Pyramid Grafting Network for One-Stage High Resolution Saliency Detection -Supplementary Material-

A. More Comparison with State-of-the-arts

A.1. Quantitative Comparison

We compare our PGNet training on different datasets with other SOTA SOD methods on ECSSD [15] with 1,000 images, PASCAL-S [5] with 850 images and HKU-IS [4] with 4,447 images. As can be seen in Tab. 3, our PGNet-D and PGNet-DH outperforms other methods by a large margin on the three low-resolution datasets, especially in terms of the metric S_m. The high performance on S_m means that our methods can generate saliency maps with high structural similarity to ground truth (GT) [2]. This is consistent with our expected results, as compared to the multi-stage approaches, our method can obtain a full view of a salient object from a high-resolution view, avoiding the problem of losing some small parts which is easy to happen at low resolution. What's more, we show the PR curves on three high-resolution datasets in Fig. 1. It's clear that the PR curve of our method is much higher than others.

A.2. Visual Comparison

We argue that visual comparisons in high-resolution tasks are more important than in low-resolution tasks. This is due to the fact that the possible loss of some minor details does not have a significant impact on the pixel-wise metrics such as MAE, as the detailed pixels represent a small percentage of the total. However, these details may be important for downstream tasks. We therefore performed more visual comparisons to demonstrate the superiority of our approach, which can be seen in Fig. 2 and Fig. 3. The representative examples in the two pictures are from HRSOD-TE [16] and UHRSD-TE respectively.

A.3. PGNet of Resnet-50 Architecture

For a fair comparison, we show the performance of PGNet with Swin [6] replaced by Resnet-50 [3] compared to other high-resolution methods shown in Tab. 1. It can be seen that even after replacing Swin used on the low-resolution branch, our method still outperforms than other methods on the high-resolution datasets with a faster speed.

Table 1. Quantitative comparisons among high-resolution methods and our PGNet of Resnet-50 architecture. The three methods are all trained on DUTS-TR [10] and HRSOD-TR [16] mixed dataset. The best two results are in red and green fonts.

		Ours-R	**HRSOD**	**DHQSOD**
HRSOD-TE	F_β^{Max}	0.931	0.905	0.922
	MAE	0.024	0.030	0.022
	E_ξ	0.940	0.934	0.947
	S_m	0.925	0.896	0.920
DAVIS-S	F_β^{Max}	0.940	0.899	0.938
	MAE	0.014	0.026	0.012
	E_ξ	0.961	0.955	0.947
	S_m	0.937	0.876	0.920
UHRSD-TE	F_β^{Max}	0.928	-	-
	MAE	0.041	-	-
	E_ξ	0.901	-	-
	S_m	0.904	-	-

Table 2. Ablation studies of Loss function.

Components of the Loss function				HRSOD-TE			
$\mathcal{L}_{bce}$	$\mathcal{L}_{IoU}$	$\mathcal{L}_{Aux}$	$\mathcal{L}_{AGL}$	F_β^{Max}	MAE	E_ξ	S_m
✓				.931	.025	.920	.932
✓	✓			.940	.023	.944	.936
✓	✓	✓		.941	.022	.942	.936
✓	✓	✓	✓	.945	.020	.946	.938

A.4. Model Size and Speed

To demonstrate the advantages of our one-stage method in terms of model size and inference speed, we show the comparison among different methods. In Tab. 4, Ours is architecture consisting of Swin Transformer and Resnet-18 as mentioned in main text. And the Ours-Res50 is the architecture replacing Swin Transformer with Resnet-50. As can be seen, our Ours-Res50 has a much smaller model size than the DHQSOD [9] which also uses Resnet-50 as encoder. Not only that, our methods maintain high inference speed with high resolution input, far exceeding other high-resolution methods.

Table 3. Quantitative comparisons with state-of-the-art SOD models on other three low-resolution benchmark datasets in terms of max F-measure, MAE , E-measure, S-measure. D: trained on DUTS-TR, HD: trained on DUTS-TR and HRSOD-TR, UH: trained on UHRSD-TR and HRSOD-TR . The best two results are in red and green fonts.

