Fig. 4. We draw the F-measure curves of our method and 18 state-of-the-art methods on 7 datasets.

Fig. 5$\sim$ **Fig. 9.** We show a large quantity of predicted saliency maps of our model and 18 state-of-the-art methods.

Table 1: Performance comparisons of different models on the recent DUT [45] dataset. The models are trained and tested on the DUT dataset using the proposed training and test sets split from [45].

Methods	LHM [44]	DESM [9]	DCMC [12]	CDCP [68]	DF [47]	CTMF [24]	MMCI [5]	PDNet [66]	PCF [3]	DMRA [45]	BBS-Net (ours)
$S_\alpha \uparrow$	.568	.659	.499	.687	.730	.834	.791	.799	.801	.888	**.912**
$F_\beta \uparrow$	.659	.668	.406	.633	.748	.792	.753	.757	.760	.883	**.904**
$E_\xi \uparrow$	.767	.733	.712	.794	.842	.884	.855	.861	.858	.927	**.942**
$M \downarrow$	.174	.280	.243	.159	.145	.097	.113	.112	.100	.048	**.038**

Table 2: Ablation study of GCM module. 'w/' and 'w/o' represent that we implement our model with and without (*i.e.*, replace GCM with a 1×1 convolution) GCM module.

Models	NJU2K [29]			NLPR [44]			STERE [42]			DES [9]			LFSD [36]			SSD [67]		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
w/o GCM	.916	.914	.037	.925	.907	.026	.901	.895	.044	.923	.914	.023	.853	.852	.075	.877	.863	.048
w/ GCM	**.921**	**.920**	**.035**	**.930**	**.918**	**.023**	**.908**	**.903**	**.041**	**.933**	**.927**	**.021**	**.864**	**.859**	**.072**	**.882**	**.859**	**.044**

Table 3: Analysis of different data augumentation strategies. We conduct experiemnts using different data augumentation strategies, *i.e.*, random flipping, border clipping and random rotating.

Strategy	NJU2K [29]			NLPR [44]			STERE [42]			DES [9]			LFSD [36]			SSD [67]		
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$M \downarrow$
None	.906	.906	.039	.919	.903	.025	.894	.887	.043	.907	.896	.026	.848	.849	.077	.860	.842	.054
Flip	.913	.914	.038	.937	.913	**.023**	.900	.893	**.041**	.909	.898	.025	.854	.852	.073	.868	.849	.048
Crop	.912	.912	.037	.922	.906	.025	.901	.896	**.041**	.924	.914	.022	.848	.846	.081	.854	.827	.056
Rotate	.916	.917	.037	.922	.906	.025	.902	.893	.042	.917	.905	.024	.859	**.859**	.073	.860	.829	.052
All	**.921**	**.920**	**.035**	**.930**	**.918**	**.023**	**.908**	**.903**	**.041**	**.933**	**.927**	**.021**	**.864**	**.859**	**.072**	**.882**	**.859**	**.044**

Table 4: Speed test of *BBS-Net*. We test the speed of only predicting the initial saliency map S_1, and the final map S_2, respectively. 'BS' denotes the batch size, of which the maximum value for a single GTX 1080Ti GPU is 10. 'io' is the time consumed by reading and writing.

#	BS:1	BS:10	w/ io	w/o io	S_1 (fps)	S_2 (fps)
1	✓		✓		17	14
2	✓			✓	19	15
3		✓	✓		56	48
4		✓		✓	133	123

Table 5: Runtime of different methods. The timings are borrowed from original paper or provided by authors.

