Fantastic Animals and Where to Find Them: Segment Any Marine Animal with Dual SAM

Supplementary Material

1. Introduction

In the main paper, we have provided quantitative comparisons with some existing methods as well as the ablation studies. In this supplementary material, we first provide details of evaluation metrics. Then, we compare our method with more methods. Afterwards, we verify the transferability and zero-shot ability of our proposed method. In addition, we further validate the effectiveness of MCP and PMS through more ablation results. Finally, we present some visual results to show the effects of key modules.

2. Evaluation Metrics

In this section, we provide details of the five evaluation metrics used to assess compared models. With these metrics, we can comprehensively and adequately demonstrate the superior performance of our model.

1) The Mean Intersection over Union (mIoU) is computed by first determining the Intersection over Union (IoU) for each individual class, and then averaging these values across all classes. It can be represented as:

$$IoU = \frac{|A \cap B|}{|A \cup B|}, mIoU = \frac{1}{C}\sum_{i=1}^{C} IoU_i, \tag{1}$$

where A represents the predicted values for a certain class and B represents the true values of that class.

2) The weighted F-measure (F_β^w) is determined by computing the F_β score for each class and then weighting each class's contribution according to its occurrence frequency in the dataset. This metric can emphasize the performance on less-represented classes. It can be represented as:

$$F_\beta = \frac{\left(1+\beta^2\right) \times \text{ Precision }^\omega \times \text{ Recall }^\omega}{\beta^2 \times \text{ Precision }^\omega + \text{ Recall }^\omega} \tag{2}$$

where $Precision$ and $Recall$ are the precision and recall scores. β is a parameter to trade-off the precision and recall. It is usually set to 0.3.

3) The structural similarity measure (S_α) [4] is a metric used to evaluate the structural similarity between two images. S_α aligns more closely with the human visual judgment of image quality.

4) The Mean Enhanced-Alignment Measure (mE_ϕ) [40] is a metric that merges local pixel information and overall image means into a single score. This metric effectively captures both the global statistics of the image and the nuances of local pixel alignments.

5) The Mean Absolute Error (MAE) quantifies the average of the absolute discrepancies between the prediction and the ground truth. It offers an overall assessment without considering class boundaries. Superior performance is reflected in lower MAE values. It can be represented as:

$$MAE(p,g) = \frac{1}{m}\sum_{i=1}^{m} |p_i - g_i| \tag{3}$$

where p is the prediction and g is the ground truth. m is the pixel number.

With the aforementioned five metrics, we can fully assess the overall completeness of mask predictions while ensuring the reliability of object boundaries. Therefore, achieving optimal results across these five metrics can sufficiently demonstrate the effectiveness of our model.

3. More Comparison Results

In the main paper, we compare most recent methods. Here, we present more comparison results corresponding to more methods. As shown in Tab. 1, Tab. 2 and Tab. 3, the experimental results fully demonstrate the effectiveness of our proposed method.

4. Transferability and Zero-shot Ability

In fact, our model can adapt to other complex tasks, such as saliency detection, camouflaged object detection and polyp segmentation. To verify this fact, we conduct zero-shot and transferability testing on other datasets with large domain gaps, i.e., DUTS [39], COD10K [5] and Kvasir [14]. As shown in Tab. 4, our method also achieves better results than other SAM-based methods and task-specific ones. These results clearly verify the generalization of our method. In addition, since we freeze SAM's encoder, it somewhat preserves the zero-shot ability. As shown in Tab. 4, our method delivers comparable results with SAM, showing an expressive zero-shot ability.

5. More Ablation Results on MCP and PMS

Experiments are conducted on MAS3K [19] for its challenging and high-quality annotations.