Method	ECSSD				PASCAL-S				HKU-IS			
	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m
CPD [13]	.939	.037	.925	.918	.864	.072	.849	.842	.925	.034	.944	.905
SCRN [14]	.950	.037	.926	.927	.877	.062	.857	.863	.934	.034	.949	.916
DASNet [17]	.950	.032	.927	.927	.876	.061	.861	.857	.941	.026	.955	.919
F3Net [11]	.945	.033	.927	.924	.878	.062	.859	.855	.937	.028	.953	.917
GCPA [1]	.948	.035	.920	.927	.876	.061	.850	.861	.938	.031	.949	.920
ITSD [19]	.947	.034	.927	.925	.876	.064	.853	.856	.934	.031	.952	.917
LDF [12]	.950	.034	.925	.924	.881	.060	.865	.856	.939	.027	.954	.919
CTD [18]	.950	.032	.925	.925	.878	.061	.861	.856	.941	.027	.955	.927
PFS [7]	.952	.031	.928	.930	.875	.063	.856	.854	.943	.026	.956	.924
HRSOD [16]	.925	.052	.916	.888	.846	.079	.844	.810	.810	.042	.934	.877
DHQSOD [9]	.953	.030	.932	.926	.878	.059	.862	.851	.944	.025	.957	.922
Our PGNet												
Ours-D	.960	.027	.932	.938	.894	.052	.875	.873	.948	.024	.961	.929
Ours-DH	.960	.027	.931	.937	.883	.056	.872	.866	.949	.024	.961	.930
Ours-UH	.948	.032	.928	.928	.869	.057	.863	.855	.937	.029	.953	.915

B. UHRSD dataset

B.1. More analysis of UHRSD

In addition to the introduction to UHRSD in the main text, we analyze some basic attributes of the salient object in our UHRSD datasets comparing to widely used SOD dataset. In Fig. 4, the first figure shows the distance of the center of salient object from the image center. And the second one shows that how far the farthest pixel of salient objects from the image center. Both of them proved that the salient objects in our dataset do not suffer from the center bias. What's more, the last figure in Fig. 4 illustrates that the distribution of salient object sizes in our UHRSD is consistent with the widely used SOD dataset.

B.2. Methods trained on UHRSD and HRSOD

In the main text we suggest that the distribution of the high-resolution dataset differs significantly from the distribution of the low-resolution dataset, so to better represent this difference, we trained and tested different methods using a mixed dataset UHRSD-TR and HRSOD-TR. As shown in Tab. 5, the performance on low-resolution datasets trained with UH is all significantly degraded. In contrast, there is a small decrease or a significant increase in the three high-resolution datasets. This indicates that a high-quality high-resolution training set is helpful and necessary for training high-resolution models.

B.3. Hard Cases from UHRSD

Due to the nature of high-resolution images with rich details, we have specially selected many challenging scenes to construct our UHRSD. As shown in Fig. 5, these salient objects all have very complex holes and edges, which are very difficult to distinguish in low-resolution scenes. We have selected these images and annotated them with a fine level and there are more such examples in our UHRSD dataset.

C. More details of loss function

C.1. IoU loss

The widely-used IoU loss [8] can be formulated as Eq. (1).

$$\mathcal{L}_{iou} = 1 - \frac{\sum\limits_{i,j}^{H,W} (G_{ij} \times P_{ij})}{\sum\limits_{i,j}^{H,W} (G_{ij} + P_{ij} - G_{ij} \times P_{ij})} \tag{1}$$

where G_{ij} and P_{ij} represent the value of pixel (i, j) on Ground-Truth map and Prediction map respectively.

C.2. Detailed ablation study for loss function

The $\mathcal{L}_{bce}$ is the most widely used loss in SOD tasks. We also adopt $\mathcal{L}_{IoU}$ to optimize the global structure like other existing SOD methods.

Different from methods that apply auxiliary supervision on each side output, we only apply $\mathcal{L}_{Aux}$ on the prediction maps RP and SP. Our purpose is to provide quality features for CMGM and to generate more accurate error weight ω_{ij} for our proposed $\mathcal{L}_{AGL}$.

We conduct the ablation study of $\mathcal{L}_{total}$ in Tab. 2 , which shows the effectiveness of our $\mathcal{L}_{AGL}$.