Method	LHM[44]	CDB[37]	DESM[9]	GP[48]	CDCP[68]	LBE[21]	DCMC[12]	MDSF[51]	SE[23]
Time (s)	2.130	0.600	7.790	12.98	>60.0	3.110	1.200	1.570	>60.0
Type	CPU	CPU	CPU	CPU	CPU	CPU	CPU	CPU	CPU

Method	DF[47]	AFNet[55]	CTMF[24]	MMCI[5]	PCF[3]	TANet[4]	CPFP[65]	BBS-Net (ours)
Time (s)	10.36	0.030	0.630	0.050	0.060	0.070	0.170	0.021
Type	GPU	GPU	GPU	GPU	GPU	GPU	GPU	GPU

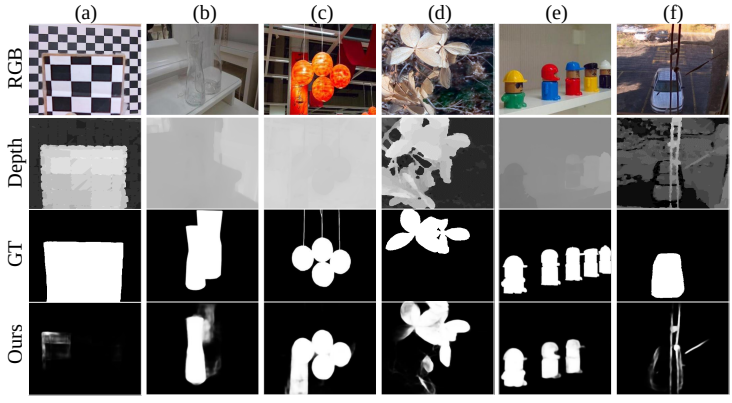

Fig. 1: Representative failure cases selected from several datasets. (a)&(b): The model may not detect the salient object or detect the salient object imperfectly. (c)&(d): The model takes the background as the salient parts. (e): The model fails to find all salient objects, especially for those objects far from the lens. (f): The model may not deal with the circumstances when the salient objects are occluded by non-salient objects.

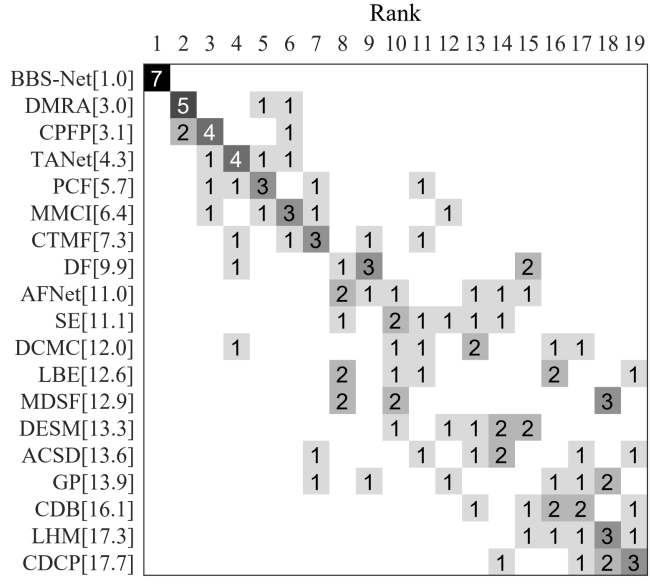

Fig. 2: Ranking 19 models in Tab. 6 with the metric of max F-measure. Element(i,j) represents the number of times that model i ranks at j^{th}. Models are ranked by the mean rank (shown in brackets) over the 7 datasets.

