Inferring Attention Shifts for Salient Instance Ranking

Avishek Siris[1] · Jianbo Jiao[2,3] · Gary K.L. Tam[1] · Xianghua Xie[1] · Rynson W.H. Lau[4]

Abstract

The human visual system has limited capacity in simultaneously processing multiple visual inputs. Consequently, humans rely on shifting their attention from one location to another. When viewing an image of complex scenes, psychology studies and behavioural observations show that humans prioritise and sequentially shift attention among multiple visual stimuli. In this paper, we propose to predict the saliency rank of multiple objects by inferring human attention shift. We first construct a new large-scale salient object ranking dataset, with the saliency rank of objects defined by the order that an observer attends to these objects via attention shift. We then propose a new deep learning-based model to leverage both bottom-up and top-down attention mechanisms for saliency rank prediction. Our model includes three novel modules: Spatial Mask Module (SMM), Selective Attention Module (SAM) and Salient Instance Edge Module (SIEM). SMM integrates bottom-up and semantic object properties to enhance contextual object features, from which SAM learns the dependencies between object features and image features for saliency reasoning. SIEM is designed to improve segmentation of salient objects, which helps further improve their rank predictions. Experimental results show that our proposed network achieves state-of-the-art performances on the salient object ranking task across multiple datasets. Code and data are available at https://github.com/SirisAvishek/Attention_Shift_Ranks.

✉ Avishek Siris
a.siris.789605@swansea.ac.uk

Jianbo Jiao
j.jiao@bham.ac.uk

Gary K.L. Tam
k.l.tam@swansea.ac.uk

Xianghua Xie
x.xie@swansea.ac.uk

Rynson W.H. Lau
Rynson.Lau@cityu.edu.hk

[1] Department of Computer Science, Swansea University, Swansea, UK

[2] Department of Computer Science, University of Birmingham, Birmingham, UK

[3] Department of Engineering Science, University of Oxford, Oxford, UK

[4] Department of Computer Science, City University of Hong Kong, Hong Kong, China

1 Introduction

In recent years, research on salient object detection has grown extensively. It aims to locate objects that attract human visual attention. Reliable prediction of human visual attention often benefits downstream high-level applications such as image parsing (Lai & Gong, 2016), image captioning (Xu et al., 2015a), person re-identification (Zhao et al., 2013) and many more (Borji, 2018). Most saliency methods (Qin et al., 2019; Zhao & Wu, 2019; Zhao et al., 2019; Wu et al., 2019a; Pang et al., 2020; Wei et al., 2020) propose to model salient object detection as a binary prediction problem, where all predicted objects are given the same value and importance. Humans, however, are shown to have the ability to sequentially select and shift attention from one region or object to another (Koch & Ullman, 1987; Itti & Koch, 2000). Such an ability allows humans to deal with multiple simultaneous visual inputs, given the limited capacity of our visual system (Neisser, 2014). Modelling this ability is important for the understanding of how humans interpret images, and helps improve the performance of relevant applications, e.g., autonomous driving (Palazzi et al., 2018) and robot-human interactions

(Schillaci et al., 2013). Here, we interpret and define the order of objects attended to by this human ability as *saliency ranking*. This new task can provide similar benefits to traditional salient object detection and further introduces rank ordering information, which can be useful in downstream applications that would leverage the knowledge of human attention shifts on objects in an image. In robotic interaction, rank order can enable the robot to assess a scene and prioritise its tasks or dynamically plan its actions (Chang et al., 2010). It would also benefit the image captioning task where enhanced captions can be generated to reflect the importance of objects (Tavakoli et al., 2017; Cornia et al., 2018). Such importance can be determined by object rank order information for an input image. We believe that salient object ranking can be applied in video applications such as VR and automated driving. In VR applications, salient object ranking can be utilised to prioritise graphics rendering power for certain objects of interest while reducing rendering for others in immersive environments (Sitzmann et al., 2018). In terms of automated driving, rank order could be used to prioritise the danger degree and action/response towards the identified objects of importance for increased situational awareness (Arvanitis et al., 2023). Progressive data transmission can be improved for target devices with limited computational resources, by prioritising specific objects for compression and others for detail retention based on rank order (Park et al., 2017).

Some early works apply *attention shift* in applications such as visual search (Itti & Koch, 2000) and scene analysis (Itti et al., 1998). They use a saliency map, which represents the visual saliency of each region in the scene, to guide the selection of the attended regions, and model attention shift as the shifting of attention from one region to another in the order of decreasing values in the saliency map (Koch & Ullman, 1987; Itti & Koch, 1999). In these early works, only low-level features (*e.g.*, colour, intensity and orientation) are computed to generate the saliency map. A more recent work of Gorji and Clark (2017) models *attentional push*, which refers to how scene actors (humans) may manipulate the attention (gaze direction/location) of observers in viewing an image. This work heavily relies on the *gaze-following* concept (Recasens et al., 2015), which constrains attention to only social image scenarios (images with at least one human actor). It also limits attention to a single shift from a person in a scene to some other region.

The research of relative ranking of salient objects was first introduced in Islam et al. (2018). Their relative ranking is inferred from the agreement of binary object saliency among multiple observers. The study (Islam et al., 2018) is motivated by the fact that observers are likely to have different views of what objects are considered salient. In their implementation, they implicitly assume that multiple objects picked by the same observer share equal saliency rank (top row of Fig. 1). However, simultaneous attention to multiple objects
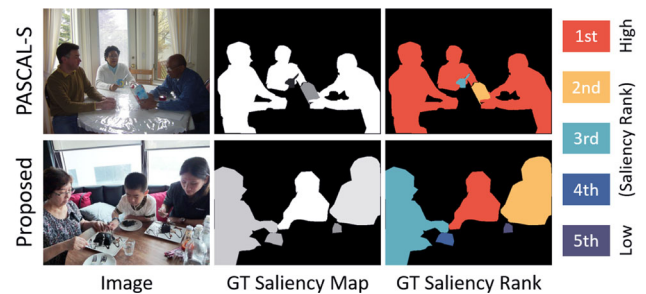

Fig. 1 The first row shows an image from the PASCAL-S dataset (Li et al., 2014b). It is used for saliency ranking in Islam et al. (2018). We can see that multiple objects can be given the same saliency rank. The second row shows an image from our proposed dataset with distinct ground-truth saliency ranks, motivated by psychological studies. The color (**red→ purple**) indicates the saliency rank (1→5)

is not supported by behavioural observations, because dividing attention among multiple objects often leads to poorer performance (Desimone & Duncan, 1995) and may not truly reflect how humans shift their attention. Multiple objects with the same rank would also make it difficult to model the order of attention shift.

The authors of Islam et al. (2018) have since extended their work and propose a new COCO-SalRank dataset (Kalash et al., 2019). Unlike the rank modified PASCAL-S dataset, their new dataset does not contain tied saliency ranks. However, their ground-truth rank generation uses hand-designed criteria for producing fixation maps, which are then applied to determine instance ranks. Similarly, Liu et al. (2021a) follow Kalash et al. (2019) for the relative salient object ranking task. They also propose a new saliency ranking dataset by utilising the fixation maps from the SALICON (Jiang et al., 2015) dataset with MS-COCO (Lin et al., 2014). Both Kalash et al. (2019) and Liu et al. (2021a) model relative saliency ranking based on fixation maps that are generated from applying a Gaussian filter onto fixation data. They do not consider the process of shifting attention that is performed by the human visual system.

Inspired by the above saliency and psychological studies (Neisser, 2014), we aim in this work to investigate saliency rank that models human attention shift. Our idea follows psychology studies that humans attend to one object at a time in a complex scene (Koch & Ullman, 1987). We define our task as detecting salient instances in an image, while inferring human attention shifts on these instances. It differs from traditional salient object detection in that our task prioritises salient object rank order rather than simply detecting the most salient objects. We first propose a new saliency ranking dataset collected based on attention shift. Different from Kalash et al. (2019) and Liu et al. (2021a), we consider that the first object attended by an individual should have the highest saliency. Subsequent attended objects should be associated with descending saliency values (*i.e.*, attention
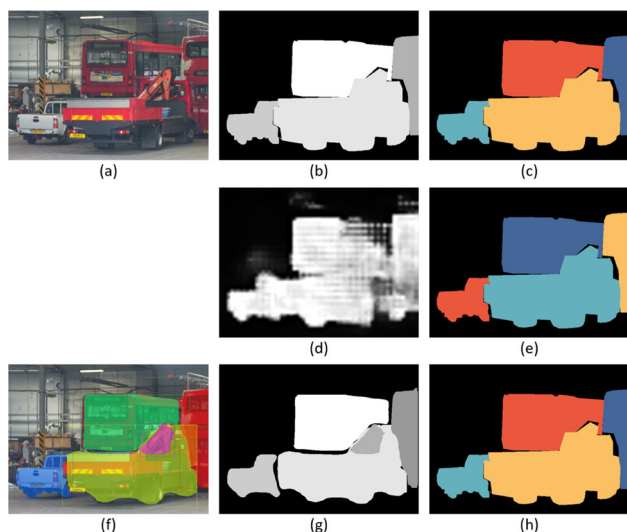
Fig. 2 Example of saliency rank prediction and comparison. **a** Example input image, **b** corresponding ground-truth (GT) saliency rank, **c** corresponding GT saliency rank (colourised), **d** saliency rank prediction by RSDNet (Islam et al., 2018), **e** corresponding saliency rank by RSDNet with only GT objects (overlaid and colourised), **f** salient object and segmentation proposed by our model, **g** our salient object rank prediction, **h** our corresponding saliency rank with only GT objects (overlaid and colourised)

shift towards objects of lower saliency values). Considering that multiple observers may have different saliency rank orders among them, we average their rank orders to obtain the ground-truth saliency rank (Sec. 3.2). We perform a user study and show that such a human attention shift on object instances correlates with the object saliency rank. Figure 1 (bottom row) shows a sample, where each object in an image is assigned a distinct saliency rank (from value 1 to 5) corresponding to the order of attention shift.

Traditional saliency models often introduce many false positive saliency to non-salient objects and the background (Fig. 2d). When the shape of the objects is not captured well (Fig. 2d), it further impacts the saliency rank prediction of the objects (Fig. 2e). Motivated by the above observations, we propose a saliency rank prediction method to infer human attention, leveraging both bottom-up and top-down attention to determine the saliency rank of individual objects. Our model utilises object proposals for object segmentation and object rank prediction. This allows our network to perform saliency reasoning on the object level and capture individual salient instances, while most prior works (*e.g.*, (Islam et al., 2018)) perform at region level and are unable to distinguish between multiple objects properly.

Furthermore, generating accurate segmentation of individual instances is important for predicting salient objects. In complex image scenes, multiple objects tend to overlap or are closely located to one another. If clear boundaries between objects are not captured, then the accuracy of segmentation

will drop and features for discriminating the saliency between objects can also become poorer (Fig. 2d–e). To tackle this issue, we further propose a Salient Instance Edge Module (SIEM) to enhance boundary and complete segmentation of salient objects. We pair SIEM with the instance mask segmentation branch in order to mutually improve the segmentation of salient instance masks and edges. Such a design is essential to further boost the performance of saliency ranking, as it enables the network to distinguish salient instances from the background and other objects better.

The main contributions of this work include:

- We propose a new research problem to infer human attention shift through salient object ranking. The task is inspired by psychological and behavioural studies. It goes beyond human-object interactions (Recasens et al., 2015) by also modelling object-object attention shift.
- We propose a new large-scale dataset for our salient object ranking task. The generation of our GT saliency rank is justified by our user study.
- We propose a deep learning approach to jointly predict saliency ranks of salient object instances and their corresponding object masks, with bottom-up and top-down attention mechanisms.
- We introduce a new Salient Instance Edge Module (SIEM) to enhance the segmentation of salient instances, which further boosts saliency ranking performances.
- We adopt a new nDCG metric for evaluation. Extensive experimental evaluations and analyses show that the proposed model outperforms SOTA methods on salient object ranking.

This manuscript extends our preliminary work that was presented in CVPR (Siris et al., 2020). In this work, we make the following four major extensions: First, we have updated Sect. 2 with the latest relevant works, and also compared our method with these works. Second, we propose a new Salient Instance Edge Module (SIEM) to improve segmentation quality of salient instances and further increase the performance of saliency ranking. Third, we have made significant modifications to our preliminary architecture in Siris et al. (2020), including the introduction of list-wise ranking loss for saliency ranking instead of formulating it as rank-id classification, top-12 object proposals for salient object and rank reasoning, and end-to-end training with warm-up and fine-tuning. Finally, we have added further experiments to evaluate the proposed method, including additional comparisons with state-of-the-arts, ablation study, introducing a new metric for saliency ranking, and evaluation on an additional dataset.

2 Related Work

2.1 Salient Object Detection

Salient object detection can be categorised into bottom-up, top-down, or a combination of both. Here, we focus on those that combine both bottom-up and top-down approaches. Early methods that combine bottom-up and top-down approaches use hand-crafted and computational-based features. Bottom-up features often come from local and global contrasts in color, intensity and orientation (Judd et al., 2009). Top-down features often relate to the specific tasks at hand. Notable examples include using high-level face features (Xu et al., 2015b), photography bias (Judd et al., 2009), person and car detector (Borji, 2012), gist features (Peters & Itti, 2007) and gaze patterns learnt from performing specific tasks (Borji et al., 2012). With the advance of Convolutional Neural Networks (CNNs), CNN features are leveraged to improve the performance of saliency detection (Gao et al., 2020). Some works (Pan et al., 2016) use a simple stack of convolution and deconvolution layers, while some others (Li & Yu, 2015; Song et al., 2018) design multi-scale networks to capture contextual information for saliency inference.

