A SALIENCY ENHANCED FEATURE FUSION BASED MULTI-SCALE RGB-D SALIENT OBJECT DETECTION NETWORK

Rui Huang[1] Qingyi Zhao[1] Yan Xing[2] Sihua Gao[1] Weifeng Xu[3] Yuxiang Zhang[1] Wei Fan[1]

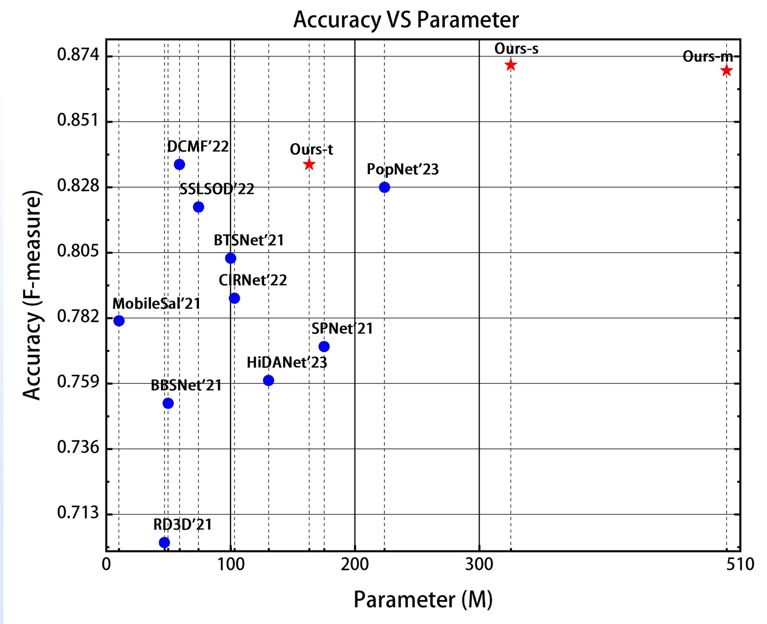
1 College of Computer Science and Technology, Civil Aviation University of China

2 College of Safety Science and Engineering, Civil Aviation University of China

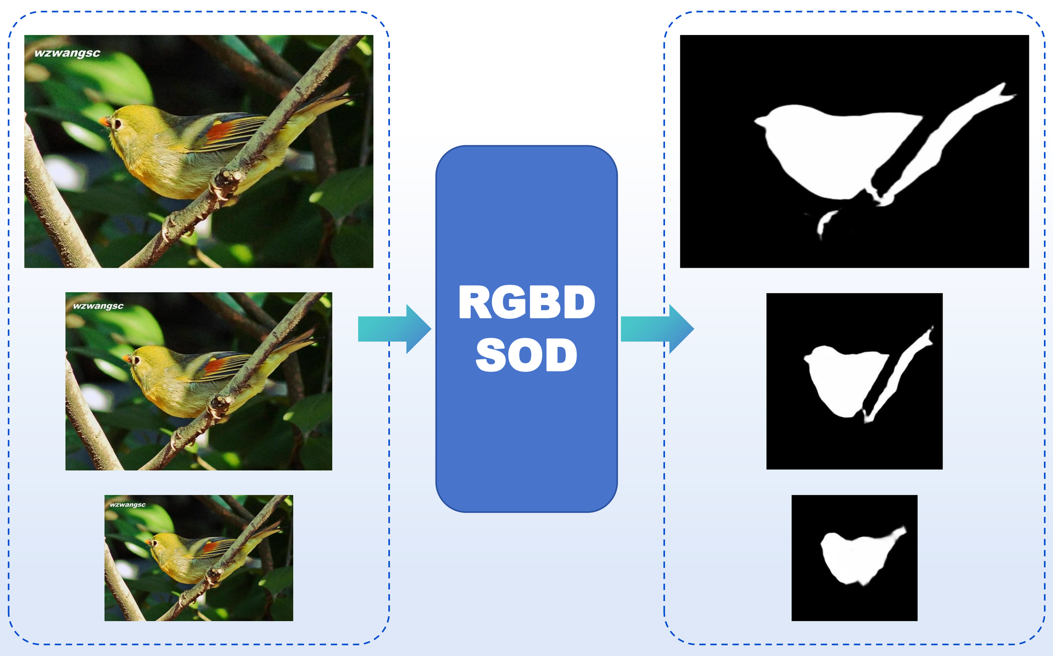
3 Department of Computer, North China Electric Power University

MOTIVATION

- Now RGB-D salient object detection (SOD) methods have proposed **complex fusion architectures** that refine RGB features and depth features simultaneously. Although these complex fusion strategies improve RGB-D SOD performance, they also **increase the size of models.**
- Recent studies have shown that **multiscale Convolutional Neural Networks** (CNNs) can achieve better performance in Super-resolution and image deblurring than single-scale CNNs. However, the use of multiscale CNNs in RGB-D saliency detection is hindered by **the large model sizes and computations required**.