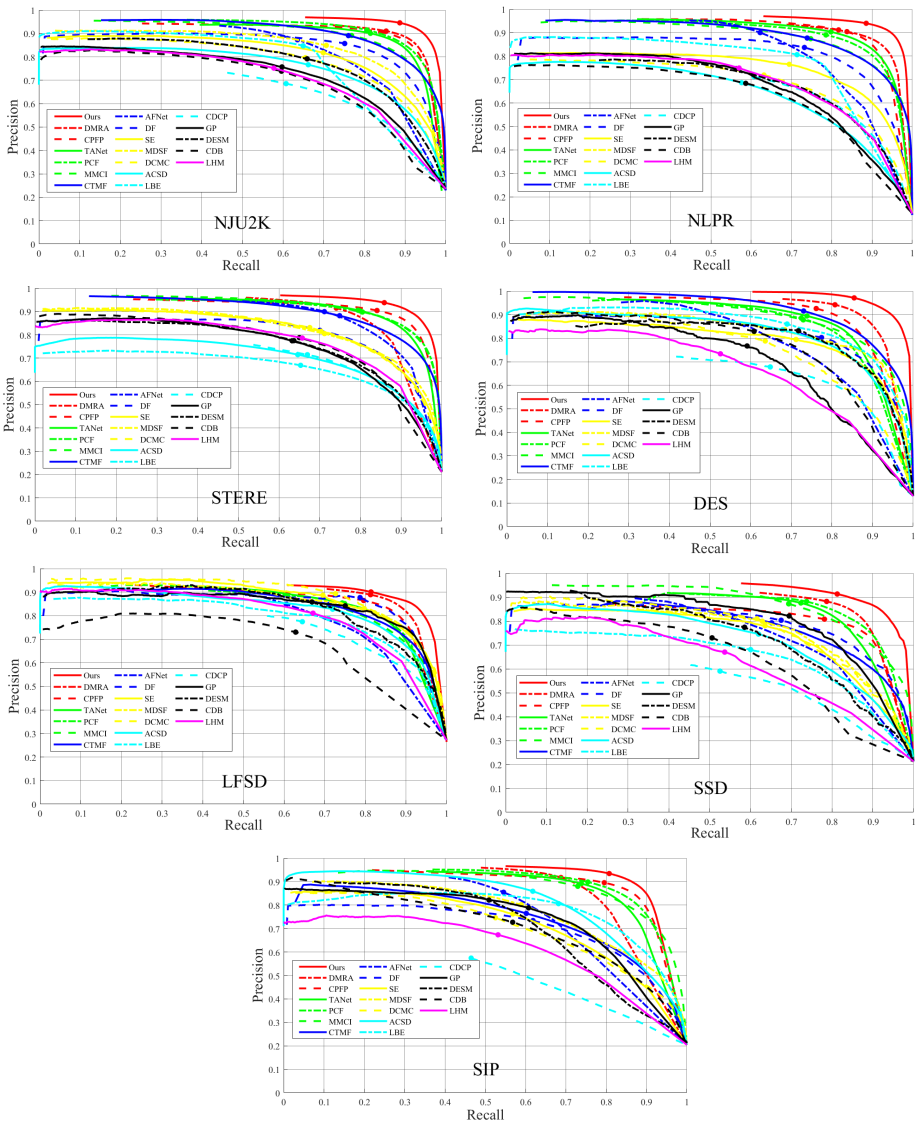

Fig. 3: The complete PR curves of our method and 18 SOTA methods on 7 datasets. Dots on curves represent the values of precision and recall at the maximum F-measure.

