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 processing simultaneously 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, we propose in this paper to predict the saliency rank by inferring human attention shift. We first construct a large 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.

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 [27], image captioning [62] and person re-identification [75, 74]. 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 [26, 22]. Such an ability is to deal with multiple simultaneous visual inputs, given the limited capacity of the human visual system [39]. Modeling this ability is important for the understanding of how humans interpret images, and helps improve performance of relevant applications, *e.g.*, autonomous driving [40] and robot-human interactions [47].

Figure 1: First row shows a sample of PASCAL-S datatset [33] which is used for saliency ranking in [2]. 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 color (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 [26, 21]. These early works estimate the saliency map only based on low-level features (*e.g.*, color, intensity and orientation). Recently, Gorji and Clark [16] 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 "gaze-following" concept [45], which limits attention to a single shift from a person in a scene to some other region. Islam *et al.* [2] introduce the problem of relative ranking of salient regions and apply them to rank on ground-truth salient ob-

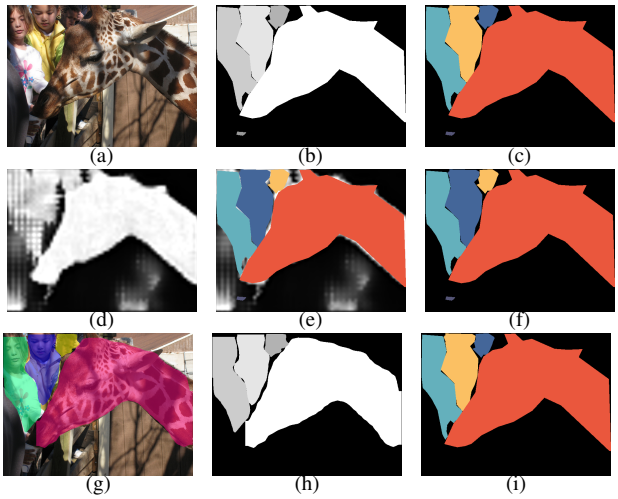

Figure 2: (a) image from our dataset, (b) corresponding ground-truth saliency rank, (c) corresponding ground-truth saliency rank (colourised), (d) saliency rank prediction by RSDNet [2], (e) corresponding saliency rank by RSDNet (gt objects overlaid and colourised), (f) corresponding saliency rank by 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).

jects from an existing PASCAL-S dataset [33]. 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, row 1). Simultaneous attention to multiple objects, however, is not supported by behavioural observation because dividing attention between multiple objects often lead to poorer performance [11] 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.

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 the first object attended by an individual should have the highest saliency. Subsequent attended objects should associate to 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 (row 2) 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 segementation and object rank prediction in one go, while the model in [2] operates 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 due to human attention shift. It is inspired by psychological and behavioural studies, goes beyond human-object interaction [45], 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 objects and their objects masks in image. It infers attention shift on multiple salient object instances 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 generally 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 approach use hand-crafted and computational based features. Bottom-up features often come from local and global contrast in colour, intensity and orientation [25]. Top-down features often relate to the specific tasks at hand. Notable examples include using high-level face features [63] for face saliency, photography bias [25], person and car detector [5], gist features [43] and gaze patterns learnt from performing specific tasks [8]. With the advance of Convolutional Neural Networks (CNNs), CNNs features are leveraged to improve the performance of saliency detection. [41, 72] use a simple stack of convolution and deconvolution layers, while [28, 49, 32] design multi-scale networks to capture contextual information for saliency inference. Recent studies further incorporate a top-down pathway [29, 55, 37, 20, 71, 10, 68]. High-level semantics in the top-layers are refined with the low-level features in the shallow-layers through side connections. It generates better

representation at each layer [19] and is thought to imitate the bottom-up (low-level stimuli) and top-down (high-level semantic) human visual process [56]. [57] follows the relationship between eye fixation and object saliency previously studied in [33, 6] and propose to use fixation maps to guide salient object detection in a top-down manner.

These methods mimic 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 from [3]). We further introduce spatial size and location of object proposals in our model. Our top-down pathway is inspired from 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 perform binary saliency prediction only, which do not provide 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*. [2], 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 [39] where multiple attentions of foci is not supported [11].

In the literature, there are work that use ranking techniques for saliency estimation. For example [64, 53, 70] use graph-based manifold ranking for saliency inference. [4, 30, 31] also incorporate rank learning to select visual feature that best distinguish salient targets from real distractors. However, all these work 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 to improve natural language processing [46, 42, 51] and many visual tasks [9, 66, 52, 58, 38]. 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 [54] because noisy and non-relevant features may impact the saliency network [36]. Motivated by this, [36] 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. [54] 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 most salient areas whilst suppress 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 attention mechanism with high-level scene semantics to guide the prediction of salient object ranks.

