This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

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

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

Abstract

Saliency ranking detection (SRD) has emerged as a challenging task in computer vision, aiming not only to identify salient objects within images but also to rank them based on their degree of saliency. Existing SRD datasets have been created primarily using mouse-trajectory data, which inadequately captures the intricacies of human visual perception. Addressing this gap, this paper introduces the first large-scale SRD dataset, SIFR, constructed using genuine human fixation data, thereby aligning more closely with real visual perceptual processes. To establish a baseline for this dataset, we propose QAGNet, a novel model that leverages salient instance query features from a transformer detector within a tri-tiered nested graph. Through extensive experiments, we demonstrate that our approach outperforms existing state-of-the-art methods across two widely used SRD datasets and our newly proposed dataset. Code and dataset are available at https://github.com/EricDengbowen/QAGNet.

Figure 1. Comparison between our proposed dataset, ASSR [30] dataset and IRSR [18] dataset, where all SRD ground truths are colorized. The SOD task learns to detect pixel-level salient objects without distinction, while the SRD task aims to assign varying degrees of saliency to each detected object (columns (b) and (c)). Our proposed dataset is curated based on eye-tracking fixations while both other two datasets are based on mouse-trajectory.

1. Introduction

Salient Object Detection (SOD) aims to identify and segment the most visually prominent objects within an image. Recent advances in deep learning models [17, 19, 40, 41] have achieved notable results in this domain. Saliency Ranking Detection (SRD), in which salient objects are also ranked on the degree of saliency (Fig. 1 (b) and (c)), is an emerging task that aims to better reflect the ability of humans to focus on multiple objects that elicit different level of interest. This provides added value to many downstream tasks, such as image captioning [36, 39], image cropping [1], and autonomous driving [23].

Recent works in SRD have leveraged the COCO-SalRank [12], ASSR [30] and IRSR [18] datasets. Saliency ranking in these datasets is computed based on mouse-movements from the SALICON [10] dataset. These mouse trajectories are captured as human observers explore blurred scenes from the MS-COCO [16] dataset. All current SRD datasets assume that mouse pointing is a suitable proxy for real eye-fixation data that can be captured from gaze measurements under controlled conditions. There are several issues which make this simplifying assumption problematic. First, mouse-pointing actions are under voluntary control, whereas a large proportion of the fixation shifts (saccadic eye movements) are reflexive and will therefore reflect different aspects of saliency [29] (Fig. 1 (c), (d) and (e)). Second, mouse-pointing actions and shifts in eye fixation are likely processed in different reference frames in the human brain. Mouse pointing behaviour will therefore also reflect biases and limitations related to transforms between these two response modalities [3]. In addition, current datasets

*Corresponding Author.

are also dependent on the accuracy of MS-COCO annotations, which can result in missing or incorrect instances.

In this paper, we present a large-scale instance-level SRD dataset derived from genuine human eye-tracking fixations aiming to train and evaluate models that are intended to capture and predict human visual attention. Using images drawn from MS-COCO, focusing exclusively on scenes containing at least three objects, we calculate human fixation duration across multiple users per scene. We then systematically improve the annotations across all images, adding missing objects, separating joined objects, and refining existing objects. Unlike other datasets, we place no limit on the number of potential salient instances in a given scene, instead referring to the number of fixations as a guide for whether an object is sufficiently interesting to be included.

We also develop a strong baseline method for SRD, evaluated against popular methods across existing datasets and our new dataset. When ranking saliency, existing methods often rank only highly salient objects proposed by a detector while neglecting those with lesser saliency [6, 18, 30], or they constrain outputs to a predetermined maximum number [6, 9, 30, 31]. We argue that less salient objects still contribute valuable information to the saliency ranking problem, and selecting a low fixed number of outputs is a simplification that does not reflect real human visual perception. Our approach builds a novel GNN [28] over a query-based transformer [2], computing relative saliency ranking of high numbers of objects per scene, and offers state-of-the-performance across all SRD datasets. Our main contributions are:

- We present the first large-scale relative saliency ranking dataset based on the natural viewing patterns of human observers. The dataset comprises only challenging multi-object scenes, with detailed instance-level annotations for all salient objects.
- We provide a strong end-to-end baseline on this dataset namely QAGNet. We structure the query features of a transformer detector within a novel tri-tiered nested GNN to calculate the relative rankings for up to 200 objects.
- Our experimental results demonstrate the effectiveness of different modules in QAGNet and our approach shows a substantial jump in performance above other competing methods on two widely used datasets and our newly created dataset.

2. Related Work

Salient Object Detection (SOD) is the problem of highlighting the most visually interesting or important objects in scenes. Similar to foreground segmentation, popular approaches [17, 19, 24, 34, 35, 40, 41] and datasets [14, 33, 37, 38] for SOD do not distinguish between different salient objects, instead producing a binary segmentation across all salient regions. Recently, the extended task

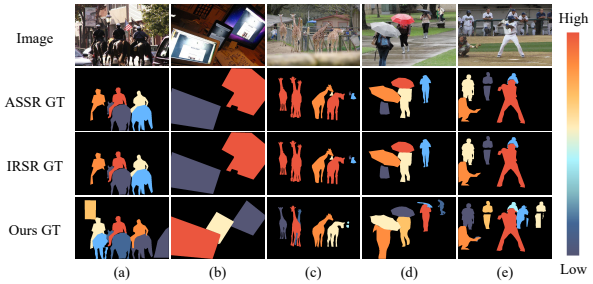

Figure 2. Limitations observed in ASSR and IRSR datasets.

of Saliency Ranking Detection (SRD) aims to identify and then rank objects based on their perceived level of saliency. This task aims to better reflect the ability of human observers to perceive relative importance in different objects within their visual field.

