Scene Context-Aware Salient Object Detection
Supplementary Material

Avishek Siris[1], Jianbo Jiao[2], Gary K.L. Tam[1], Xianghua Xie[1], Rynson W.H. Lau[3]
Department of Computer Science, Swansea University[1]
Department of Engineering Science, University of Oxford[2] and City University of Hong Kong[3]
a.siris.789605@swansea.ac.uk, jianbo@robots.ox.ac.uk,
{k.l.tam, x.xie}@swansea.ac.uk, rynson.lau@cityu.edu.hk

In this supplementary material, we provide more details on:

- Construction of our proposed dataset (Sec. 1).
- Further statistics of the proposed dataset (Sec. 2).
- Experimental evaluations on existing popular datasets (Sec. 3).
- Additional qualitative results on the proposed dataset with comparison to state-of-the-arts (Sec. 4).

1. Details of the Proposed Dataset

As mentioned in the main paper, existing salient object detection datasets mainly contain images with simple foreground and background, and often consist of very few objects. These images do not well represent the real-world scenes that are generally more complex, containing multiple objects in foreground and background along with rich scene context. To address this problem, we propose a new dataset that contains more challenging images with rich semantic context and multiple objects. These images are closer to the real-world scenarios.

The dataset is built on the MS-COCO [6] dataset, which supplies complex image scenes and semantic segmentation annotations of instances (*Things*) and regions (*Stuff*). We then use the SALICON [3] dataset which provides mouse-based fixation sequence of respective images. These sequences determine the ground-truth salient objects. Our dataset construction process consists of two phases. In the first phase, ground truth salient objects are automatically generated. In the second phase, a manual filtering process is applied to ensure these ground truth salient objects generated from phase one are consistent with respective fixation maps offered by SALICON. This process improves the quality of the dataset.

(1) Automatic phase: Although the SALICON dataset provides mouse-fixation sequence data up to five seconds, we only utilise the first three seconds based on the observations in [2]. Fosco *et al.* [2] show that humans generally gaze from people to other objects (*Things*) during the range of 0-3s. After which more fixations fall onto regions (*Stuff*). We automate the process to define ground-truth salient objects on all 15k images using SALICON fixation data. For each image $I \in \mathbb{R}^{W \times H}$ with spatial dimensions $W \times H$, there are N number of observer fixation data. For each observer $i \in [1, N]$, we augment the fixation sequence to obtain a new fixation sequence F^i. The augmentation includes a) cropping the fixation sequence to at most 3s and b) removing repeated fixations on the same object. We assign a saliency score s_o to an object o, if the j-th fixation $f^i_j \in F^i$ lands on that object.

$$s_o = \sum_i^N \sum_j^T g(f^i_j), \qquad g(f^i_j) = \begin{cases} 1 & \text{if } f^i_j \in P_o, \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where T denotes the number of fixations in F^i and P_o refers to the set of pixels belonging to the segmentation of object

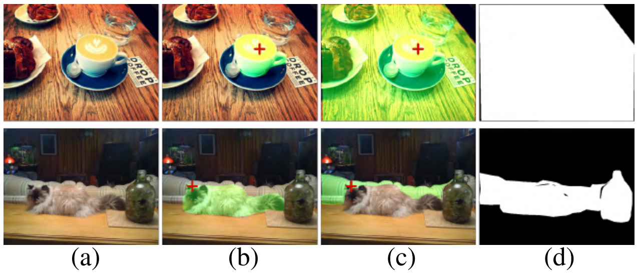


Figure 1: Example images of overlapping segmentations that often cause issues in the automatic ground-truth saliency generation. (a) image, (b) foreground object, (c) background object, (d) resulting incorrect saliency map. Red cross in (b) and (c) correspond to an observer's fixation. Best viewed zoomed in.