Recent studies further incorporate a top-down pathway (Zhang et al., 2017; He et al., 2017b; Zhang et al., 2018). High-level semantics in the top layers are refined with the low-level features in the shallow layers through side connections (Zhao & Wu, 2019; Pang et al., 2020; Zhao et al., 2020; Tang et al., 2021). The refinement generates a better representation at each layer (Hou et al., 2017) and is thought to imitate the bottom-up (low-level stimuli) and top-down (visual understanding) human visual process (Wang et al., 2019a). Wang et al. (2018b) follow the relationship between eye fixation and object saliency previously studied in Li et al. (2014b) and propose to use fixation maps to guide saliency in a top-down manner. In Zhao et al. (2019) and Zhou et al. (2020), saliency features are complemented with edge information at various resolutions, in order to improve the accuracy of salient object segmentation and their boundaries. Wei et al. (2020) further propose to decompose saliency maps into body and detail maps. The body map contains the central area of objects, while the detail map focuses on the boundaries of objects. Li et al. (2021) pair salient object and camouflaged object detection to learn contradictory information and enhance the performance of the two tasks. Siris et al. (2021) proposes to exploit semantic segmentation of things and stuff categories for extracting high-level scene context features.

The above methods mimic the human visual process using both bottom-up and top-down pathways. Our network is also CNN-based and contains both bottom-up and top-down pathways. However, our bottom-up mechanism comes from salient object proposals (inspired by Anderson et al. (2018), which focuses on the captioning task). These salient object proposals produce features that largely capture the whole object, while traditional methods usually capture patches or parts of an object. We further introduce spatial size and location of object proposals in our model to model the relationship between the objects and scene for boosting saliency ranking. Our top-down mechanism considers the operation of explicit object-level features generated from object proposals, and further integrates high-level image semantics obtained from a backbone network. Note that most salient object detection methods only perform binary saliency prediction, not providing clear segmentation between salient instances. Further, they do not consider different saliency values between individual objects. To the best of our knowledge, we are the first to model salient object rank order according to the attention shift with bottom-up and top-down mechanisms.

2.2 Ranking in Saliency

Ranking of salient objects is a relatively new problem. It is introduced by Islam et al. (2018), in which they define object ranks as the *degree of agreement* among multiple observers who consider if objects are salient. In our work, we define the saliency rank differently as the sequential order of distinct objects attended by an observer one at a time through attention shift. The order is related to a *descending level of saliency values of the objects* that attract human attention. Our definition is closer to human visual attention and is motivated by psychological studies and behavioural observations (Neisser, 2014), where multiple attentions of foci are not supported (Desimone & Duncan, 1995). Both Islam et al. (2018) and the extended work (Kalash et al., 2019) use the same patch-based network architecture. In contrast, our network is built on object proposals that provide instance-level features.

A network proposed in Liu et al. (2021a) follows closely to Islam et al. (2018) and Kalash et al. (2019), and exploits the same backbone as ours Siris et al. (2020) for generating object proposals. The difference between ours and Liu et al. (2021a) is that they use a graph-based module for learning to rank with a pair-wise ranking loss. Instead, we adopt a Transformer (Vaswani et al., 2017) to dynamically learn the relations between objects and image features for ranking. We also use a simpler network to predict the final object ranks, and train it with a list-wise ranking loss.

Fang et al. (2021) recently proposed a Position-Preserved Attention module to incorporate positional information with object features. They concatenate absolute positional coordinate maps to the feature maps before ROI pooling. In contrast, we propose a Spatial Mask Module to capture object positional information and concatenate it with object features after ROI pooling. The main difference between ours and Fang et al. (2021) is how we extract and embed our posi-

tional information (more details in Sect. 4.2). In addition, our module allows explicit learning of the relationship between object position and scene for saliency ranking.

There are other works that use ranking techniques for saliency estimation. For example, Zhang et al. (2016) use graph-based manifold ranking for saliency inference, and Li et al. (2012) use rank learning to select visual features that best distinguish salient targets from real distractors. However, these works use ranking as a formulation to output a final binary saliency prediction. They do not predict saliency rank order as in our work. In contrast to our saliency ranking task, Lv et al. (2021) use fixation detection to help segment and rank camouflaged objects.

2.3 Attention Mechanism

Attention mechanism has been shown to be effective in improving natural language processing (Vaswani et al., 2017) and many vision tasks (Ma et al., 2018; Jiao et al., 2019). The attention mechanism discussed here can be considered as top-down attention. However, simple concatenation or element-wise operations on multi-level features may not improve saliency prediction (Wang et al., 2018a) as noisy and non-relevant features may impact the saliency network (Lin et al., 2018). To address this problem, Lin et al. (2018) compute attention weights using convolutional layers on the local pixel neighbourhood. Zhang et al. (2018) consider message passing to capture rich contextual information from multi-level feature maps and use a gating function to control the rate of message passing. Wang et al. (2018a) introduces a recurrent mechanism to gather multi-scale contextual information and iteratively refine convolutional features. Liu et al. (2021b) propose a Transformer-based network to unify RGB and RGB-D salient object detection.

All these object saliency techniques apply attention mechanisms on region or patch-level features to find the most salient areas, while suppressing areas that do not contribute to saliency. In our case, we compute attention explicitly on the object-level and determine which objects (not regions) are most relevant. We further use an attention mechanism with high-level scene semantics to guide the prediction of salient object ranks.

Wang et al. (2020) also apply attention onto object-level features obtained from an object detector. However, they take a video as input and generate object, noun, and verb features from three distinct sub-networks. They then perform attention between object and noun features, and between object and verb features to classify noun and verb, respectively, for action recognition. Both Recasens et al. (2015) and Gorji and Clark (2017) employ the *gaze-following* concept to find objects or regions that are likely gazed by humans. They incorporate a gaze-pathway that takes human head regions and locations to generate a mask. The mask indicates the

likely locations that humans would gaze towards. Combining with a saliency map, they produce the final gaze saliency. Unlike these works, our method is not limited to only social scenes. We explore attention shift among multiple generic objects, which is more challenging as objects that influence attention shift may not be present especially when the objects in a scene have limited interactions.

Fan et al. (2019a) introduce saliency shift across temporal video frames. The objective is to predict the shift of salient objects in dynamic scenes (inter-frame), but they do not capture ranking or attention shifts among objects within a frame (intra-frame).

2.4 Edge-Aware Saliency Methods

To further improve salient object detection, edge information is introduced to enhance object boundary segmentation quality (Zhang et al. 2017; Li et al. 2018). Two approaches have been considered to enforce object boundary quality. The first approach uses a hybrid loss, where a term for the edge loss is combined with the binary segmentation loss (Feng et al., 2019; Qin et al., 2019; Zhao & Wu, 2019) for saliency training. The second approach uses an edge detection module, and jointly trains it with salient object detection (Lin et al., 2019; Wang et al., 2019b; Zhao et al., 2019). Wu et al. (2019b) develop a module to bi-directionally pass messages and refine features between the salient object detection and edge detection tasks. Su et al. (2019) investigate the selectivity-invariance dilemma for salient object detection. They build a three-stream network consisting of a boundary localisation stream for effectively selecting salient boundaries, interior perception stream to capture object features that are invariant to appearance changes, and a transition compensation stream for completing the segmentation between object interiors and boundaries. With some similarities, Wei et al. (2020) propose to decompose ground-truth saliency label into body and detail maps. The detailed map contains the edges of salient objects and nearby pixels, while the body map concentrates on pixels near the body center of salient objects.

In this work, we couple an edge detection module with the mask segmentation branch. Unlike existing edge-based methods, our edge network performs edge detection on the object level, where the boundaries and mask of salient objects are segmented individually. This enables the network to focus segmentation on individual instances while reducing distracting features from areas outside the region of interest. Additionally, we implement the coupling of our mask segmentation and edge detection with a simpler design, where the interaction operations between the two streams are implemented by an addition and a subtraction operation. Addition operation improves the overall shape of object segmentation whilst the subtraction operation enhances boundary features.

3 New Attention Shift Dataset

3.1 Data Collection

To our knowledge, there are no large-scale datasets available for salient object ranking based on *attention shift*. Hence, we propose a new large-scale salient object ranking dataset, by combining the widely used MS-COCO dataset (Lin et al., 2014) with the SALICON dataset (Jiang et al., 2015). MS-COCO contains complex images with ground-truth object segmentation, while SALICON is built on top of MS-COCO to provide mouse-trajectory-based fixations. The SALICON dataset provides two sources of fixation data: 1) fixation point sequences and 2) fixation maps for each image. We exploit these two sources and consider three approaches to generate our ground-truth saliency rank annotations. The first approach awards higher saliency values to objects fixated early in a fixation sequence. The second approach uses the pixels intensity values from a fixation map. The third approach focuses on the order of distinct objects that were fixated without repetition. The first and second approaches are each expanded into four methods. In total, we consider nine methods (summarised in Table 1) to generate possible ground-truth annotations, which we will discuss below. We do not know which methods would reflect the way that humans rank multiple objects in terms of saliency. We carry out a user study in Sect. 3.2, and provide some analysis on our dataset in Sect. 3.3.

We consider up to top-10 objects in the user study, but use top-5 for saliency ranking prediction. We believe that top-5 ranks are a good setting, as they contain clear and easy to define ranks of the top-5 objects. In addition, it is challenging enough, and the saliency differences among the lower ranks (ranks beyond 5) often become minuscule and ambiguous. Further, humans can consistently identify up to 4-6 salient objects at a glance (He et al., 2017b; Kaufman et al., 1949). Hence, top-5 salient objects should be a reasonable choice for the number of GT ranks.

Approach 1: For each image, we follow the fixation points in a fixation sequence and assign descending saliency scores to the fixated image pixels. We repeat this scoring of pixels over all observers' fixation data. The saliency rank of an object can be computed by aggregating these saliency scores that the object contains (*i.e.*, the higher the aggregated score, the more salient the object and the higher the rank). The number of fixation points varies among observers, leading to a large difference in scores.

We first assign scores to pixel values using fixation points from the SALICON (Jiang et al., 2015) dataset. We then obtain the score for each object based on the values of pixels belonging to the object. Specifically, for every image $I \in \mathbb{R}^{W \times H}$ of dimension $W \times H$, there are N observers. Let F^j be the fixation sequence obtained from one of the N observers $j \in [1, N]$ and a fixation f_i^j with index order $i \in [1, t]$ that represents the ith fixation in the sequence F^j of length t. We assign a score to image pixel p if the fixation f_i^j falls on p using:

$$v_p = \sum_j^N \sum_i^t g(f_i^j), \quad \text{if } f_i^j = p, \qquad (1)$$

$$g(f_i^j) = 1 - \frac{i}{t}, \qquad (2)$$

where v_p denotes the score of pixel $p \in I$ aggregating from all N observers' fixation data. Function g takes the temporal order ith of a fixation point in the sequence into account, and assigns a lower value to a fixation point if it is latter in the sequence.

To build our dataset on the idea of attention shift for saliency ranking, we focus on the order of fixation points. By doing so, we are able to closely define ground-truth saliency rank based on the sequential shift of attention, as observed by humans (Koch & Ullman, 1987). We thus do not take into account the duration of the fixation points in our formulation for two main reasons. First, there are large variances in the duration of fixations among different observers. Second, it is difficult (if not impossible) to obtain the exact duration of

Table 1 Comparison of nine different methods for generating ground-truth saliency rank order

Approach	Method	Source Data	Process
1	*FixSeq-avg*	Fixation sequence	Average rank score
1	*FixSeq-max*	Fixation sequence	Max rank score
1	*FixSeq-avgPmax*	Fixation sequence	Average + maximum score
1	*FixSeq-avgMmax*	Fixation sequence	Average × maximum score
2	*FixMap-avg*	Fixation map	Average rank score
2	*FixMap-max*	Fixation map	Max rank score
2	*FixMap-avgPmax*	Fixation map	Average + maximum score
2	*FixMap-avgMmax*	Fixation map	Average × maximum score
3	*DistFixSeq*	Fixation sequence	First T unique fixation sequence

each fixation point whilst the fixations are obtained from a re-sampling process (Jiang et al., 2015). In contrast, using the order of fixation points would ensure that there is a consistent gap between the scores of each pair of consecutive fixation points, leading to a higher stability in the final object scoring.

