

Supplementary Materials: Multi-Scale and Detail-Enhanced Segment Anything Model for Salient Object Detection

Paper ID 1376

1 INTRODUCTION

In this supplementary material, we provide further ablation studies and additional comparison experiments to demonstrate the effectiveness of our methods. Specifically, in Sec. 2, we analyze the impact of different hyperparameter designs and conduct further ablation studies to demonstrate the superiority of our modules. Sec. 3 will present an attribute-based analysis experiment. Sec. 4 supplements the quantitative and qualitative comparison under SOD and COD tasks mentioned in the main paper.

2 MORE ABLATION STUDY

In our proposed Lightweight Multi-Scale Adapter (LMSA), we use Global Average Pooling (GAP) to obtain different scales of the features. We also introduce local information to features to address the shortcomings of the Vision Transformer. In this Section, we conduct experiments on different scale sizes of LMSA and whether to introduce local information in MDSAM. The experiments were conducted under the fully designed MDSAM and the input resolution of the model is set to 512×512 .

Size of Scale. As shown in Table 1, indicated by (a) and (b), as well as (d) and (e), under the same local information conditions, the change in scale size has a very small impact on the results as long as it is multi-scale. Table 1's (c) and (e) demonstrate that multi-scale performs better than single-scale. Figure 1 demonstrates that, compared to the single-scale setting (c), the multi-scale settings (d) and (e) can more accurately detect objects in complex scenes. Therefore, the scale size setting of LMSA only needs to maintain multi-scale, without the need for specific values.

Effectiveness of Local Information. We conduct experiments on using the original design of PPM from PSPNet [5], which is (a) in Table 1. Compared to the original PPM, it can be observed that with the introduction of local information, the model's performance has significantly improved. Additionally, as shown in Table 1'(b) and (e), the presence of local information, MDSAM exhibits better performance with the same scale size. Figure 1 illustrates that local information enables the model to extract more precise features, resulting in better segmentation.

3 ATTRIBUTE-BASED ANALYSIS

In this section, we provide an attribute-based analysis by evaluating our proposed method on the challenging SOC [3] dataset. The SOC test dataset is divided into 9 major categories, which are Appearance Change (AC, 79 images), Big Object (BO, 24 images), Clutter (CL, 92 images), Heterogeneous Object (HO, 153 images), Motion Blur (MB, 32 images), Occlusion (OC, 157 images), Out-of-View (OV, 155 images), Shape Complexity (SC, 116 images), and Small Object (SO, 389 images). We compare our MDSAM with 17 methods, including Amulet [16], DSS [17], NLDF [30], SRM [20], BMPM [11],

Table 1: Ablation studies of different scale sizes and the effectiveness of local information.

Method	Scale Size	Local	DUTS-TE			DUT-OMRON		
			MAE	F_{β}^{max}	S_m	MAE	F_{β}^{max}	S_m
(a)	1,2,3,6	×	0.026	0.927	0.910	0.044	0.865	0.869
(b)	3,6,9,12	×	0.026	0.928	0.912	0.042	0.872	0.873
(c)	9,9,9,9	✓	0.025	0.930	0.916	0.041	0.875	0.873
(d)	3,5,7,9	✓	0.024	0.936	0.921	0.039	0.882	0.881
(e) MDSAM	3,6,9,12	✓	0.024	0.937	0.920	0.039	0.887	0.878

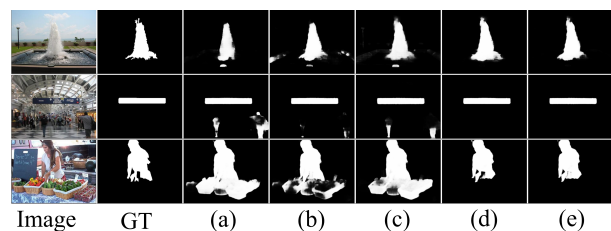


Figure 1: Visual comparison of the impact of scale size and effectiveness of local information.

C2SNet [22], DGRL [21], RANet [18], CPD [28], EGNNet [7], PoolNet [8], SCRNet [29], BANet [10], MINet [24], ICON-R [14], DC-Net-R [9], and full fine-tuned original SAM [1]. As shown in Table 2, our MDSAM demonstrates superior performance in most scenarios at resolutions of 512×512 and 384×384 . And they perform averagely only in the categories AC and BO, which have a small amount of data. The visualization results of the proposed MDSAM and six representative state-of-the-art methods are shown in Figure 2. This demonstrates that current methods struggle to accurately localize both large and small objects, and the results lack fine-grained details. Our MDSAM can accurately locate multi-scale targets, and both edges and details are highly precise.

4 MORE COMPARISON RESULTS

In the main paper, we compared our MDSAM with other methods by widely used metrics. Here, in Figure 3 and Figure 4, we present the precision-recall curves and F-measure curves compared to CAGNet-L [19], TE7 [12], MENet [23], VST [15], SELFREFORMER [27], ICONS [14], BBRF [13], DC-Net-S [9] and full fine-tuned original SAM [1] on five SOD datasets. And we provide more visual comparisons of them in Figure 5, Figure 6, and Figure 7. In Figure 8 and Figure 9, we display curves compared to SINet-v2 [4], BSA-Net [6], BGNet [26], ZoomNet [25], FEDER [2], FSPNet [31] on three COD datasets. More visual comparisons are shown in Figure 10.

Table 2: Performance comparison between our method and other 17 methods on SOC test datasets in terms of MAE, F_β^ω , S_m , and E_m . The best, second best, and third best results are highlighted in red, green, and blue, respectively. MDSAM is with a 512×512 input. MDSAM* is with a 384×384 input.