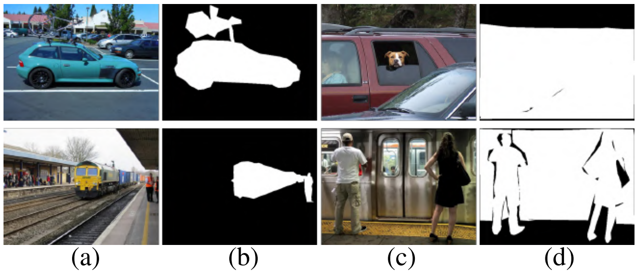


Figure 2: Examples of certain object categories (*e.g.*, car and train) correctly and incorrectly identified as salient in separate scenes, with different object sizes. (a) and (b) are image and corresponding generated saliency map, with those object categories correctly defined as salient, during phase 1 of automatic saliency generation. (c) and (d) are image and resulting saliency map generated with incorrect saliency. The same object categories (car and train) are relatively large and should be considered as background instead.

o. An object is then considered salient in the ground-truth, if its saliency score is greater than half the number of observers for a given image.

We find that the segmentation data in MS-COCO sometimes overlap but no depth information is available to determine their depth orders. It leads to the problem that a saliency score can be assigned to two different objects when a fixation lands on the overlapping regions of the segmentation of both objects. Fig. 1 shows examples of such overlapping segmentations and the resulting incorrect saliency maps. Further, there are object categories that are salient in certain scenarios but are background in others. We observe that the size of these objects are often relatively large when they are background objects. Fig. 2 shows examples (*e.g.*, cars, trains) of such cases and the incorrect saliency maps produced.

To tackle the first issue, we manually go through the dataset and define a set of object categories that are often considered as background objects. We then exclude these objects categories from the dataset. For the second issue, we manually identify these object categories and exclude these objects only if their relative size is greater than a threshold of 60% of the images size. We determined this threshold empirically and observe that the threshold effectively determines if the object is a foreground or background object in many scenarios. We utilize these two conditions to filter through the annotation data and exclude objects before performing the automatic saliency scoring.

After we have generated the ground-truth saliency for all images, we apply another automated filtering step to ensure that the majority of the images contain complex scenes. We observe empirically that a minimum of 4 objects and 2 different object categories produces a good compromise of obtaining complex images whilst maintaining a high number of image count.

(2) Manual phase: This phase ensures that we establish a more consistent ground-truth saliency throughout our dataset. We perform manual filtering by comparing the generated ground-truth saliency map with the saliency fixation maps provided by SALICON. Specifically, we ensure that the salient objects defined in phase-(1) have strong fixation intensities in the corresponding SALICON fixation maps. This allows us to find and remove images with inconsistent saliency, as well as images with high fixation intensity on objects that are not available in the MS-COCO dataset. Examples can be found in Fig. 3 of the main paper.