As a new task, there exist only a limited number of datasets for SRD. Islam *et al*. [9] repurposed the Pascal-S [15] salient object detection dataset to serve as a training set for a saliency ranking task. This dataset has seen limited use for SRD, as it comprises only 850 images, many of which (40.4%) contain only a single salient object. COCO-SalRank [12], and the widely used ASSR [30] and IRSR [18] all combine mouse-tracking with MS-COCO polygons into ranked salient instances. COCO-SalRank uses complex hand-crafted rules to produce saliency rankings. ASSR uses human attention shift, ranking instances examined earlier than others as more salient. IRSR utilises the maximum value within each instance based on the fixation heatmaps. In all cases, the underlying fixation maps are generated based on mouse-trajectory information from SALICON, which is fundamentally different from real human gaze. The datasets are also dependent on the accuracy of MS-COCO annotations, for which saliency was not considered as a downstream task. Scenes may miss annotations for key salient objects (e.g. advertising boards and a screen in Fig. 2 (a) and (b)). Objects may be incorrectly grouped together and assigned the same saliency ranking (Fig. 2 (c)), or incorrect objects may be annotated (Fig. 2 (d)). In some scenes (e.g. Fig. 2 (e)), imposing an arbitrary limit on the number of salient objects may limit the ability to model human perception.

A number of approaches to SRD have been developed and trained on the above datasets. Islam *et al*. [9] modelled SRD as a pixel-level relative saliency problem, returning a 2D map of salient areas. SRD is more naturally considered as an object-level problem, and more recent methods have taken this object-centric approach. Siris *et al*. [30] proposed a model leveraging bottom-up and top-down attention to infer attention shift. Liu *et al*. [18] employed improved Mask R-CNN [8] and graph reasoning [28] to solve the SRD task. Fang *et al*. [6] embedded coordinates of objects as posi-

tional information in an end-to-end framework. Tian *et al*. [31] proposed a bi-directional object-context method involving both spatial attention and object-level attention. Although progress has been made, most existing approaches consider only a limited number of potential salient objects or constrain outputs to a predetermined maximum number.

3. Proposed Dataset

In this section we present our novel dataset we call Saliency Instance Fixation Ranking (SIFR).

3.1. Image Selection

We draw images from the MS-COCO dataset, which contains annotations for a variety of classes in diverse scenes. We first select only images that contain three or more annotated foreground objects, rather than two in ASSR and IRSR. We restrict our dataset in this way to ensure a sufficiently varied and complex visual environment to test the performance of saliency ranking models.

3.2. Gaze Recording and Fixation Filtering

We next proceed with a task of 'freeviewing' using an eye-tracking system, observing the viewing habits of a group of eight participants. Gaze tracking is performed using a Tobii Pro Nano (Tobii, SE) set to a sampling frequency of 60 Hz. Each image is resized proportionally to fit full screen resolution, which is presented to subjects for a fixed duration of 3 seconds. To maintain eye-tracking accuracy, a recalibration procedure is performed every 200 images. To relieve eye fatigue, participants are offered the opportunity to rest or pause the gaze recording process at these same 200-image intervals. It should be noted that this procedure is not a swift one; the process of gaze recording was spread over six months. During data collection, we monitored levels of attention over recording sessions and between sessions to ensure consistency.

Raw gaze information from each participant is then processed to identify areas of focus. We use a velocity-based method [27] to group points into distinct fixation events, in which an observer is focused on an object [25]. Saccades, rapid eye movements between objects, are removed automatically by our approach. We remove all fixations with a duration of less than 200ms, based on the average duration of fixations presented in [25]. We also remove the first captured fixation from each participant to reduce centre bias. Once filtered, we append the fixation points for all eight participants for each image.

3.3. Salient Objects Threshold and Annotation

We aim to identify and annotate each observed salient instance using the best combination of existing MS-COCO annotations, Mask-RCNN and human annotation. Within each scene, MS-COCO will provide annotations of some objects, however these may not be salient, or salient objects may not be annotated. We remove, add or refine annotations as appropriate to ensure that all observed salient objects possess high quality annotations. We first calculate the mean number m of fixation points lying within existing polygons and the corresponding standard deviation σ. We then apply a threshold, $N_{\text{fixation}} \geq m - \sigma$, where N_{fixation} denotes the number of fixation points inside a polygon. Existing polygons that contain a sufficient number of fixation points are automatically included within the dataset. We cluster all remaining fixation points using DBSCAN [4], and if these clusters meet the same threshold criteria above, we mark these as a potentially unannotated object. For images in the MS-COCO test set, with no annotations, we preform the same process, but using initial estimates for polygons from Mask R-CNN.

