Advancing Saliency Ranking with Human Fixations: Dataset, Models and Benchmarks Supplementary Material

Bowen Deng[1], Siyang Song[2], Andrew P. French[1], Denis Schluppeck[3], Michael P. Pound[1]*
[1]School of Computer Science, University of Nottingham, UK
[2]School of Computing and Mathematical Sciences, University of Leicester, UK
[3]School of Psychology, University of Nottingham, UK
{bowen.deng, andrew.p.french, denis.schluppeck, michael.pound}@nottingham.ac.uk, ss2796@cam.ac.uk

In this supplementary material, we provide additional detail of our proposed dataset SIFR and proposed method QAGNet. These include:

- Additional statistical analysis of our proposed SIFR dataset in Sec. 1.
- Further implementation details and ablation analysis of our proposed QAGNet in Sec. 2.
- More qualitative comparison images between our proposed QAGNet and other saliency ranking methods in Sec. 3.

1. Proposed SIFR Dataset

We propose SIFR, which to the best of our knowledge is the first large-scale saliency ranking dataset using real human fixations. Here, we provide further details of our proposed dataset.

1.1. Fixation Stability Over Time

In our proposed dataset, we conduct a 'freeviewing' task using an eye-tracking system to observe the viewing habits of a group of eight participants. As the process of gaze recording was spread over six months, it is important to ensure that fixations across the participants were consistent during this time. Each individual participant completed their recording sessions within 3 weeks. Analysis of common eye movement metrics such as average fixations per second are stable across the 3 week period. Fig. 1 shows fixation rates are stable both within a subset of 200 images, and across the data collection period.

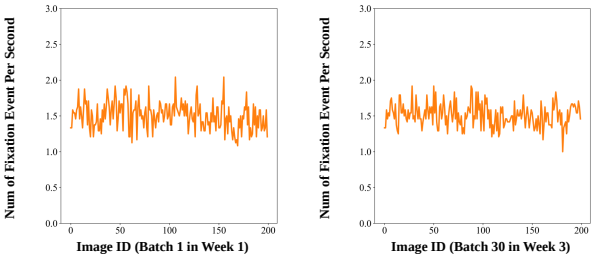

Figure 1. Average number of effective fixation events per second in batch 1 (week 1) and batch 30 (week 3) across 8 participants.

1.2. Statistics on Instance Number

In Tab. 1, we present summary statistics on the instance number for the ASSR dataset [6], IRSR dataset [5] and our proposed dataset. As can be seen, our proposed dataset has the highest average instance number (6.22) per image and the same median instance number (5) per image as the ASSR dataset. The more salient instances per scene potentially bring more challenges to saliency ranking models.

1.3. Statistics on Category Distribution

In Fig. 2, we provide the category distribution statistics on our proposed SIFR dataset and ASSR dataset in 12 supercategories out of 80 classes in MS-COCO dataset [4]. It can be found the 'person' category occupies the highest percentage in both datasets, followed by the 'animal' class in our dataset and 'vehicle' class in ASSR dataset respectively. In our dataset, there are many challenging scenes containing crowds of people and groups of animals, raising the percentage of 'person' and 'animal' categories above that of than other classes. Differences in salient objects between our SIFR dataset and ASSR dataset may also be due to the different data modalities: real human fixations captured in our dataset rather than mouse-trajectory data utilized in ASSR, which may yield distinct salient object distributions across categories.

*Corresponding Author.

Datasets	Avg Instance # Per Image	Median Instance # Per Image
ASSR [6]	4.30	5
IRSR [5]	3.36	3
Ours	6.22	5

Table 1. Statistics for three SRD datasets on average and median instance number per image.