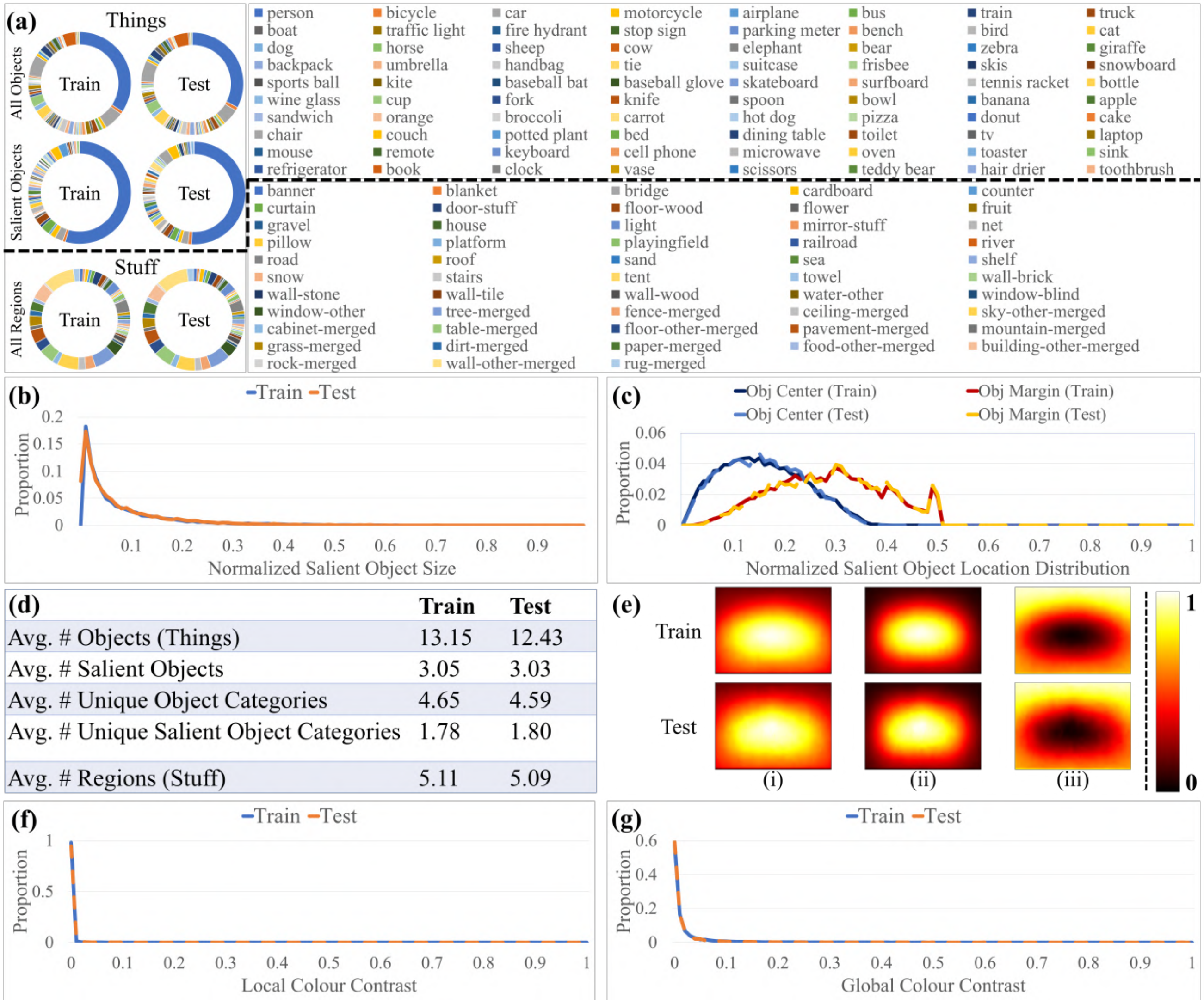
2. Additional Statistics of the Proposed Dataset

Here we provide further statistical information of the proposed dataset. Fig. 3 (a) shows the distribution of object and Stuff region categories in the proposed dataset. We can see that the "person" category is most prevalent throughout the dataset, which is quite expected as photos are commonly taken with people as one of the main targets. In terms of stuff region categories, there is a balance and it is not dominated by a single category. The distribution of the sizes of salient object shown in Fig. 3 (b) indicates that our salient objects are generally of smaller scale with respect to the image size. Fig. 3 (c) displays the distribution of distances between salient objects and the image centre. Fig. 3 (d) reports the statistics of the objects, object categories and stuff regions. The statistics show that our dataset contains images with complex image scenes. Fig. 3(e) visualises the overlay map from locations of all objects (i), salient objects (ii) and Stuff regions (iii), respectively in an intensity map. We also report the contrast of foreground salient objects and surrounding background in local (f) and global (g) views (following [5]). The local contrast compares the colour contrast between foreground (salient object) and background in the local boundary of each salient object. Conversely, global contrast compares the colour contrast of foreground (salient object) and background from the entire image for all salient objects. These graphs show that our GT salient objects have lower colour contrast to their surroundings. This makes the GT salient objects more challenging to detect. This suggests that top-down factors may be more useful for our dataset, while simple low-level contrast is unlikely to be effective.

3. Further Experiments and Generalisability

3.1. Runtime of Training

Table 1 illustrates the runtime for training our model and state-of-the-art methods. All experiments were conducted on the same system (CPU: i7-7700, GPU: GTX 1080Ti, RAM: 32GB). Note, we do not include training runtime for CapSal [15] as we are not able to train their network on our dataset (as explained in the main paper).