Next, we try to accommodate the varying sizes of objects in an image. Larger objects may collect more fixations from observers and be considered more salient with higher ranks. However, small objects that are rare may also be more salient even if there are fewer fixations. As we are unsure which methods best model how human ranks, we develop four methods to aggregate scores for subsequent object saliency ranks, namely: (1st) *FixSeq-avg* (average score), (2nd) *FixSeq-max* (maximum score), (3rd) *FixSeq-avgPmax* (average + maximum score) and (4th) *FixSeq-avgMmax* (average × maximum score). Let o be one of the objects in an image I, $|o|$ be the number of pixels in o, and v_p^o be the score of a pixel $p \in o$ inside an object. We define:

$$FixSeq\text{-}avg(o, I) = \frac{1}{|o|} \sum_{p \in o} v_p^o, \qquad (3)$$

$$FixSeq\text{-}max(o, I) = \max_{p \in o}(v_p^o), \qquad (4)$$

$$FixSeq\text{-}avgPmax(o, I) = FixSeq\text{-}avg(o, I)$$
$$+FixSeq\text{-}max(o, I), \qquad (5)$$

$$FixSeq\text{-}avgMmax(o, I) = FixSeq\text{-}avg(o, I)$$
$$\times FixSeq\text{-}max(o, I). \qquad (6)$$

For a given image, *FixSeq-avg* (Eq. 3) calculates the final score of an object by taking the average values of pixels belonging to the object. It takes into account the size differences between objects. In *FixSeq-max* (Eq. 4), the final score of an object is the maximum value v_p^o of all its pixels. It ranks objects higher if they are observed earlier in the fixation sequence. It does not consider the object size. For the methods *FixSeq-avgPmax* (Eq. 5) and *FixSeq-avgMmax* (Eq. 6), we consider weighting the final scores by performing addition or multiplication with the results from Eqs. 3 and 4, respectively. The use of addition in *FixSeq-avgPmax* is a shorthand of averaging the effect of both *FixSeq-avg* and *FixSeq-max* values. *FixSeq-avgMmax* considers to weight *FixSeq-avg* by multiplying *FixSeq-max*. We then sort all objects in descending order of the saliency score, and each object is given a distinct rank.

Approach 2 We use the fixation maps in this approach as the source for the saliency score. We directly take intensity values from the fixation map as pixel scores v_p. Similar to **Approach 1**, we define four methods to generate the final scores for each object. Accordingly, we have (5th) *FixMap-avg* (average score), (6th) *FixMap-max* (maximum score), (7th) *FixMap-avgPmax* (average + maximum score) and
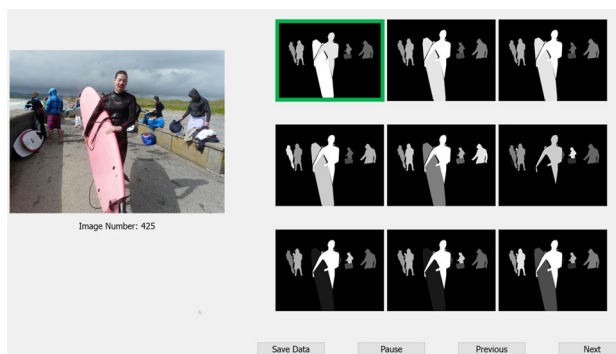

Fig. 3 Screenshot of the annotation tool used by the participants during the user study. Participants are not told how the maps are generated. They are asked to go through all maps and pick a map that best respects the "order of attractiveness". The green box indicates the map picked by one of the participants (Color figure online)

(8th) *FixMap-avgMmax* (average × maximum score). These four methods compute the final object scores in the same way as their counterparts in **Approach 1** (*i.e.*, Eq. 3-6). Again, we consider the first distinct T objects, and assign the saliency rank in the order of descending scores.

Approach 3 This approach also considers temporal order. However, we only focus on the first T distinct objects and ignore repeated fixations on already visited objects. In addition, we directly assign a score to the whole object if a fixation point resides in its segmentation. We term this method as (9th) *DistFixSeq*. Specifically, we define a new sequence $\hat{f}_i^n$ by removing fixations that fall on objects already visited by earlier fixations in f_i^n. We then define *DistFixSeq*, for each object o in an image I, as:

$$DistFixSeq(o, I) = \frac{1}{N} \sum_j^N \sum_i^T h(\hat{f}_i^j), \quad \text{if } \hat{f}_i^n \in o \qquad (7)$$

$$h(\hat{f}_i^n) = T - i, \qquad (8)$$

where $T = 10$. Function h assigns higher scores to objects if they are observed earlier. Equation 7 takes into account only the first T objects and averages the final scores across all N observers. We obtain the object ranks in the order of descending scores.

3.2 User Study

We perform a user study with 11 participants to find out which of these nine methods produce more consistent ground-truth attention shift order based on human judgment. For each image, the participants were presented with the image and the nine corresponding saliency rank maps arranged in a grid. Figure 3 shows an example screenshot of the annotation tool used in the user study. After a briefing session on how to use the annotation tool, each participant is told to observe
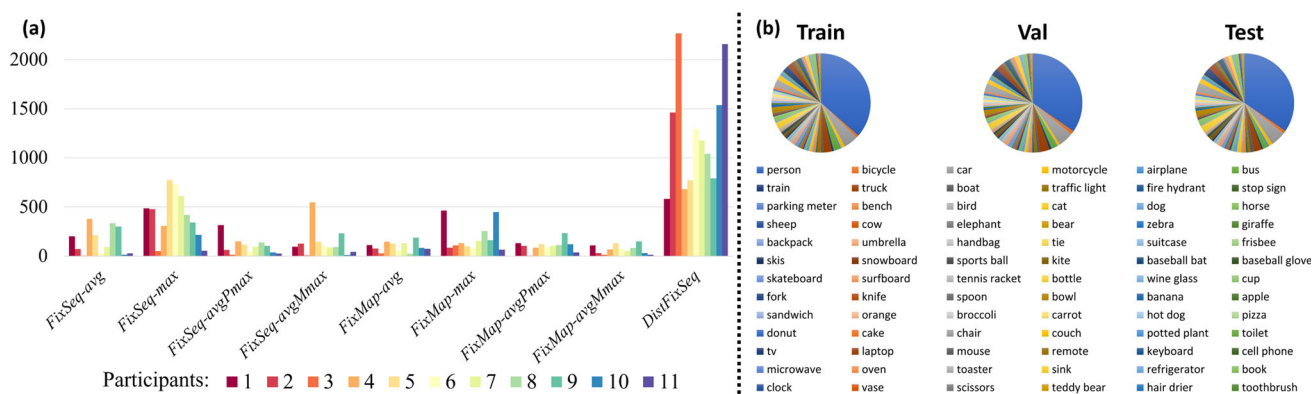
Fig. 4 **a** Pick rates of maps by 11 participants in our user study across 2500 images. These maps are generated by nine methods that we experimented with in Sect. 3.1. **b** Distribution of ground-truth salient instances of all object categories in each data split of our dataset
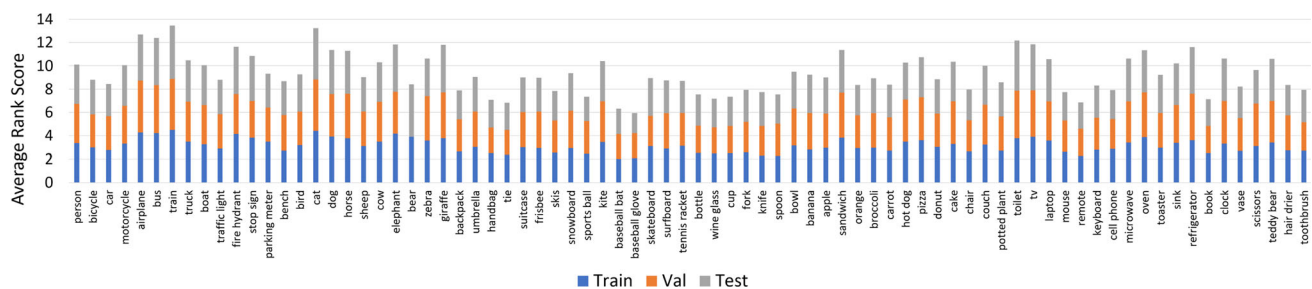

Fig. 5 Average rank score of each object category in the proposed dataset

the image first, and then pick a map that shows objects with "order of decreasing attractiveness". No other instructions or information were given, to avoid task-based bias and prior knowledge. The participants were considered to be in a neutral state condition, with no other effects (*e.g.*, emotions) influencing their visual attention and judgment. Participants are not told how the maps are generated. Each participant was asked to annotate a set of 2500 images. These images are randomly sampled from our dataset. Participants annotate them in 5 sessions (500 images each). Each annotation session lasts under an hour on average. It takes around 11 s on average for participants to select one of the nine generated maps for a single image. The user study took a total of 52.5 h to complete. After the annotation task, participants were rewarded with a £25 Amazon gift voucher.

Figure 4**a** shows that, on average, the map generated by (9th) *DistFixSeq* has the highest number of picks. The map aligns most with the order of attractiveness of objects. It suggests that the temporal order of fixated objects (attention shift) is vital for determining the strength of attractiveness among multiple objects. Attractiveness of objects is considered as attracting attention towards the objects and reflects their saliency strength (Zhang et al., 2008).

We can further see that there are more picks of the methods from **Approach 1** (maps generated from temporal fixation) than those from **Approach 2** (maps generated from fixation

map only, without temporal data). This suggests that ignoring the temporal fixation order, or using the order by fixation intensity alone, does not always capture the expected order of saliency (attractiveness of objects). These results correlate to the idea of attention shift by descending saliency values in Itti and Koch (2000), and prompt our definition of saliency rank order via attention shift. It supports us to use (9th) *DistFixSeq* from **Approach 3** to generate the ground-truth saliency ranking for the development of our rank prediction technique.

3.3 Dataset Analysis

Our dataset is adapted from MS-COCO (Lin et al., 2014) and SALICON (Jiang et al., 2015), and thus share similar characteristics. All existing popular datasets (*e.g.*, ECSSD (Yan et al., 2013), DUTS-OMRON (Yang et al., 2013), PASCAL-S (Li et al., 2014b), HKU-IS (Li & Yu, 2015), DUTS (Wang et al., 2017b)) target binary salient object detection while ours focuses on **salient object ranking**. Our dataset contains more complex images and is the largest in size. Note that all other datasets do not include individual object labels, making them ill-suited for our task.

We report that the average number of objects per image in our dataset is around 11 (maximum of 68). The "person" object category occurs most frequently in the dataset. This
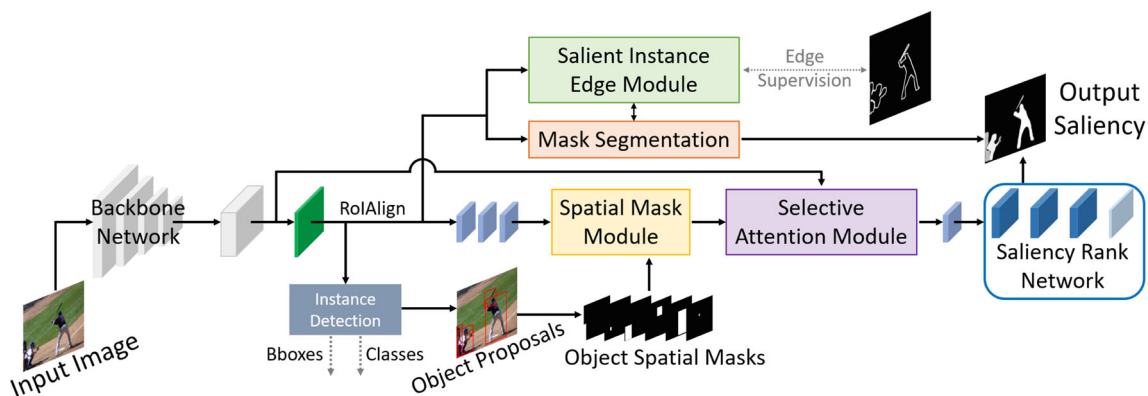
Fig. 6 Architecture Overview. Our model consists of a backbone network, Selective Attention Module (SAM), Spatial Mask Module (SMM), Salient Instance Edge Module (SIEM), and network for salient object ranking. We use Mask R-CNN (He et al., 2017a) as our bottom-up backbone to provide object proposals with FPN (Lin et al., 2017), and object mask segmentation from the segmentation branch. The mask segmentation is coupled with SIEM to exchange mask and edge information for boosting instance segmentation. The bottom-up SMM extracts low-level features of the proposed objects, while the top-down SAM extracts high-level contextual attention features

is expected as most photo images target people as the subject. Additionally, many images contain crowd of people with small individual annotations, causing the total count to be 4-16 times greater than other categories. Correspondingly, "person" objects receive the most instances of ground-truth saliency, which aligns with previous observations that humans usually attract attention (Judd et al., 2009). Figure 4b shows the distribution of ground-truth salient instances of each object category in our dataset. Figure 5 shows the average rank of each object category based on instances, given the ground-truth saliency. We can see that large objects (*e.g.*, "train", "airplane") have fewer instances per image, and some animal categories (*e.g.*, "cat", "dog", "elephant") have a larger rank average score than the "person" object. We also find that object categories relating to appliances (*e.g.*, "refrigerator", "microwave") have quite high scores, which mainly come from indoor scenes with no other object(s) of interest.

4 Proposed Approach

We propose a CNN-based model to predict saliency rank with a bottom-up bias stimuli (Itti et al., 1998; Borji & Itti, 2012), which we find useful in picking up the most salient objects in the scene. The saliency rank, especially on those less salient objects, may be related to the scene structure and observer interpretation (Einhäuser et al., 2008). Hence, saliency rank modelling requires higher-level cues and prior knowledge (Gao et al., 2009).

The proposed network architecture consists of five modules, namely, a backbone network based on Mask R-CNN (He et al., 2017a), a Spatial Mask Module (SMM), Selec-

tive Attention Module (SAM), Salient Instance Edge Module (SIEM) and a Saliency Rank Network, as illustrated in Fig. 6.

