Inferring Attention Shift Ranks of Objects for Image Saliency

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

Abstract

Psychology studies and behavioural observation show that humans shift their attention from one location to another when viewing an image of a complex scene. This is due to the limited capacity of the human visual system in simultaneously processing multiple visual inputs. The sequential shifting of attention on objects in a non-task oriented viewing can be seen as a form of saliency ranking. Although there are methods proposed for predicting saliency rank, they are not able to model this human attention shift well, as they are primarily based on ranking saliency values from binary prediction. Following psychological studies, in this paper, we propose to predict the saliency rank by inferring human attention shift. Due to the lack of such data, we first construct a large-scale salient object ranking dataset. The saliency rank of objects is defined by the order that an observer attends to these objects based on attention shift. The final saliency rank is an average across the saliency ranks of multiple observers. We then propose a learning-based CNN to leverage both bottom-up and top-down attention mechanisms to predict the saliency rank. Experimental results show that the proposed network achieves state-of-the-art performances on salient object rank prediction. Code and dataset are available at https://github.com/SirisAvishek/Attention_Shift_Ranks.

1. Introduction

Research in saliency detection has grown extensively in recent years, with the aim of locating objects or regions that attract human visual attention. A good saliency detection technique benefits many high-level applications such as image parsing [28], image captioning [63] and person re-identification [74, 75]. Many methods are proposed that model salient object detection as a binary prediction problem. Very few works explicitly model human attention shift from one object to another.

Humans, however, are shown to have the ability to sequentially select and shift attention from one region/object to another [22, 27]. Such an ability is to deal with multiple simultaneous visual inputs, given the limited capacity of the human visual system [40]. Modelling this ability is important for the understanding of how humans interpret images, and helps improve performance of relevant applications, *e.g.*, autonomous driving [41] and robot-human interactions [48].

Figure 1: First row shows a sample of PASCAL-S datatset [34] which is used for saliency ranking in [1]. Note that multiple objects can be given the same saliency rank. Second row shows a sample from our proposed dataset with distinct ground-truth saliency ranks motivated by psychological studies. The colour (orange→purple) indicates the saliency rank 1→5.

Some early applications of attention shift include visual search [22] and scene analysis [23]. The attended regions are guided by a saliency map representing the conspicuity of each region in a scene. Attention shift is then modelled as shifting of attention from one region to another in an order of decreasing values in the saliency map [21, 27]. These early works estimate the saliency map only based on low-level features (*e.g.*, colour, intensity and orientation). Recently, Gorji and Clark [15] model "Attentional Push", which refers to how scene actors (humans) may manipulate the attention (gaze direction and location) of observers in viewing an image. The work heavily relies on the "gaze-

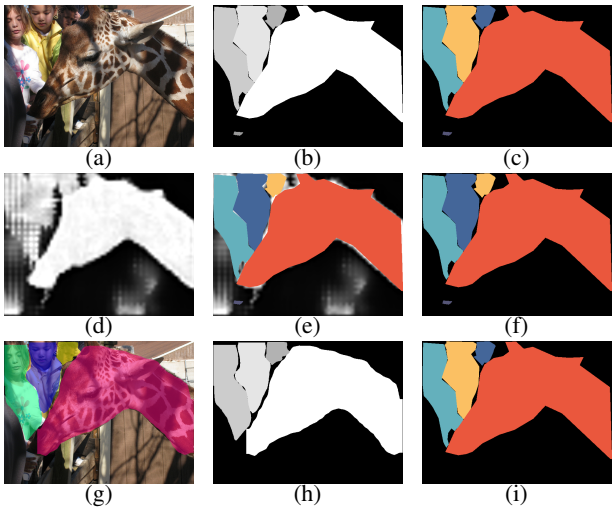

Figure 2: (a) image from our dataset, (b) corresponding ground-truth (GT) saliency rank, (c) corresponding GT saliency rank (colourised), (d) saliency rank prediction by RSDNet [1], (e) corresponding saliency rank by RSDNet (GT objects overlaid and colourised), (f) corresponding saliency rank by RSDNet with only GT objects (overlaid and colourised), (g) salient object and segmentation proposed from our model, (h) our salient object rank prediction, (i) our corresponding saliency rank with only GT objects (overlaid and colourised).

following" concept [46], which limits attention to a single shift from a person in a scene to some other region. Islam *et al*. [1] introduce the problem of relative ranking of salient regions and apply them to rank on ground-truth salient objects from an existing PASCAL-S dataset [34]. The relative ranking is inferred from the agreement of binary object saliency among multiple observers. The study 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 (Fig. 1, top row). Simultaneous attention to multiple objects, however, is not supported by behavioural observation because dividing attention between multiple objects often lead to poorer performance [10] and may not truly reflect how humans shift their attentions. Multiple objects with the same rank would also make it hard to model the order of attention shift.