Attr	Metrics	Amulet [16]	DSS [17]	NLDF [30]	SRM [20]	BMPM [11]	C2SNet [22]	DGRL [21]	RANet [18]	CPD [28]	EGNet [7]	PoolNet [8]	SCRN [29]	BANet [10]	MINet [1]	ICON-R [14]	DC-Net-R [9]	SAM [24]	MDSAM*	MDSAM
AC	MAE	0.120	0.113	0.119	0.096	0.098	0.109	0.081	0.132	0.089	0.085	0.093	0.078	0.086	0.079	0.062	0.076	0.078	0.079	0.078
	F_β^ω	0.620	0.629	0.620	0.690	0.680	0.647	0.718	0.603	0.721	0.731	0.713	0.723	0.739	0.930	0.784	0.768	0.757	0.759	0.764
	S_m	0.752	0.753	0.737	0.791	0.780	0.755	0.791	0.709	0.799	0.806	0.795	0.809	0.806	0.802	0.835	0.824	0.821	0.815	0.819
	E_m	0.790	0.787	0.793	0.824	0.815	0.806	0.853	0.765	0.852	0.854	0.846	0.848	0.858	0.843	0.891	0.867	0.860	0.863	0.868
BO	MAE	0.334	0.343	0.341	0.294	0.292	0.257	0.207	0.440	0.236	0.358	0.339	0.217	0.261	0.175	0.200	0.278	0.232	0.264	0.231
	F_β^ω	0.625	0.628	0.635	0.679	0.683	0.739	0.794	0.469	0.755	0.602	0.625	0.784	0.729	0.828	0.794	0.699	0.749	0.709	0.758
	S_m	0.589	0.577	0.583	0.628	0.619	0.667	0.696	0.437	0.679	0.546	0.578	0.707	0.657	0.743	0.714	0.637	0.684	0.658	0.687
	E_m	0.566	0.554	0.556	0.630	0.635	0.674	0.736	0.423	0.699	0.547	0.572	0.716	0.663	0.769	0.740	0.641	0.703	0.681	0.709
CL	MAE	0.141	0.153	0.159	0.134	0.123	0.144	0.119	0.188	0.112	0.139	0.134	0.113	0.117	0.108	0.113	0.112	0.098	0.092	0.090
	F_β^ω	0.763	0.721	0.713	0.758	0.760	0.742	0.769	0.633	0.786	0.757	0.760	0.795	0.784	0.783	0.791	0.798	0.818	0.819	0.823
	S_m	0.763	0.721	0.713	0.758	0.760	0.742	0.769	0.633	0.786	0.757	0.760	0.795	0.784	0.783	0.791	0.798	0.818	0.819	0.823
	E_m	0.788	0.763	0.764	0.792	0.801	0.789	0.824	0.715	0.823	0.789	0.800	0.819	0.824	0.819	0.832	0.834	0.847	0.851	0.855
HO	MAE	0.119	0.124	0.126	0.115	0.116	0.123	0.104	0.143	0.098	0.106	0.100	0.096	0.094	0.089	0.091	0.092	0.087	0.083	0.080
	F_β^ω	0.688	0.660	0.661	0.696	0.684	0.668	0.722	0.626	0.736	0.720	0.739	0.743	0.753	0.759	0.767	0.761	0.793	0.788	0.795
	S_m	0.790	0.767	0.755	0.794	0.781	0.768	0.791	0.713	0.807	0.802	0.815	0.823	0.819	0.821	0.823	0.818	0.837	0.835	0.842
	E_m	0.809	0.796	0.798	0.819	0.813	0.805	0.833	0.777	0.838	0.829	0.845	0.842	0.850	0.858	0.865	0.848	0.864	0.869	0.872
MB	MAE	0.142	0.132	0.138	0.115	0.105	0.128	0.113	0.139	0.104	0.109	0.121	0.100	0.104	0.105	0.100	0.109	0.091	0.074	0.065
	F_β^ω	0.561	0.577	0.551	0.619	0.651	0.593	0.655	0.576	0.655	0.649	0.642	0.690	0.670	0.676	0.699	0.676	0.758	0.754	0.779
	S_m	0.712	0.719	0.685	0.742	0.762	0.719	0.744	0.696	0.753	0.762	0.751	0.792	0.764	0.761	0.774	0.757	0.810	0.820	0.834
	E_m	0.762	0.760	0.755	0.780	0.799	0.784	0.808	0.718	0.818	0.798	0.800	0.800	0.808	0.821	0.828	0.787	0.845	0.859	0.870
OC	MAE	0.143	0.144	0.149	0.129	0.119	0.130	0.116	0.169	0.106	0.121	0.118	0.111	0.112	0.102	0.106	0.102	0.089	0.085	0.089
	F_β^ω	0.607	0.595	0.593	0.630	0.644	0.622	0.659	0.527	0.679	0.658	0.659	0.673	0.677	0.686	0.683	0.708	0.732	0.728	0.733
	S_m	0.735	0.719	0.709	0.749	0.752	0.738	0.747	0.641	0.773	0.754	0.756	0.775	0.766	0.771	0.771	0.787	0.808	0.798	0.802
	E_m	0.762	0.760	0.755	0.780	0.799	0.784	0.808	0.718	0.818	0.798	0.800	0.800	0.808	0.821	0.817	0.824	0.843	0.847	0.845
OV	MAE	0.173	0.180	0.184	0.150	0.360	0.159	0.125	0.217	0.125	0.146	0.148	0.126	0.119	0.117	0.120	0.126	0.110	0.100	0.097
	F_β^ω	0.637	0.622	0.616	0.682	0.701	0.671	0.733	0.529	0.724	0.707	0.697	0.723	0.751	0.738	0.749	0.738	0.769	0.780	0.790
	S_m	0.721	0.700	0.688	0.745	0.751	0.728	0.762	0.611	0.765	0.752	0.747	0.774	0.779	0.775	0.779	0.771	0.796	0.804	0.811
	E_m	0.750	0.737	0.736	0.778	0.806	0.789	0.828	0.664	0.809	0.802	0.795	0.807	0.835	0.822	0.834	0.814	0.842	0.857	0.860
SC	MAE	0.098	0.098	0.101	0.090	0.081	0.100	0.087	0.110	0.076	0.083	0.075	0.078	0.078	0.077	0.080	0.072	0.068	0.062	0.065
	F_β^ω	0.608	0.599	0.593	0.638	0.677	0.611	0.669	0.594	0.701	0.678	0.695	0.691	0.705	0.711	0.714	0.749	0.757	0.746	0.772
	S_m	0.768	0.761	0.746	0.783	0.799	0.756	0.772	0.724	0.807	0.793	0.807	0.809	0.807	0.808	0.808	0.826	0.831	0.825	0.839
	E_m	0.793	0.798	0.787	0.812	0.840	0.805	0.837	0.791	0.848	0.843	0.856	0.843	0.850	0.859	0.871	0.869	0.866	0.882	0.882
SO	MAE	0.119	0.109	0.115	0.099	0.096	0.116	0.092	0.113	0.084	0.098	0.087	0.082	0.091	0.820	0.079	0.084	0.079	0.072	0.070
	F_β^ω	0.523	0.524	0.526	0.561	0.567	0.531	0.602	0.518	0.613	0.594	0.626	0.614	0.617	0.624	0.660	0.656	0.685	0.676	0.694
	S_m	0.718	0.713	0.703	0.737	0.732	0.707	0.736	0.682	0.756	0.749	0.768	0.767	0.755	0.759	0.778	0.774	0.791	0.786	0.798
	E_m	0.744	0.755	0.747	0.769	0.779	0.751	0.802	0.758	0.806	0.784	0.814	0.796	0.801	0.806	0.835	0.813	0.834	0.844	0.847