The ASSR and IRSR datasets only consider objects already correctly annotated in MS-COCO. We instead re-annotate any poor or missing segmentation. Ten participants were involved in this labelling task, applying the following steps: **i)** Refine the boundaries of existing salient objects and assign classes to these objects. As this dataset is specifically for the SRD task, we utilise 12 superclasses out of the 80 available ones in MS-COCO. **ii)** Create instance-level polygons for any unannotated objects. **iii)** Confirm that three or more instance-level polygons remain, and if not, the images were re-checked to be either preserved or removed. Salient objects are ranked according to the number of fixation points in each instance. Following [18], we uniformly assign different saliency values to salient objects based on ranking order. Fig. 4 shows three examples of our dataset. The final dataset comprises 8389 images, with a split of 6701 training images and 1688 testing images.

3.4. Statistical Analysis

Tab. 1 shows numerical comparison between ASSR, IRSR and our proposed dataset. Every image within our dataset contains three or more salient objects without any arbitrary cap, and our dataset contains the most instances among current SRD datasets. We also consider instance scale, defined as the percentage of an image comprising each instance. Our proposed dataset contains more small objects, potentially bringing more complexity to SRD models.

Relative Saliency Ranking By Category: Fig. 3 (a) shows relative saliency ranking across different categories. This is derived by normalizing each instance's rank (where a higher value indicates greater saliency) by the total number of instances within each image, category-wise averaged across all images. In our dataset, 'person' is the most salient category, which matches the findings in [11, 18]. We also observe differences in other categories, which are potentially caused by the different data sources used (mouse-tracking

Datasets	#Images	#Instances	Images % of Various Salient Instance Numbers Per Image									Instances % of Scales		
			2	3	4	5	6	7	8	9	10+	Large	Medium	Small
ASSR [30]	11500	49445	9.8	14.0	12.5	63.6	-	-	-	-	-	3.7	26.2	70.1
IRSR [18]	8988	30176	34.1	29.9	17.5	9.4	5.0	2.5	1.7	-	-	5.1	39.2	55.7
Ours	8389	52173	-	16.0	18.7	17.0	13.1	10.3	7.4	5.1	12.5	0.2	13.3	86.5

Table 1. Statistics for three SRD datasets regarding image numbers, instance numbers, salient instance numbers per image and instance scales. Large: (instance size $\geq$ 30%), Medium: (5% $\leq$ instance size $\leq$ 30%), Small: (instance size $\leq$ 5%).

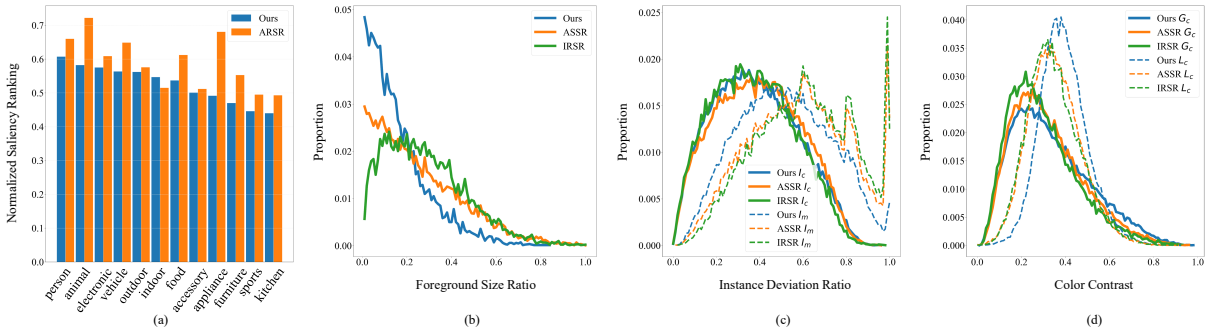

Figure 3. Numerical comparison of three SRD datasets. The relative saliency ranking among 12 superclasses of 80 COCO categories is shown in (a). Note there is no category information for each instance in published IRSR dataset. We also present the proportion of images with various foreground sizes in (b) and the percentage of instances regarding location and color contrast in (c) and (d) respectively.

Figure 4. Examples of our proposed dataset. Our fixation maps are highlighted with points representing thresholded salient objects (red), clusters of fixations on unannotated object (green) and clusters not considered salient (yellow).

data in ASSR and our real fixations) and dataset modelling (attention shift in ASSR and fixation duration in ours).

Foreground Size: Considering salient objects as foreground, Fig. 3 (b) visualises the frequency of images containing different foreground/background ratios. Our dataset contains a higher frequency of images with smaller foreground, presenting a potential challenge.

Instance Location: We conduct instance location analysis in Fig. 3 (c). Following [5, 15], two quantities l_c and l_m are introduced, denoting the distances from the instance center and farthest boundary point to the image center respectively. Distances are normalized by dividing half the image diagonal length. All three datasets exhibit similar centre bias, with our dataset showing smaller distances between the l_c and l_m distributions due to smaller objects.

Instance Contrast: Global contrast G_c and local contrast L_c are analyzed in Fig. 3 (d). Here, G_c is calculated by χ^2 distance between the RGB histograms of foreground and background for each instance. Following [15], L_c is derived from cropping 5x5 image patches at boundary points in salient instances, followed by χ^2 calculation on separate RGB histograms. Our proposed dataset contains a larger proportion of instances with higher global contrast and local contrast, suggesting that our dataset contains more visually striking objects and well-defined boundaries.

4. Proposed Method

In this section, we propose a novel Query as Graph Network (QAGNet) for the SRD task, which serves as a strong baseline for the proposed SIFR dataset.