Saliency ranking of objects is impactful to many vision areas and beyond. It is useful for fine-grained saliency detection and current applications utilizing traditional salient object detection. Saliency ranks provides the priorities of objects attended, where such rank priorities would benefit tasks that require the understanding of human visual processing (*e.g*., compression and streaming of 8k video).

Inspired by the aforementioned saliency and psychological studies, we aim to investigate saliency rank that models human attention shift in this paper. We first propose a new saliency ranking dataset collected based on attention shift. Our idea follows psychology studies that humans attend one object at a time in a complex scene. 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 shift towards objects of lower saliency values). Since different observers may have different saliency ranks on objects, we take the average of the saliency ranks from multiple observers to obtain the ground-truth saliency rank (Sec. 3.2). We show, with a user study, that such human attention shift on object instances correlates with object saliency rank. Fig. 1 (bottom row) shows one sample. Each object in an image is assigned a distinct saliency rank (1-5) that associates to the order of attention shift. Traditional saliency models often introduce many false positive saliency to non-salient objects and background (see Fig. 2 d-f). When the shape of the objects is not well captured, it further impacts the saliency rank prediction of the objects (*e.g*., "person" in Fig. 2 d-f). Motivated by the above observations, we propose a saliency rank prediction method inferring human attention, using bottom-up and top-down attention mechanisms. Our model carries out object proposal, object segmentation and object rank prediction in one go, while prior work (*e.g*., [1]) performs on region-level and makes no object proposal. The main contributions of this work include:

- We propose a new research problem to predict objects' saliency ranks according to human attention shift. It is inspired by psychological and behavioural studies, goes beyond human-object interaction [46], and shows that object-object attention shift can also be modelled.
- We propose a new large-scale dataset for the problem of salient object ranking, justified by our user study.
- We propose a deep learning architecture to jointly predict saliency ranks of multiple salient object instances and their corresponding objects masks, with bottom-up and top-down attention mechanisms.
- Extensive evaluations show that the proposed model outperforms existing methods for salient object ranking and achieves state-of-the-art performance.

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 colour, intensity and orientation

[26]. Top-down features often relate to the specific tasks at hand. Notable examples include using high-level face features [64], photography bias [26], person and car detector [4], gist features [44] and gaze patterns learnt from performing specific tasks [7]. With the advance of Convolutional Neural Networks (CNNs), CNN features are leveraged to improve the performance of saliency detection. [42, 72] use a simple stack of convolution and deconvolution layers, while [29, 33, 50] design multi-scale networks to capture contextual information for saliency inference. Recent studies further incorporate a top-down pathway [9, 18, 20, 30, 38, 56, 68, 71]. High-level semantics in the top-layers are refined with the low-level features in the shallow-layers through side connections. The refinement generates better representation at each layer [19] and is thought to imitate the bottom-up (low-level stimuli) and top-down (visual understanding) human visual process [57]. [58] follows the relationship between eye fixation and object saliency previously studied in [5, 34] and proposes to use fixation maps to guide saliency in a top-down manner.

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 [2]). We further introduce spatial size and location of object proposals in our model. Our top-down pathway is inspired by the operation of explicit object-level features generated from object proposals, with 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 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*. [1], 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 *descending level of saliency values* that relates to the order of distinct objects attended by an observer, one at a time. Our definition is closer to human visual attention and is motivated from past psychological studies and behavioural observations [40] where multiple attentions of foci is not supported [10].

In the literature, there are works that use ranking techniques for saliency estimation. For example, [54, 65, 70] use graph-based manifold ranking for saliency inference. [3, 31, 32] also incorporate rank learning to select visual features that best distinguish salient targets from real distractors. However, all 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.

2.3. Attention Mechanism

Attention mechanism has been proven to be effective in improving natural language processing [43, 47, 52] and many visual tasks [8, 25, 39, 53, 59, 66]. 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 [55] because noisy and non-relevant features may impact the saliency network [37]. To solve this, [37] computes attention weights using convolutional layers on the local neighbourhood of pixels. [68] considers message passing to capture rich contextual information from multi-level feature maps and uses a gate function to control the rate of message passing. [55] introduces a recurrent mechanism to gather multi-scale contextual information and iteratively refine convolutional features. A recurrent mechanism is also included in [73]. However, they learn to weight features spatially and in a channel-wise manner.