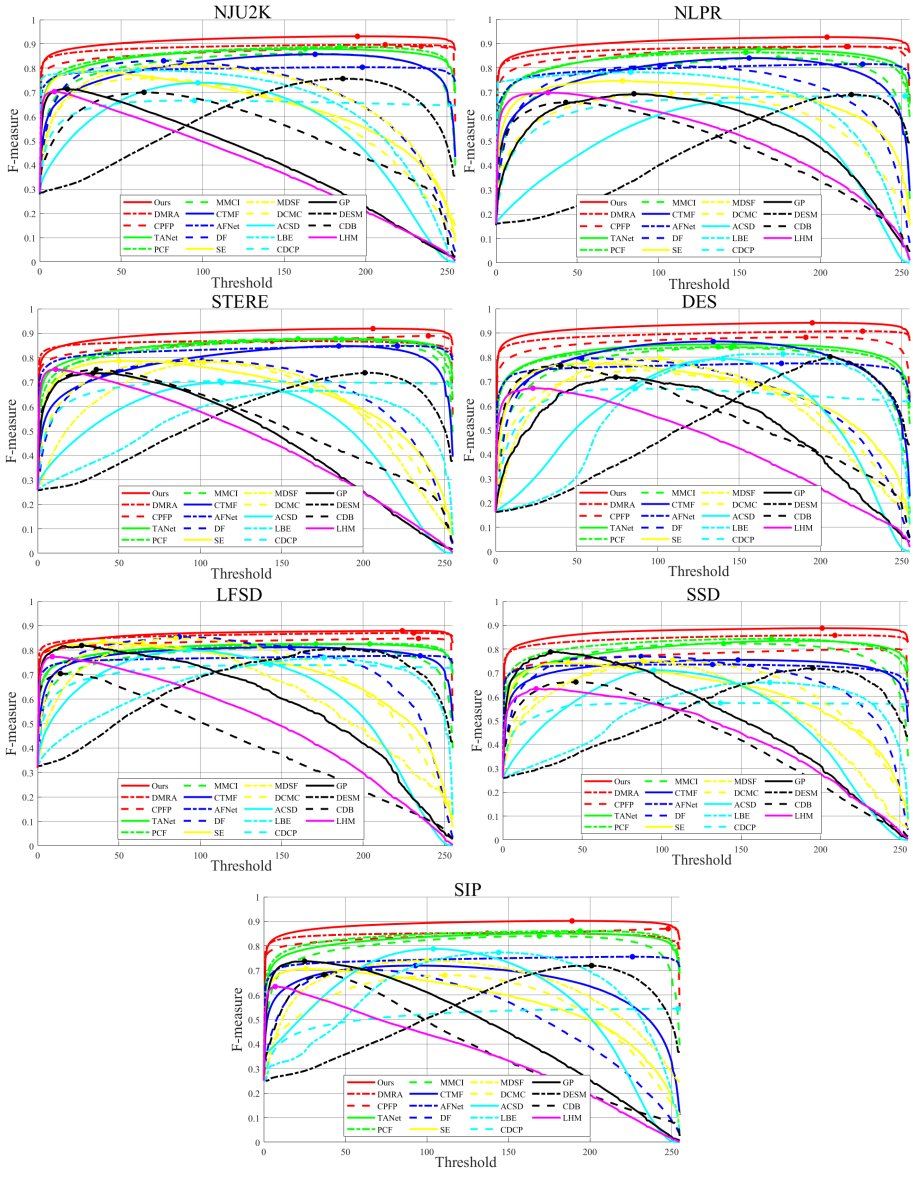

Fig. 4: The F-measure results of our method and 18 SOTA methods on 7 dataset. Dots on the curves represent the values of the maximum F-measure.

Table 6: Quantitative comparison of models using S-measure (S_α), adaptive F-measure ($adaF_\beta$), mean F-measure ($meanF_\beta$), max F-measure ($maxF_\beta$), adaptive E-measure ($adpE_\xi$), mean E-measure ($meanE_\xi$), max E-measure ($maxE_\xi$) and MAE (M) scores on 7 public datasets. ↑ (↓) denotes that the higher (the lower) the better. The best score and the best score of the compared methods in each row are highlighted in **red** and **blue**. From left to right: 10 models based on hand-crafted features and 8 CNNs-based models. Besides, 'S1' and 'S2' denote the initial and final saliency outputs results of the proposed method, respectively.