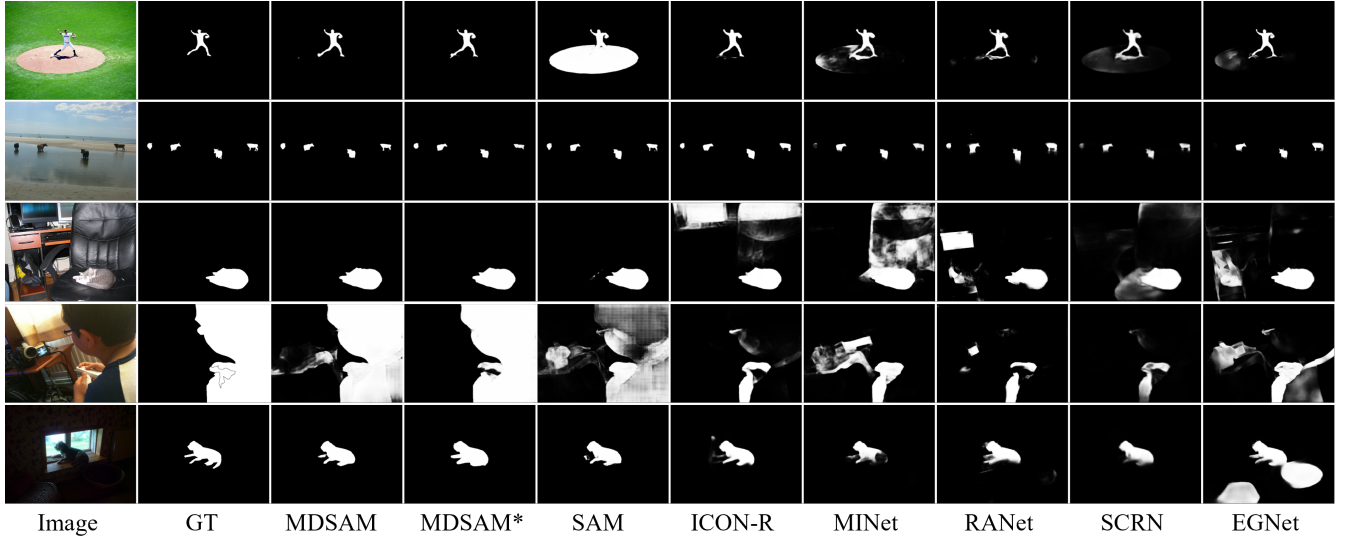


Figure 2: Visual comparison of output from our model with 6 representative methods. MDSAM is with a 512×512 input resolution, MDSAM* is with a 384×384 input resolution.