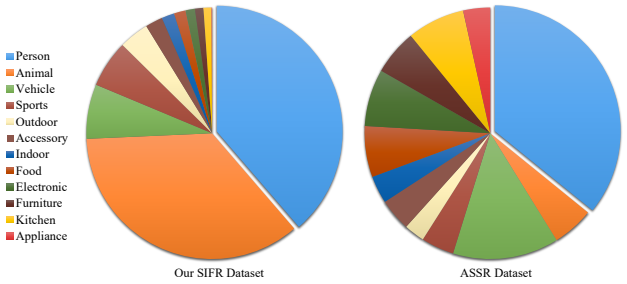

Figure 2. Category distributions of ground-truth salient objects on our proposed SIFR dataset and ASSR dataset [6]. Note there is no category information on publicly available IRSR dataset [5].

1.4. Statistics on Foreground Size

In Tab. 2, We present additional detailed foreground size statistics to supplement to Fig.3 (b) in the main paper. We define foreground size as the total percentage of pixels within all salient objects for each image. Our dataset comprises more images with small foreground size, potentially bringing more complexity to saliency ranking models. We also show both the maximum and minimum foreground size ratios across each dataset. Both ASSR and IRSR dataset contain several images with the max foreground size ratio being 1.00. This phenomenon typically arises due to images that are contaminated by the background being labeled as a salient object, as depicted in Fig. 3. In these cases, the salient object segmentation is not limited to the actual foreground objects but also includes the background that contributes to the ranking ground truth. This influences the task of correctly identifying and ranking the actual salient objects and may confound the learning of relative saliency ranking models, potentially undermining their generalizability and performance in real-world applications. Such cases also reflect that both ASSR dataset and IRSR dataset highly rely on the accuracy of MS-COCO [4] and SALICON [3] datasets. In our SIFR dataset, we remove, add or refine annotations as appropriate to ensure that all observed salient objects possess high-quality annotations without background.

2. Proposed QAGNet

We provide a strong baseline, QAGNet, for our proposed SIFR dataset. The source code will be made publicly available after the review process. Here, we provide further ablation studies on QAGNet.

Datasets	#Images	#Images of Foreground Sizes			Max Ratio	Min Ratio
		Large	Medium	Small		
ASSR [6]	11500	4062	5855	1583	1.00	0.01
IRSR [5]	8988	3818	4535	635	1.00	0.01
Ours	8389	1292	5202	1895	0.84	0.01

Table 2. Statistics for three SRD datasets on foreground size and foreground size ratio. Large: (foreground size $\geq$ 30%), Medium: (5% $\leq$ foreground size $\leq$ 30%), Small: (foreground size $\leq$ 5%).

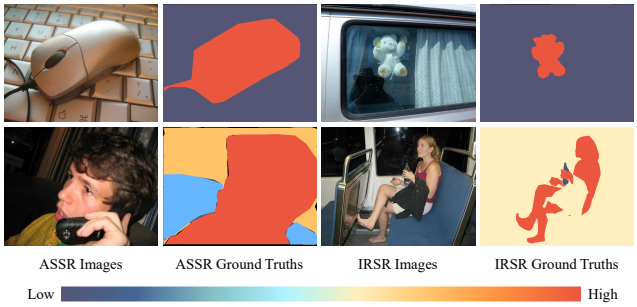

Figure 3. Examples in ASSR dataset [6] and IRSR dataset [5], where that ground truths are contaminated by the background.

2.1. GAT Attention Head Number in GRG

We apply GAT [8] for edge calculations and feature aggregation in our proposed tri-tiered nested graph. In Tab. 3, we present the analysis of the GAT attention head number K within the Global Relationship Graph (GRG) in hidden and output layers, where the GRG is responsible for incorporating multi-scale features and providing robust ranking-aware information to the ranking head. In each Single Scale Graph (SSG) and Multi-Scale Graph (MSG), we conduct a single GAT layer, while in GRG, GAT layers are designed to independently replicate K times to capture enriched features with the outputs feature-wise aggregated. Considering SOR metric does not penalize the missing objects, we choose one-head and eight-head GAT in our hidden layers and output layer respectively (setting III), which demonstrate the best performance in SA-SOR and MAE. We note here, however, that our network design is robust to this choice of parameters, and shows good performance across the range.