All these object saliency techniques apply attention mechanism 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 are most relevant (not region for object saliency). We further use an attention mechanism with high-level scene semantics to guide the prediction of salient object ranks.

Both [46] and [15] employ "gaze-following" concept to find objects or regions 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 be gazing towards in a scene. Combining with a saliency map, they produce the final gaze saliency. Unlike both works, our technique does not limit to social scenes only and we explore attention shift among multiple generic objects. It is more challenging as objects that influence on attention shift may not present when there is little interaction between the objects in a scene.

3. Saliency Rank Dataset from Attention Shift

3.1. Data Collection

In this paper, we propose, to our knowledge, the first large-scale salient object ranking dataset by combining the widely used MS-COCO dataset [36] with the SALICON dataset [24]. 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 [24] provides two sources of fixation data: 1) fixation point sequences and 2) fixa-

tion maps for each image. We exploit these two sources and consider three main 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 focuses only on the order of distinct objects that were fixated without repetition. The third approach uses the pixel intensity values from a fixation map. Both the first and third approaches are further extended, leading to nine strategies to generate ground-truth annotations. We consider up to top-10 objects in the user study, but use top-5 for saliency ranking prediction. These approaches are elaborated on below, with more details in the supplementary material.

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 observer fixation data. The saliency rank of an object can be computed by aggregating these saliency scores that each 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 try four methods to generate the final saliency score for each object.

FixSeq-avg (average score): The final score for each object is the average score of all its pixels.

FixSeq-max (maximum score): The final score is the maximum score of all its pixels.

FixSeq-avgPmax (average + maximum scores): It considers soft weighting of object scores by adding the average and maximum pixel scores in an object. It tries to consistently assign higher scores to objects that are more regularly fixated among observers.

FixSeq-avgMmax (average $\times$ maximum scores): Hard weighting of object scores through multiplication of the average and maximum pixel score values.

Approach 2: Next, we focus on distinct objects fixated in a sequence but ignore any repeating objects. We assign descending scores to objects based on the order of fixation and average them across all observers (*i.e.*, the higher the score of an object, the higher its rank). This is the *DistFixSeq*.

Approach 3: We use the pixel values from the fixation maps as the scores for pixels. Similar to *Approach 1*, we extend this approach into four methods, namely, *FixMap-avg* (average score), *FixMap-max* (maximum score), *FixMap-avgPmax* (average + maximum scores) and *FixMap-avgMmax* (average $\times$ maximum scores).

3.2. User Study and Analysis

We perform a user study with 11 participants to find out which of these methods produce more consistent ground-truth attention shift order based on human judgment. Participants were instructed to observe an image first, and then select one of nine corresponding maps that represents the order of attractiveness of objects (see the Supplemental).

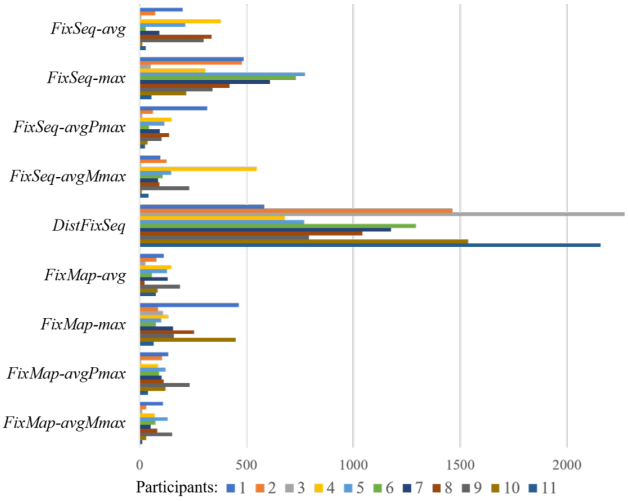

Figure 3: Pick rates of maps from 11 participants in our user study across 2500 images. These maps are generated by nine methods that we experimented with in Sec. 3.1.

Fig. 3 shows that, on average, the map generated by *DistFixSeq* has the highest number of picks from participants. The map aligns most to the order of attractiveness of objects. This 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 thus indicating their saliency [69].

We can further see that there are more picks of the methods from *Approach 1* (maps generated from temporal fixation) than those of *Approach 3* (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 [22], and prompt our definition of saliency rank order via attention shift. It supports us to use *DistFixSeq* to generate the ground-truth saliency ranking for the development of our rank prediction technique.

4. Proposed Network Architecture

4.1. Network Architecture Overview

We propose a CNN model to predict saliency rank with a bottom-up bias stimuli [6, 23], which we find useful to pick up the most salient objects in the scene. The saliency rank, especially on those less salient objects, may relate to the scene structure and observer interpretation [11]. As a result, the saliency rank modelling requires higher-level cues and prior knowledge [14].