Mask R-CNN generates object proposals as a bottom-up approach similar to Anderson et al. (2018). This provides us with individual object features and allows us to learn semantics information on the object level in subsequent modules. We make a small modification to the existing instance detection branch in Mask R-CNN. Specifically, we modify the final object class detection layer to only predict two classes (salient or background). The instance detection branch is fine-tuned to detect salient objects on our dataset. We use spatial masks from SMM as a low-level cue, which embeds the relative size and location of each object in the image. SAM then compares the features of each object to the global semantic image features in order to determine relevant target salient objects. This module provides a top-down attention mechanism and is motivated by psychophysical findings that humans frequently gaze towards interesting objects. It encapsulates important scene semantics (Xu et al., 2014) and interpretation due to eye gazes (Einhäuser et al., 2008). We adopt the segmentation branch of Mask R-CNN to produce segmentation for the object instances. The segmentation branch is coupled with SIEM to jointly enhance the prediction of salient instance segmentation and edges. The pairing of instance segmentation and edges cooperatively boost their prediction. The joint supervision helps the network distinguish instances and their saliency rank from other objects and background. Such design is essentially different from existing edge-based methods (Sect. 2.4). Finally, we infer saliency rank of object instances with a small ranking network.
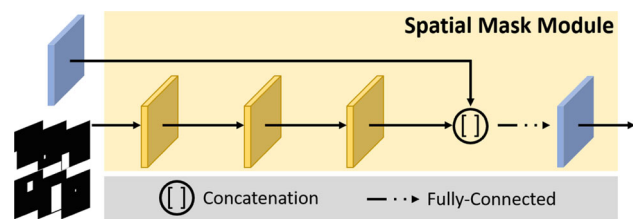
Fig. 7 Details of the spatial mask module (SMM)

4.1 Backbone Network

Objectness and object proposals for binary salient object detection have been explored (Feng et al. 2011; Siva et al. 2013; Zhang et al. 2016a). Feng et al. (2011) extend the global rarity principle (rare and less frequently occurring objects are likely to be salient) to derive object saliency. It uses a sliding-window mechanism to determine if the features inside the windows contain foreground or background features. Feng et al. (2011) and Zhang et al. (2016a) further extend it to many sliding windows of various scales. Fan et al. (2019b) present a model architecture much like the Mask R-CNN (He et al., 2017a). They produce object proposals by adopting the Feature Pyramid Network (FPN) (Lin et al., 2017) and propose a salient instance segmentation branch that extends the segmentation branch in the Mask R-CNN. The purpose of their network is to perform salient-instance segmentation, while we investigate salient object ranking based on attention shift order. Inspired by these works, we adopt Mask R-CNN as the backbone of our model and to provide efficient object proposals and segmentation. The FPN serves as a bottom-up attentive mechanism (Anderson et al., 2018).

To model saliency in the object-level, we apply RoIAlign (He et al., 2017a) and three fully connected layers (FCs) to extract object-level features, $o_i \in \mathbb{R}^{448}$, for each object proposal, leading to a set of object features $O = \{o_1, o_2, \ldots, o_M\}$, where $M = 12$ is the maximum number of object proposals. We further take the pyramid features "$P5$" from the FPN as the high-level features input to SAM for top-down attention. The segmentation branch generates pixel-wise segmentation of objects for a clearer final saliency map.

4.2 Spatial Mask Module (SMM)

Understanding the relationship between object properties and scene context can help select relevant targets in a complex scenario (Torralba, 2003). For example, very small objects in a scene may not attract human attention. Objects close to the centre of the image may be more salient due to the "center bias" concept (Yang et al., 2013; Judd et al., 2009). These motivate us to include low-level objects properties (*e.g.*, size

and locations) to learn contextual features that model the relationship between objects and scene.

Using the bounding boxes of object proposals, we generate a spatial mask for each object. Spatial masks embed the size and location of the proposed objects in relation to the visual scene. We capture such information with a binary mask (*i.e.*, assigning a value of 1 to pixels within a bounding box, and 0 otherwise). We pass the spatial masks through three convolutional layers to compress each of them into a 64-D feature vector. Each set of spatial features is then combined with the corresponding object features via a concatenation layer (Fig. 7). It is similar to the procedure of positional encoding in the Transformer (Vaswani et al., 2017) before the attention. This module can be considered as a process of combining bottom-up and semantic attributes of objects (Xu et al., 2014).

4.3 Selective Attention Module (SAM)

A straightforward choice to model how humans attend one object to another would be a recurrent strategy. Such a strategy is computation and memory-intensive, especially when there are a lot of objects in an image (like those in our proposed dataset). To model all relationships of objects and their associated attention shift probabilities in a potential sequence, it would easily lead to an exponential growth problem as the number of proposals increases. Instead of using recurrent strategy to model attention shift, we get inspirations from recent task-based techniques (Cao et al., 2015; Yang et al., 2016; Wang et al., 2017a; Vaswani et al., 2017; Ma et al., 2018; Wang et al., 2018c), which were greatly benefited from some forms of attention mechanisms. These mechanisms are often designed to dynamically weight relevant features or entities tailored to certain tasks while suppressing the distractors. Here, we consider that an attention mechanism would be useful to infer the way observers shift their attentions because it encapsulates important scene semantics (Xu et al., 2014) and interpretation due to eye gaze (Einhäuser et al., 2008). In addition, though human actors in an image would affect observers to shift their gazes (Gorji & Clark, 2017), we consider that individual generic objects may not necessarily have such a strong influence on attention shift. For generic images (*e.g.*, non-human scenes and images with little interactions among objects), we consider that the scene structure and relationship between objects may have a stronger influence on attention shift (Peters & Itti, 2007). We thus develop a Selective Attention Module (SAM) to compute top-down attention by comparing object features individually to the image scene features.

We build the attention module using Scaled Dot-Product Attention (Vaswani et al., 2017) (Fig. 8) with image and object features. We use the pyramid features, "$P5$", from the backbone network as the image features. A (1×1) convolu-
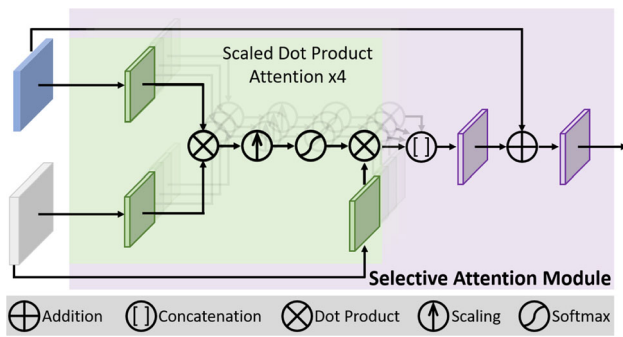
Fig. 8 Details of the Selective Attention Module (SAM)

Fig. 9 Details of the salient instance edge module (SIEM) and Mask Segmentation. M-Fuse and E-Fuse represent the interaction operations specific to the mask segmentation and edge module

tion and global average pooling are applied onto the pyramid features to obtain our high-level image representation.

Before computing the dot-product, we first project the object and image features into a 512-D space (Vaswani et al., 2017). Here, we embed the features of each object into a separate feature vector using a shared FC layer. Two separate feature vectors are generated with separate FC layers, both taking the pooled image features as input. The sets of new features from the pooled image features are further repeated M times. The attention mechanism then uses these embeddings to perform dot product similarity of individual object features with the image features. We add scaling factor (Vaswani et al., 2017), and apply softmax activation to obtain the attention score. Our attention module computes attention scores with multiple heads (4 heads) in parallel. The idea is that each attention head learns different high-level information to guide scoring/weighting for salient targets. The outputs from multiple attention heads are concatenated and then sent through a FC layer. Finally, we add a residual connection and a FC layer for the module output.

4.4 Salient Instance Edge Module (SIEM)

Our dataset contains images with many objects and noisy background features. In these images, salient objects may be very close to other objects and may have features similar to the background. This introduces noise to the boundaries of salient objects and can make it difficult to distinguish salient objects from other objects or the background. As a result, accurately segmenting and predicting salient ranks of multiple salient instances can be challenging. This can also be seen in Fig. 2d–e, where traditional methods do not explicitly capture individual objects and can have difficulties in differentiating between multiple objects. Based on these observations, we propose the Salient Instance Edge Module (SIEM), and jointly train it with the mask segmentation branch. The coupling of the instance mask and SIEM refines their predictions mutually. This further propagates enhancement to the salient instance segmentation and ranking tasks.
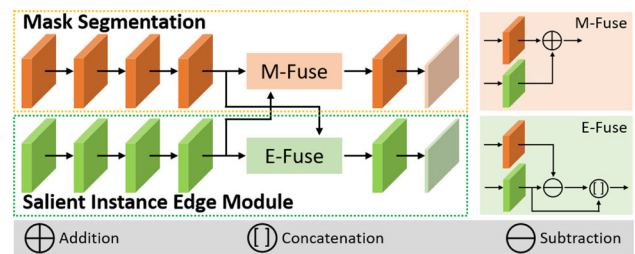
Both mask segmentation and SIEM have the same structure, but the connection between the networks differs. Figure 9 shows the structures of SIEM and the mask segmentation branch. Mask segmentation first contains four convolutional layers. We then add a fusion layer to combine the mask features with the edge features by addition. The final mask segmentation is then generated with a convolutional layer and a prediction layer. Likewise, SIEM contains four convolutional layers, a fusion layer, a convolutional layer and a prediction layer. Here, the fusion layer differs as we subtract the edge features from the mask segmentation features. The resulting features are then concatenated with a residual connection. The two fusion layers enable the networks to use features from the other network to effectively focus on their particular task. For example, the addition of edge features to the mask features (M-Fuse) allows the mask segmentation network to accurately capture the shape of instances, while the subtraction in the E-Fuse forces the network to focus around the boundary regions of the mask. Such overall design is essentially different from existing edge-based methods (Sect. 2.4).

4.5 Saliency Rank Network

We employ a simple ranking network to predict rank scores for salient instances. Our rank network consists of three fully connected layers and a scoring layer.

During inference, we combine the saliency rank scores with object segmentation (from the mask segmentation branch) to generate the final salient object rank map. Like (Islam et al., 2018), which determines object saliency rank by the descending average pixel saliency value of each object. We consider the top-5 saliency rank order of objects from their descending score values.

4.6 Loss Function

We define our training loss function as:

$$L = L_{inst} + L_{edge} + L_{rank} \qquad (9)$$

where L_{inst} includes the loss functions for bounding boxes, classification and segmentation as in the Mask R-CNN (He et al., 2017a). L_{edge} is the boundary loss as in (Cheng et al., 2020). Inspired by the learning to rank problem (Cao et al., 2007), we adopt the list-wise loss ListMLE (Xia et al., 2008), as our ranking loss L_{rank}. The loss computes the probability of the ideal permutation (ground-truth saliency rank order) and is defined according to the Plackett-Luce model (Marden, 1996).

$$L_{rank} = -\log P(y|x; g), \quad (10)$$

$$P(y|x; g) = \prod_{i=1}^{n} \frac{\exp(g(x_{y(i)}))}{\sum_{k=i}^{n} \exp(g(x_{y(k)}))}, \quad (11)$$

where the probability $P(y|x; g)$ decomposes the ideal permutation to the product of step-wise conditional probability. y is the ground-truth saliency rank order (ideal permutation). $g(x_{y(i)})$ provides the rank score function for the i-th conditional probability that an object is ranked at the i-th order, given the top $i - 1$ objects have been ordered correctly.

L_{inst} and L_{edge} range from [0-1], while L_{rank} range from greater than 1. The larger scale of L_{rank} can be considered as naturally adding more weighting to the loss, when compared with the other losses. Each loss is used to optimise a specific task (*e.g.*, instance detection, edge detection and saliency ranking) and they share some layers of the network. The summation of the losses leads to the optimisation of all tasks jointly. Due to the larger scale of L_{rank}, it provides more contribution to the overall loss and can push the other tasks (*e.g.*, instance detection and edge detection) to improve the detection of high rank salient objects. In this study, we leave the losses to their natural scale and do not adjust the weighting of the losses.

5 Experiments

5.1 Experimental Setup

Implementation Details: We fine-tune our backbone components of the Mask R-CNN on salient objects before training our final model on salient object ranking. A pre-trained ResNet-101 (He et al., 2016) is used to initialise the convolutional layers of the Mask R-CNN. All images for training and testing are resized to 1024×1024 before feeding into the network. During inference, we resize the output saliency map back to the original size of 640×480. Our model is built on top of the matterport Mask R-CNN framework (Abdulla, 2017) with TensorFlow and trained on an NVIDIA RTX 3090 GPU. During saliency rank training, we use a warm-up strategy by freezing the backbone layers and training the rest. We then fine-tune all the layers together. We set the mini-batch

size to 8 during warm-up and 2/4 for fine-tuning depending on memory limitations. We train variations of the network for up to 30 epochs for warm-up and 10 epochs for fine-tuning. We use the SGD optimizer with gradient norm clipping set to 5. The learning rate is set to 10^{-3} for warm-up. For fine-tuning, we set the learning rate to 10^{-8} for the backbone layers and 10^{-6} for the rest. Momentum and weight decay are configured as 0.9 and 10^{-4}, respectively.

Datasets: Our dataset employs the same set of images and fixation sequence from SALICON (Jiang et al., 2015), and contains object segmentation masks from MS-COCO (Lin et al., 2014). The SALICON dataset consists of 10K training, 5K validation and testing images, with no annotations for the test set. We use the training and validation sets to build our dataset. We consider saliency ranking based on the fixation sequence of the first 5 distinct objects visited without repetition (*DistFixSeq*, Sect. 3). The choice of the method is supported by our user study. We discard images with no object annotations, and images containing smaller objects that are completely enclosed by larger ones. Finally, we use images containing at least two salient objects (*i.e.*, at least two ranks) to ensure that we have attention shift for our ranking task. The dataset is randomly split into 7646 training, 1436 validation and 2418 test images. All images in our dataset are of size 480x640 (height, width).