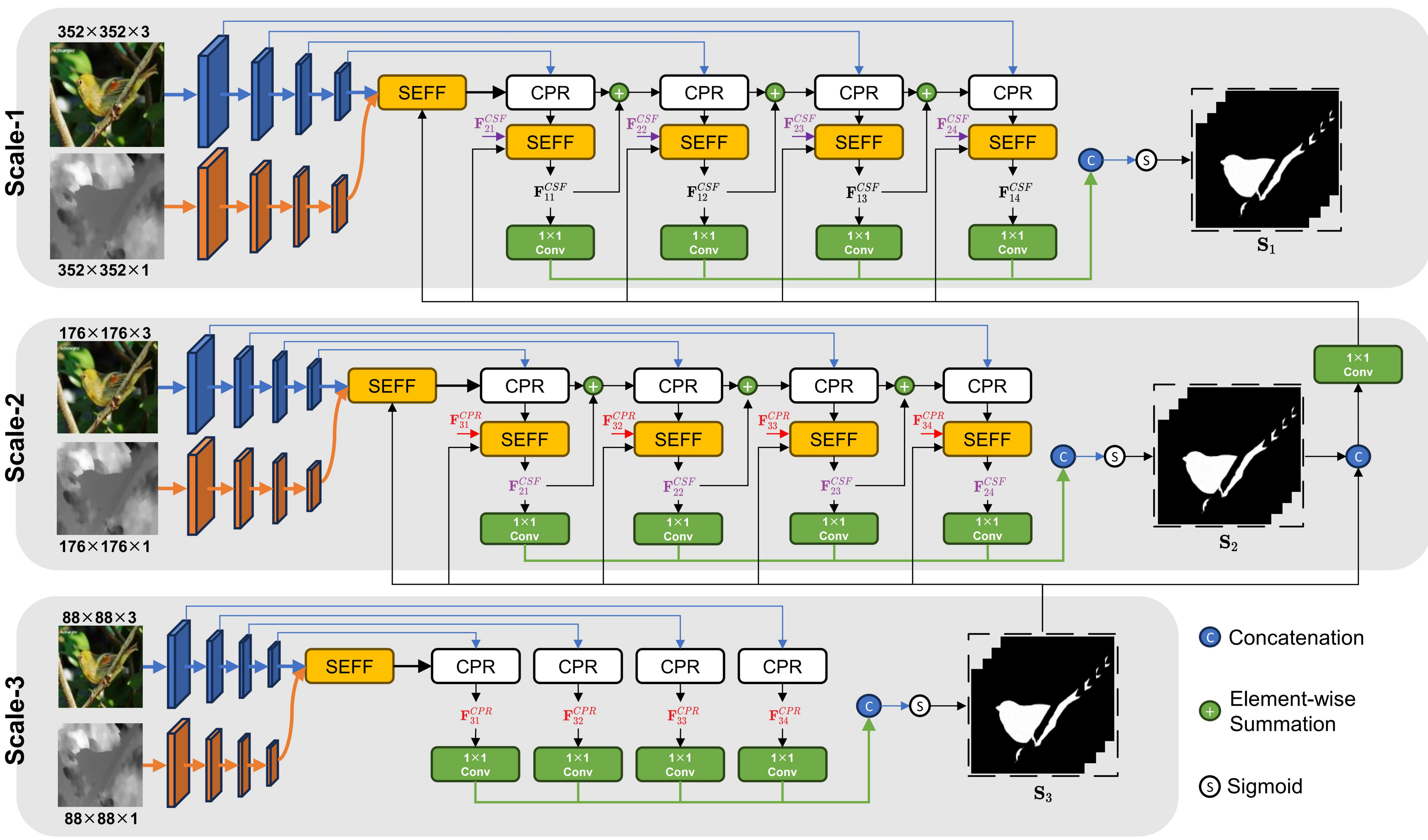


Comparision with different methods

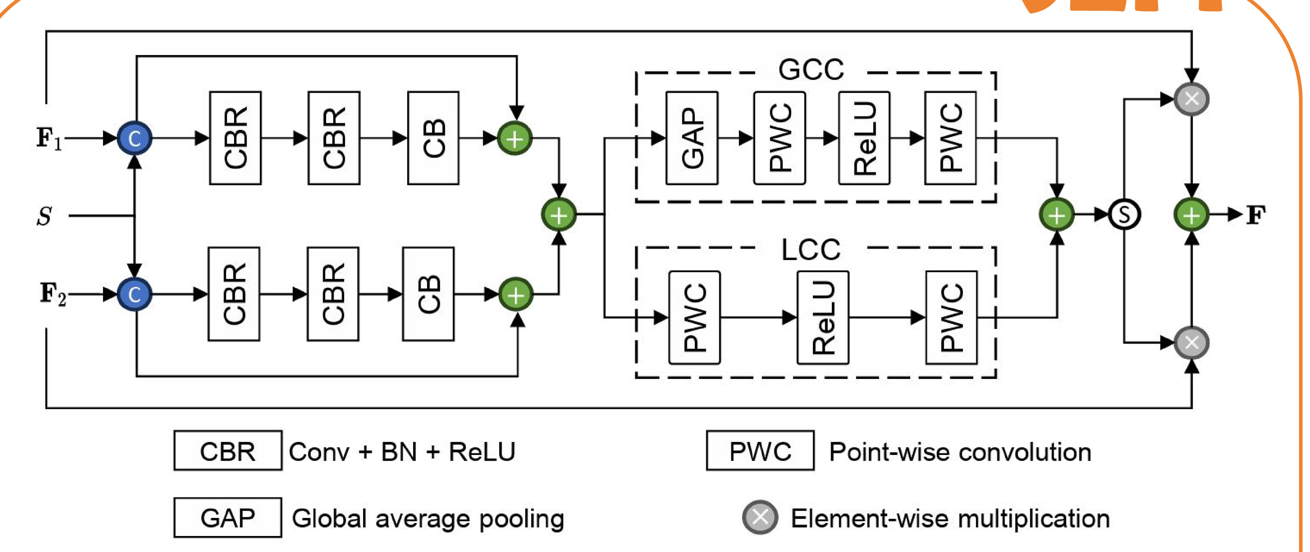


Different features can be observed at different scales, allowing more comprehensive information to be extracted.

FRAMWORK



SEFF



The whole process of SEFF is formalized as:

$$\mathbf{F} = \Phi(\mathbf{F}_1, \mathbf{F}_2, \mathbf{S})$$

We use SEFF to integrate the features of RGB and depth images:

$$\mathbf{F}_{34}^{fusion} = \Phi(\mathbf{F}_{34}^{R}, \mathbf{F}_{34}^{D}, \mathbf{Z}),$$
$$\mathbf{F}_{24}^{fusion} = \Phi(\mathbf{F}_{24}^{R}, \mathbf{F}_{24}^{D}, \mathbf{S}_3),$$
$$\mathbf{F}_{14}^{fusion} = \Phi(\mathbf{F}_{14}^{R}, \mathbf{F}_{14}^{D}, Conv_1(Cat(\mathbf{S}_2, \mathbf{S}_3), 4))$$

Similarly, we use SEFF to fuse the decode features of adjacent scale as follows:

$$\mathbf{F}_{2j}^{CSF} = \Phi(\mathbf{F}_{2j}^{CPR}, \mathbf{F}_{3j}^{CPR}, \mathbf{S}_3),$$
$$\mathbf{F}_{1j}^{CSF} = \Phi(\mathbf{F}_{1j}^{CPR}, \mathbf{F}_{2j}^{CSF}, Conv_1(Cat(\mathbf{S}_2, \mathbf{S}_3), 4))$$

We have developed a multiscale network called **SEFFSal** to detect salient objects.

01. SEFFSal takes **RGB** and **depth** images at 3 different scales and employs FasterNet as the fundamental feature extractor to extract features.