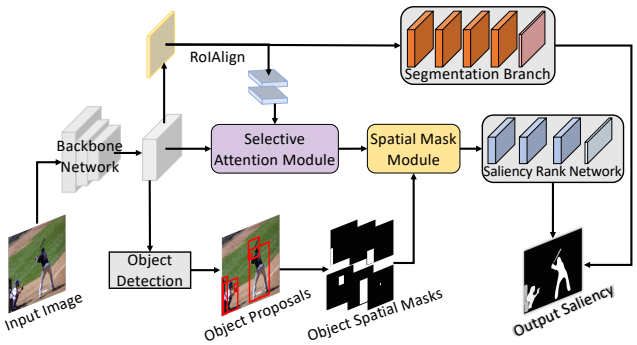

Figure 4: Architecture Overview. The model consists of a backbone network, Selective Attention Module (SAM), Spatial Mask Module (SMM) and a classification network for salient object ranking. We utilise Mask-RCNN [16] as our bottom-up backbone to provide object proposals with the FPN [35], and object segmentation from the segmentation branch. The bottom-up SMM extracts low-level features of the proposed objects while the top-down SAM considers high-level contextual attention features.

The proposed network architecture consists of four modules, namely, a backbone network based on Mask-RCNN [16], a Selective Attention Module (SAM), Spatial Mask Module (SMM) and a saliency rank network, as illustrated in Fig. 4. They are arranged to provide alternate bottom-up and top-down attention mechanisms.

Mask-RCNN generates object proposals as a bottom-up approach similar to [2]. This provides us individual object features and allows us to learn semantics information on the object-level in subsequent modules. Next, the SAM 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 [62] and interpretation due to eye gazes [11]. We then combine the features output by SAM with spatial masks in the SMM. We use spatial masks as a low-level cue, which embeds the relative size and location of each object in the image. Finally, we infer saliency rank of object instances with a small classification network. We adopt the segmentation branch of Mask-RCNN to produce segmentation for the object instances.

4.2. Backbone Network

Objectness and object proposals for binary salient object detection have been explored in [13, 49, 67]. Feng *et al*. [13] 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. [13] and [67] further extend it to many sliding windows of various scales. Fan *et al*. [12] present a model architecture much like the Mask-RCNN [16]. They produce object proposals by adopting the Feature Pyramid Network (FPN) [35] and propose a salient instance segmentation branch that extends the segmentation branch in Mask-RCNN. 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 work, we adopt Mask-RCNN as the backbone of our model and to provide efficient object proposals and segmentation. The FPN serves as a bottom-up attentive mechanism [2].

To model saliency in the object-level, we apply RoIAlign [16] and two fully connected layers (FCs) to extract object-level features, $o_i \in \mathbb{R}^{1024}$, for each object proposal, leading to a set of object features $O = \{o_1, o_2, \ldots, o_M\}$, where $M = 30$ is the maximum number of object proposals. We further take the pyramid features "$P5$" from the FPN as the high-level features input to the SAM module for top-down attention. The segmentation branch generates pixel-wise segmentation of objects for a clearer final saliency map. Different from [13, 49, 67], we do not output bounding boxes of salient objects. Instead, we predict a saliency map that indicates the pixel-wise segmentation and the saliency ranks of object instances. In contrast to [12], we exploit components of Mask-RCNN to build our bottom-up and top-down model for salient object ranking.

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 expensive, 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 [8, 39, 52, 53, 59, 66] 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 [62] and interpretation due to eye gazes [11].

Furthermore, though human actors in an image would affect observers to shift their gazes [15], we consider that individual generic objects may not necessarily have such strong influence on attention shift. For generic images (*e.g*., non-human scenes and images with little interactions between objects), we consider that the scene structure and relationship between objects may have a stronger influence

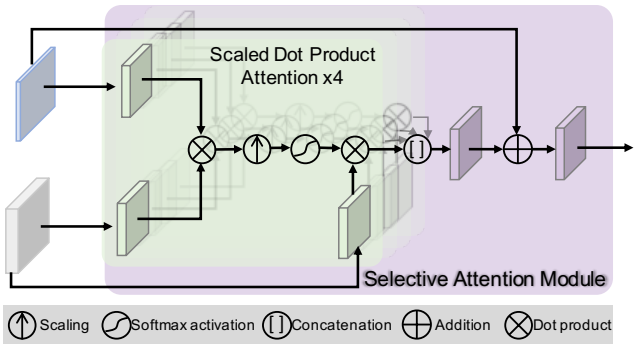

Figure 5: Details of the Selective Attention Module (SAM).

on attention shift [44]. We thus develop a Selective Attention Module 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 [52] (Fig. 5) with image and object features. We use the pyramid feature, "$P5$", from the backbone network as the image feature. A (1×1) convolution and global average pooling is 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 [52]. Here we embed the features of each object into individual 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 use these embeddings to perform dot product similarity of individual object features with the image features. We add scaling factor [52], 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 would learn different high-level information to guide scoring/weighting for salient targets. The outputs from multiple attention heads are concatenated and is sent through a FC layer. Finally, we add a residual connection and a FC layer for the module output.