(d)	Train	Test
Avg. # Objects (Things)	13.15	12.43
Avg. # Salient Objects	3.05	3.03
Avg. # Unique Object Categories	4.65	4.59
Avg. # Unique Salient Object Categories	1.78	1.80
Avg. # Regions (Stuff)	5.11	5.09

Figure 3: Statistics of the proposed dataset. (a-c) are enlarged graphs from the main paper. (a) presents the distribution of all objects and Stuff regions categories in the dataset. (b) and (c) reports the distribution of size and distance from image centre of salient objects, respectively. (d) average number of all objects (Things), salient objects and regions (Stuff). (e) shows an intensity map from overlays of all individual objects (i), salient objects (ii) and Stuff regions (iii). Local (f) and global (g) colour contrasts of salient objects.

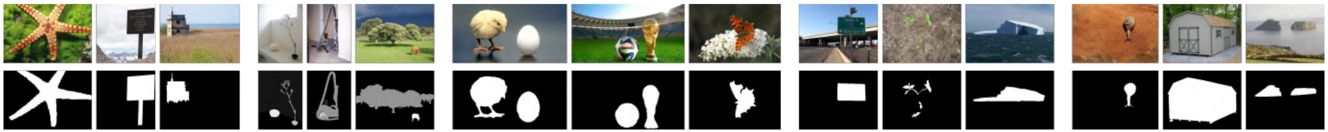
3.2. Existing Datasets

In the main paper, we only evaluate our model and state-of-the-arts on our proposed dataset, because existing popular datasets are not well suited for training our model (*e.g.*, no object instance and semantic segmentation annotations) and mostly do not contain images with more complex real-world scenes. Here we provide further experiments to show the generalisability of our model on these existing datasets.

We carry out evaluation on five common benchmark datasets: ECSSD [13], PASCAL-S [5], HKU-IS [4], DUT-OMRON [14] and DUTS [9]. For fair comparison we use the training set (5534 images) *from our proposed dataset* to train all comparison models and *directly test* on the five datasets. Furthermore, the test images for each of the five datasets are filtered into a new subset that mainly include images that contain object categories defined in our dataset. These are the object categories our model is able to generate saliency prediction. The resulting ECSSD, PASCAL-S, HKU-IS, DUT-OMRON and DUTS thus respectively contains 928, 807, 4177, 3228 and 4338 test images.

3.3. Quantitative Evaluation on Existing Datasets

Table 2 evaluates our technique on five common datasets in comparison to existing state-of-the-arts. We would like to note that the results reported in the table are established



ECSSD [13] PASCAL-S [5] HKU-IS [4] DUT-OMRON [14] DUTS [9]

Figure 4: Example images in common datasets that contain simple scenes. There are only one or very few objects. The object categories are not defined in MS-COCO [6]. Images (top) and corresponding ground-truth saliency (bottom).

Table 1: Runtime for training our model and state-of-the-art methods on our dataset.

Method	Training Runtime
BASNet [8]	36hrs (early stop)
CPD-R [11]	6hrs 40min
PFANet [17]	3hrs 20min
S4Net [1]	1hrs 20min
EGNet [16]	36hrs
SCRN [12]	13hrs
ITSD [18]	2hrs 13min
LDF [10]	5hrs 15min
MINet [7]	10hrs 40min
Ours	38hrs

from the five subsets defined earlier. Fig. 4 shows examples of simple image scenes with salient object categories that are not defined in our dataset. We also note here that CapSal [15] is omitted in this experiment as CapSal is unable to train on our dataset.

The results show that our proposed model is able to outperform state-of-the-arts for certain metrics and datasets with good margin. Particularly, we perform the best on the PASCAL-S for average F-measure and a very close second for E-measure. Our method also outperforms on DUT-OMRON in terms of average F-measure. For the rest, our model is able to produce quite comparable results to the best method for each dataset-metric combination.