Both [45] and [16] 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 a human is 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 object influences 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

To our knowledge, there is no large dataset available for salient object ranking based on attention shift. In this paper, we propose a new large-scale salient object ranking dataset by combining the widely used MS-COCO dataset [35] with the SALICON dataset [24]. MS-COCO contains complex images with ground-truth object segmentation, whilst 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) fixation 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 methods for generating ground-truth annotations. We consider up to top-10 objects in the user study, but use top-5 for the saliency ranking prediction. We summarise these methods below and refer reads to supplementary for details.

Approach 1: For a given image, we follow each of 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 the object contains (*i.e*. the higher the aggregated scores, the more salient the object, the higher the rank). The number of fixation points varies among observers and leads 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 scores of all its pixels.

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

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

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

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, then select one of nine corresponding maps that represents the order of attractiveness of objects (see supplementary material).

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 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 order by fixation intensity alone, does not always capture the expected order of saliency (attractiveness of objects) of the participants.

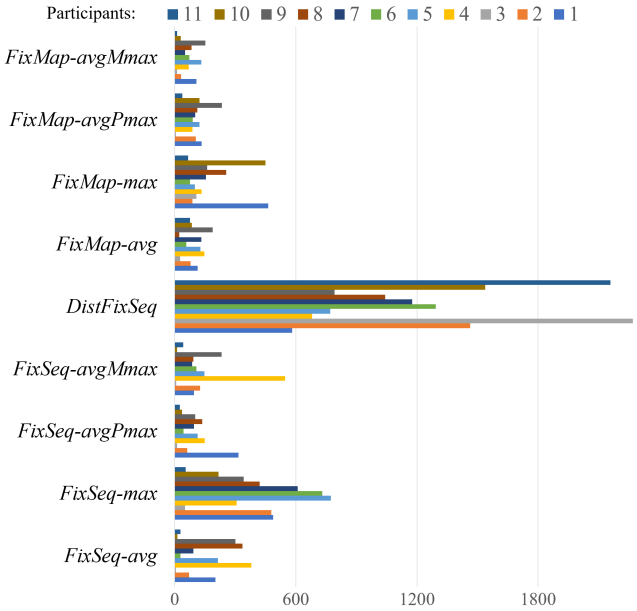

Figure 3: Pick rate of maps from 11 participants in our user study across 2500 images. These maps are generated by nine methods we experienced in Sec. 3.1.

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

In this section we start with an overview (Sec. 4.1) of the proposed model architecture (Fig. 4). It consists of the backbone network (Sec. 4.2), Selective Attention Module (Sec. 4.3), Spatial Mask Module (Sec. 4.4) and salient object rank inference (Sec. 4.5). We discuss the details below.

4.1. Network Architecture Overview

Specifically, we propose a CNN model to predict saliency rank with a bottom-up bias stimuli [23, 7], which we find useful to pick up the most salient objects in the scene. The saliency rank, especially those less salient objects, may relates to the scene structure and observer interpretation [12]. As a result, the saliency rank modeling requires higher-level cues and prior knowledge [15].

The proposed network architecture consists of 4 modules, namely, a backbone network based on Mask-RCNN [17], 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

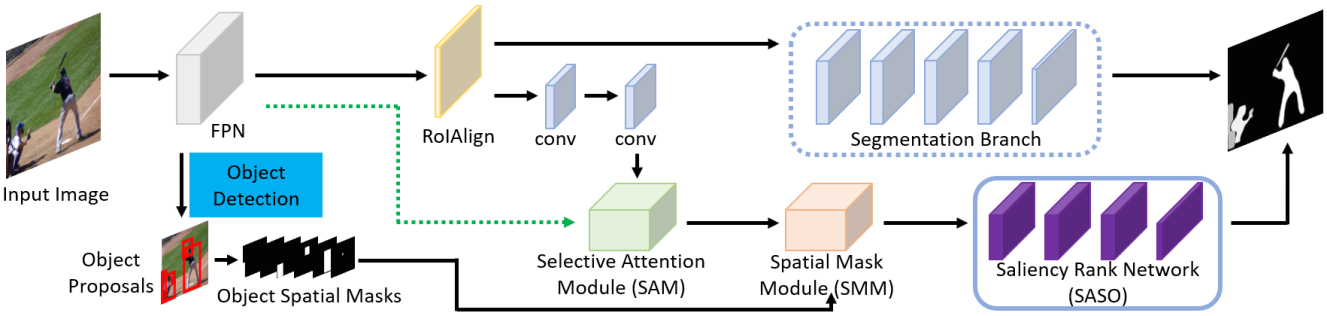

Figure 4: Architecture Overview. The model consists of a backbone network, Selective Attention Module (SAM), Spatial Mask Module (SMM) and a classification network for salient object ranking. We utilize Mask-RCNN [17] as our bottom-up backbone to provide object proposals with the FPN [34], and object segmentation from the segmentation branch. The bottom-up SMM extracts low-level features of the proposed objects whilst the top-down SAM considers high-level contextual attention features. The bottom-up and top-down modules are alternatively arranged to support the prediction of saliency rank.

and top-down attention mechanisms.