Effects of MCP. For MCP, we first enhance the features through a self-attention mechanism, and then integrate the features extracted from SAM by using a cross-attention mechanism. In Tab. 5, we compare the effectiveness of these internal components of MCP. In the second

Method	MAS3K					RMAS				
	mIoU	S_α	F_β^w	mE_ϕ	MAE	mIoU	S_α	F_β^w	mE_ϕ	MAE
UNet++ [60]	0.506	0.726	0.552	0.790	0.083	0.558	0.763	0.644	0.835	0.046
BASNet [34]	0.677	0.826	0.724	0.862	0.046	0.707	0.847	0.771	0.907	0.032
PFANet [54]	0.405	0.690	0.471	0.768	0.086	0.556	0.767	0.582	0.810	0.051
SCRN [44]	0.693	0.839	0.730	0.869	0.041	0.695	0.842	0.731	0.878	0.030
U2Net [35]	0.654	0.812	0.711	0.851	0.047	0.676	0.830	0.762	0.904	0.029
SINet [5]	0.658	0.820	0.725	0.884	0.039	0.684	0.835	0.780	0.908	0.025
PFNet [29]	0.695	0.839	0.746	0.890	0.039	0.694	0.843	0.771	0.922	0.026
RankNet [27]	0.658	0.812	0.722	0.867	0.043	0.704	0.846	0.772	0.927	0.026
C2FNet [37]	0.717	0.851	0.761	0.894	0.038	0.721	0.858	0.788	0.923	0.026
ECDNet [20]	0.711	0.850	0.766	0.901	0.036	0.664	0.823	0.689	0.854	0.036
OCENet [21]	0.667	0.824	0.703	0.868	0.052	0.680	0.836	0.752	0.900	0.030
ZoomNet [31]	0.736	0.862	0.780	0.898	0.032	0.728	0.855	0.795	0.915	0.022
MASNet [9]	0.742	0.864	0.788	0.906	0.032	**0.731**	**0.862**	**0.801**	0.920	0.024
SETR [57]	0.715	0.855	0.789	0.917	0.030	0.654	0.818	0.747	0.933	0.028
TransUNet [1]	0.739	0.861	0.805	0.919	0.029	0.688	0.832	0.776	**0.941**	0.025
H2Former [10]	**0.748**	0.865	**0.810**	**0.925**	**0.028**	0.717	0.844	0.799	0.931	**0.023**
SAM [16]	0.566	0.763	0.656	0.807	0.059	0.445	0.697	0.534	0.790	0.053
SAM-Adapter[2]	0.714	0.847	0.782	0.914	0.033	0.656	0.816	0.752	0.927	0.027
SAM-DADF [17]	0.742	**0.866**	0.806	0.925	0.028	0.686	0.833	0.780	0.926	0.024
Dual-SAM	**0.789**	**0.884**	**0.838**	**0.933**	**0.023**	**0.735**	**0.860**	**0.812**	**0.944**	**0.022**

Table 1. Performance comparison on MAS3K and RMAS. The best and second results are in red and blue, respectively.

Method	UFO120					RUWI				
	mIoU	S_α	F_β^w	mE_ϕ	MAE	mIoU	S_α	F_β^w	mE_ϕ	MAE
UNet++ [60]	0.412	0.459	0.433	0.451	0.409	0.586	0.714	0.678	0.790	0.145
BASNet [34]	0.710	0.809	0.793	0.865	0.097	0.841	0.871	0.895	0.922	0.056
PFANet [54]	0.677	0.752	0.723	0.815	0.129	0.773	0.765	0.811	0.867	0.096
SCRN [44]	0.678	0.783	0.760	0.839	0.106	0.830	0.847	0.883	0.925	0.059
U2Net [35]	0.680	0.792	0.709	0.811	0.134	0.841	0.873	0.861	0.786	0.074
SINet [5]	0.767	0.837	0.834	0.890	0.079	0.785	0.789	0.825	0.872	0.096
PFNet [29]	0.570	0.708	0.550	0.683	0.216	0.864	0.883	0.870	0.790	0.062
RankNet [27]	0.739	0.823	0.772	0.828	0.101	0.865	0.886	0.889	0.759	0.056
C2FNet [37]	0.747	0.826	0.806	0.878	0.083	0.840	0.830	0.883	0.924	0.060
ECDNet [20]	0.693	0.783	0.768	0.848	0.103	0.829	0.812	0.871	0.917	0.064
OCENet [21]	0.605	0.725	0.668	0.773	0.161	0.763	0.791	0.798	0.863	0.115
ZoomNet [31]	0.616	0.702	0.670	0.815	0.174	0.739	0.753	0.771	0.817	0.137
MASNet [9]	0.754	0.827	0.820	0.879	0.083	0.865	0.880	0.913	0.944	0.047
SETR [57]	0.711	0.811	0.796	0.871	0.089	0.832	0.864	0.895	0.924	0.055
TransUNet [1]	0.752	0.825	0.827	0.888	0.079	0.854	0.872	0.910	0.940	0.048
H2Former [10]	**0.780**	**0.844**	**0.845**	**0.901**	**0.070**	0.871	0.884	0.919	0.945	0.045
SAM [16]	0.681	0.768	0.745	0.827	0.121	0.849	0.855	0.907	0.929	0.057
SAM-Adapter [2]	0.757	0.829	0.834	0.884	0.081	0.867	0.878	0.913	**0.946**	0.046
SAM-DADF [17]	0.768	0.841	0.836	0.893	0.073	**0.881**	**0.889**	**0.925**	0.940	**0.044**
Dual-SAM	**0.810**	**0.856**	**0.864**	**0.914**	**0.064**	**0.904**	**0.903**	**0.939**	**0.959**	**0.035**