Dataset	Metric	Hand-crafted-features-Based Models										CNNs-Based Models								BBS-Net	
		LHM [44]	CDB [37]	DESM [9]	GP [48]	CDCP [68]	ACSD [29]	LBE [21]	DCMC [12]	MDSF [51]	SE [23]	DF [47]	AFNet [55]	CTMF [24]	MMCI [5]	PCF [3]	TANet [4]	CPFP [65]	DMRA [45]	Ours (S1)	Ours (S2)
NJU2K [29]	S_α ↑	.514	.624	.665	.527	.669	.699	.695	.686	.748	.664	.763	.772	.849	.858	.877	.878	.879	**.886**	.914	**.921**
	$adpF_\beta$ ↑	.638	.648	.632	.655	.624	.696	.740	.717	.757	.734	.804	.768	.788	.730	.844	.844	.837	**.872**	.889	**.902**
	$meanF_\beta$ ↑	.328	.482	.550	.357	.595	.512	.606	.556	.682	.583	.784	.764	.779	.737	.840	.841	.850	**.873**	.889	**.902**
	$maxF_\beta$ ↑	.632	.648	.717	.647	.621	.711	.748	.715	.775	.748	.650	.775	.845	.852	.872	.874	.877	**.886**	.911	**.920**
	$adpE_\xi$ ↑	.708	.745	.682	.716	.747	.786	.791	.791	.812	.772	.864	.846	.864	.872	.896	.893	.895	**.908**	.917	**.924**
	$meanE_\xi$ ↑	.447	.565	.590	.446	.706	.593	.655	.619	.677	.624	.835	.826	.846	.841	.895	.895	.910	**.920**	.930	**.938**
	$maxE_\xi$ ↑	.724	.742	.791	.703	.741	.803	.803	.799	.838	.813	.696	.853	.913	.915	.924	.925	.926	**.927**	.948	**.949**
	M ↓	.205	.203	.283	.211	.180	.202	.153	.172	.157	.169	.141	.100	.085	.079	.059	.060	.053	**.051**	.040	**.035**
NLPR [44]	S_α ↑	.630	.629	.572	.654	.727	.673	.762	.724	.805	.756	.802	.799	.860	.856	.874	.886	.888	**.899**	.923	**.930**
	$adpF_\beta$ ↑	.664	.613	.563	.659	.608	.535	.736	.614	.665	.692	.744	.747	.724	.730	.795	.796	.823	**.854**	.867	**.882**
	$meanF_\beta$ ↑	.427	.422	.430	.451	.609	.429	.626	.543	.649	.624	.664	.755	.740	.737	.802	.819	.840	**.864**	.881	**.896**
	$maxF_\beta$ ↑	.622	.618	.640	.611	.645	.607	.745	.648	.793	.713	.778	.771	.825	.815	.841	.863	.867	**.879**	.892	**.918**
	$adpE_\xi$ ↑	.813	.809	.698	.804	.800	.742	.855	.786	.812	.839	.868	.884	.869	.872	.916	.916	.924	**.941**	.947	**.952**
	$meanE_\xi$ ↑	.560	.565	.541	.571	.781	.578	.719	.684	.745	.742	.755	.851	.840	.841	.887	.902	.918	**.940**	.942	**.950**
	$maxE_\xi$ ↑	.766	.791	.805	.723	.820	.780	.855	.793	.885	.847	.880	.879	.929	.913	.925	.941	.932	**.947**	.959	**.961**
	M ↓	.108	.114	.312	.146	.112	.179	.081	.117	.095	.091	.085	.058	.056	.059	.044	.041	.036	**.031**	.028	**.023**
STERE [42]	S_α ↑	.562	.615	.642	.588	.713	.692	.660	.731	.728	.708	.757	.825	.848	.873	.875	.871	**.879**	.835	.899	**.908**