2.2. Dropout Rate in GAT

We explore different dropout rate settings in GAT layers in Tab. 4 for a one hidden layer QAGNet. We choose dropout rate 0.2 (setting III) in our proposed QAGNet considering it provides slightly higher performance in SASOR.

Setting	GAT Head Number K in GRG		SASOR ↑	SOR ↑	MAE ↓
	Hidden Layer	Output Layer			
I	1	1	0.6057	0.7715	0.0440
II	1	4	0.6042	0.7721	0.0443
III	1	8	**0.6086**	0.7736	**0.0439**
IV	1	16	0.6019	0.7654	0.0446
V	4	8	0.6066	0.7737	0.0442
VI	8	8	0.6073	**0.7751**	0.0441

Table 3. Ablation analysis on the GAT [8] attention head number in GRG based on one hidden layer QAGNet.

Setting	GAT Dropout Rate	SASOR ↑	SOR ↑	MAE ↓
I	0	0.6069	0.7701	0.0441
II	0.1	0.6070	0.7714	0.0441
III	0.2	**0.6086**	**0.7736**	0.0439
IV	0.3	0.6078	0.7722	**0.0438**
V	0.4	0.6049	0.7684	0.0442

Table 4. Ablation analysis on the dropout rate in GAT [8] based on one hidden layer QAGNet.

Setting	Initialization Method	SASOR ↑	SOR ↑	MAE ↓
I	Random	0.6106	0.7872	0.0438
II	Average	**0.6119**	**0.7899**	**0.0437**

Table 5. Ablation analysis on the initialization method for feature representatives based on two hidden layer QAGNet.

2.3. Representatives Initialization Method

We leverage different query features from a transformer detector into a novel graph architecture for saliency ranking detection. In this process, our QAGNet gradually learns the feature representatives $\{R^{32}, R^{64}, R^{128}\}$ in Single Scale Graphs (SSGs) and Z in Multi-Scale Graphs (MSGs). In Tab. 5, we explore the different initialization methods for these feature representatives, where setting I adopts a random initialization method while setting II applies the average method as used in our model. Different initialization methods have relatively similar performance, where setting II is slightly better. This demonstrates the robustness of our proposed method using different initialization methods.

3. Further Qualitative Comparison

In the main paper, we have presented the qualitative comparison between our proposed QAGNet and other saliency ranking methods (RSDNet [2], ASSR [6], IRSR [5], SOR [1] and OCOR [7]) on our proposed SIFR dataset. Here, we show further qualitative comparison on IRSR [5] dataset and ASSR [6] dataset in Fig. 4 and Fig. 5 respectively. Across these datasets we see a variety of diverse scenarios, and we can see that the performance of our proposed QAGNet remains strong. It can be seen that our proposed method adapts well to the varied saliency ranking approaches used across the datasets, where the ASSR dataset uses human attention shift, and the ISSR dataset uses maximum value within the saliency map per instance.

References

[1] Hao Fang, Daoxin Zhang, Yi Zhang, Minghao Chen, Jiawei Li, Yao Hu, Deng Cai, and Xiaofei He. Salient object ranking with position-preserved attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16331–16341, 2021. 3

[2] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7142–7150, 2018. 3

[3] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015. 2

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2

[5] Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):8321–8337, 2021. 1, 2, 3

[6] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Inferring attention shift ranks of objects for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12133–12143, 2020. 1, 2, 3

[7] Xin Tian, Ke Xu, Xin Yang, Lin Du, Baocai Yin, and Rynson WH Lau. Bi-directional object-context prioritization learning for saliency ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5882–5891, 2022. 3