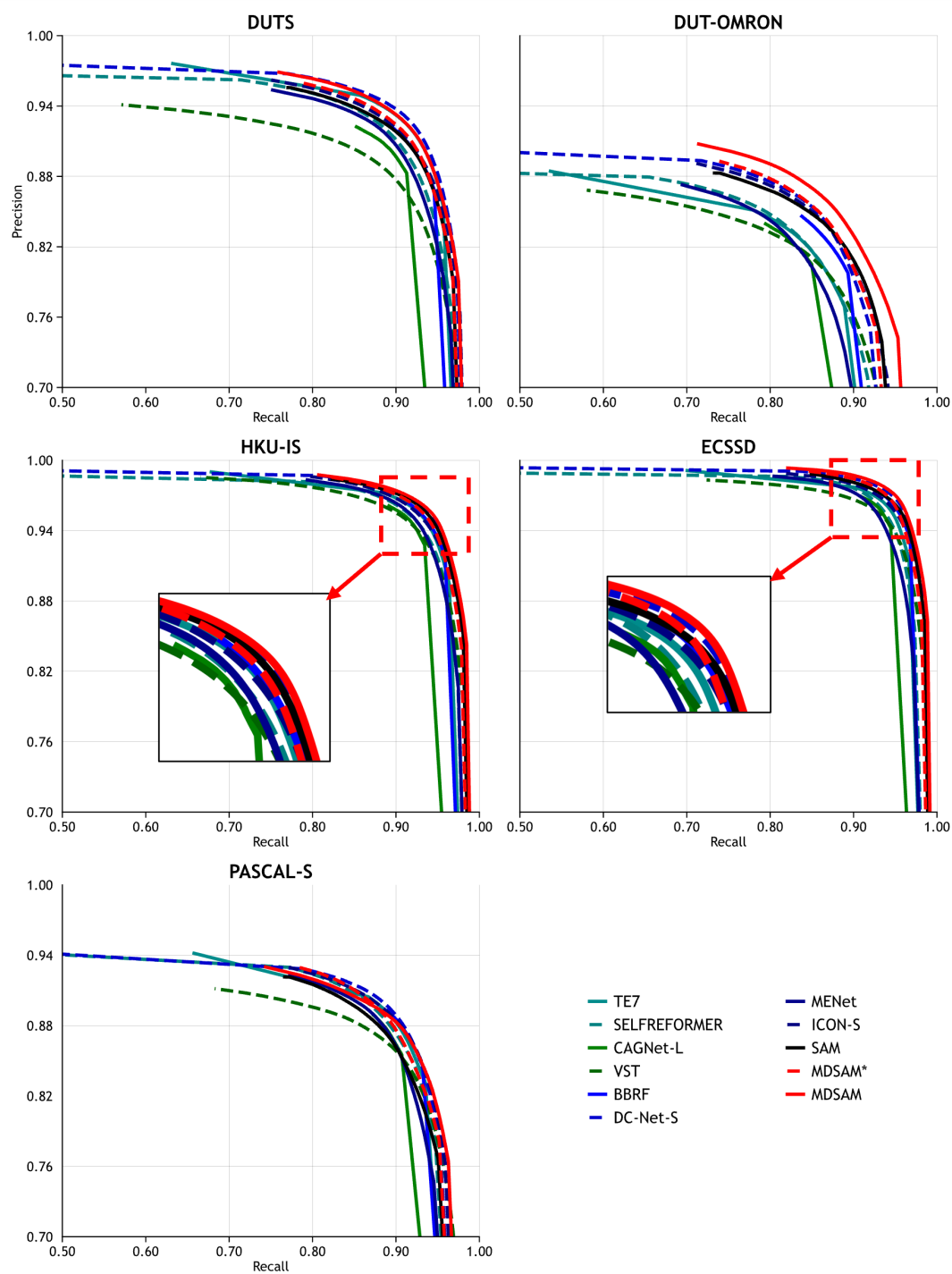


Figure 3: Precision-Recall curves comparison on five SOD datasets.