We find three reasons for our method not always outperforming: a) Our method requires instance data for training. However, such data is not available in existing datasets. b) Our method mainly focuses on multiple salient object detection in complex scenes. As the other datasets contain mostly images of very few objects, it is difficult for our model to explore/leverage object and scene context relationships. c) Some test images also contain GT salient objects of categories not defined in our dataset. Our method thus may not recognise such objects. Despite the above constraints, our method still show some ability to capture and predict parts of those undefined objects, albeit not always outperforms.

Overall, our model can generalise quite well on common datasets, even when it is directly tested on those datasets. We believe that given sufficient data (*i.e.*, object instance and image segmentation), our model could be trained on existing datasets and thus potentially perform better.

3.4. Qualitative Evaluation on Existing Datasets

In addition to the quantitative results on existing datasets, we present qualitative visual comparisons in Fig. 5. It demonstrates that our method is able to utilise scene context information, separate salient objects from surrounding background and reduce false saliency from distractors.

4. Additional Results on Our Dataset

Additional qualitative comparison with state-of-the-arts on our dataset can be found in Fig. 6 and Fig. 7.

References

[1] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, pages 6103–6112, 2019.

[2] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4473–4482, 2020.

[3] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015.

[4] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[5] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014.

[6] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

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

[8] 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.

[9] 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 *CVPR*, pages 136–145, 2017.

Table 2: Quantitative comparison with state-of-the-art methods on five existing datasets to show *generalisability*. Note that we *train all comparison methods on our proposed dataset*, and *test directly* on these existing datasets for fair comparison. Furthermore, the test images for each dataset are filtered into a new subset that mainly include images that contain object categories defined in the proposed dataset. The number of images of the new subsets and the original test sets are respectively indicated next to the dataset. avgF refers to the average F-measure taken and E_m refers to E-measure. Red, Blue and Magenta respectively indicate the top 3 performance.

	ECSSD (928/1000)			PASCAL-S (807/850)			HKU-IS (4177/4447)			DUT-OMRON (3228/5168)			DUTS (4338/5019)		
Method	avgF ↑	E_m ↑	MAE ↓	avgF ↑	E_m ↑	MAE ↓	avgF ↑	E_m ↑	MAE ↓	avgF ↑	E_m ↑	MAE ↓	avgF ↑	E_m ↑	MAE ↓
BASNet [8]	0.748	0.834	0.092	0.740	0.827	0.092	0.800	0.884	0.073	0.736	0.859	0.071	0.694	0.825	0.085
CPD-R [11]	0.872	0.894	0.058	0.837	0.859	0.069	0.862	0.925	0.049	0.764	0.876	0.060	0.781	0.872	0.058
PFANet [17]	0.769	0.848	0.111	0.730	0.803	0.120	0.746	0.859	0.098	0.595	0.756	0.111	0.609	0.758	0.112
S4Net [1]	0.789	0.853	0.085	0.738	0.790	0.111	0.780	0.882	0.069	0.613	0.766	0.109	0.647	0.782	0.101
EGNet [16]	0.819	0.855	0.070	0.826	0.851	0.066	0.756	0.839	0.072	0.632	0.748	0.058	0.713	0.814	0.060
SCRN [12]	0.856	0.888	0.068	0.819	0.846	0.072	0.849	0.918	0.058	0.741	0.866	0.066	0.759	0.861	0.063
ITSD [18]	0.857	0.899	0.051	0.815	0.864	0.061	0.872	0.933	0.044	0.769	0.880	0.058	0.786	0.880	0.053
LDF [10]	0.875	0.888	0.057	0.848	0.858	0.062	0.875	0.927	0.046	0.749	0.860	0.068	0.787	0.869	0.057
MINet [7]	0.882	0.901	0.051	0.847	0.868	0.059	0.875	0.930	0.045	0.768	0.877	0.059	0.806	0.889	0.049
Ours	0.859	0.880	0.070	0.860	0.866	0.067	0.858	0.899	0.059	0.788	0.873	0.063	0.804	0.866	0.066

[10] 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.

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

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

[13] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

[14] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013.