[8] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 2, 3

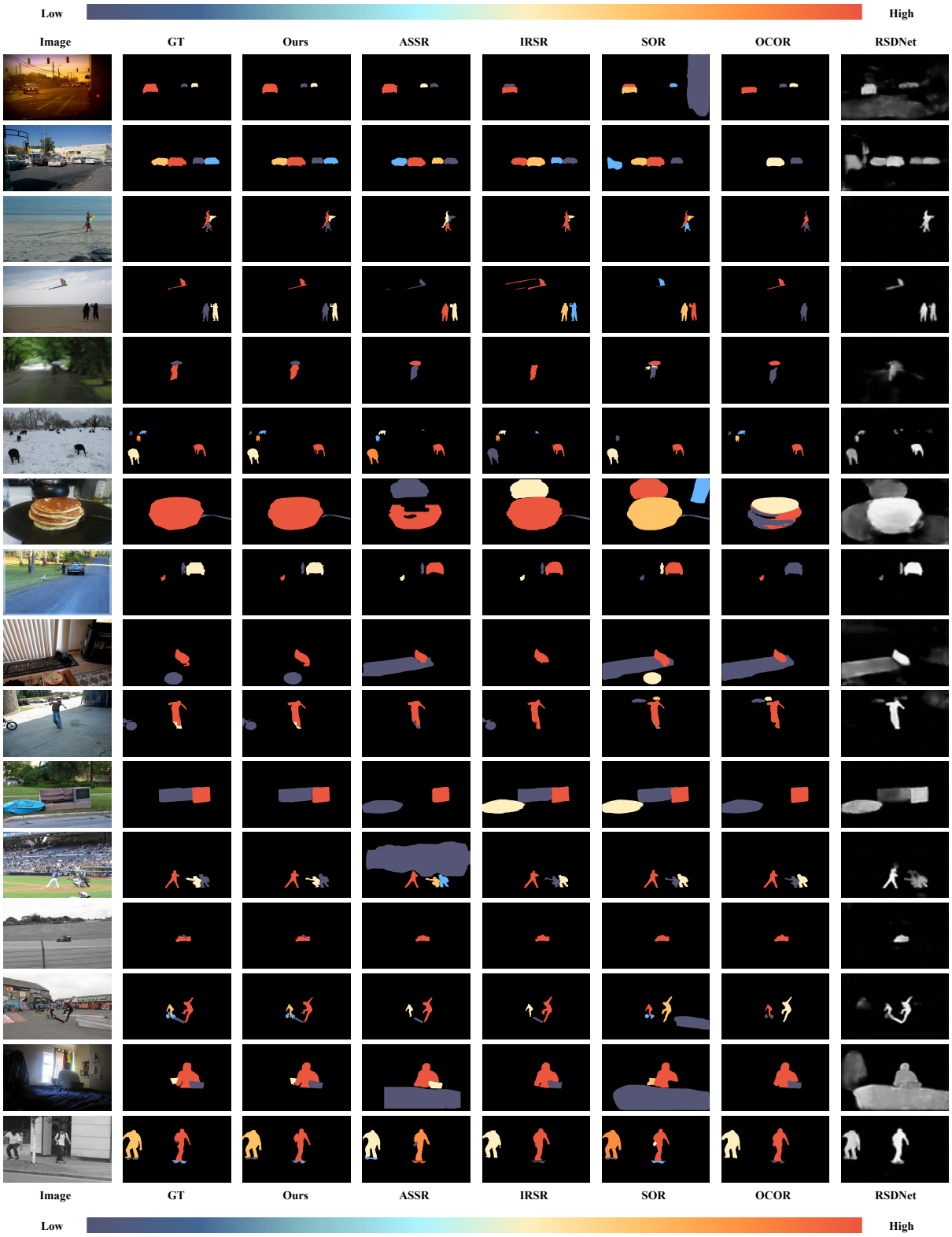

Figure 4. Qualitative comparison on IRSR dataset.