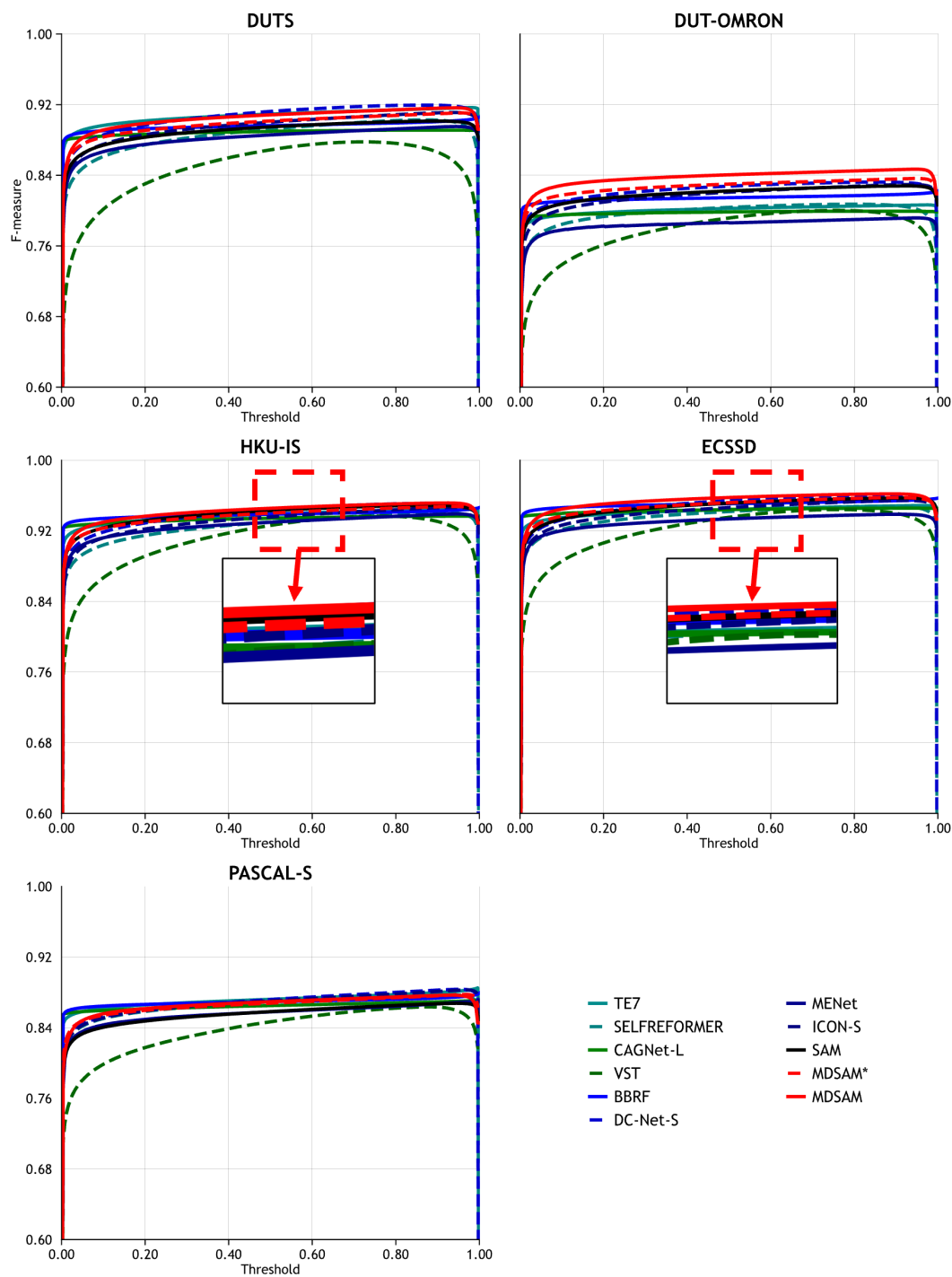


Figure 4: F-measure curves comparison on five SOD datasets.



Figure 5: Visual comparison of output from our model with 9 representative methods on five SOD datasets.

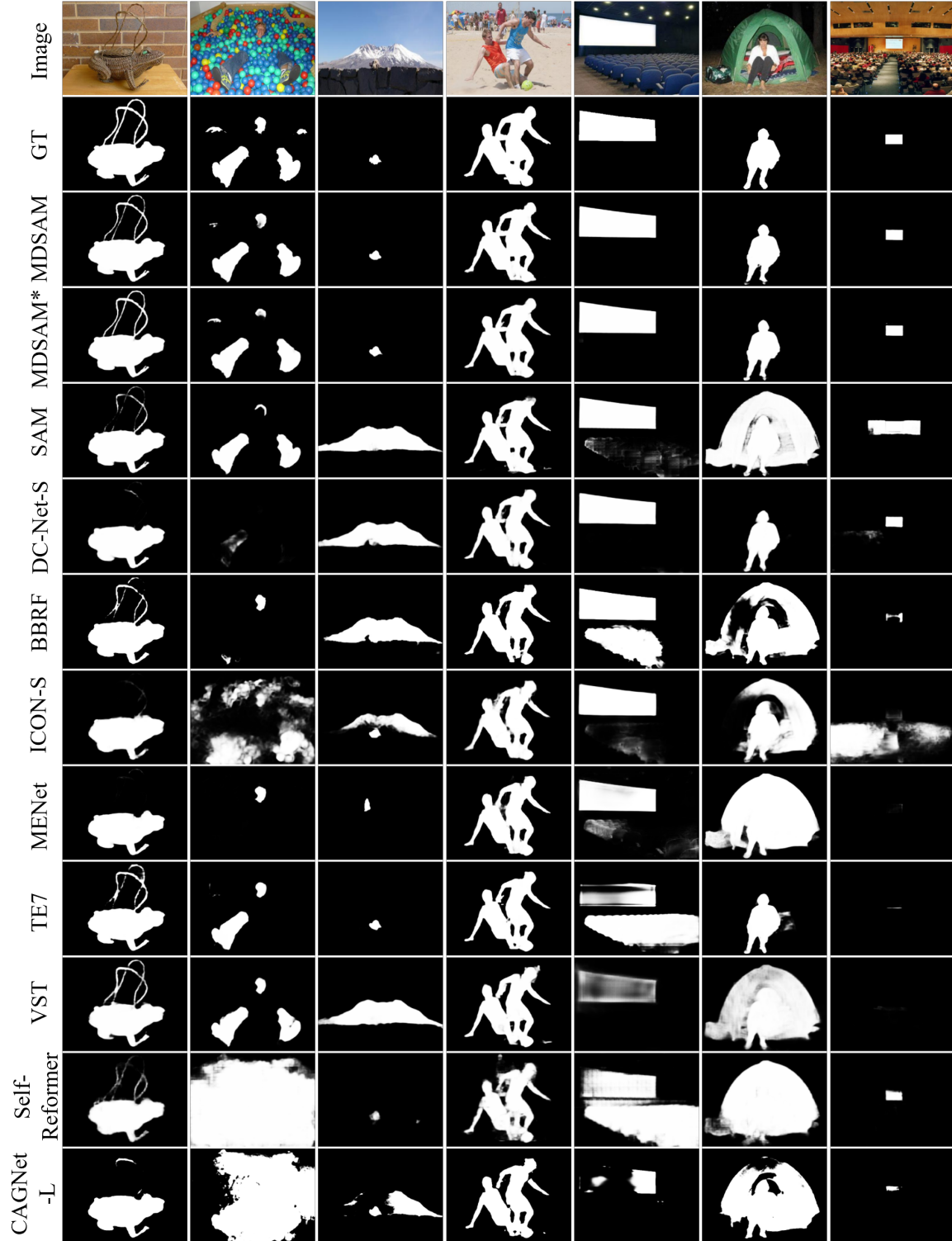


Figure 6: Visual comparison of output from our model with 9 representative methods on five SOD datasets.