Table 2. Performance comparison on UFO120 and RUWI. The best and second results are in red and blue, respectively.

Method	USOD10k			
	S_α	mE_ϕ	maxF	MAE
Itti [13]	.6112	.6670	.4676	.1798
RCRR [47]	.6449	.6898	.5592	.1831
DF [36]	.6410	.7576	.5589	.1400
CPD [43]	.9076	.9484	.8991	.0290
DMRA [32]	.8746	.9274	.8682	.0422
SAMNet [54]	.8875	.9382	.8739	.0396
PoolNet [22]	.9152	.9562	.9105	.0283
BASNet [34]	.9075	.9378	.8849	.0352
EGNet [53]	.9125	.9488	.9040	.0291
FC-SOD [49]	.7036	.7004	.6231	.0852
LDF [42]	.9135	.9574	.9173	.0260
F3Net [41]	.9140	.9599	.9171	.0251
PFPN [38]	.9090	.9547	.9055	.0302
MINet [30]	.9105	.9501	.9072	.0287
DASNet [52]	.9204	.9603	.9212	.0245
JL-DCF [8]	.9062	.9485	.8978	.0300
UCNet [50]	.8997	.9463	.8968	.0301
S2MA [23]	.8664	.9208	.8530	.0558
BBSNet [7]	.9061	.9512	.9056	.0337
DANet [55]	.9006	.9449	.8934	.0279
SGL-KRN [45]	.9214	.9633	**.9245**	.0237
DCF [15]	.9116	.9541	.9045	.0312
SPNet [58]	.9075	.9554	.9069	.0280
HAINet [18]	.9123	.9552	.9116	.0279
VST [25]	.9136	.9614	.9108	.0267
TriTransNet [26]	.7889	.8479	.7501	.0659
CSNet [3]	.8595	.9178	.8462	.0548
D3Net [6]	.8931	.9413	.8807	.0374
SVAM-Net [12]	.7465	.7649	.6451	.0915
BTS-Net [51]	.9093	.9542	.9104	.0291
CDINet [48]	.7049	.8644	.7362	.0904
CTDNet [56]	.9085	.9531	.9073	.0285
MFNet [33]	.8425	.9146	.8193	.0512
PFSNet [28]	.8983	.9421	.8966	.0370
PSGLoss [46]	.8640	.9078	.8508	.0417
TC-USOD [11]	**.9215**	**.9683**	.9236	**.0201**
SAM [16]	.8543	.9095	.8812	.0380
SAM-Adapter [2]	.8952	.9533	.9153	.0276
SAM-DADF [17]	.9051	.9552	.9154	.0250
Dual-SAM	**.9238**	**.9684**	**.9311**	**.0185**

Table 3. Performance comparison on USOD10k. The best and second results are in red and blue, respectively.

and third rows, we list the results of using the self-attention mechanism (S_{only}MCP) and the cross-attention mechanism (C_{only}MCP), respectively. Compared with the whole MCP structure in the last row, it indicates that both mechanisms have a positive effect.