4.1. Overview of the pipeline

As illustrated in Fig. 5, our pipeline starts with a **Multi-scale Salient Instance Query Extraction (SQE)** module, which inherits the same architecture as the widely-used Mask2Former [2]. For a given image, multi-scale feature maps are captured by a pixel decoder [42]. A randomly initialized salient instance query $Q_0 \in \mathbb{R}^{N \times D}$ is then fed to the transformer decoder through its 9 sequential layers interacting with these multi-scale features obtained by pixel decoder, where N represents the number of instance

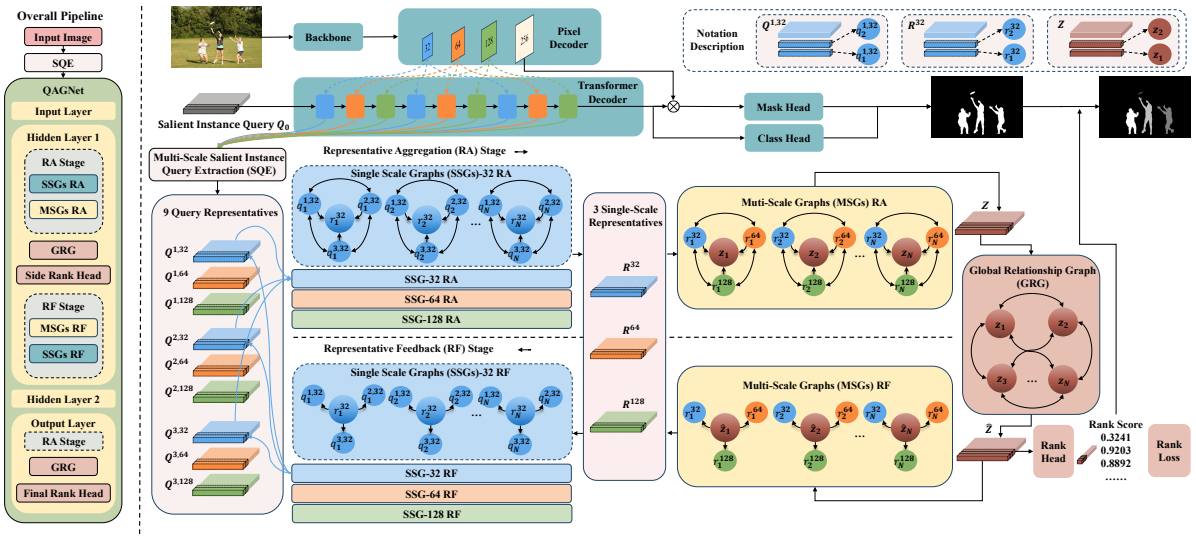

Figure 5. The architecture of QAGNet. The overall pipeline is shown on the left, while the detailed structure is presented on the right. The network aggregates query representations from a transformer detector in RA stage to produce saliency ranking features for each detected object. These features are fed back to the query representations during RF stage. Our proposed method is trained end-to-end.

queries, D denotes the feature dimension. As a result, 9 updated query representatives are extracted from the corresponding 9 decoder layers as:

$$Q^{l,s} = \text{SQE}(Q_0), l \in \{1,2,3\}, s \in \{32,64,128\} \quad (1)$$

where each query representative $Q^{l,s} \in \mathbb{R}^{N\times D}$ consists of N salient instance queries $\{q_1^{l,s}, q_2^{l,s}, \ldots, q_N^{l,s}\} \in Q^{l,s}$ (each is $1 \times D$ dimension) corresponding to N potential salient instances, including both prominent and less-prominent instances. Here, s denotes the scale of the feature maps from the pixel decoder, drawn by different layers of the transformer decoder via cross-attention, and l indicates the relative position of the layers during decoding. These learnable query representatives $Q^{l,s}$ play a similar role to the regional proposals produced by the region proposal network [26]. The obtained 9 query representatives $Q_{\text{all}} = \{Q^{l,s} | l \in \{1,2,3\}, s \in \{32,64,128\}\}$ encode the information of all N potential salient objects across varying depths in the decoding phase at multiple scales.

A novel QAGNet consisting of multiple QAG layers is proposed to capture ranking-aware features that describe each salient instance by considering not only the multi-scale features extracted from multiple pixel decoder layers, but also its relationships with other instances in the given image. The QAGNet takes the obtained multi-scale query representatives Q_{all} as input, and produces a feature representative $\hat{Z}_{\text{final}} \in \mathbb{R}^{N\times D}$ describing multi-scale and ranking-aware cues of all N instances in the input image as:

$$\hat{Z}^{\text{final}} = \text{QAGNet}(Q_{\text{all}}) \quad (2)$$

where $\hat{Z}^{\text{final}} = \{\hat{z}_1^{\text{final}}, \hat{z}_2^{\text{final}}, \cdots, \hat{z}_N^{\text{final}}\}$.

Finally, the $\hat{Z}^{\text{final}}$ is fed to a **rank head** including a single linear layer, predicting the relative saliency ranking scores of all N instances.

4.2. QAGNet and QAG layers

The proposed QAGNet is made up of an input layer, multiple hidden layers and an output layer.