02. Our **Saliency Enhanced Feature Fusion (SEFF) module** is responsible for fusing features of RGB and depth images.

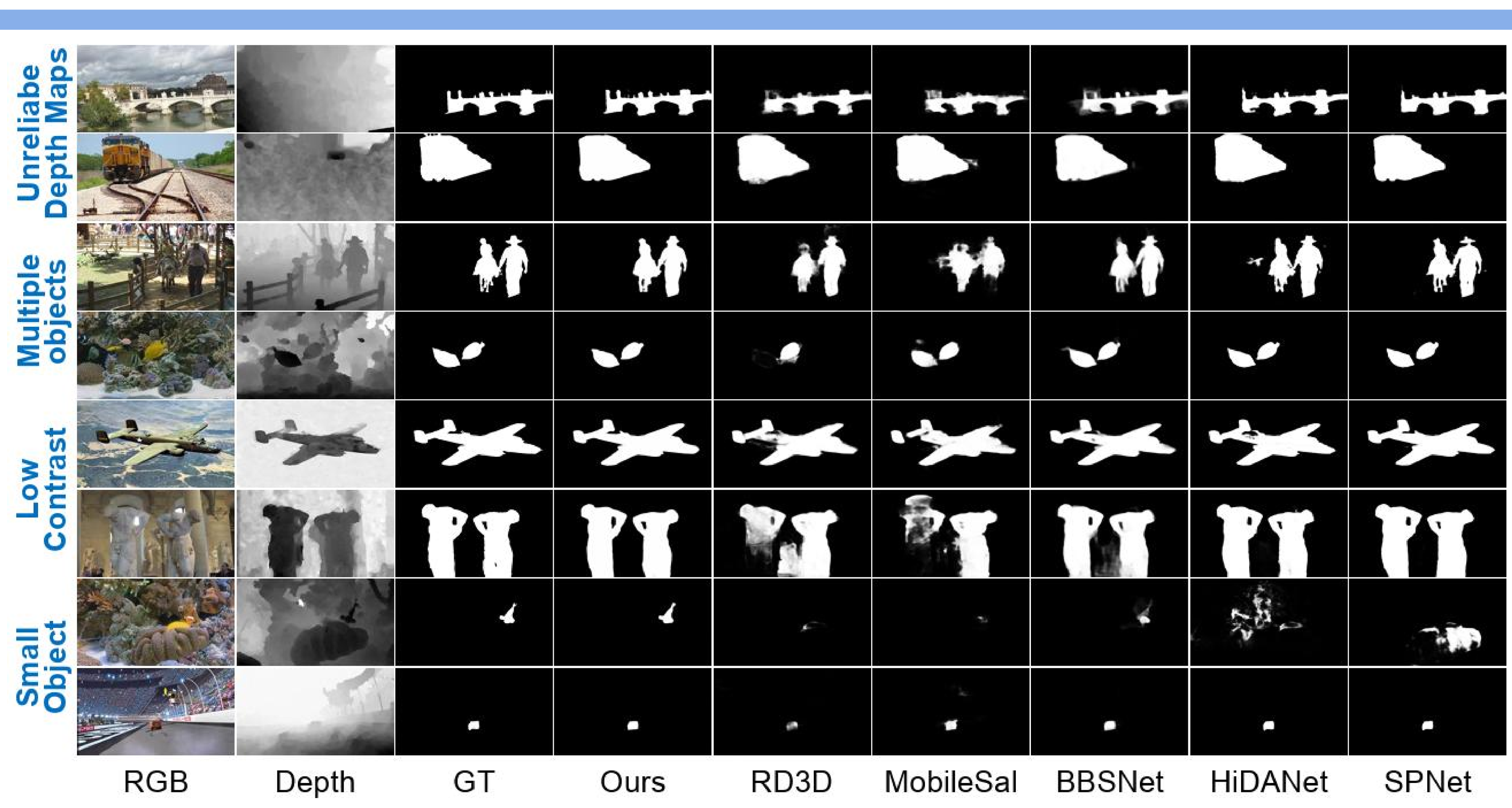
03. To improve the features, we have incorporated **Compact Pyramid Refinement (CPR)** as the decoder module. Then, SEFF is used to fuse the features of the decoders of different scales.