	$adpF_\beta$ ↑	.703	.713	.594	.711	.666	.661	.595	.742	.744	.748	.742	.807	.771	.829	.826	.835	.830	**.844**	.867	**.885**
	$meanF_\beta$ ↑	.378	.489	.519	.405	.638	.478	.501	.590	.527	.610	.617	.806	.758	.813	.818	.828	**.841**	.837	.863	**.883**
	$maxF_\beta$ ↑	.683	.717	.700	.671	.664	.669	.633	.740	.719	.755	.757	.823	.831	.863	.860	.861	**.874**	.847	.892	**.903**
	$adpE_\xi$ ↑	.770	.808	.675	.784	.796	.793	.749	.831	.830	.825	.838	.886	.864	.901	.897	**.906**	.903	.900	.918	**.925**
	$meanE_\xi$ ↑	.484	.561	.579	.509	.751	.592	.601	.655	.614	.665	.691	.872	.841	.873	.887	.893	**.912**	.879	.917	**.928**
	$maxE_\xi$ ↑	.771	.823	.811	.743	.786	.806	.787	.819	.809	.846	.847	.887	.912	**.927**	.925	.923	.925	.911	.938	**.942**
	M ↓	.172	.166	.295	.182	.149	.200	.250	.148	.176	.143	.141	.075	.086	.068	.064	.060	**.051**	.066	.048	**.041**
DES [9]	S_α ↑	.562	.645	.622	.636	.709	.728	.703	.707	.741	.741	.752	.770	.863	.848	.842	.858	.872	**.900**	.929	**.933**
	$adpF_\beta$ ↑	.631	.729	.698	.686	.625	.717	.796	.702	.744	.726	.753	.730	.778	.762	.782	.795	.829	**.866**	.895	**.906**
	$meanF_\beta$ ↑	.345	.502	.483	.412	.585	.513	.576	.542	.523	.617	.604	.713	.756	.735	.765	.790	.824	**.873**	.896	**.910**
	$maxF_\beta$ ↑	.511	.723	.765	.597	.631	.756	.788	.666	.746	.741	.766	.728	.844	.822	.804	.827	.846	**.888**	.919	**.927**
	$adpE_\xi$ ↑	.761	.868	.795	.785	.816	.855	.911	.849	.869	.852	.877	.874	.911	.904	.912	.919	.927	**.944**	.966	**.967**
	$meanE_\xi$ ↑	.477	.572	.565	.503	.748	.612	.649	.632	.621	.707	.684	.810	.826	.825	.838	.863	.889	**.933**	.940	**.949**
	$maxE_\xi$ ↑	.653	.830	.868	.670	.811	.850	.890	.773	.851	.856	.870	.881	.932	.928	.893	.910	.923	**.943**	.965	**.966**
	M ↓	.114	.100	.299	.168	.115	.169	.208	.111	.122	.090	.093	.068	.055	.065	.049	.046	.038	**.030**	.024	**.021**
LFSD [36]	S_α ↑	.553	.515	.716	.635	.712	.727	.729	.753	.694	.692	.783	.738	.788	.787	.786	.801	.828	**.839**	.859	**.864**
	$adpF_\beta$ ↑	.714	.678	.676	.752	.695	.751	.705	.812	.795	.774	.802	.738	.778	.775	.788	.790	.809	**.845**	.850	**.858**
	$meanF_\beta$ ↑	.395	.374	.611	.516	.679	.562	610.	.652	.518	.636	.676	.732	.752	.718	.757	.767	.807	**.841**	.828	**.843**
	$maxF_\beta$ ↑	.708	.677	.762	.783	.702	.763	.722	.817	.779	.786	.817	.744	.787	.771	.775	.796	.826	**.852**	.854	**.859**
	$adpE_\xi$ ↑	.730	.696	.701	.776	.773	.794	.763	.835	.810	.777	.836	.802	.844	.832	.835	.838	.859	**.892**	.886	**.889**