QAG input layer: The QAG input layer groups the obtained multi-scale query representatives Q_{all} into three subsets ($Q^{32} = \{Q^{1,32}, Q^{2,32}, Q^{3,32}\}$, $Q^{64} = \{Q^{1,64}, Q^{2,64}, Q^{3,64}\}$ and $Q^{128} = \{Q^{1,128}, Q^{2,128}, Q^{3,128}\}$), where each subset contains three query representatives of the same scale but from different decoding layers. These features serve as the input of the hidden layers, upon which the graph is built.

QAG hidden layer: A set of QAG hidden layers are stacked to model not only multi-scale relationship cues of each instance but also the relationship among instances. Each QAG hidden layer builds a tri-tiered nested graph (illustrated in Fig. 6) from the features generated or updated from the previous input or hidden layer. We use three-level graphs: **Single Scale Graphs (SSGs)** that contain single-scale instance-level features, **Multi-Scale Graphs (MSGs)** that contain multi-scale instance-level features and a **Global Relationship Graph (GRG)** that contains not only multi-scale instance-level cues but also the relationships among salient instance proposals. Each QAG hidden layer is designed to have two stages: (i) a **Representative Aggregation (RA)** stage that gradually refines and aggregates all multi-scale features to be fed into a GRG; and (ii) a **Representative Feedback (RF)** stage that feeds back ranking-

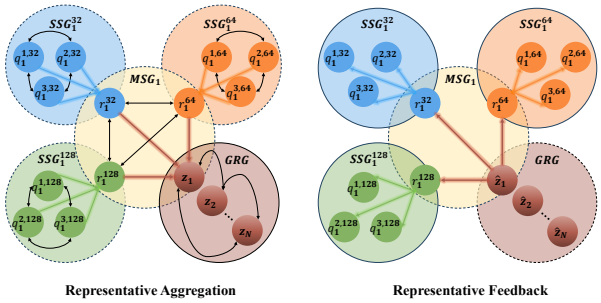

Figure 6. Illustration of the tri-tiered nested graph. Representative aggregation combines query features from the same scale (SSG), then different scales (MSG) and finally across instances (GRG). Representative feedback reverses this process.

aware information encoded in the GRG to further update MSGs and SSGs, allowing the query representatives in SSGs to derive ranking-related cues from not only other scales but also other instances.

In detail, the **RA** stage first constructs N SSGs for each subset (Q^{32}, Q^{64}, Q^{128}), where each SSG_n^s represents the single-scale cues of the nth salient instance. Each SSG_n^s contains three regular nodes $\{q_n^{1,s}, q_n^{2,s}, q_n^{3,s}\} \in \mathbb{R}^{1\times D}$ and one representative node $r_n^s \in R^s$ that is initialized as the average of three regular nodes as:

$$\tilde{r}_n^s = \text{Avg}(q_n^{1,s}, q_n^{2,s}, q_n^{3,s}) \tag{3}$$

In total $3N$ SSGs are constructed representing N instances across three scales. All regular nodes within each SSG are then connected via a GNN layer as:

$$\hat{q}_n^{1,s}, \hat{q}_n^{2,s}, \hat{q}_n^{3,s} = f(q_n^{1,s}, q_n^{2,s}, q_n^{3,s}|\mathcal{A}_1^{\text{SSG}}) \tag{4}$$

where f denotes the edge and node feature updating processes, which can be customized from any existing GNNs; $\mathcal{A}_1^{\text{SSG}}$ denotes the initial adjacency matrix of the SSG in RA stage, which fully connects all regular nodes in each SSG. Each initialized SSG representative node $\tilde{r}_n^s$ is then updated as r_n^s by combining cues contained in the corresponding three updated regular nodes as:

$$r_n^s = f(\hat{q}_n^{1,s}, \hat{q}_n^{2,s}, \hat{q}_n^{3,s}, \tilde{r}_n^s|\mathcal{A}_2^{\text{SSG}}) \tag{5}$$

where $\mathcal{A}_2^{\text{SSG}}$ denotes the adjacency matrix that only contains directed edges from updated regular nodes to its representative node, i.e., this allows the messages only pass from updated nodes to the SSG representative node. This way, each representative node r_n^s draws features representing the corresponding instance at a single-scale s. After this, every three obtained representative nodes $r_n^{32}, r_n^{64}, r_n^{128}$ that summarise three scales of the n_{th} instance are further combined as an MSG_n, in which each is treated as a regular MSG node. Applying the same rule, an MSG representative node $z_n \in Z$ comprising multi-scale cues of the n_{th} instance can be computed by combining cues contained in the corresponding three regular MSG nodes as:

$$\begin{aligned} \tilde{z}_n &= \text{Avg}(\tilde{r}_n^{32}, \tilde{r}_n^{64}, \tilde{r}_n^{128}) \\ \hat{r}_n^{32}, \hat{r}_n^{64}, \hat{r}_n^{128} &= f(r_n^{32}, r_n^{64}, r_n^{128}|\mathcal{A}_1^{\text{MSG}}) \\ z_n &= f(\hat{r}_n^{32}, \hat{r}_n^{64}, \hat{r}_n^{128}, \tilde{z}_n|\mathcal{A}_2^{\text{MSG}}) \end{aligned} \tag{6}$$

where $\mathcal{A}_1^{\text{MSG}}$ is the adjacency matrix defining fully connected edges for regular nodes of an MSG, while $\mathcal{A}_2^{\text{MSG}}$ denotes the adjacency matrix contains directed edges from the updated MSG regular nodes to its representative node. As a result, each MSG representative node represents multi-scale features of an instance.