Effects of PMS. In Tab. 6, we compare the impact of using mutual supervision at different decoder layers. "1 PMS" refers to the incorporation of the mutual supervision module in the first layer of the decoder, and the other definitions follow similarly. As the number of layers increases, the performance gradually improves. We can observe that mutual supervision has a positive effect. With mutual supervision between the two branches, the objects' details are adequately complemented.

Method	DUTS (SOD)		COD10K (COD)		Kvasir (Medical)	
	F^w_β	MAE	F^w_β	MAE	F^w_β	MAE
VST [24]	0.828	0.037	—	—	—	—
PFNet [29]	—	—	0.660	0.040	—	—
FAPNet [59]	—	—	—	—	0.894	0.027
SAM [16]	0.764	0.058	0.633	0.050	0.769	0.062
SAM-Adapter [2]	0.878	0.029	0.801	0.025	0.876	0.029
Ours (zero-shot)	0.783	0.048	0.677	0.044	0.696	0.082
Ours	**0.885**	**0.025**	**0.889**	**0.012**	**0.909**	**0.025**

Table 4. Performance comparison on other complex tasks.

Method	mIoU	S_α	F^w_β	mE_ϕ	MAE
no MCP	0.778	0.877	0.825	0.929	0.026
S_{only} MCP	0.779	0.878	0.828	0.931	0.026
C_{only} MCP	0.783	0.879	0.832	0.931	0.025
MCP	0.789	0.884	0.838	0.933	0.023

Table 5. Performance comparisons of MCP.

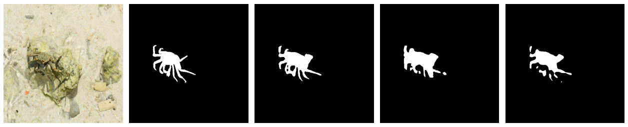
Method	mIoU	S_α	F^w_β	mE_ϕ	MAE
no PMS	0.771	0.874	0.820	0.923	0.029
1 PMS	0.776	0.876	0.823	0.926	0.027
2 PMS	0.779	0.878	0.827	0.927	0.026
3 PMS	0.783	0.880	0.830	0.932	0.025
4 PMS	0.789	0.884	0.838	0.933	0.023

Table 6. Performance comparisons with different layers of PMS.

6. More Visual Results

In the main paper, we have already presented a visual comparison of typical methods. In this supplementary material, we provide more visual results to verify the effects of our proposed key modules.

Visual Results with Key Modules. In Fig. 1, we show the visual effect of our C^3P module. One can observe that our C^3P module helps to obtain a better overall shape of underwater targets. The binary cross-entropy loss and nearby connectivity prediction are not good at predicting the animal boundaries In Fig. 2, we show the visual effect of our PMS module. By employing dual branches for mutual supervision, the segmentation maps have comprehensive information, effectively removing redundant information. In Fig. 3, we show the visual effect of our MCP module. With the multi-level coupled guidance, SAM has gained enhanced representational capabilities for animals and suppressed the cluttered backgrounds. In Fig. 4, we show the

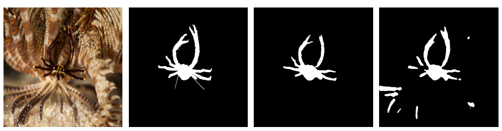


(a) Image (b) GT (c) C^3P (d) Binary (e) Nearby

Figure 1. Visualizing the effect of our C^3P module.

(a) Image (b) GT (c) PMS (d) Single (e) Dual

Figure 2. Visualizing the effect of our PMS module.

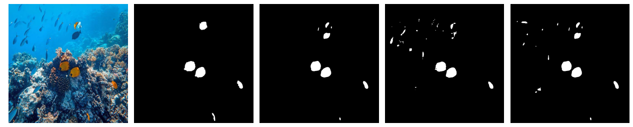


(a) Image (b) GT (c) MCP (d) no MCP

Figure 3. Visualizing the effect of our MCP module.

(a) Image (b) GT (c) DFAM (d) no DFAM

Figure 4. Visualizing the effect of our DFAM module.