[15] Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *CVPR*, pages 6024–6033, 2019.

[16] 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.

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

[18] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, pages 9141–9150, 2020.

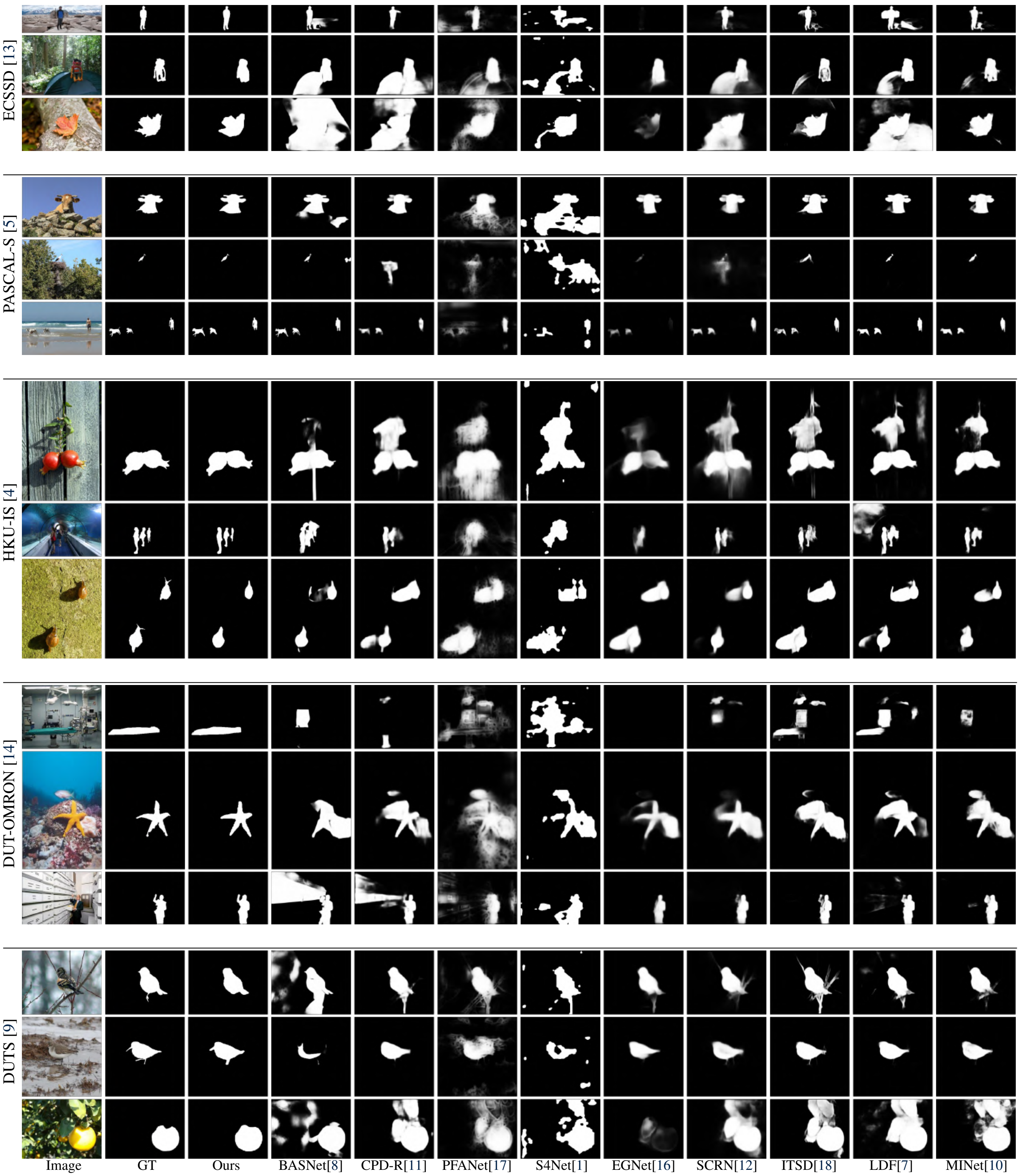


Figure 5: Visual comparison of our proposed method with state-of-the-arts on existing popular datasets.