Finally, GRG is obtained by combining all MSG representative nodes z_n within a graph. We introduce short connections between z_n and the one in previous hidden layer to reduce information loss. These are then fed to a GNN layer to update all GRG nodes in the context of other instances, allowing them to be ranking-aware. This can be defined as:

$$\hat{z}_1, \cdots, \hat{z}_N = f(z_1, \cdots, z_N|\mathcal{A}^{\text{GRG}}) \tag{7}$$

where $\mathcal{A}^{\text{GRG}}$ denotes that GRG is a fully connected graph. This way, the final obtained instance representatives $\hat{Z} = \{\hat{z}_1, \hat{z}_2, \cdots, \hat{z}_N\}$ describe multi-scale and ranking-aware cues for all salient instances in the image. We feed $\hat{Z}$ to side rank head for better supervising the learning process.

The **RF** stage gradually passes the cues encoded in the $\hat{Z}$ to SSGs, allowing these SSG node features (9 query representatives in 3 subsets) to be updated by considering ranking-aware and relationship cues among instances as well as each instance's multi-scale features. This stage starts with feeding back representative nodes $\hat{z}_n$ to the corresponding MSG regular nodes and treating each updated MSG regular node as the representative node of the corresponding SSG to further update the SSG regular nodes. The RF stage conducts the reverse process of RA stage with the edges only directing from each representative node to all of its corresponding regular nodes.

QAG output layer: The output layer utilises an RA stage with a GRG, which takes the updated query representatives produced by the previous hidden layer as input and produces the final feature representative $\hat{Z}^{\text{final}} \in \mathbb{R}^{N\times D}$ (i.e., updated GRG node features), followed by a final rank head.

5. Experiment

5.1. Experimental Setup

Datasets and Metrics: We conduct experiments on two publicly available datasets, ASSR [30], IRSR [18], and our proposed dataset with three widely-used metrics in SRD: Salient Object Ranking (SOR) [9], Segmentation-Aware SOR (SA-SOR) [18] and Mean Absolute Error (MAE).

Method	Backbone	ASSR			IRSR			Ours			#Para.(M)
		SASOR ↑	SOR ↑	MAE ↓	SASOR ↑	SOR ↑	MAE ↓	SASOR ↑	SOR ↑	MAE ↓	
ResNet and VoVNet											
RSDNet (TPAMI 2019)	ResNet-101	0.6313	0.7758	0.1236	0.4232	0.7096	0.1175	0.4791	0.7239	0.0772	42.7
ASSR (CVPR 2020)	ResNet-101	0.5400	0.7920	0.1010	0.3207	0.6521	0.1098	0.3281	0.5843	0.0624	44.2
IRSR-U (TPAMI 2021)	ResNet-50	0.7051	0.8314	0.0923	0.5647	**0.8143**	0.0953	**0.5585**	0.7487	**0.0465**	128.1
IRSR-L (TPAMI 2021)	ResNet-50	0.7090	0.8283	0.0914	0.5648	**0.8141**	0.0953	**0.5585**	0.7487	**0.0465**	128.1
SOR (CVPR 2021)	VoVNet-39	0.6371	0.8330	0.0799	0.5171	0.7909	0.0988	0.3820	**0.7554**	0.0580	119.0
Ours-Res50-U	ResNet-50	**0.7545**	**0.8514**	**0.0619**	**0.6110**	0.8108	**0.0845**	**0.6119**	**0.7899**	**0.0437**	47.3
Ours-Res50-L	ResNet-50	**0.7658**	**0.8469**	**0.0609**	**0.6107**	0.8106	**0.0845**	**0.6119**	**0.7899**	**0.0437**	47.3
Swin											
OCOR-U (CVPR 2022)	Swin-L	0.6413	**0.8843**	0.0786	0.5183	0.8149	0.1003	0.4392	0.7436	0.0488	401.7
OCOR-L (CVPR 2022)	Swin-L	0.6474	**0.8937**	0.0863	0.5058	0.8184	0.1052	0.4426	0.7462	0.0531	401.7
Ours-SwinB-U	Swin-B	0.7741	0.8583	0.0538	0.6252	0.8152	0.0792	**0.6167**	**0.7933**	**0.0409**	110.2
Ours-SwinB-L	Swin-B	**0.7809**	0.8529	0.0528	0.6252	0.8151	0.0792	**0.6167**	**0.7933**	**0.0409**	110.2
Ours-SwinL-U	Swin-L	0.7793	0.8591	**0.0492**	**0.6466**	**0.8241**	**0.0768**	**0.6206**	**0.7982**	**0.0416**	218.8
Ours-SwinL-L	Swin-L	**0.7873**	0.8535	**0.0478**	**0.6468**	**0.8240**	**0.0767**	**0.6206**	**0.7982**	**0.0416**	218.8

Table 2. Quantitative comparison with other saliency ranking methods, with backbones shown, e.g. ResNet [7], VoVNet [13] and Swin [20]. We evaluate our method using ResNet-50, Swin-Base and Swin-Large. -L and -U represent limited and unlimited model variants. ↑ indicates the higher the better, while ↓ denotes the lower the better. The best two results have been marked as red and blue.