4.4. Spatial Mask Module (SMM)

Understanding the relationship between object properties and scene context can help select relevant targets in a complex scenario [51]. 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 [26, 65]. These motivate us to include low-level objects properties (*e.g.*, size and locations) to learn contextual features that model relationship between objects and scene.

Using the bounding boxes of object proposals, we generate a spatial mask for each object. Spatial masks embed

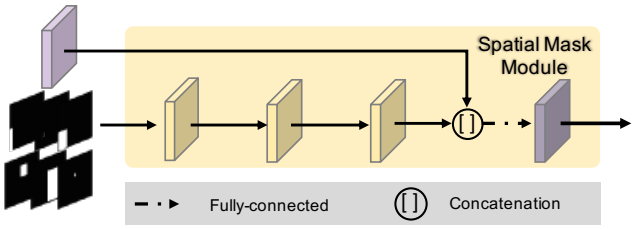

Figure 6: Details of the Spatial Mask Module (SMM).

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 spatial features are then combined with their corresponding object features with a concatenation layer and followed by a FC layer. It reduces the feature dimension to a fixed size of 512 [52]. This module can be considered as a process of combining bottom-up and semantic attributes of objects [62].

4.5. Saliency Rank Network

Our initial attempt to model salient object detection and attention shift order ranking is to cast it into a classification problem. In our setting, we consider $C = 5$ ranks and leave exploring higher ranks as future work. With one additional background class for non-salient objects, our classification has $6 = 5 + 1$ classes. Saliency and rank are then predicted with a small classification network consisting of three convolution layers and one classification layer. During inference, we combine the saliency rank classification with object segmentation (from the segmentation branch) to generate the final salient object rank map. However, a classification formulation cannot ensure that the detected salient objects would be assigned distinct saliency ranks.

To address this problem, we instead use the softmax rank classification probabilities in a scoring mechanism. For each object, we first take the probability of its predicted saliency rank as the initial score. We then add and multiply the initial score with a value relative to the predicted rank. Objects that are supposedly of higher ranks will accumulate higher scores. This is inspired by [1], which determines object saliency rank by the descending average pixel saliency value of each object. By doing so, we can ensure distinct saliency rank to be predicted for each object. Finally, we consider the top-5 saliency rank order of objects from their descending score values.

5. Experiments

5.1. Experimental Setup

Implementation Details: We fine-tune our backbone components of Mask-RCNN on salient objects before train-

ning our final model on salient object ranking. A pre-trained ResNet-101 [17] is used to initialise the convolutional layers of the Mask-RCNN. All images during 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 implemented by the Tensorflow framework and trained on an Nvidia GTX 1080 Ti GPU. We set the mini-batch size to 8. We train variations of the network for 40 epochs each, taking a maximum of 6 hours for one model training. We use the SGD optimizer with gradient norm clipping set to 5. Learning rate is set to 10^{-3}, with momentum and weight decay configured as 0.9 and 10^{-4}, respectively.

Datasets: Our dataset employs the same set of images and fixation sequence from SALICON [24], and contains object segmentation masks from MS-COCO [36]. The SALICON dataset consists of 10K training, 5K validation and testing images. There are 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*, Sec. 3). The choice of the method is supported by our user study. We discard images with no object annotations, and those 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 salient object ranking task. The dataset is randomly split into 7646 training, 1436 validation and 2418 test images, respectively.

Evaluation Metrics: We use the Salient Object Ranking (SOR) metric [1] 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 with $[-1, 1]$ indicating negative to positive correlation. However it does not 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. Therefore, we further report how many images were used to calculate the average SOR for the whole test set, where the more images used the more reliable the SOR is. The reported SOR measurement is all normalised to [0,1].

We also do a comparison 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.

Table 1: Comparison with state-of-the-art methods on our dataset. Note that RSDNet scores are based on direct prediction with pre-trained weights from their dataset. ↑(↓) means the higher(lower) the better. Top two scores are shown in red and blue, respectively.