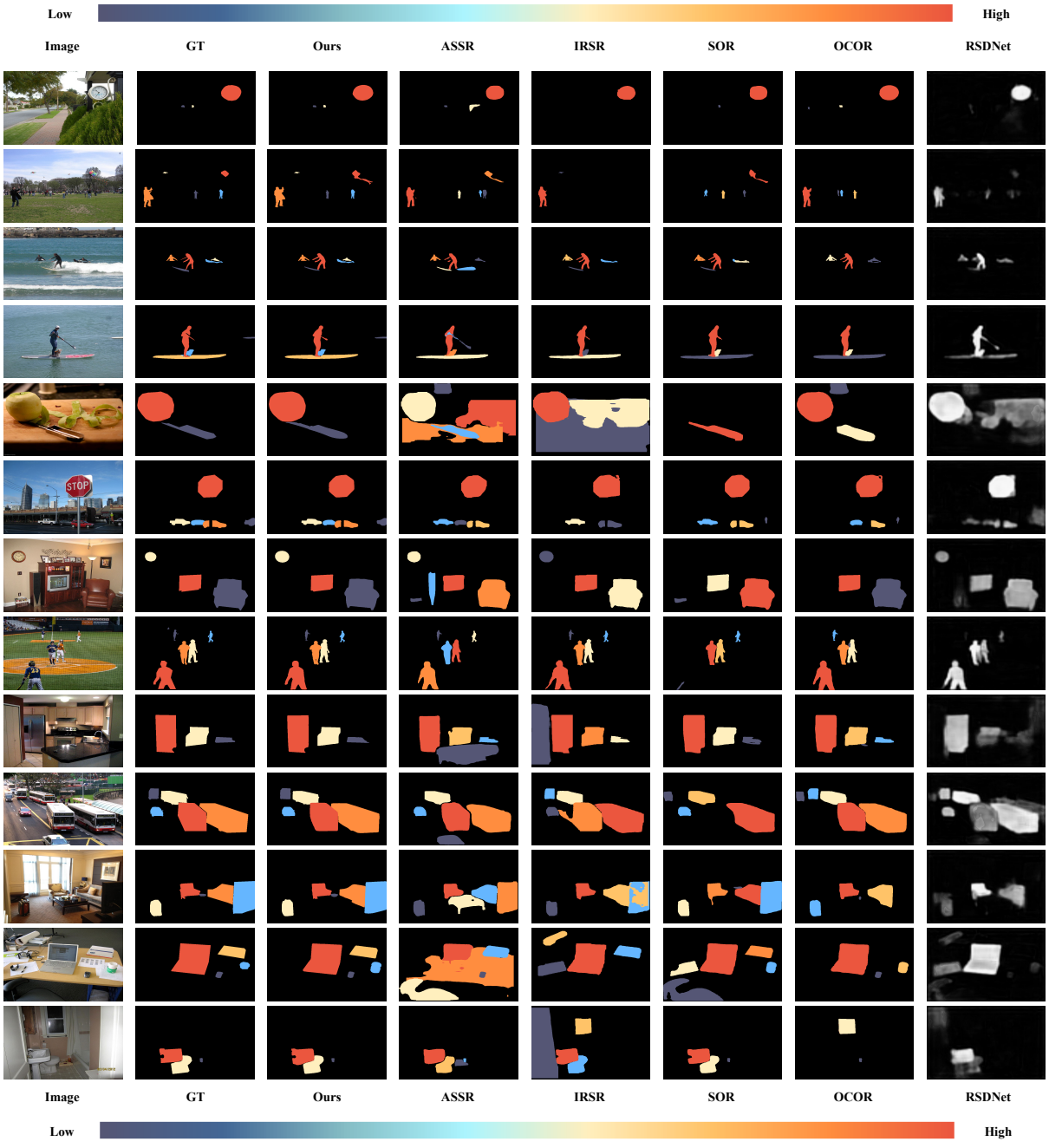

Figure 5. Qualitative comparison on ASSR dataset.