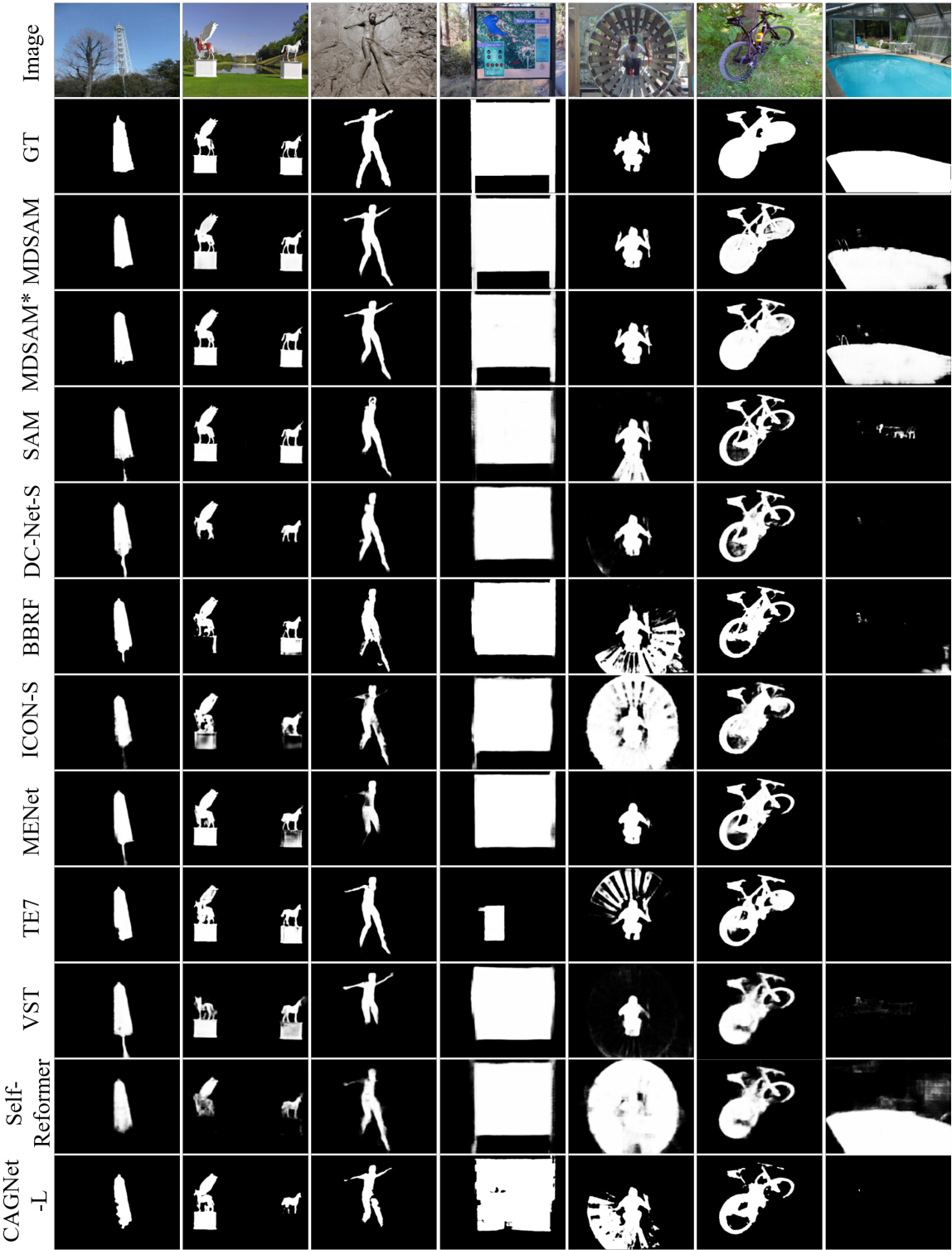


Figure 7: Visual comparison of output from our model with 9 representative methods on five SOD datasets.

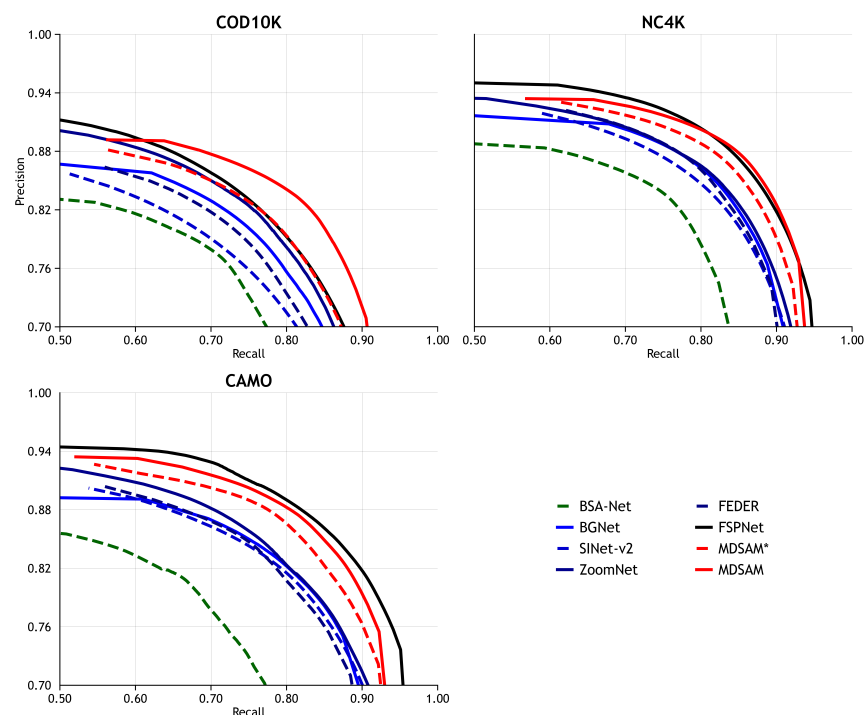


Figure 8: Precision-Recall curves comparison on three COD datasets.