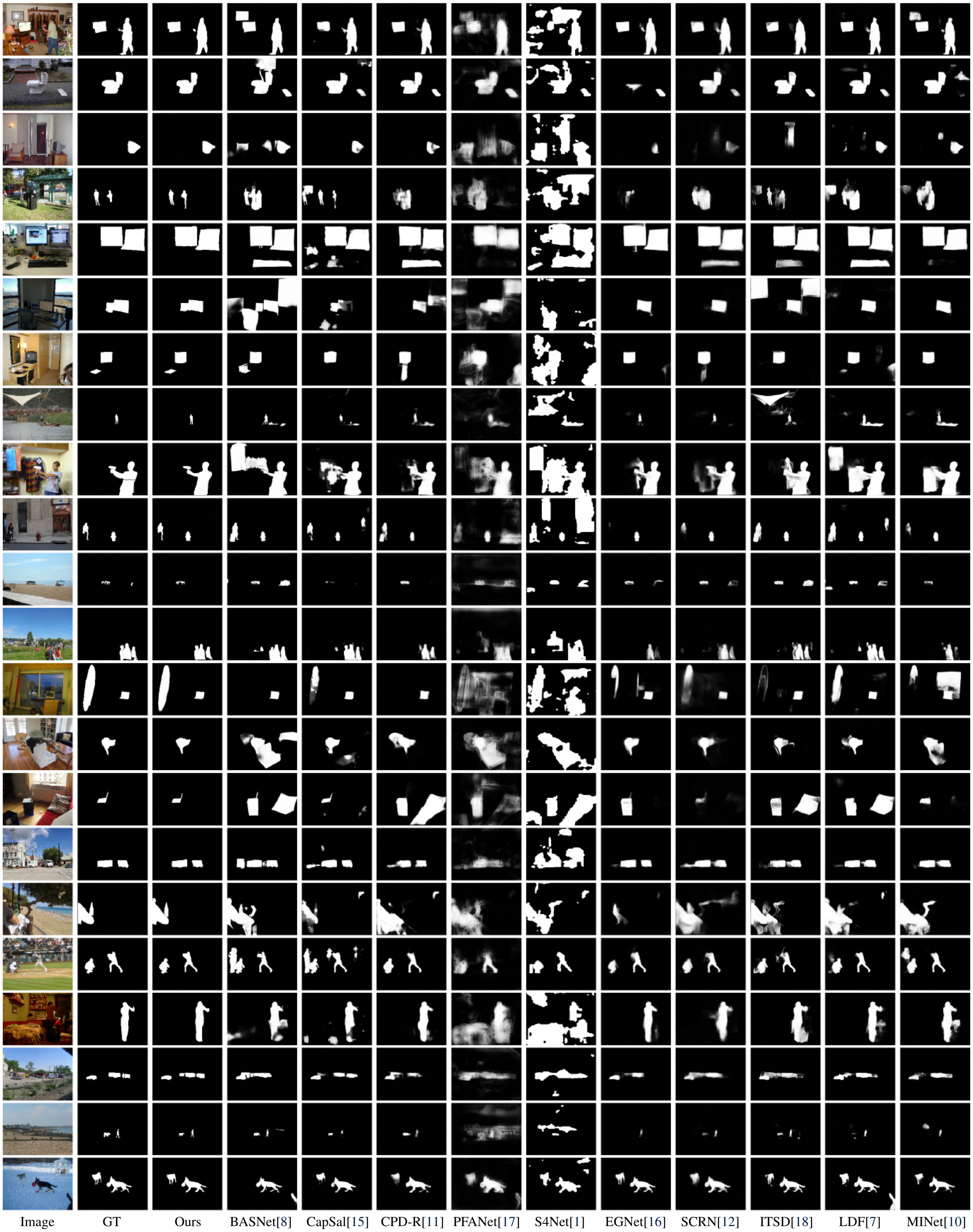


Figure 6: Additional qualitative comparison of our method with ten other state-of-the-art saliency methods on our proposed dataset.