(a) Image (b) GT (c) Ours (d) LoRA (e) Adapter

Figure 5. Visualizing the effect of different adapter mechanisms.

visual effect of our DFAM module. We integrate the features extracted from both the encoder and decoder through the DFAM module, and select more important feature channels. The design can adaptively aggregate more contextual information and significantly improve the segmentation results. In Fig. 5, we show the visual effect of our adapter mechanism. One can observe that our method effectively injects underwater domain information into the SAM backbone. Furthermore, the use of our dual adapter mechanisms continues to have a positive impact on the performance.

Visualization of Failed Results. In Fig. 6, we present some failure cases. Due to the similarity between the animal and its environment, it is challenging for our model to capture it accurately. However, other existing methods also result in significant segmentation errors. Therefore, distinguishing such organisms has become a focus of our further efforts.

References

[1] Jieneng Chen, Yongyi Lu, Qihang Yu, Xiangde Luo, Ehsan Adeli, Yan Wang, Le Lu, Alan L Yuille, and Yuyin Zhou. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv*, 2021. 2

[2] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Shangzhan Zhang, Yan Wang, Zejian Li, Lingyun Sun, Papa Mao, and Ying Zang. Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. *arXiv*, 2023. 2, 3

[3] Ming-Ming Cheng, Shang-Hua Gao, Ali Borji, Yong-Qiang Tan, Zheng Lin, and Meng Wang. A highly efficient model to study the semantics of salient object detection. *PAMI*, 44 (11):8006–8021, 2021. 3

[4] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint arXiv:1805.10421*, 2018. 1

[5] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020. 1, 2

[6] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *TNNLS*, 32(5):2075–2089, 2020. 3

[7] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *ECCV*, pages 275–292. Springer, 2020. 3

[8] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In *CVPR*, pages 3052–3062, 2020. 3

[9] Zhenqi Fu, Ruizhe Chen, Yue Huang, En Cheng, Xinghao Ding, and Kai-Kuang Ma. Masnet: A robust deep marine animal segmentation network. *IEEE Journal of Oceanic Engineering*, 2023. 2

[10] Along He, Kai Wang, Tao Li, Chengkun Du, Shuang Xia, and Huazhu Fu. H2former: An efficient hierarchical hybrid transformer for medical image segmentation. *TMI*, 2023. 2

[11] Lin Hong, Xin Wang, Gan Zhang, and Ming Zhao. Usod10k: a new benchmark dataset for underwater salient object detection. *TIP*, 2023. 3

[12] Md Jahidul Islam, Ruobing Wang, and Junaed Sattar. Svam: saliency-guided visual attention modeling by autonomous underwater robots. *arXiv*, 2020. 3

[13] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *PAMI*, 20(11):1254–1259, 1998. 3

[14] Debesh Jha, Pia H Smedsrud, Michael A Riegler, Pål Halvorsen, Thomas De Lange, Dag Johansen, and Håvard D

Figure 6. Visualizing failure segmentation cases.

Johansen. Kvasir-seg: A segmented polyp dataset. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 451–462. Springer, 2020. 1

[15] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated rgb-d salient object detection. In *CVPR*, pages 9471–9481, 2021. 3

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv*, 2023. 2, 3

[17] Yingxin Lai, Zhiming Luo, and Zitong Yu. Detect any deepfakes: Segment anything meets face forgery detection and localization. *arXiv*, 2023. 2, 3

[18] Gongyang Li, Zhi Liu, Minyu Chen, Zhen Bai, Weisi Lin, and Haibin Ling. Hierarchical alternate interaction network for rgb-d salient object detection. *TIP*, 30:3528–3542, 2021. 3

[19] Lin Li, Eric Rigall, Junyu Dong, and Geng Chen. Mas3k: An open dataset for marine animal segmentation. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 194–212. Springer, 2020. 1

[20] Lin Li, Bo Dong, Eric Rigall, Tao Zhou, Junyu Dong, and Geng Chen. Marine animal segmentation. *TCSVT*, 32(4): 2303–2314, 2021. 2

[21] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *WACV*, pages 1445–1454, 2022. 2