Table 4. Comparisons of model size and FPS with state-of-the-art SOD models. Ours is the architecture consisting of Swin Transformer and Resnet-18 as used in main paper. The Ours-Res50 is the architecture replacing Swin Transformer with Resnet50.

	Ours	**Ours-Res50**	**DHQNet**	**HRNet**	**CPD**	**SCRN**
Model Size(MB)	278	142	310	130	48	25
FPS	19	38	5	3	62	32
Input Size	1024 × 1024	1024 × 1024	1024 × 1024	1024 × 1024	352 × 352	352 × 352
	PFSNet	**LDF**	**F3Net**	**ITSD**	**CTD**	**GCPA**
Model Size(MB)	119	96	98	27	25	67
FPS	36	50	33	43	110	50
Input Size	352 × 352	352 × 352	352 × 352	288 × 288	352 × 352	320 × 320

Table 5. Quantitative comparisons with state-of-the-art SOD models trained on DUTS-TR dataset and UHRSD-TR+HRSOD-TR dataset. D: trained on DUTS-TR, UH: trained on UHRSD-TR + HRSOD-TR . The improved metrics are in red and decreased in green fonts.

Method	**HRSOD-TE**				**DAVIS-S**				**UHRSD-TE**				**DUT-OMRON**				**DUTS-TE**			
	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m	F_β^{Max}	MAE	E_ξ	S_m
LDF-D	0.904	0.032	0.919	0.904	0.911	0.019	0.947	0.922	0.913	0.047	0.891	0.888	0.820	0.051	0.873	0.838	0.898	0.034	0.910	0.892
LDF-UH	0.898	0.035	0.908	0.903	0.933	0.016	0.959	0.932	0.924	0.036	0.898	0.913	0.803	0.059	0.858	0.822	0.884	0.040	0.894	0.879
F3Net-D	0.900	0.035	0.913	0.897	0.915	0.020	0.940	0.914	0.909	0.046	0.887	0.890	0.813	0.053	0.871	0.838	0.891	0.035	0.902	0.888
F3Net-UH	0.883	0.041	0.891	0.890	0.893	0.023	0.922	0.907	0.917	0.039	0.896	0.910	0.799	0.065	0.849	0.821	0.871	0.046	0.880	0.871
CTDNet-D	0.905	0.032	0.921	0.905	0.904	0.019	0.938	0.911	0.917	0.043	0.898	0.897	0.826	0.052	0.875	0.844	0.897	0.034	0.909	0.893
CTDNet-UH	0.901	0.031	0.917	0.905	0.905	0.019	0.939	0.917	0.928	0.033	0.902	0.917	0.819	0.054	0.873	0.837	0.888	0.065	0.846	0.846
PFSNet-D	0.911	0.033	0.922	0.906	0.916	0.019	0.946	0.923	0.918	0.043	0.896	0.897	0.823	0.055	0.875	0.842	0.896	0.036	0.902	0.892
PFSNet-UH	0.902	0.036	0.908	0.901	0.946	0.013	0.967	0.942	0.933	0.034	0.904	0.921	0.815	0.061	0.863	0.833	0.886	0.042	0.890	0.881
Our PGNet																				
Ours-D	0.931	0.021	0.944	0.930	0.936	0.015	0.947	0.935	0.931	0.037	0.904	0.912	0.835	0.045	0.887	0.855	0.917	0.027	0.922	0.911
Ours-UH	0.945	0.020	0.946	0.938	0.957	0.010	0.979	0.954	0.949	0.026	0.916	0.935	0.772	0.058	0.884	0.786	0.871	0.038	0.897	0.859

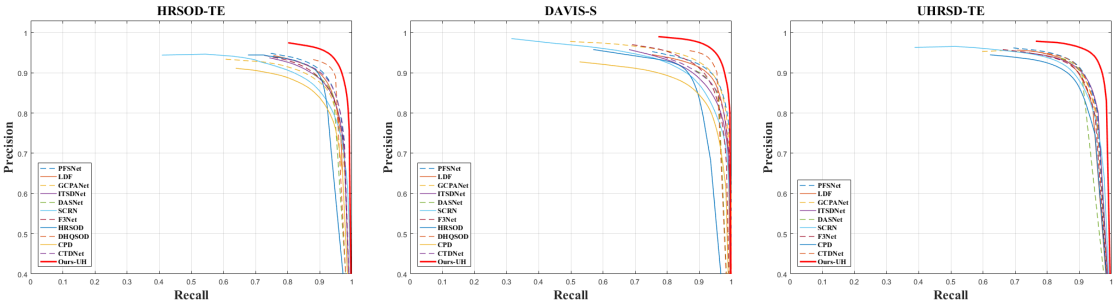

Figure 1. Comparison of PR curves across three high-resolution datasets. Ours-UH is our method trained on mixed high-resolution dataset HRSOD-TR+UHRSD-TR.