Method	MAE ↓	SOR ↑	#Images used ↑
RSDNet [1]	0.139	0.728	2418
S4Net [12]	0.150	0.891	1507
BASNet [45]	0.115	0.707	2402
CPD-R [60]	0.100	0.766	2417
SCRN [61]	0.116	0.756	2418
Ours	0.101	0.792	2365

5.2. Comparison with State-of-the-Arts

Quantitative Evaluation: We compare against five state-of-the-art methods, namely the RSDNet [1], S4Net [12], BASNet [45], CPD-R [60] and SCRN [61], in which RSDNet first introduces saliency rank. Note that all these methods do not predict object segmentation and instead only provide a single binary saliency map.

The S4Net is chosen since it has a similar structure to our backbone and outputs object instance segmentation. We modify the S4Net code in order to predict up to 6 classes for each object instead of the binary prediction as in their original paper [12], for a fair comparison. We then apply our method of inference to obtain distinct saliency ranks. For all the rest compared models and RSDNet, the predicted saliency ranks of ground-truth objects is obtained by averaging the pixel saliency values. Object rank is determined by descending order of such averages.

The experimental results are shown in Table 1, which shows that our method outperforms other methods on the proposed dataset, achieving the best overall performance with better scores among all measurements (MAE, SOR and Images used). Note that RSDNet uses all images during the SOR calculation, due to its single binary saliency maps often containing many false saliency. Noise or very weak saliency is often propagated throughout the image and reach parts of the objects. This allows RSDNet to obtain saliency rank by averaging object pixel values to cover most objects.

S4Net shows the highest SOR score; however, it is only able to calculate the score in under two thirds of the test images. The rest is not used as it cannot predict any objects matching the ground-truth for those images. In general, good rank prediction that covers all objects should translate to both high SOR and low MAE simultaneously. Though S4Net has the highest SOR, it also has the highest (worst) MAE. It means that S4Net only performs well to predict a small subset but not all salient objects and their ranks. SOR excludes any missing objects and does not penalise such

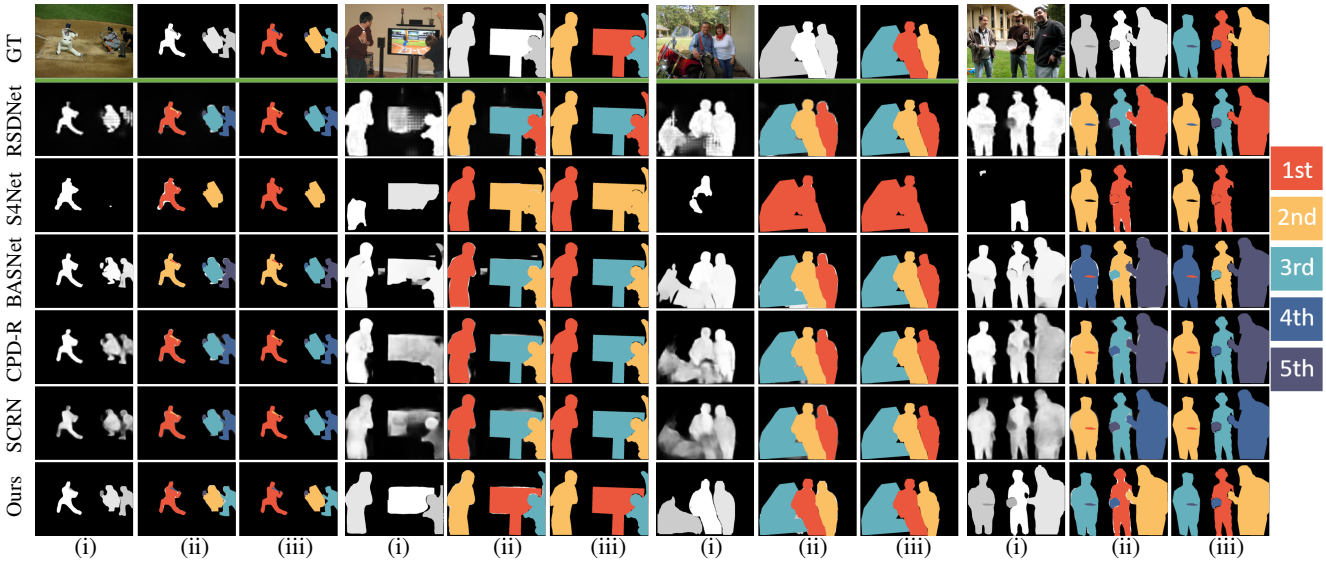

Figure 7: Comparison of the proposed method with state-of-the-art methods: RSDNet [1], S4Net [12], BASNet [45], CPD-R [60] and SCRN [61]. Each example in the top row shows the input image, ground-truth saliency map and ground-truth ranks, while for the following rows: (i) saliency prediction map, (ii) saliency prediction map with predicted rank of ground-truth object segments colourised on top, and (iii) corresponding map that contains only the predicted rank of ground-truth objects. The result in (iii) is leveraged to obtain the predicted saliency ranks for quantitative evaluation.

missing prediction. The high MAE of S4Net indicates both incorrect prediction of saliency maps and object ranks.