	$meanE_\xi$ ↑	.488	.461	.632	.580	.748	.620	.664	.677	.583	.648	.719	.788	.802	.767	.810	.813	.856	**.885**	.871	**.883**
	$maxE_\xi$ ↑	.763	.871	.811	.824	.780	.829	.797	.856	.819	.832	.857	.815	.857	.839	.827	.847	.863	**.893**	.896	**.901**
	M ↓	.218	.225	.253	.190	.172	.195	.214	.155	.197	.174	.145	.133	.127	.132	.119	.111	.088	**.083**	.079	**.072**
SSD [67]	S_α ↑	.566	.562	.602	.615	.603	.675	.621	.704	.673	.675	.747	.714	.776	.813	.841	.839	.807	**.857**	.878	**.882**
	$adpF_\beta$ ↑	.580	.628	.614	.749	.522	.656	.613	.679	.674	.693	.724	.694	.710	.748	.791	.767	.726	**.821**	.829	**.849**
	$meanF_\beta$ ↑	.367	.347	.502	.453	.515	.469	.489	.572	.470	.564	.624	.672	.689	.721	.777	.773	.747	**.828**	.829	**.843**
	$maxF_\beta$ ↑	.568	.592	.680	.740	.535	.682	.619	.711	.703	.710	.735	.687	.729	.781	.807	.810	.766	**.844**	.853	**.859**
	$adpE_\xi$ ↑	.730	.737	.683	.795	.705	.765	.729	.786	.772	.778	.812	.803	.838	.860	.886	.879	.832	**.892**	.903	**.912**
	$meanE_\xi$ ↑	.498	.477	.560	.529	.676	.566	.574	.646	.576	.631	.690	.762	.796	.796	.856	.861	.839	**.897**	.894	**.904**
	$maxE_\xi$ ↑	.717	.698	.769	.782	.700	.785	.736	.786	.779	.800	.828	.807	.865	.882	.894	.897	.852	**.906**	**.922**	.919
	M ↓	.195	.196	.038	.180	.214	.203	.278	.169	.192	.165	.142	.118	.099	.082	.062	.063	.082	**.058**	.050	**.044**
SIP [18]	S_α ↑	.511	.557	.616	.588	.595	.732	.727	.683	.717	.628	.653	.729	.716	.833	.842	.835	**.850**	.806	.875	**.879**
	$adpF_\beta$ ↑	.592	.624	.644	.699	.495	.727	.733	.645	.694	.662	.673	.705	.684	.795	**.825**	.809	.819	.819	.862	**.872**
	$meanF_\beta$ ↑	.287	.341	.496	.411	.482	.542	.571	.499	.568	.515	.464	.702	.608	.771	.814	.803	**.821**	.811	.855	**.868**
	$maxF_\beta$ ↑	.574	.620	.669	.687	.505	.763	.751	.618	.698	.661	.657	.712	.694	.818	.838	.830	**.851**	.821	.877	**.883**
	$adpE_\xi$ ↑	.719	.771	.742	.774	.722	.827	.841	.786	.805	.756	.794	.815	.824	.886	**.899**	.893	**.899**	.863	.913	**.916**
	$meanE_\xi$ ↑	.437	.455	.564	.511	.683	.614	.651	.598	.645	.592	.565	.793	.705	.845	.878	.870	**.893**	.844	.898	**906**
	$maxE_\xi$ ↑	.716	.737	.770	.768	.721	.838	.853	.743	.798	.771	.759	.819	.829	.897	.901	.895	**.903**	.875	.921	**.922**
	M ↓	.184	.192	.298	.173	.224	.172	.200	.186	.167	.164	.185	.118	.139	.086	.071	.075	**.064**	.085	.060	**.055**