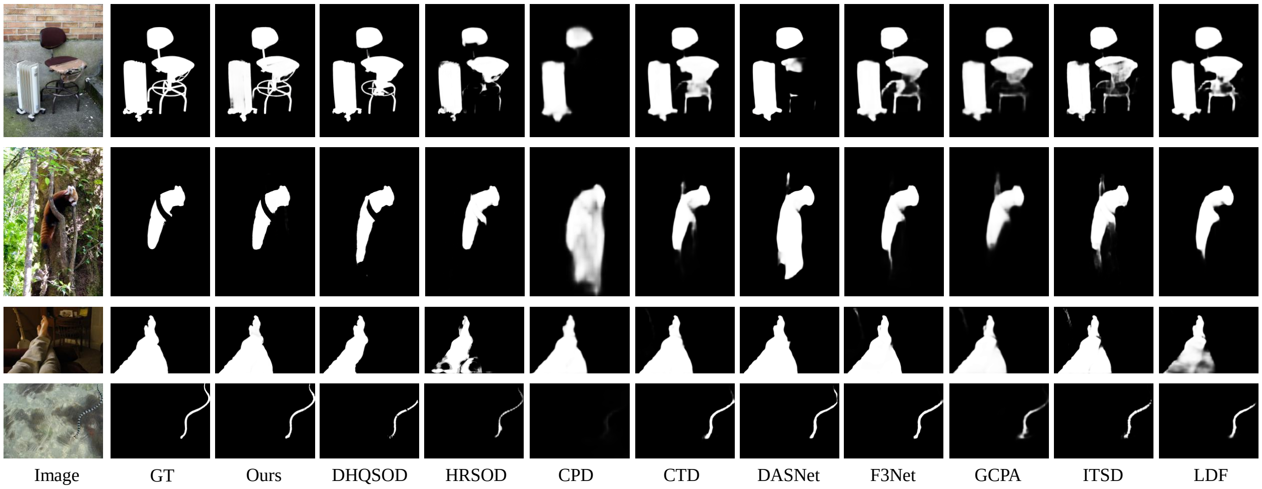

Figure 2. More visual comparison between our method and SOTA methods from HRSOD-TE datset. Best viewed by zooming in.