04. We generate the saliency maps from the features of the **3-rd scale j-th CPR** module:

$$\mathbf{S}_{3j} = \sigma(Conv_1(\mathbf{F}_{3j}^{CPR}, 1))$$

We generate the **saliency maps** from the features of SEFF modules by:

$$\mathbf{S}_{ij} = \sigma(Conv_1(\mathbf{F}_{ij}^{CSF}, 1))$$



RESULTS

Method	LFSD				NJU2K				NLPR				SIP				STERE				AVG				Parameter	Speed
	$\mathcal{M}\downarrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha^{max}\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha^{max}\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha^{max}\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha^{max}\uparrow$	$\mathcal{M}\downarrow$	$F_\beta^{max}\uparrow$	$E_\phi^{max}\uparrow$	$S_\alpha\uparrow$	M	fps
RD3D [9]	.134	.703	.780	.739	.035	.918	.953	0.921	.033	.874	.936	.899	.094	.768	.851	.779	.050	.875	.924	.886	.069	.827	.889	.845	46.90	54.62
BBSNet [20]	.122	.752	.820	.766	.022	.963	.980	.954	.023	.922	**.965**	.933	.083	.816	.883	.818	.048	.880	.930	.890	.059	.867	.915	.872	49.80	26.06
MobileSal [18]	.099	.781	.839	.801	.034	.924	.960	.918	.029	.893	.946	.906	.080	.815	.885	.815	.047	.880	.929	.889	.058	.859	.912	.866	10.24	69.11
BTSNet [12]	.098	.803	.855	.824	.023	.961	.980	.952	.029	.898	.952	.919	.057	.878	.919	.872	.049	.888	.935	.896	.051	.886	.928	.892	100.17	23.22
SPNet [21]	.118	.772	.843	.771	.016	.965	.981	.958	.020	.923	.964	.930	.096	.772	.866	.782	.043	.892	.939	.896	.059	.865	.919	.867	175.29	12.42
DCMF [22]	.085	.836	.881	.842	.018	.968	.983	.960	.026	.910	.956	.921	.067	.856	.900	.850	.038	.905	.948	.911	.047	.895	.934	.897	58.94	20.63
SSLSOD [23]	.083	.821	.868	.838	.026	.962	.969	.950	.032	.862	.906	.900	.085	.822	.862	.837	.047	.884	.917	.897	.055	.870	.904	.884	74.17	52.41
CIRNet [11]	.118	.789	.845	.785	.029	.953	.976	.945	.024	.920	.961	.930	.086	.820	.886	.824	.053	.886	.933	.887	.062	.873	.920	.874	103.15	30.91
HiDANet [10]	.121	.760	.829	.771	.018	.962	.980	.952	.021	.925	.964	.931	.093	.793	.883	.788	.046	.885	.933	.889	.060	.865	.918	.866	130.64	9.42
PopNet [24]	.079	.828	.875	.844	**.014**	.970	.984	.961	.020	.921	.964	.928	.051	.888	.924	.878	.033	.914	.951	.916	.040	.904	.939	.905	223.88	10.06
Ours-m	**.062**	**.869**	**.906**	**.870**	**.014**	**.974**	**.986**	**.965**	**.019**	**.927**	**.965**	**.937**	**.047**	**.898**	**.926**	**.885**	**.032**	**.917**	**.953**	**.921**	**.035**	**.917**	**.947**	**.915**	498.94	12.41
Ours-s	.064	.871	.906	.865	.016	.966	.984	.959	.020	.923	.963	.934	.061	.862	.900	.849	**.032**	.914	**.953**	.917	.039	.907	.941	.905	325.35	14.80
Ours-t	.080	.836	.886	.837	.022	.953	.971	.943	.023	.918	.960	.924	.078	.819	.868	.810	.038	.904	.942	.902	.048	.886	.925	.883	163.27	16.56
Ours-scale1	.085	.819	.861	.828	.016	.968	.984	.959	.020	.919	.960	.930	.049	.893	.922	.878	.035	.910	.945	.914	.041	.902	.934	.902	145.93	42.04
Ours-scale2	.071	.854	.892	.857	.015	.971	.985	.963	.020	.926	.964	.935	.050	.889	.920	.877	.033	.916	.951	.919	.038	.911	.942	.910	322.44	18.79
w/o SEFF	.091	.844	.875	.846	.039	.946	.972	.929	.034	.896	.952	.910	.077	.846	.897	.836	.054	.893	.941	.894	.059	.885	.927	.883	-	-

CONCLUTION

- We proposed a **multiscale** RGB-D salient object detection network based on a novel and effective feature fusion module, **Saliency Enhanced Feature Fusion(SEFF)**.
- We utilize SEFF to **fuse the features** of RGB and depth images,**as well as the features** of decoders at different scales.
- Through extensive experiments on **five benchmark datasets**, we have demonstrated that our method outperforms **ten** state of-the-art saliency detectors.
- We plan to explore a **light weight multiscale network** for RGB-D SOD in future work.

Contact e-mail : qingy_zhao@163.com

Paper Address

Code Address