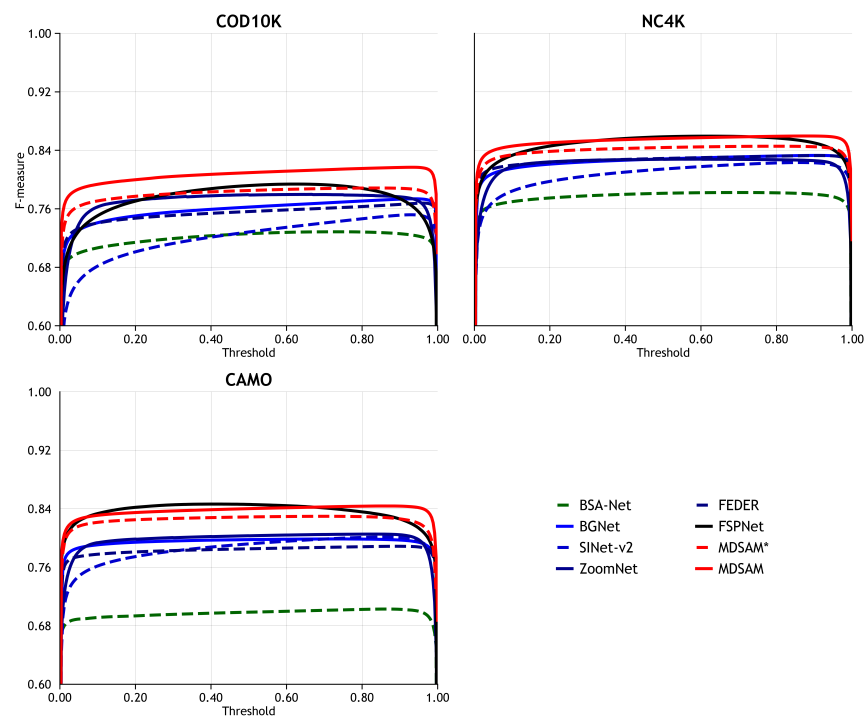
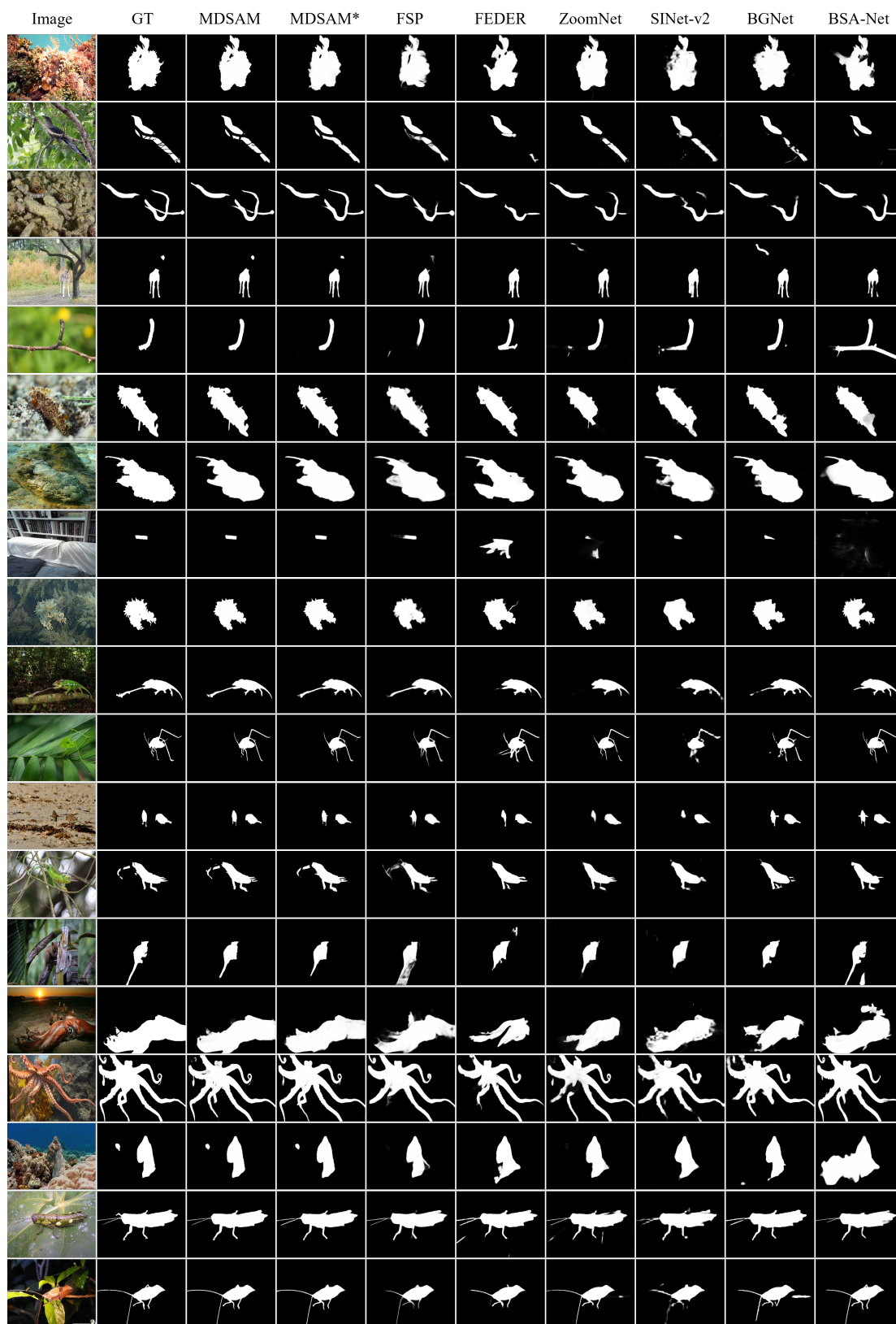


Figure 9: F-measure curves comparison on three COD datasets.