Fig. 5: Quantitative visual results of our method and 18 SOTA methods. 'RGB', 'Depth', 'GT' denote the input RGB image, depth map and ground-truth, respectively. 'S1' and 'S2' represent the intial and final saliency output of the proposed method.

Fig. 6: Quantitative visual results of our method and 18 SOTA methods. 'RGB', 'Depth', 'GT' denote the input RGB image, depth map and ground-truth, respectively. 'S1' and 'S2' represent the intial and final saliency output of the proposed method.

Fig. 7: Quantitative visual results of our method and 18 SOTA methods. 'RGB', 'Depth', 'GT' denote the input RGB image, depth map and ground-truth, respectively. 'S1' and 'S2' represent the intial and final saliency output of the proposed method.

Fig. 8: Quantitative visual results of our method and 18 SOTA methods. 'RGB', 'Depth', 'GT' denote the input RGB image, depth map and ground-truth, respectively. 'S1' and 'S2' represent the intial and final saliency output of the proposed method.

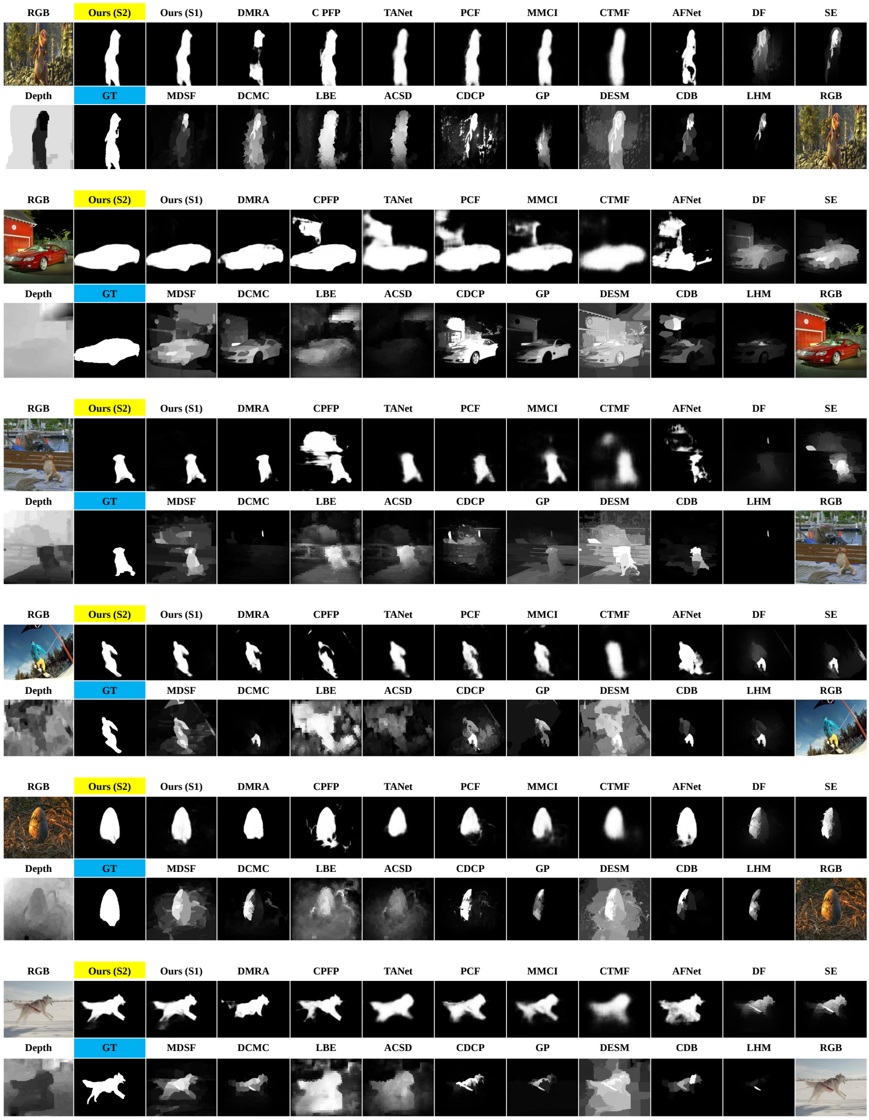

Fig. 9: Quantitative visual results of our method and 18 SOTA methods. 'RGB', 'Depth', 'GT' denote the input RGB image, depth map and ground-truth, respectively. 'S1' and 'S2' represent the intial and final saliency output of the proposed method.