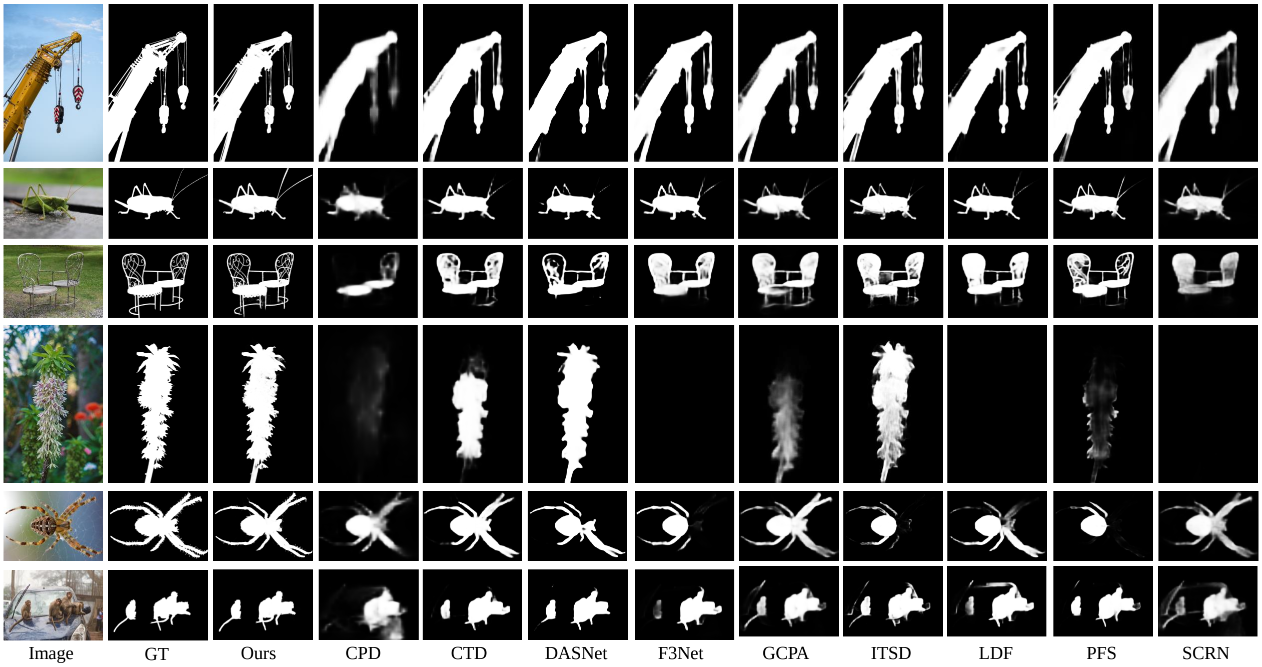

Figure 3. More visual comparison between our method and SOTA methods from UHRSD-TE dataset. Best viewed by zooming in.

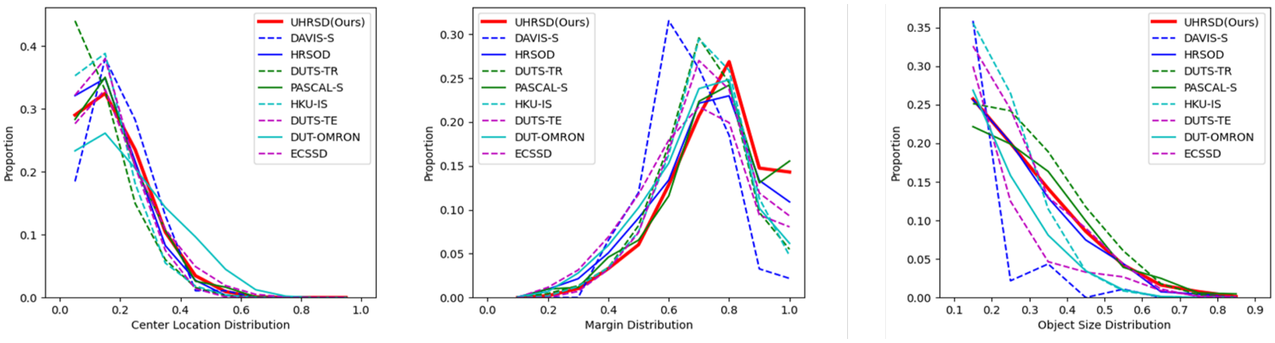

Figure 4. Basic attributes of UHRSD. Best viewed by zoom in.