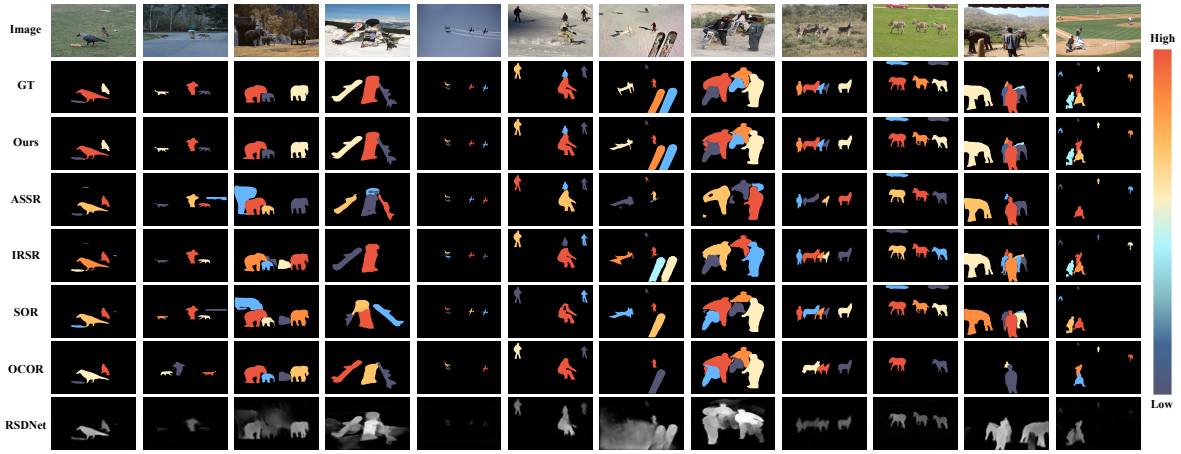

Figure 7. Qualitative comparison between our proposed QAGNet and other SRD methods on our proposed dataset.

SOR computes the Spearman's rank-order correlation between the prediction and ground truth. This presupposes the predicted instances match the ground truth and only considers the rank orders. SA-SOR utilize the Intersection over Union (IoU) to choose the matched instances and then computes the Spearman's rank-order. We follow [18] to set the IoU threshold to 0.5. We directly calculate the pixel-level difference between generated saliency ranking maps and ground truth for MAE.

Implementation Details: We utilise a Graph Attention Network [32] for both edge calculations and node aggregation with a feature dimension of 256 and a dropout rate of 0.2 in QAG layers. Following [2], we use binary cross-entropy loss and dice loss [22] as mask loss. For rank prediction, we utilize the pair-wise SRD loss [18] and set weights to 3.0 for side rank loss and 5.0 for final rank loss. The final loss is a combination of mask loss, saliency classification loss and rank loss. In inference, we determine the final confidence score by multiplying the saliency class confidence with the mask confidence. Only salient instance predictions exceeding a confidence threshold of 0.7 are retained. We train our model for 30,000 iterations with AdamW [21] optimizer with weight decay set to 1×10^{-4}. The learning rate starts from 2.5×10^{-5} and is reduced by a factor of 10 at 22,000 and 26,000 iterations. We resize all the input images to 1024×1024 and do not apply additional preprocessing. We use 4 A6000 GPUs and set the batch size to 4 for the Swin-L backbone.

5.2. Comparisons with the State-of-the-Art

We compare our proposed method with 5 existing SRD methods: RSDNet [9], ASSR [30], IRSR [18], SOR [6] and OCOR [31]. We prioritize using the pre-trained models provided by the authors, if available, otherwise we retrain models from source code with the recommended settings from the original papers. For models that output a fixed number of salient objects: RSDNet, ASSR, SOR, and OCOR, we adjust this fixed number to match each dataset's maxi-

Settings	Specific Configuration				SASOR ↑	SOR ↑	MAE ↓
I (Baseline)	Last query + Linear layer				0.5623	0.7292	0.0469
II (Baseline)	Average 9 queries + Linear layer				0.5807	0.7381	0.0456
	RA Stage		GRG	RF Stage			
	SSG	MSG					
III	✓				0.5837	0.7423	0.0451
IV		✓			0.5944	0.7582	0.0442
V			✓		0.5932	0.7599	0.0445
VI		✓	✓		0.5989	0.7623	0.0441
VII	✓	✓	✓		0.6016	0.7653	0.0442
VIII	✓	✓	✓	✓	**0.6086**	**0.7736**	**0.0439**

Table 3. Ablation analysis of different modules in QAGNet.

Settings	Hidden Layer Number	Short Connection	SASOR ↑	SOR ↑	MAE ↓
I	1	✓	0.6086	0.7736	0.0439
II	2	✓	**0.6119**	**0.7899**	**0.0437**
III	2		0.6084	0.7803	0.0442
IV	3	✓	0.6089	0.7794	0.0440

Table 4. Ablation analysis of layer number and short connection.

Settings	Query Number	SASOR ↑	SOR ↑	MAE ↓
I	50	0.6061	0.7702	0.0435
II	100	0.6119	**0.7899**	0.0437
III	200	**0.6128**	0.7886	**0.0426**

Table 5. Ablation analysis of query number N.

mum instance count. Within RSDNet, we follow [12] and use the stacked representation of the ground truth to regress the saliency values. We average the predicted saliency values within each instance to get the saliency scores. RSDNet cannot predict instance-level masks, we follow [18] to utilize the instance masks from the IRSR to calculate SA-SOR. For IRSR and our proposed method, both models use a confidence score during inference to select the salient instances. This approach might produce instances surpassing the prescribed limits in ASSR and IRSR datasets. We follow [18] to present top 5 and top 8 ranked instances in these datasets as the limited version results. Results acquired directly post-confidence thresholding are reported as the unlimited version. OCOR can generate multiple instances for each rank. We directly report this result as the unlimited version, and report the limited version by only choosing the highest scoring instance for different rank predictions.