Evaluation Metrics: We use the Salient Object Ranking (SOR) metric (Islam et al., 2018) for evaluation. It is formulated as the Spearman's Rank-Order correlation between the rank order of the predicted salient objects and the ground truth. The correlation metric measures the strength and direction of the monotonic relationship between two rank order lists. The measure in $[-1, 1]$ indicates the negative to positive correlation. We make a simple modification to the SOR metric, where during the calculation we make sure that there exists at least two objects predicted as salient with ranks and those objects are in the ground-truth. This ensures that there are at least two objects to correlate the rank order otherwise if only one object exists, then a full score is awarded even when there are missing salient objects. As a result of this modification, the SOR results in Table 2 differs from our previous results (Siris et al., 2020), which does not employ this minimum of two objects condition. Moreover, the metric is unable to cater for the case when there are no common objects between the two rank variables. For example, when one technique predicts a completely different set of objects from the ground truth, SOR is not defined. Hence, we further report the number of images used to compute the average SOR for the whole test set, where the more images used the more reliable the SOR is. The reported SOR measurements are normalised to [0,1].

Although the SOR metric provides a good evaluation for relative rank order, it does not penalise missing rank predictions and incorrect rank positions. Liu et al. (2021a) tries to

Table 2 Comparison with the state-of-the-art methods on our dataset

Method	MAE↓	SOR↑	#Images used↑	nDCG↑	Input Image Size	Train Time (hours)	Test Time (minutes)
S4Net (Fan et al., 2019b)	0.150	0.681	1661	0.674	320	~1.9	~3
BASNet (Qin et al., 2019)	0.115	0.696	2321	0.787	224	~60.0	~3
CPD-R (Wu et al., 2019a)	0.100	0.763	2394	0.837	352	~16.0	~5
SCRN (Wu et al., 2019b)	0.116	0.756	2416	**0.843**	352	~13.6	~5
PFANet (Zhao & Wu, 2019)	0.156	0.741	**2418**	0.834	256	~3.4	~3
EGNet (Zhao et al., 2019)	0.097	0.764	2413	*0.842*	480x640	~55.0	~12
ITSD (Zhou et al., 2020)	0.098	0.729	2416	0.823	288	~4.8	~3
MINet (Pang et al., 2020)	0.099	0.706	2415	0.814	320	~20.9	~25
LDF (Wei et al., 2020)	*0.093*	0.734	2413	0.828	256-352	~7.4	~2
CSNet (Gao et al., 2020)	0.136	0.738	**2418**	0.831	224	~45.0	~9
GateNet (Zhao et al., 2020)	0.094	0.719	*2417*	0.820	384	~20.3	~4
VST (Liu et al., 2021b)	*0.093*	*0.766*	2411	0.841	224	~10.0	~2
SCAS (Siris et al., 2021)	0.100	0.717	2118	0.758	640-800	~86.0	~10
RSDNet (Islam et al., 2018)	0.139	0.728	**2418**	0.827	321	~5.3	~6
IL-RSR (Liu et al., 2021a)	**0.091**	0.726	2239	0.785	480x640	~13.3	~4
SOR-PPA (Fang et al., 2021)	0.098	0.755	2351	0.801	480x640	~12.9	~2
ASSR (Siris et al., 2020)	0.101	0.764	2333	0.822	1024	~6.0	~12
Ours	**0.091**	**0.800**	2371	**0.843**	1024	~43.0	~12

The top section consists of salient object detection methods and middle section contains saliency ranking methods. The bottom section contains our methods, with ASSR (Siris et al., 2020) being our earlier CVPR model. Note that RSDNet scores are based on direct prediction with pre-trained weights from their datasets. #Images used refers to the number of predicted images usable for SOR calculation. Input Image Size: single value = same height and width, hyphen = range of image sizes, otherwise defined (height x width). The reported resolutions are the default input image size of each network (for fair comparison), which have been re-scaled from the original image size (480x640). Best scores are in bold, while second best scores are in italics

Fig. 10 Demonstration of the true rank position order being ignored by the SOR metric (Islam et al., 2018). The newly adopted nDCG metric provides a more reliable measure as it penalises both missing predictions and incorrect true rank position

resolve missing rank predictions by setting them to a value of 0 in their modified SA-SOR metric. However, if multiple rank predictions are missing and instead assigned a value of 0, they will have tied ranks and the SOR will not calculate accurate relative rank order. In terms of incorrect rank positions, this case is mainly found when the number of predicted ranks is less than the number of ground-truth ranks. The SOR metric ignores true rank positions and only considers the relative rank order, in the sense of whether an object is correctly ranked above or below other objects. Figure 10 demonstrates this issue of not penalising true rank positions. To tackle these problems, we adopt the nDCG (normalised Discounted Cumulative Gain) (Wang et al., 2013) to measure the consistency between the predicted and GT ranks. We define the metric as:

$$nDCG = \frac{DCG}{Ideal(DCG)}, \tag{12}$$

$$DCG = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}, \tag{13}$$

$$rel_i = 5 - abs(r_{GT} - r_p), \tag{14}$$

where rel_i is the relevance score of the ith object. k is the maximum number of GT objects per image. r_{GT} and r_p are the GT rank and predicted rank of object i. $Ideal(DCG)$ is the best score when the rank order is perfect, which is a relevance score of 5 for each object.

We also compare with the mean absolute error (MAE), which measures the average per-pixel difference between the prediction and ground truth. We calculate MAE between the original predicted saliency map and the ground-truth map, before any post-processing of saliency prediction to obtain the saliency rank. It is an alternative measure for the quality of both predicted saliency maps and ranks. It also works even when a technique predicts a completely different set of objects from the ground truth.

5.2 Quantitative Evaluation

We compare against three relative ranking methods: RSD-Net (Islam et al., 2018), IL-RSR (Liu et al., 2021a) and SOR-PPA (Fang et al., 2021). We also compare with thirteen state-of-the-art salient object detection methods: S4Net (Fan et al., 2019b), BASNet (Qin et al., 2019), CPD-R (Wu et al., 2019a), SCRN (Wu et al., 2019b), PFANet (Zhao & Wu, 2019), EGNet (Zhao et al., 2019), ITSD (Zhou et al., 2020), MINet (Pang et al., 2020), LDF (Wei et al., 2020), CSNet (Gao et al., 2020), GateNet (Zhao et al., 2020), VST (Liu et al., 2021b) and SCAS (Siris et al., 2021). Note that these methods (except S4Net and SCAS) do not predict object segmentation and instead only provide a single binary saliency map.

S4Net and SCAS have a similar structure to our backbone and output object instance segmentation. We modify both S4Net and SCAS in order to predict up to 6 classes (5 Ranks + 1 BG) for each object instead of the binary prediction as in their original papers (Fan et al., 2019b; Siris et al., 2021) for a fair comparison. We then use the predicted rank classification and descending score probabilities to obtain distinct saliency ranks. Similarly for IL-RSR, we re-train and test their network on our dataset, so that we can obtain rank predictions for calculating the nDCG score. We only make modification for the network to train/predict up to 5 ranks and make no other changes to the rest of the network and original source code. We note that there are discrepancies in the MAE and SOR scores between our results and the results reported in Liu et al. (2021a), even after using their original provided source code. For the rest of the salient object detection models and RSDNet, the predicted saliency ranks of ground-truth objects are obtained by averaging the pixel saliency values. The object rank is determined by descending order of such averages. We clarify that RSDNet is directly evaluated on our dataset using its pre-trained weights. When we try to adapt and train their model on our dataset (using their available source code), the model does not converge. We thus use their model with the pre-trained weights to evaluate on our dataset.

Table 2 shows the experimental results. It shows that our method outperforms other methods on the proposed dataset, achieving the best overall performance with better scores among all performance measurements (MAE, SOR and nDCG). Note that RSDNet, PFANet and CSNet use all images during the SOR calculation, as their single binary saliency maps often contain many false saliency. Noise or very weak saliency is often propagated throughout the image and reaches parts of objects. This allows RSDNet, PFANet and CSNet to obtain saliency rank by averaging object pixel values to cover most objects. However, their MAE and ranking performance is near the lower end. We have the best MAE performance tied with IL-RSR, although IL-RSR has much lower scores for SOR and nDCG rank metrics. One reason
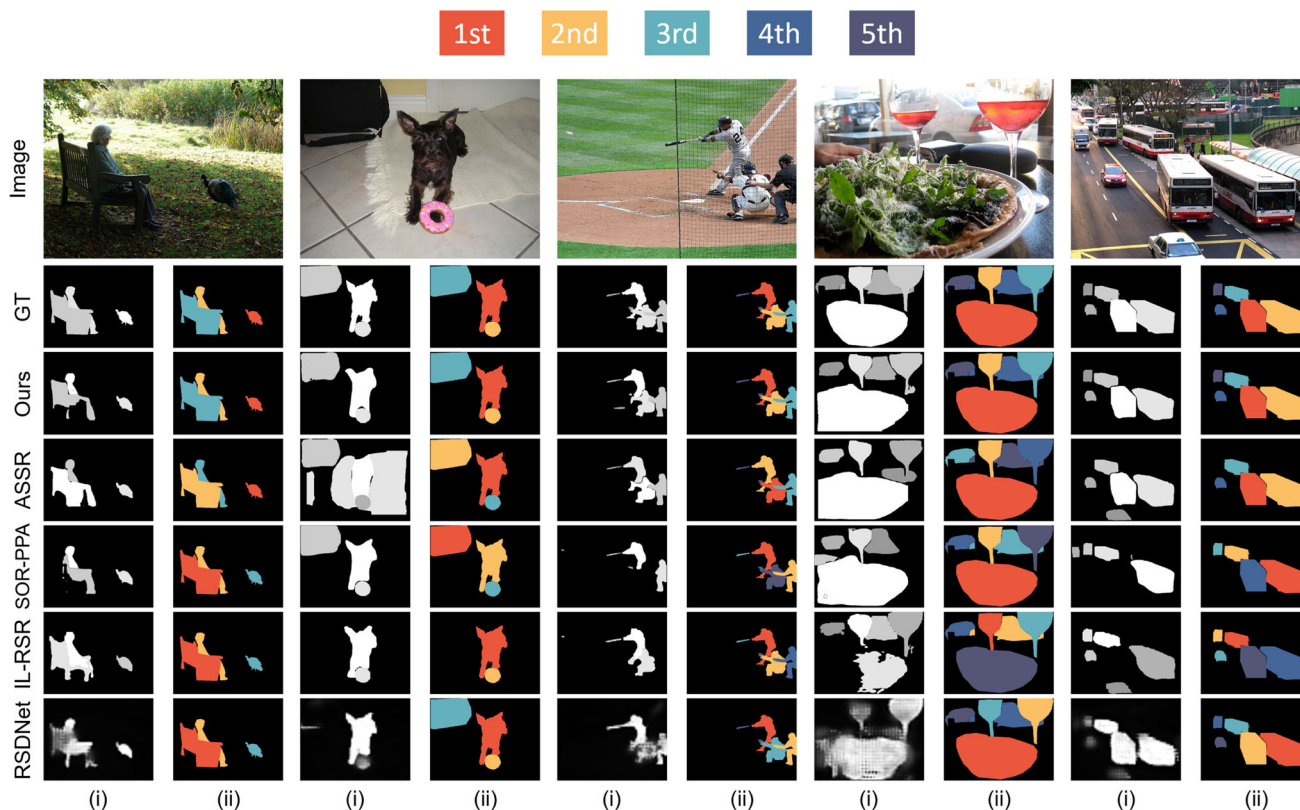
Fig. 11 Comparison of the proposed method against state-of-the-art methods designed for saliency ranking: ASSR (Siris et al., 2020), SOR-PPA (Fang et al., 2021), IL-RSR (Liu et al., 2021a), and RSDNet (Islam et al., 2018). The first row shows the input images, while the second row shows: (i) ground-truth saliency maps, and (ii) ground-truth saliency ranks. The following rows are predictions from the compared methods: (i) saliency prediction maps, and (ii) corresponding maps containing only the predicted ranks of ground-truth objects. Results in (ii) are used to obtain the predicted saliency ranks for our quantitative evaluation

for this is that IL-RSR is able to capture the overall segmentation of most GT objects, resulting in a good MAE score, but fails to predict the rank order between objects accurately, resulting in lower rank scores.

Our extended method proposed in this work has gained a significant improvement on the SOR rank metric from our initial work (Siris et al., 2020). It produces the best results with a 4.43 % performance gain over the second best method.

SCRN has the highest nDCG score also tied with our method, as nDCG awards higher scores for correct predictions of top ranks. Nonetheless, SCRN has much lower MAE and SOR scores than ours. This suggests that SCRN is mainly able to predict the top ranks well, but is unable to correctly predict the lower ranks.

In general, good rank predictions should translate into both high SOR and nDCG scores but low MAE score simultaneously.

As our network is based on the matterport Mask R-CNN framework (Abdulla, 2017), our default input image size is also 1024x1024 (Table 2). For fairness, we leave the network parameters of state-of-the-arts at their default. Our prelim-

inary (ASSR) and the new architecture both use the same default input size.