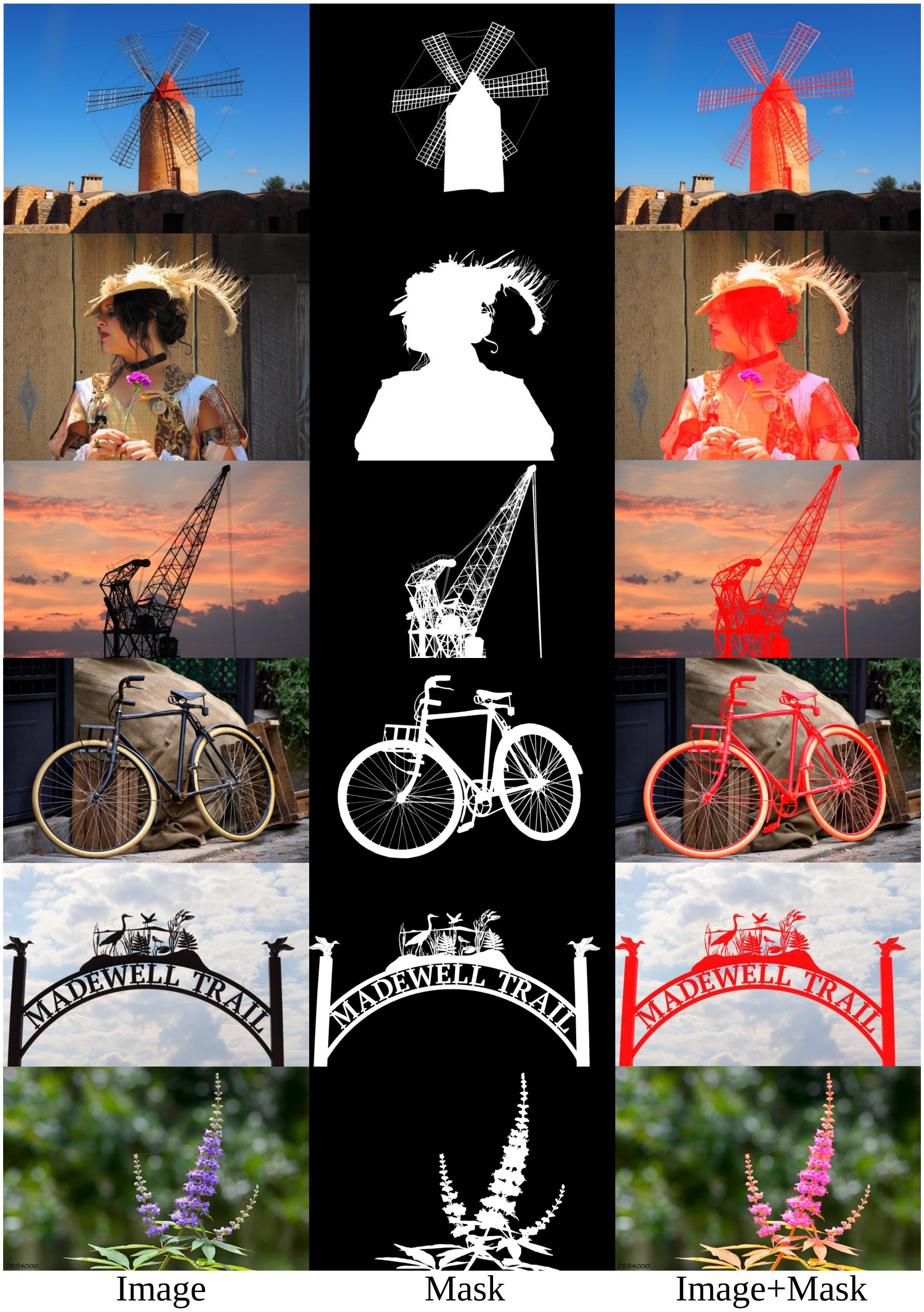

Figure 5. Examples of hardcase images and corresponding annotations from our UHRSD. It can be clearly seen that our UHRSD has challenging salient objects with rich details, and meanwhile has a high level of annotation fineness.

References

[1] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. *arXiv preprint arXiv:2003.00651*, 2020. 2

[2] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017. 1

[3] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. 2016 ieee conf comput vispattern recognit. 2016: 770-778 https://doi. org/10.1109. CVPR, 2016. 1

[4] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 1

[5] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014. 1

[6] 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. *arXiv preprint arXiv:2103.14030*, 2021. 1

[7] Mingcan Ma, Changqun Xia, and Jia Li. Pyramidal feature shrinking for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 2311–2318, 2021. 2

[8] Gellért Máttyus, Wenjie Luo, and Raquel Urtasun. Deeproadmapper: Extracting road topology from aerial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3438–3446, 2017. 2

[9] Lv Tang, Bo Li, Yijie Zhong, Shouhong Ding, and Mofei Song. Disentangled high quality salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3580–3590, 2021. 1, 2

[10] 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

[11] Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: Fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12321–12328, 2020. 2

[12] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13025–13034, 2020. 2

[13] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019. 2

[14] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7264–7273, 2019. 2

[15] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013. 1

[16] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7234–7243, 2019. 1, 2

[17] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1745–1754, 2020. 2

[18] Zhirui Zhao, Changqun Xia, Chenxi Xie, and Jia Li. Complementary trilateral decoder for fast and accurate salient object detection. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4967–4975, 2021. 2

[19] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9141–9150, 2020. 2