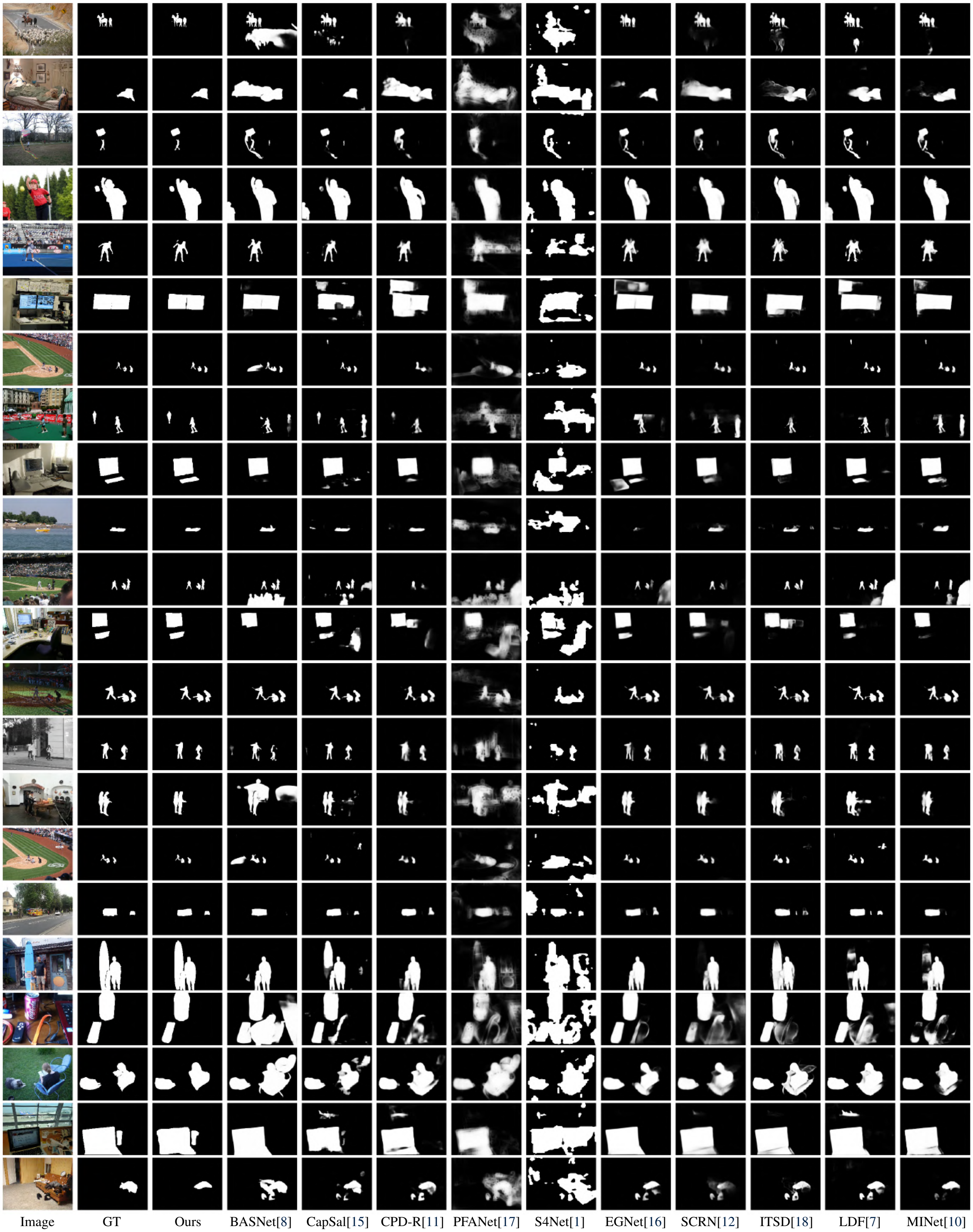


Figure 7: Additional qualitative comparison of our method with ten other state-of-the-art saliency methods on our proposed dataset.