Table 2 also reports the training and testing run times. Our network is relatively large as the backbone is based on Mask R-CNN. This, together with the input size, causes our network to have longer training and inference times, although they are not the longest. Lightweight and quick inference speed can be considered in future work.

5.3 Qualitative Evaluation

Figure 11 showcases qualitative comparison results. We compare our method with state-of-the-art methods that are specifically designed for saliency ranking. We can see that the saliency maps obtained from RSDNet (Islam et al., 2018) often do not capture all the GT objects well, which can lead to incorrect rank predictions. Even if an instance is captured well by the instance detection method, it is still difficult to correctly rank the GT objects. In contrast, our method is able to segment the overall shape of most GT objects and correctly rank them. It is also able to rank multiple objects in

Table 3 Comparison with instance-based SOTA methods on generalization ability, with the dataset from Liu et al. (2021a)

Method	MAE↓	SOR↑	#Images used↑	nDCG↑
IL-RSR (Liu et al., 2021a)	*0.110*	**0.782**	2695	*0.903*
SOR-PPA (Fang et al., 2021)	0.119	0.532	*2744*	0.859
Ours	**0.099**	*0.739*	**2759**	**0.915**

"#Images used" refers to the number of predicted images usable for SOR calculation. Best and second best scores are in bold and italics

complex image scenarios, where there are cluttered and similar visual features among multiple objects (*e.g.*, fourth and fifth columns).

5.4 Evaluation on Additional Data

In addition to the experiments on our dataset, we also evaluate on the relative saliency ranking dataset from Liu et al. (2021a). Note that this dataset is not based on attention shift. They define saliency rank based on the order of maximum saliency intensity value from a fixation map. The fixation map does not consider the sequential order of fixations, but rather simply encapsulates the density of fixation points around objects. This does not follow the idea of saliency ranking from the order of sequential shift between multiple objects. We test our method on the dataset for generalisability. For this experiment, we modify both our network and SOR-PPA (Fang et al., 2021) to train and predict up to 8 GT saliency ranks. Table 3 compares between ours and two instance-based state-of-the-arts designed for saliency rank prediction. Note that the evaluation results are based on 2787 test images from the original 2929 test images. We find that the dataset in Liu et al. (2021a) contains 142 test images, where the rank data consists of multiple instances with the same rank and/or more than 8 GT ranks. Table 3 shows that our method achieves the best MAE and nDCG scores and second best on the SOR score. This indicates that our method can predict most of the ground-truth salient objects well. It is also able to predict correct ranks for top-ranked objects more consistently, and only experiences some difficulties for ordering lower-ranked objects. Overall, our method generalizes well to the dataset from Liu et al. (2021a), even though the dataset is not built according to our definition of attention shift for saliency ranking. Given that our method shows strong results on MAE and nDCG, we believe our method should have room for improvements (especially on SOR) with further adjustment on network configuration and training parameters.

Our task of salient object ranking requires the knowledge of objects. In addition, our network works best if it is trained and tested on a dataset with given object information. Our dataset and the dataset from Liu et al. (2021a) can be considered as a closed dataset, since the test sets mainly contain seen objects. Here, we test our network in a more open-
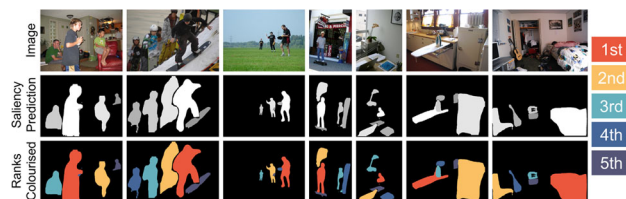

Fig. 12 Direct testing of our ranking network on the VRD (Lu et al., 2016) dataset

world setting with the VRD (Lu et al., 2016) dataset to show the generalisability of our method. We only show qualitative examples in Fig. 12 with direct testing, as the dataset does not provide GT saliency ranking data for quantitative evaluation. Figure 12 shows that our method is able to identify salient objects and rank them quite well (*e.g.*, first three images from left), especially if the image contains objects that are defined in our dataset. For object categories not defined in our dataset, our method also shows potential to capture them. For example, the shop sign and poster board (fourth image), lamp (fifth image), iron and ironing board (sixth image) and guitar (last image).

To our knowledge, there are no video datasets that contain object segmentation and human-eye fixation sequence data. To explore salient object ranking on the video level, we directly test our network onto the video VSPW (480p) (Miao et al., 2021) dataset. Again, we only show the qualitative results. Figure 13 shows that our network is able to capture and rank the main objects of interest within the video/frame. As the video progresses in time, the ranking of objects changes dynamically with the change of interest and actions developing in the scene. These examples show that improving our model for video application is a promising direction (Table 3).

5.5 Evaluation on Salient Instance Segmentation

Like S4Net (Fan et al., 2019b), our network is able to generate individual segmentation for each salient object instance. Hence, we further compare our network, IL-RSR (Liu et al., 2021a), SCAS (Siris et al., 2021), SOR-PPA (Fang et al., 2021) and ASSR (Siris et al., 2020) with S4Net on the salient instance segmentation task. We omit the other state-of-the-art methods from this experiment, as they are unable to
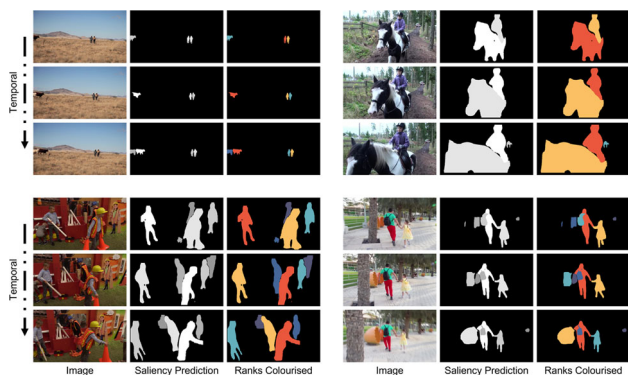
Fig. 13 Direct testing of our ranking network on the VSPW (480p) (Miao et al., 2021) dataset
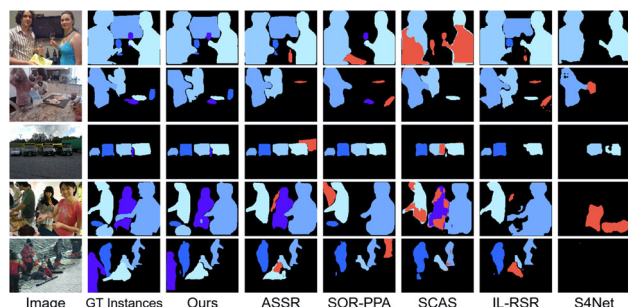

Fig. 14 Qualitative comparison for instance-based salient object segmentation with ASSR (Siris et al., 2020), SOR-PPA (Fang et al., 2021), SCAS (Siris et al., 2021), IL-RSR (Liu et al., 2021a), and S4Net (Fan et al., 2019b). Instances with *5 different shades of blue* are predicted instances that match with their corresponding GT instances. **Red** instances represent false predictions (Color figure online)

produce salient object instances. We use mean Average Precision (mAP^r, $r = 0.5/0.7$) to measure the performance, as in Fan et al. (2019b). Table 4 reports the results on our dataset. It shows that our network outperforms S4Net by a large margin. S4Net predicts very few salient objects when compared to our network (as shown in Fig. 14). It usually predicts only one or two positive salient instances over the images in the test set. In contrast, Fig. 14 shows that our network can predict multiple instances with good segmentation accuracy. It can capture smaller objects that are difficult to distinguish from the sur-

roundings. Other state-of-art methods tend to introduce false salient instances and fail to segment all GT instances. Our network also improves upon our previous work Siris et al. (2020) and achieves the best performance.

5.6 Ablation Study

We perform an ablation study to evaluate each of the proposed components, as shown in Table 5. The full model has the best overall performance across all metrics. It produces the highest SOR and nDCG scores. MAE is also tied best. Addition of each component to the base model of the proposed method generally improves the performance across the metrics. We see a more substantial gain in SOR performance when we add SIEM. This suggests that explicitly using edge information enables the network to improve object segmentation, consequently boosting rank order prediction. This also shows that SIEM helps to enhance the captured object features to be more distinctive from those of other close objects, and thus enable better saliency reasoning.

Table 5 also shows the numbers of parameters (#Parameters) and calculations (#FLOPS) for each ablated model. When adding our three modules to the base architecture (BbSR+SMM+SAM+SIEM vs BbSR), the number of parameters increases only by around ∼11.14%. For the number of calculations (#FLOPS), the increase is around ten-fold. This is largely due to the Spatial Mask Module (SMM) producing binary bounding box masks for each object.

Figure 15 compares the salient object segmentation accuracy between our full model with the Salient Instance Edge Module against the full model without the edge module. It shows that the proposed edge module coupled with mask segmentation enhances the capture of salient instances. Segmentation of both the body and instance boundaries is improved.

We also perform ablation of the interaction operations specific to the mask segmentation and edge detection streams. Table 6 shows that the interaction operations between the two streams do contribute to a small overall performance

Table 4 Quantitative comparison with S4Net for the salient instance segmentation task on our dataset

Method	mAP^r @0.5 (%)↑	mAP^r @0.7 (%)↑
S4Net (Fan et al., 2019b)	16.7	10.6
IL-RSR (Liu et al., 2021a)	48.2	38.3
SCAS (Siris et al., 2021)	38.6	27.6
SOR-PPA (Fang et al., 2021)	55.1	47.1
ASSR (Siris et al., 2020)	*60.6*	*51.0*
Ours	**64.4**	**53.8**

Note that we do not compare with other state-of-the-arts since they are unable to perform this task. Best scores are in bold, second best scores are in italics

Table 5 Ablation study of the proposed model

Method	MAE↓	SOR↑	#Images used↑	nDCG↑	#Parameters (M)	#FLOPS (G)
BbSR	0.096	0.787	*2373*	0.835	∼68.78	∼0.22
BbSR+SMM	*0.092*	0.793	**2377**	*0.842*	∼69.64	∼2.33
BbSR+SAM	0.093	0.788	2366	0.839	∼71.27	∼0.30
BbSR+SIEM	*0.092*	0.798	2372	*0.842*	∼73.17	∼0.23
BbSR+SMM+SAM	*0.092*	0.794	2366	*0.842*	∼72.04	∼2.40
BbSR+SMM+SIEM	**0.091**	*0.799*	*2373*	*0.842*	∼74.04	∼2.34
BbSR+SAM+SIEM	*0.092*	0.795	2371	0.841	∼75.67	∼0.31
BbSR+SMM+SAM+SIEM	**0.091**	**0.800**	2371	**0.843**	∼76.44	∼2.41

BbSR refers to the backbone network plus the small saliency rank network. **M** = million. **G** = giga. Best scores are in bold, second best scores are in italics
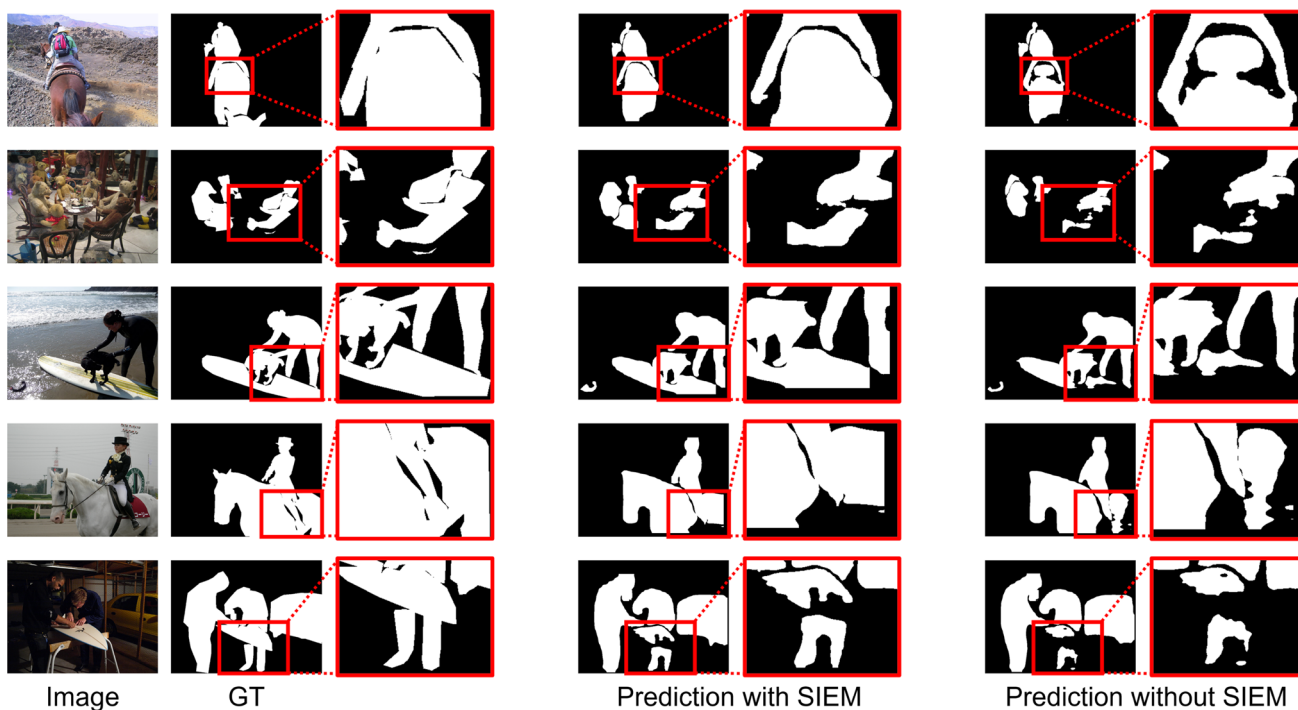
Fig. 15 Qualitative comparison of salient object segmentation between our full model with Salient Instance Edge Module (SIEM) versus full model without SIEM