REFERENCES

- [1] Nikhila Ravi Hanzi Mao Chloe Rolland-Laura Gustafson Tete Xiao Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollar Alexander Kirillov, Eric Mintun and Ross Girshick. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4015–4026.
- [2] Yachao Zhang Longxiang Tang Yulun Zhang-Zhenhua Guo Chunming He, Kai Li and Xiu Li1. 2023. Camouflaged object detection with feature decomposition and edge reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22046–22055.
- [3] Gang Xu Ming-Ming Cheng Deng-Ping Fan, Jing Zhang and Ling Shao. 2022. Salient objects in clutter. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 2 (2022), 2344–2366.
- [4] Ming-Ming Cheng Deng-Ping Fan, Ge-Peng Ji and Ling Shao. 2021. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2021), 6024–6042.
- [5] Xiaojuan Qi Xiaogang Wang Hengshuang Zhao, Jianping Shi and Jiaya Jia. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2881–2890.
- [6] Haoran Xie Xuefeng Yan Dong Liang-Dapeng Chen Mingqiang Wei Hongwei Zhu1, Peng Li1 and Jing Qin. 2022. I can find you! boundary-guided separated attention network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 3608–3616.
- [7] Deng-Ping Fan Yang Cao Jufeng Yang Jia-Xing Zhao, Jiang-Jiang Liu and Ming-Ming Cheng. 2019. EGNNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8779–8788.
- [8] Ming-Ming Cheng Jiashi Feng Jiang-Jiang Liu, Qibin Hou and Jianmin Jiang. 2019. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3917–3926.
- [9] Xuebin Qin Jiayi Zhu and Abdulmotaleb Elsadik. 2023. DC-Net: Divide-and-Conquer for Salient Object Detection. *arXiv preprint arXiv:2305.14955* (2023).
- [10] Yu Zhang-Changqun Xia Jinming Su, Jia Li and Yonghong Tian. 2019. Selectivity or invariance: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3799–3808.
- [11] Huchuan Lu-You He Lu Zhang, Ju Dai and Gang Wang. 2018. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1741–1750.
- [12] WooSeok Shin Min Seok Lee and Sung Won Han. 2022. TRACER: Extreme attention guided salient object tracing network. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 12993–12994.
- [13] Chenxi Xie Mingcan Ma, Changqun Xia and Xiaowu Chen. 2023. Boosting broader receptive fields for salient object detection. *IEEE Transactions on Image Processing* 32 (2023), 1026–1038.
- [14] Nian Liu Dingwen Zhang Dong Xu Mingchen Zhuge, Deng-Ping Fan and Ling Shao. 2022. Salient object detection via integrity learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 3 (2022), 3738–3752.
- [15] Kaiyuan Wan Ling Shao Nian Liul1, Ni Zhang and Junwei Han. 2021. Visual saliency transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4722–4732.
- [16] Huchuan Lu Hongyu Wang Pingping Zhang, Dong Wang and Xiang Ruan. 2017. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 202–211.
- [17] Xiaowei Hu Ali Borji Zhuowen Tu Qibin Hou, Ming-Ming Cheng and Philip H. S. Torr. 2017. Deeply supervised salient object detection with short connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3203–3212.
- [18] Ben Wang Huchuan Lu Xuelong Hu Shuhan Chen, Xiuli Tan and Yun Fu. 2020. Reverse attention-based residual network for salient object detection. *IEEE Transactions on Image Processing* 29 (2020), 3763–3776.
- [19] Ali Bahria Sina Ghofrani Majelana Sina Mohammadia, Mehrdad Nooria and Mohammad Havaeib. 2020. CAGNet: Content-aware guidance for salient object detection. *Pattern Recognition* 103 (2020), 107303.
- [20] Lihe Zhang Pingping Zhang Tiantian Wang, Ali Borji and Huchuan Lu. 2017. A stagewise refinement model for detecting salient objects in images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4019–4028.
- [21] Shuo Wang Huchuan Lu Gang Yang Xiang Ruan Tiantian Wang, Lihe Zhang and Ali Borji. 2018. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3127–3135.
- [22] Hong Cheng Wei Liu Xin Li, Fan Yang and Dinggang Shen. 2018. Contour knowledge transfer for salient object detection. In *Proceedings of the European Conference on Computer Vision*. 355–370.
- [23] Xin Fan Tianzhu Wang Yi Wang, Ruili Wang and Xiangjian He. 2023. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10031–10040.
- [24] Lihe Zhang Youwei Pang, Xiaoqi Zhao and Huchuan Lu. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9413–9422.
- [25] Tian-Zhu Xiang Lihe Zhang1 Youwei Pang1, Xiaoqi Zhao and Huchuan Lu. 2022. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2160–2170.
- [26] Chenglizhao Chen Yujia Sun, Shuo Wang and Tian-Zhu Xiang. 2022. Boundary-guided camouflaged object detection. *arXiv preprint arXiv:2207.00794* (2022).
- [27] Yi Ke Yun and Weisi Lin. 2022. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283* (2022).
- [28] Li Su Zhe Wu and Qingming Huang. 2019. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3907–3916.
- [29] Li Su Zhe Wu and Qingming Huang. 2019. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7264–7273.
- [30] Andrew Achkar Justin Eichel Shaozi Li Zhiming Luo, Akshaya Mishra and Pierre-Marc Jodoin. 2017. Non-local deep features for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6609–6617.
- [31] Tian-Zhu Xiang Shuo Wang Huai-Xin Chen Jie Qin Zhou Huang, Hang Dai and Huan Xiong. 2023. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5557–5566.