[22] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019. 3

[23] Nian Liu, Ni Zhang, Ling Shao, and Junwei Han. Learning selective mutual attention and contrast for rgb-d saliency detection. *TPAMI*, 44(12):9026–9042, 2021. 3

[24] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4722–4732, 2021. 3

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 3

[26] Zhengyi Liu, Yuan Wang, Zhengzheng Tu, Yun Xiao, and Bin Tang. Tritransnet: Rgb-d salient object detection with a triplet transformer embedding network. In *ACMMM*, pages 4481–4490, 2021. 3

[27] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, pages 11591–11601, 2021. 2

[28] Mingcan Ma, Changqun Xia, and Jia Li. Pyramidal feature shrinking for salient object detection. In *AAAI*, pages 2311–2318, 2021. 3

[29] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8772–8781, 2021. 2, 3

[30] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020. 3

[31] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022. 2

[32] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 3

[33] Yongri Piao, Jian Wang, Miao Zhang, and Huchuan Lu. Mfnet: Multi-filter directive network for weakly supervised salient object detection. In *ICCV*, pages 4136–4145, 2021. 3

[34] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 2, 3

[35] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *PR*, 106:107404, 2020. 2

[36] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *TIP*, 26(5):2274–2285, 2017. 3

[37] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. *arXiv*, 2021. 2

[38] Bo Wang, Quan Chen, Min Zhou, Zhiqiang Zhang, Xiaogang Jin, and Kun Gai. Progressive feature polishing network for salient object detection. In *AAAI*, pages 12128–12135, 2020. 3

[39] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 136–145, 2017. 1

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[41] Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *AAAI*, pages 12321–12328, 2020. 3

[42] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13025–13034, 2020. 3

[43] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019. 3

[44] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, pages 7264–7273, 2019. 2

[45] Binwei Xu, Haoran Liang, Ronghua Liang, and Peng Chen. Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection. In *AAAI*, pages 3004–3012, 2021. 3

[46] Sheng Yang, Weisi Lin, Guosheng Lin, Qiuping Jiang, and Zichuan Liu. Progressive self-guided loss for salient object detection. *TIP*, 30:8426–8438, 2021. 3

[47] Yuchen Yuan, Changyang Li, Jinman Kim, Weidong Cai, and David Dagan Feng. Reversion correction and regularized random walk ranking for saliency detection. *TIP*, 27(3): 1311–1322, 2017. 3

[48] Chen Zhang, Runmin Cong, Qinwei Lin, Lin Ma, Feng Li, Yao Zhao, and Sam Kwong. Cross-modality discrepant interaction network for rgb-d salient object detection. In *ACMMM*, pages 2094–2102, 2021. 3

[49] Dingwen Zhang, Haibin Tian, and Jungong Han. Few-cost salient object detection with adversarial-paced learning. *ANIPS*, 33:12236–12247, 2020. 3

[50] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *CVPR*, pages 8582–8591, 2020. 3

[51] Wenbo Zhang, Yao Jiang, Keren Fu, and Qijun Zhao. Bts-net: Bi-directional transfer-and-selection network for rgb-d salient object detection. In *ICME*, pages 1–6. IEEE, 2021. 3

[52] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *ACMMM*, pages 1745–1754, 2020. 3

[53] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019. 3

[54] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, pages 3085–3094, 2019. 2, 3

[55] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time rgb-d salient object detection. In *ECCV*, pages 646–662. Springer, 2020. 3

[56] Zhirui Zhao, Changqun Xia, Chenxi Xie, and Jia Li. Complementary trilateral decoder for fast and accurate salient object detection. In *ACMMM*, pages 4967–4975, 2021. 3

[57] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890, 2021. 2

[58] Tao Zhou, Huazhu Fu, Geng Chen, Yi Zhou, Deng-Ping Fan, and Ling Shao. Specificity-preserving rgb-d saliency detection. In *ICCV*, pages 4681–4691, 2021. 3

[59] Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature aggregation and propagation network for camouflaged object detection. *IEEE Transactions on Image Processing*, 31:7036–7047, 2022. 3

[60] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *MICCAI*, pages 3–11. Springer, 2018. 2