Table 6 Ablation study of the interaction operations (E-Fuse/M-Fuse) between the coupled mask segmentation and edge detection streams

Method	MAE↓	SOR↑	#Images used↑	nDCG↑	#Parameters (M)	#FLOPS (G)
BbSR+SIEM(w/o E-Fuse/M-Fuse)	**0.0923**	0.7927	*2373*	0.8416	∼71.40	∼0.227
BbSR+SIEM(E-Fuse)	*0.0924*	0.7968	**2375**	**0.8432**	∼72.58	∼0.229
BbSR+SIEM(M-Fuse)	0.0926	**0.7989**	**2375**	0.8418	∼71.99	∼0.228
BbSR+SIEM(E-Fuse + M-Fuse)	**0.0923**	*0.7986*	2372	*0.8427*	∼73.17	∼0.230

BbSR refers to the backbone network plus the saliency rank network. **M** = million. **G** = giga. Best scores are in bold, second best scores are in italics
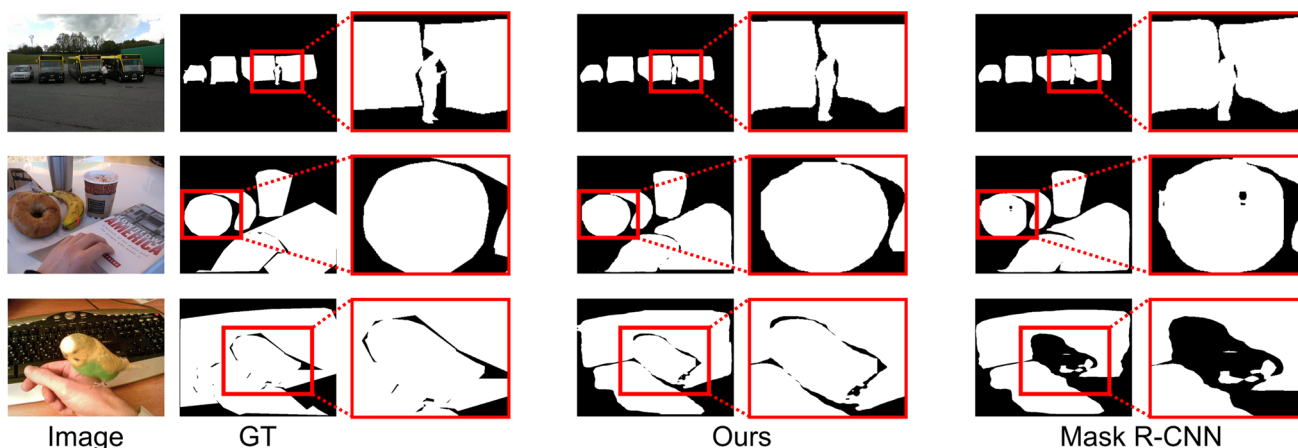
Fig. 16 Qualitative comparison of the predicted segmentation between our network and Mask R-CNN. Note that the predictions of Mask R-CNN have been filtered to show only predicted segmentation of objects that match with the ground truth

gain. We believe that further performance gain may potentially be possible if we adopt more cooperative features with deeper interactive designs between the mask segmentation branch and the edge detection module (Chen et al., 2020; Ke et al., 2021; Kim et al., 2021). We leave this to future work.

Considering that we utilise Mask R-CNN as our backbone architecture, we compare the segmentation quality between Mask R-CNN and our new network. Figure 16 shows that our network can segment individual objects well (top image), capture the whole body of objects (middle image), and do not mistakenly group different object segments together (bird parts mistaken as hand in bottom image).

Table 7 compares the performance between our baseline network (BbSR) with multi-class classification loss versus baseline network with list-wise ListMLE loss. The table shows that the preliminary architecture with multi-class classification loss has a better MAE score. However, this is largely due to the architecture predicting fewer salient objects, resulting in fewer false saliency values and lower overall average error from missing predictions of small low ranked salient objects. The preliminary architecture performs worse on the SOR ranking metric, which enforces correct relative rank order between multiple objects. The performance on the nDCG metric is also poor, as the metric penalises missing saliency and incorrect rank positions. Our new architecture obtains consistent and significant performance gains, especially in the ranking metrics

(SOR and nDCG) when we use the list-wise ListMLE loss instead.

Overall, these results show the effectiveness of the proposed components.

5.7 Saliency Ranking on Different Contexts

Our study proposes the first deep network to model human attention shift. Our approach is based on bottom-up and top-down inference, which aligns closely with human visual processing. In the design, we have not fully explored scene context (we have only used spatial context and global image features), yet the results are promising. Spatial context corresponds to the size and spatial location of objects in relation to the image scene. The global image context features correspond to prominent features in the image, which establish the scene setting.

Our network learns to reason the saliency rank of individual object features against the global features of an image scene. Such learning can also capture relationships between separate image features and corresponding saliency ranked objects. Figure 17 showcases examples of different image scenes containing "vehicles". The vehicle can be of different sizes, at different locations, or adjacent to other vehicles. Our network learns the relationships between the "vehicle" objects and context (spatial and global image features), and determines the correct saliency rank for each object accordingly.

Table 7 Ablation study of the baseline network with multi-class classification loss (our preliminary architecture) versus the baseline network with list-wise ListMLE loss (current architecture)

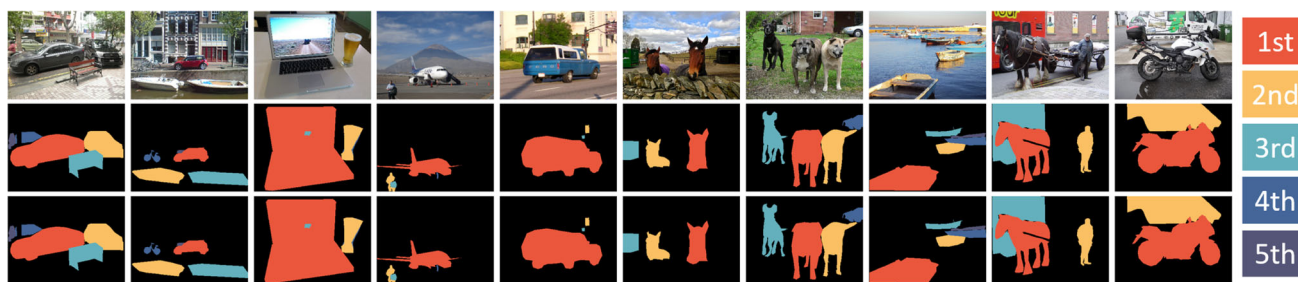| Method	MAE↓	SOR↑	#Images used↑	nDCG↑
BbSR(with multi-class classification loss)	**0.089**	*0.740*	*2253*	*0.798*
BbSR(with ListMLE loss)	*0.096*	**0.787**	**2373**	**0.835**

Best and second best scores are in bold and italics

Fig. 17 Example scenes containing objects in the *vehicle* category, with input images from our dataset (Top row), GT Ranks (Middle row), and rank predictions from our method (Last row)

6 Discussion

We have conducted a user study to determine the best method for saliency rank order. The user study involved participants viewing an image, then selecting one of nine maps that best mirrored the sequence of decreasing object attractiveness. The participants were only instructed to view the image freely without other influences (*e.g.*, prior knowledge (Li et al., 2014)) or task (Peters & Itti, 2007), other than to choose the best map after viewing the image. The method that leads to the highest-picked map reflects the idea of sequential attention shift. As a result, our saliency ranking is based on attention shift and our task definition is derived from a free-viewing task setting. For future work, it is interesting to explore task-specific settings and other factors which would influence saliency. For example, alternative approaches like task-based saliency ranking (*e.g.*, for driving simulation) can be investigated to develop new GT saliency ranks.

Our dataset is based on the MS-COCO dataset and can be considered a closed dataset. The dataset contains a wide range of object categories which are derived from 80 different object categories that correspond to MS-COCO object and segmentation data. Therefore, generalising our problem to cover unseen object categories is an interesting but also challenging problem for future work.

The purpose of this research is to investigate image-based salient object ranking based on attention shifts. For videos, there is currently no corresponding fixation sequence and object segmentation data. An interesting problem for future work would be to investigate salient object ranking at the video level.

In this study, we have introduced a list-wise ranking loss to improve our ranking method from our preliminary architecture. In our design, we keep a simplified saliency ranking network and the ranking loss contributes to significant performance gain. We believe that there is room to design new and powerful ranking methods (*e.g.*, complex ranking network, alternative ranking losses).

We present two examples of failure cases in Fig. 18. The first cause of failure can develop from our model relying
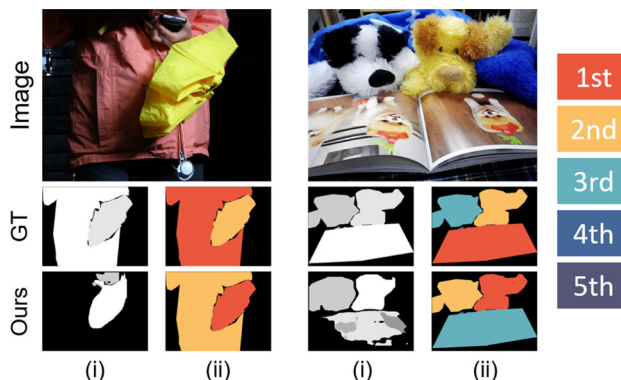

Fig. 18 Failure cases. (i) GT/predicted saliency maps, and (ii) GT/predicted maps corresponding only to the ranks of GT objects

on recognising objects in the image and then inferring the saliency rank on them. When an object is too close to the camera, our model may fail to recognise what it is (the person in the left image) and hence fail to infer correct rank order. Another failure case occurs from printed objects appearing inside a book (or picture) in the image, which may also be detected as separate objects by our model (the two dogs in the book of the right image).

7 Conclusion

In this paper, we have proposed to study a novel problem and presented the first saliency rank dataset, based on human attention shift. The dataset is motivated by psychological studies and behavioural observations. It is further supported by our user study that humans attend salient objects one at a time and in an order of decreasing values of saliency. We have also proposed a new saliency rank prediction approach to infer attention shift order. The proposed approach performs favourably against several state-of-the-art methods on the proposed saliency rank dataset as well as other existing datasets.

We find that our proposed method can correctly predict the top saliency ranks, but it does experience difficulties in

predicting lower ranks, especially on the dataset from Liu et al. (2021a), which is not constructed by means of human attention-shift. Correctly predicting the rank order for lower-ranks is very challenging as the differences among them can be subtle, and we leave it as a future work to explore.

Data Availability The datasets generated during and/or analysed during the current study are available in the GitHub repository, https://github.com/SirisAvishek/Attention_Shift_Ranks.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this manuscript.

References

Abdulla, W. (2017). Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN.

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. *In CVPR*, pp. 6077–6086.

Arvanitis, G., Stagakis, N., Zacharaki, E. I., & Moustakas, K. (2023). Cooperative saliency-based obstacle detection and ar rendering for increased situational awareness. arXiv preprint arXiv:2302.00916.

Borji, A. (2012). Boosting bottom-up and top-down visual features for saliency estimation. *In CVPR*, pp. 438–445.

Borji, A. (2018). Saliency prediction in the deep learning era: Successes, limitations, and future challenges. arXiv preprint arXiv:1810.03716.

Borji, A. & Itti, L. (2012). Exploiting local and global patch rarities for saliency detection. *In CVPR*, pp. 478–485.

Borji, A., Sihite, D. N., & Itti, L. (2012). Probabilistic learning of task-specific visual attention. *In CVPR*, pp. 470–477.

Cao, C., Liu, X., Yang, Y., Yu, Y., Wang, J., Wang, Z., Huang, Y., Wang, L., Huang, C., Xu, W., et al. (2015). Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. *In ICCV*, pp. 2956–2964.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., & Li, H. (2007). Learning to rank: from pairwise approach to listwise approach. *In ICML*, pp. 129–136.

Chang, C.-K., Siagian, C., & Itti, L. (2010). Mobile robot vision navigation & localization using gist and saliency. *In IROS*, pp. 4147–4154. IEEE.

Chen, X., Lian, Y., Jiao, L., Wang, H., Gao, Y., & Lingling, S. (2020). Supervised edge attention network for accurate image instance segmentation. *In ECCV 2020*, pp. 617–631. Springer.

Cheng, T., Wang, X., Huang, L., & Liu, W. (2020). Boundary-preserving mask r-cnn. *In ECCV*, pp. 660–676. Springer.

Cornia, M., Baraldi, L., Serra, G., & Cucchiara, R. (2018). Paying more attention to saliency: Image captioning with saliency and context attention. *TOMM, 14*(2), 1–21.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience, 18*(1), 193–222.

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision, 8*(14), 18–18.

Fan, D.-P., Wang, W., Cheng, M.-M., & Shen, J. (2019a). Shifting more attention to video salient object detection. *In CVPR*, pp. 8554–8564.