Quantitative Comparison: Tab. 2 presents a quantitative comparison of our method against other SRD approaches. Our QAGNet outperforms all other saliency ranking methods on SA-SOR and MAE by a large margin, where our best model surpasses the second-best model in each dataset by 11.0%, 14.5% and 11.1% respectively for SA-SOR, despite using fewer parameters. IRSR and OCOR can score highly on the SOR metric, however high SOR scores can be achieved if the identified salient instances maintain the correct ranking even with missing, redundant, or low-quality segmentation. These methods show a drop in performance on SA-SOR, which also penalizes missing salient objects.

Qualitative Comparison: We conduct qualitative comparison in Fig. 7 with multiple challenging images including low-contrast, difficult illumination, small objects and high instance numbers. Our proposed method generates saliency ranking maps with clear boundaries and correct rank orders.

5.3. Ablation Studies

Module Analysis: We explore the effectiveness of different modules in QAGNet on our proposed dataset in Tab. 3. Baseline I regresses saliency ranks using a linear layer on the last query representation of the Mask2Former. Baseline II averages all 9 query representations before applying the linear layer. Setting II outperforms I, suggesting all query features contribute. We demonstrate the effectiveness of QAGNet components from setting III to VII. Setting VIII integrates the RF stage to feedback ranking-aware cues and is then followed by a RA stage with GRG to predict the final rankings. This combination forms a complete QAGNet with 1 hidden layer and delivers the best results. By feeding the ranking-aware information back, the query feature representatives can be refined into the next RA stage, helping to learn ranking-aware cues in a bi-directional manner.

Layer Number and Short Connection Analysis: In Tab. 4, we explore the effect of varied hidden layers and short connections in QAGNet. Setting I, II, IV demonstrate the effectiveness of using 2 hidden layers. The effectiveness of the short connection can be observed in settings II and III, where it facilitates information passing from different hidden layers to the rank head.

Query Number Analysis: Tab. 5 shows the impact of varied numbers of salient instance queries N. Setting I, which employs 50 queries, noticeably underperforms setting II and III with 100 and 200 queries. This supports our hypothesis that considering not only the most salient objects but also those with less saliency is important for accurate ranking of complex scenes. Setting III generates the best scores on the more discriminative metrics of SASOR and MAE. The performance difference between settings II and III is marginal, thus we employ 100 queries in our lightweight Resnet-50 and Swin-B backbone QAGNet, while allocating 200 queries to the Swin-L backbone QAGNet.

6. Conclusion

In this paper we have proposed the first large-scale instance-level dataset, SIFR, for saliency ranking based on human eye-tracking. We argue that true human fixations are a more realistic measure of visual attention than previous work utilising mouse movement. We also propose a strong baseline method, QAGNet, that leverages query features from a transformer detector within a novel graph architecture. Our network successfully generates high-quality saliency ranking maps even in challenging scenes. Experimental results show that our method provides strong results across two existing datasets and our new dataset.

References

[1] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 507–515, 2016. 1

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1290–1299, 2022. 2, 4, 7

[3] Frederic R Danion and J Randall Flanagan. Different gaze strategies during eye versus hand tracking of a moving target. *Sci Rep*, 8(1):10059, 2018. 1

[4] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, pages 226–231, 1996. 3

[5] Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao. Salient objects in clutter. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):2344–2366, 2022. 4

[6] 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. 2, 7

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 2

[9] 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. 2, 6, 7

[10] 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. 1

[11] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 3

[12] Mahmoud Kalash, Md Amirul Islam, and Neil DB Bruce. Relative saliency and ranking: Models, metrics, data and benchmarks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):204–219, 2019. 1, 2, 8

[13] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13906–13915, 2020. 7

[14] 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. 2

[15] 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. 2, 4

[16] 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

[17] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3917–3926, 2019. 1, 2

[18] 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, 4, 6, 7, 8

[19] 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. 1, 2

[20] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[21] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[22] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 7

[23] Andrea Palazzi, Davide Abati, Francesco Solera, Rita Cucchiara, et al. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018. 1

[24] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7479–7489, 2019. 2

[25] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124 (3):372, 1998. 3

[26] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 5

[27] Dario D Salvucci and Joseph H Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78, 2000. 3

[28] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE transactions on neural networks*, 20(1): 61–80, 2008. 2

[29] Alexander C Schütz, Doris I Braun, and Karl R Gegenfurtner. Eye movements and perception: a selective review. *J Vis*, 11(5), 2011. 1

[30] 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, 4, 6, 7

[31] 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. 2, 3, 7

[32] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017. 7

[33] 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. 2

[34] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3907–3916, 2019. 2

[35] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7264–7273, 2019. 2

[36] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015. 1

[37] 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. 2

[38] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 2

[39] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 1

[40] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1741–1750, 2018. 1, 2

[41] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8779–8788, 2019. 1, 2

[42] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4