CPD-R produces the bets MAE score. However, the saliency maps produced are usually not as smooth as ours, and non-salient areas often filled with false saliency values. Its ranking score, SOR, is also inferior to ours.

Overall, the proposed method performs the best, with the best SOR using most images while maintaining a low MAE.

Qualitative Evaluation: We showcase results in Fig. 7 for qualitative comparison. The proposed network directly generates a saliency rank map that segments each object instance and predicts their respective ranks simultaneously. The saliency maps obtained from RSDNet [1] often contain many false saliency and with incomplete object prediction. S4Net [12] often predicts wrong and fewer object proposals than ours. Fewer object proposals lead to less available objects for SOR calculation and thus unreliable SOR score. BASNet [45] produces cleaner results. However, BASNet, RSDNet, CPD-R [60] and SCRN [61] often mix up the respective object ranks. This validates the effectiveness of our saliency rank approach that infers attention shift order.

5.3. Ablation Study

Here we perform an ablation study to evaluate each of the proposed components, in Table 2. The full model has the best overall performance. It provides the highest SOR score using large number of images. The MAE is also tied as best. These show the effectiveness of the proposed components.

Table 2: Ablation study of the proposed model. BbSR refers to the backbone network and the small saliency rank network.

Method	MAE ↓	SOR ↑	#Images used ↑
BbSR	0.109	0.773	2353
BbSR+SAM	**0.101**	0.782	**2373**
BbSR+SMM	0.111	0.769	2361
BbSR+SAM+SMM	**0.101**	**0.792**	2365

6. Conclusion

In this paper, we proposed to our knowledge the first saliency rank dataset based on attention shift order. The dataset is motivated by psychological studies and behavioural observations, and is supported by our user study, that humans attend salient objects one at a time and in an order of decreasing values of saliency. We also proposed a novel saliency rank prediction approach that infers attention shift order. The proposed approach performs favourably against several state-of-the-art methods on the proposed saliency rank dataset.

Acknowledgement: Avishek Siris is supported by the Swansea Science DTC Postgraduate Research Scholarship. Jianbo Jiao is supported by the EPSRC Programme Grant Seebibyte EP/M013774/1. The user study was supported by the College of Science, Swansea University. We thank all participants involved in the user study.

References

[1] Md Amirul Islam, Mahmoud Kalash, and Neil D. B. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*, pages 7142–7150, 2018. 1, 2, 3, 6, 7, 8

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, pages 6077–6086, 2018. 3, 5

[3] David Baldwin and Michael Mozer. Controlling attention with noise: The cue-combination model of visual search. In *Proc. Annual Meeting of the Cognitive Science Society*, volume 28, 2006. 3

[4] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *CVPR*, pages 438–445, 2012. 3

[5] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *TIP*, 24(12):5706–5722, 2015. 3

[6] Ali Borji and Laurent Itti. Exploiting local and global patch rarities for saliency detection. In *CVPR*, pages 478–485, 2012. 4

[7] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *CVPR*, pages 470–477, 2012. 3

[8] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, et al. Look and think twice: Capturing top-down visual attention with feedback convolutional neural networks. In *ICCV*, pages 2956–2964, 2015. 3, 5

[9] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, pages 234–250, 2018. 3

[10] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, 1995. 2, 3

[11] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 2008. 4, 5

[12] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, pages 6103–6112, 2019. 5, 7, 8

[13] Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun. Salient object detection by composition. In *ICCV*, pages 1028–1035, 2011. 5

[14] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *TPAMI*, 31(6):989–1005, 2009. 4

[15] Siavash Gorji and James J Clark. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In *CVPR*, pages 2510–2519, 2017. 1, 3, 5

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017. 5

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 7

[18] Shengfeng He, Jianbo Jiao, Xiaodan Zhang, Guoqiang Han, and Rynson WH Lau. Delving into salient object subitizing and detection. In *ICCV*, pages 1059–1067, 2017. 3

[19] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017. 3

[20] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *CVPR*, pages 2300–2309, 2017. 3