Fan, R., Cheng, M.-M., Hou, Q., Mu, T.-J., Wang, J., & Hu, S.-M. (2019b). S4net: Single stage salient-instance segmentation. *In CVPR*, pp. 6103–6112.

Fang, H., Zhang, D., Zhang, Y., Chen, M., Li, J., Hu, Y., Cai, D., and He, X. (2021). Salient object ranking with position-preserved attention. *In ICCV*, pp. 16331–16341.

Feng, J., Wei, Y., Tao, L., Zhang, C., & Sun, J. (2011). Salient object detection by composition. *In ICCV*, pp. 1028–1035.

Feng, M., Lu, H., & Ding, E. (2019). Attentive feedback network for boundary-aware salient object detection. *In CVPR*, pp. 1623–1632.

Gao, D., Han, S., & Vasconcelos, N. (2009). Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *TPAMI, 31*(6), 989–1005.

Gao, S.-H., Tan, Y.-Q., Cheng, M.-M., Lu, C., Chen, Y., & Yan, S. (2020). Highly efficient salient object detection with 100k parameters. *In ECCV*, pp. 702–721. Springer.

Gorji, S. & Clark, J. J. (2017). Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. *In CVPR*, pp. 2510–2519.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017a). Mask R-CNN. *In ICCV*, pp. 2980–2988.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In CVPR*, pp. 770–778.

He, S., Jiao, J., Zhang, X., Han, G., & Lau, R. W. (2017b). Delving into salient object subitizing and detection. *In ICCV*, pp. 1059–1067.

Hou, Q., Cheng, M.-M., Hu, X., Borji, A., Tu, Z., & Torr, P. H. (2017). Deeply supervised salient object detection with short connections. *In CVPR*, pp. 3203–3212.

Islam, M. A., Kalash, M., & Bruce, N. D. B. (2018). Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. *In CVPR*, pp. 7142–7150.

Itti, L. & Koch, C. (1999). Comparison of feature combination strategies for saliency-based visual attention systems. *In Human Vision and Electronic Imaging IV*, vol. 3644, pp. 473–482. International Society for Optics and Photonics.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research, 40*(10–12), 1489–1506.

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *TPAMI, 20*(11), 1254–1259.

Jiang, M., Huang, S., Duan, J., & Zhao, Q. (2015). Salicon: Saliency in context. *In CVPR*, pp. 1072–1080.

Jiao, J., Wei, Y., Jie, Z., Shi, H., Lau, R. W., & Huang, T. S. (2019). Geometry-aware distillation for indoor semantic segmentation. *In CVPR*, pp. 2869–2878.

Judd, T., Ehinger, K., Durand, F., & Torralba, A. (2009). Learning to predict where humans look. *In ICCV*, pp. 2106–2113.

Kalash, M., Islam, M. A., & Bruce, N. D. (2019). Relative saliency and ranking: Models, metrics, data and benchmarks. *TPAMI, 43*(1), 204–219.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *AJP, 62*(4), 498–525.

Ke, L., Tai, Y.-W., & Tang, C.-K. (2021). Deep occlusion-aware instance segmentation with overlapping bilayers. *In CVPR*, pp. 4019–4028.

Kim, M., Woo, S., Kim, D., & Kweon, I. S. (2021). The devil is in the boundary: Exploiting boundary representation for basis-based instance segmentation. *In WACV*, pp. 929–938.

Koch, C. & Ullman, S. (1987). Shifts in selective visual attention: towards the underlying neural circuitry. *In Matters of Intelligence*, pp. 115–141. Springer.

Lai, B. & Gong, X. (2016). Saliency guided dictionary learning for weakly-supervised image parsing. *In CVPR*, pp. 3630–3639.

Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., & Dai, Y. (2021). Uncertainty-aware joint salient object and camouflaged object detection. *In CVPR*, pp. 10071–10081.

Li, G. & Yu, Y. (2015). Visual saliency based on multiscale deep features. *In CVPR*, pp. 5455–5463.

Li, J., Tian, Y., & Huang, T. (2014). *Visual saliency with statistical priors. IJCV, 107*, 239–253.

Li, J., Xu, D., & Gao, W. (2012). Removing label ambiguity in learning-based visual saliency estimation. *TIP, 21*(4), 1513–1525.

Li, X., Yang, F., Cheng, H., Liu, W., & Shen, D. (2018). Contour knowledge transfer for salient object detection. *In ECCV*, pp. 355–370.

Li, Y., Hou, X., Koch, C., Rehg, J. M., & Yuille, A. L. (2014b). The secrets of salient object segmentation. *In CVPR*, pp. 280–287.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. *In CVPR*, pp. 2117–2125.

Lin, T.-Y., Maire, M., Belongie, S. J., Bourdev, L. D., Girshick, R. B., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. *In ECCV*, pp. 740–755.

Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J., & Jiang, J. (2019). A simple pooling-based design for real-time salient object detection. *In CVPR*, pp. 3917–3926.

Liu, N., Han, J., & Yang, M.-H. (2018). Picanet: Learning pixel-wise contextual attention for saliency detection. *In CVPR*, pp. 3089–3098.

Liu, N., Li, L., Zhao, W., Han, J., & Shao, L. (2021a). Instance-level relative saliency ranking with graph reasoning. *TPAMI*.

Liu, N., Zhang, N., Wan, K., Shao, L., & Han, J. (2021b). Visual saliency transformer. *In ICCV*, pp. 4722–4732.

Lu, C., Krishna, R., Bernstein, M., & Fei-Fei, L. (2016). Visual relationship detection with language priors. *In ECCV*, pp. 852–869. Springer.

Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., & Fan, D.-P. (2021). Simultaneously localize, segment and rank the camouflaged objects. *In CVPR*, pp. 11591–11601.

Ma, C.-Y., Kadav, A., Melvin, I., Kira, Z., AlRegib, G., & Peter Graf, H. (2018). Attend and interact: Higher-order object interactions for video understanding. *In CVPR*, pp. 6790–6800.

Marden, J. I. (1996). *Analyzing and modeling rank data*. CRC Press.

Miao, J., Wei, Y., Wu, Y., Liang, C., Li, G., & Yang, Y. (2021). Vspw: A large-scale dataset for video scene parsing in the wild. *In CVPR*, pp. 4133–4143.

Neisser, U. (2014). *Cognitive Psychology: Classic Edition*. Psychology Press.

Palazzi, A., Abati, D., Solera, F., & Cucchiara, R. (2018). Predicting the driver's focus of attention: the dr (eye) ve project. *TPAMI, 41*(7), 1720–1733.

Pan, J., Sayrol, E., Giro-i Nieto, X., McGuinness, K., & O'Connor, N. E. (2016). Shallow and deep convolutional networks for saliency prediction. *In CVPR*, pp. 598–606.

Pang, Y., Zhao, X., Zhang, L., & Lu, H. (2020). Multi-scale interactive network for salient object detection. *In CVPR*, pp. 9413–9422.

Park, J. H., Gutenko, I., & Kaufman, A. E. (2017). Transfer function-guided saliency-aware compression for transmitting volumetric data. *IEEE Transactions on Multimedia, 22*(9), 2262–2277.

Peters, R. J. & Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. *In CVPR*, pp. 1–8.

Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., & Jagersand, M. (2019). Basnet: Boundary-aware salient object detection. *In CVPR*, pp. 7479–7489.

Recasens, A., Khosla, A., Vondrick, C., & Torralba, A. (2015). Where are they looking? *In NeurIPS*, pp. 199–207.

Schillaci, G., Bodiroža, S., & Hafner, V. V. (2013). Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics, 5*(1), 139–152.

Siris, A., Jiao, J., Tam, G. K., Xie, X., & Lau, R. W. (2020). Inferring attention shift ranks of objects for image saliency. *In CVPR*, pp. 12133–12143.

Siris, A., Jiao, J., Tam, G. K., Xie, X., & Lau, R. W. (2021). Scene context-aware salient object detection. *In ICCV*, pp. 4156–4166.

Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., & Wetzstein, G. (2018). Saliency in vr: How do people explore virtual environments? *TVCG, 24*(4), 1633–1642.

Siva, P., Russell, C., Xiang, T., & Agapito, L. (2013). Looking beyond the image: Unsupervised learning for object saliency and detection. *In CVPR*, pp. 3238–3245.

Song, H., Wang, W., Zhao, S., Shen, J., & Lam, K.-M. (2018). Pyramid dilated deeper convlstm for video salient object detection. *In ECCV*, pp. 715–731.

Su, J., Li, J., Zhang, Y., Xia, C., & Tian, Y. (2019). Selectivity or invariance: Boundary-aware salient object detection. *In ICCV*, pp. 3799–3808.

Tang, L., Li, B., Zhong, Y., Ding, S., & Song, M. (2021). Disentangled high quality salient object detection. *In ICCV*, pp. 3580–3590.

Tavakoli, H. R., Shetty, R., Borji, A., & Laaksonen, J. (2017). Paying attention to descriptions generated by image captioning models. *In ICCV*, pp. 2487–2496.

Torralba, A. (2003). Modeling global scene factors in attention. *JOSA A, 20*(7), 1407–18.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *In NeurIPS*, pp. 5998–6008.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., & Tang, X. (2017a). Residual attention network for image classification. *In CVPR*, pp. 3156–3164.

Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., & Ruan, X. (2017b). Learning to detect salient objects with image-level supervision. *In CVPR*, pp. 136–145.

Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., & Borji, A. (2018a). Detect globally, refine locally: A novel approach to saliency detection. *In CVPR*, pp. 3127–3135.

Wang, W., Shen, J., Cheng, M.-M., & Shao, L. (2019a). An iterative and cooperative top-down and bottom-up inference network for salient object detection. *In CVPR*, pp. 5968–5977.

Wang, W., Shen, J., Dong, X., & Borji, A. (2018b). Salient object detection driven by fixation prediction. *In CVPR*, pp. 1711–1720.

Wang, W., Zhao, S., Shen, J., Hoi, S. C., & Borji, A. (2019b). Salient object detection with pyramid attention and salient edges. *In CVPR*, pp. 1448–1457.

Wang, X., Girshick, R., Gupta, A., & He, K. (2018c). Non-local neural networks. *In CVPR*, pp. 7794–7803.

Wang, X., Zhu, L., Wu, Y., & Yang, Y. (2020). Symbiotic attention for egocentric action recognition with object-centric alignment. *TPAMI*.

Wang, Y., Wang, L., Li, Y., He, D., & Liu, T.-Y. (2013). A theoretical analysis of ndcg type ranking measures. *In COLT*, pp. 25–54. PMLR.

Wei, J., Wang, S., Wu, Z., Su, C., Huang, Q., & Tian, Q. (2020). Label decoupling framework for salient object detection. *In CVPR*, pp. 13025–13034.

Wu, Z., Su, L., & Huang, Q. (2019a). Cascaded partial decoder for fast and accurate salient object detection. *In CVPR*, pp. 3907–3916.

Wu, Z., Su, L., & Huang, Q. (2019b). Stacked cross refinement network for edge-aware salient object detection. *In ICCV*, pp. 7264–7273.

Xia, F., Liu, T.-Y., Wang, J., Zhang, W., & Li, H. (2008). Listwise approach to learning to rank: theory and algorithm. *In ICML*, pp. 1192–1199.

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision, 14*(1), 28–28.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015a). Show, attend and tell: Neural image caption generation with visual attention. *In ICML*, pp. 2048–2057.

Xu, M., Ren, Y., & Wang, Z. (2015b). Learning to predict saliency on face images. *In ICCV*, pp. 3907–3915.

Yan, Q., Xu, L., Shi, J., & Jia, J. (2013). Hierarchical saliency detection. *In CVPR*, pp. 1155–1162.

Yang, C., Zhang, L., Lu, H., Ruan, X., & Yang, M.-H. (2013). Saliency detection via graph-based manifold ranking. *In CVPR*, pp. 3166–3173.

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. *In CVPR*, pp. 21–29.

Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., & Mech, R. (2016a). Unconstrained salient object detection via proposal subset optimization. *In CVPR*, pp. 5733–5742.

Zhang, L., Dai, J., Lu, H., He, Y., & Wang, G. (2018). A bi-directional message passing model for salient object detection. *In CVPR*, pp. 1741–1750.

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision, 8*(7), 32–32.

Zhang, L., Yang, C., Lu, H., Ruan, X., & Yang, M.-H. (2016). Ranking saliency. *TPAMI, 39*, 1892–1904.

Zhang, P., Wang, D., Lu, H., Wang, H., & Ruan, X. (2017). Amulet: Aggregating multi-level convolutional features for salient object detection. *In ICCV*, pp. 202–211.

Zhao, J.-X., Liu, J.-J., Fan, D.-P., Cao, Y., Yang, J., & Cheng, M.-M. (2019). Egnet: Edge guidance network for salient object detection. *In ICCV*, pp. 8779–8788.

Zhao, R., Ouyang, W., & Wang, X. (2013). Person re-identification by salience matching. *In ICCV*, pp. 2528–2535.

Zhao, T. & Wu, X. (2019). Pyramid feature attention network for saliency detection. *In CVPR*, pp. 3085–3094.

Zhao, X., Pang, Y., Zhang, L., Lu, H., & Zhang, L. (2020). Suppress and balance: A simple gated network for salient object detection. *In ECCV*, pp. 35–51. Springer.

Zhou, H., Xie, X., Lai, J.-H., Chen, Z., & Yang, L. (2020). Interactive two-stream decoder for accurate and fast saliency detection. *In CVPR*, pp. 9141–9150.