[21] Laurent Itti and Christof Koch. Comparison of feature combination strategies for saliency-based visual attention systems. In *Human Vision and Electronic Imaging IV*, volume 3644, pages 473–482. International Society for Optics and Photonics, 1999. 1

[22] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, 2000. 1, 4

[23] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, (11):1254–1259, 1998. 1, 4

[24] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, pages 1072–1080, 2015. 3, 7

[25] Jianbo Jiao, Yunchao Wei, Zequn Jie, Honghui Shi, Rynson WH Lau, and Thomas S Huang. Geometry-aware distillation for indoor semantic segmentation. In *CVPR*, pages 2869–2878, 2019. 3

[26] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, pages 2106–2113, 2009. 3, 6

[27] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer, 1987. 1

[28] Baisheng Lai and Xiaojin Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In *CVPR*, pages 3630–3639, 2016. 1

[29] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015. 3

[30] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, pages 478–487, 2016. 3

[31] Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Multitask rank learning for visual saliency estimation. *TCSVT*, 21(5):623–636, 2011. 3

[32] Jia Li, Dong Xu, and Wen Gao. Removing label ambiguity in learning-based visual saliency estimation. *TIP*, 21(4):1513–1525, 2012. 3

[33] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *ECCV*, pages 355–370, 2018. 3

[34] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 1, 2, 3

[35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 5

[36] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft

coco: Common objects in context. In *ECCV*, pages 740–755, 2014. 3, 7

[37] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018. 3

[38] Zhiming Luo, Akshaya Kumar Mishra, Andrew Achkar, Justin A Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, pages 6609–6617, 2017. 3

[39] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *CVPR*, pages 6790–6800, 2018. 3, 5

[40] Ulric Neisser. *Cognitive Psychology: Classic Edition*. Psychology Press, 2014. 1, 3

[41] Andrea Palazzi, Davide Abati, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: the dr (eye) ve project. *TPAMI*, 41(7):1720–1733, 2018. 1

[42] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *CVPR*, pages 598–606, 2016. 3

[43] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv:1606.01933*, 2016. 3

[44] Robert J Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *CVPR*, pages 1–8, 2007. 3, 6

[45] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 7, 8

[46] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *NeurIPS*, pages 199–207, 2015. 2, 3

[47] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv:1509.00685*, 2015. 3

[48] Guido Schillaci, Saša Bodiroža, and Verena Vanessa Hafner. Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *International Journal of Social Robotics*, 5(1):139–152, 2013. 1

[49] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *CVPR*, pages 3238–3245, 2013. 5

[50] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, pages 715–731, 2018. 3

[51] Antonio Torralba. Modeling global scene factors in attention. *JOSA A*, 20 7:1407–18, 2003. 6

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pages 5998–6008, 2017. 3, 5, 6

[53] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *CVPR*, pages 3156–3164, 2017. 3, 5

[54] Qiaosong Wang, Wen Zheng, and Robinson Piramuthu. Grab: Visual saliency via novel graph model and background priors. In *CVPR*, pages 535–543, 2016. 3

[55] Tiantian Wang, Lihe Zhang, Shuo Wang, Huchuan Lu, Gang Yang, Xiang Ruan, and Ali Borji. Detect globally, refine locally: A novel approach to saliency detection. In *CVPR*, pages 3127–3135, 2018. 3

[56] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *TIP*, 27(5):2368–2378, 2017. 3

[57] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, pages 5968–5977, 2019. 3

[58] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *CVPR*, pages 1711–1720, 2018. 3

[59] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, pages 7794–7803, 2018. 3, 5

[60] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019. 7, 8

[61] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, pages 7264–7273, 2019. 7, 8

[62] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28–28, 2014. 5, 6

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

[64] Mai Xu, Yun Ren, and Zulin Wang. Learning to predict saliency on face images. In *ICCV*, pages 3907–3915, 2015. 3

[65] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, pages 3166–3173, 2013. 3, 6

[66] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *CVPR*, pages 21–29, 2016. 3, 5

[67] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *CVPR*, pages 5733–5742, 2016. 5

[68] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, pages 1741–1750, 2018. 3

[69] Lingyun Zhang, Matthew H Tong, Tim K Marks, Honghao Shan, and Garrison W Cottrell. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32–32, 2008. 4

[70] Lihe Zhang, Chuan Yang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Ranking saliency. *TPAMI*, 39:1892–1904, 2016. 3

[71] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017. 3

[72] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221, 2017. 3

[73] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, pages 714–722, 2018. 3

[74] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *ICCV*, pages 2528–2535, 2013. 1

[75] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *CVPR*, pages 3586–3593, 2013. 1