Mask-RCNN generates object proposals as a bottom-up approach similar to [3]. This provides us individual object features and allows us to learn semantics information on the object-level in subsequent modules. Next, the SAM compares each object feature to a global semantic image feature in order to determine relevant target salient objects. This module provides top-down attention mechanism and is motivated by psychophysical findings that humans frequently gaze towards interesting objects. It encapsulates important scene semantics [61] and interpretation due to eye gazes [12]. 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 objects 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 proposal for binary salienct object detection has been explored in [14, 48, 67]. Feng *et al*. [14] extend global rarity principle (rare and less frequently occuring objects are likely to be salient) to derive object saliency. They use a sliding-window mechanism to determine if the features inside the windows contain foreground or background features. [14] and [67] further extends with many sliding windows of various scales. Fan *et al*. [13] present a model architecture much like the Mask-RCNN [17]. They produce object proposals by adopting Feature Pyramid Network (FPN) [34] 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 [3].

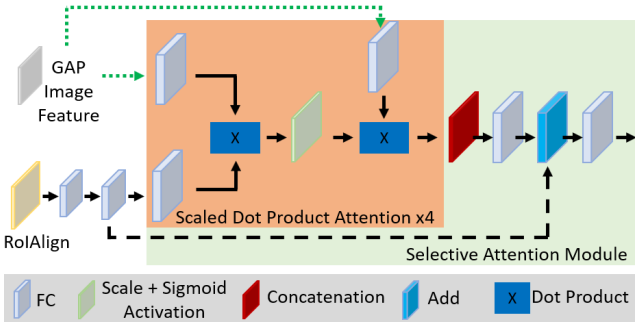

Figure 5: Design of the Selective Attention Module (SAM).

To model saliency in the object-level, we apply RoIAlign [17] 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 is the maximum number of object proposals (we set $M = 30$).

We further take the pyramid features "P5" from the FPN as the high-level feature input to the SAM module for top-down attention. The segmentation branch allows us to generate pixel-wise segmentation of objects for a clearer final saliency map. Our work differs from [14, 48, 67] that 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 [13], 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 straight-forward choice to model how humans attend one object to another would be a recurrent strategy. Such strategy is computation and memory expensive, especially when there are a lot of objects in a image (like those in our proposed dataset). To model all relationships of objects and their associated attention shift probabilities in a poten-

tial 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 [9, 66, 52, 51, 38, 58] which were greatly benefited from some forms of attention mechanisms. These attention mechanisms are often designed to dynamically weight relevant features or entities tailored to certain tasks whilst suppress distractors. In our case, we consider that attention mechanism would be useful to infer the way observers shift their attentions because it encapsulates important scene semantics [61] and interpretation due to eye gazes [12].

Furthermore, though human actors in an image would affect observers to shift their gazes [16], 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 on attention shift [43]. 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 [51] (Fig. 5) with image and object features. We use the pyramid feature, "$P5$", from the backbone network as the image feature. A (1x1) convolution and global average pooling is applied onto the pyramid feature 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, as in [51]. Here we embed each object feature into individual feature vector using a shared FC Layer. Two separate feature vectors are generated with separate FC layers, both taking the pooled image feature as input. The two new features from the pooled image feature are further repeated M times, where M is the number of objects in the set O. The attention mechanism then use these embeddings to perform dot product similarity of individual object features with the image feature. We add scaling factor as in [51], and apply softmax activation to obtain attention score. Our attention module compute 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. Output from multiple attention heads are concatenated and is sent through a FC layer. Finally, we add a residual connection and a FC layer for module output.

4.4. Spatial Mask Module (SMM)

Understanding the relationship between object properties and scene context can aid to select relevant targets in a complex scenario [50]. For example, very small objects in a scene may not attract human attention. Objects close to the centre of scene may be salient due to "center bias" concept [65, 25]. 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.

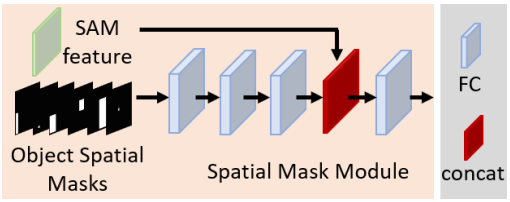

Figure 6: Design of the Spatial Mask Module (SMM).

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 it into a 64-D feature vector. The spatial features are then combined with the related object feature using one concatenation and FC layer. It reduces the feature dimension to a fixed size of 512 [51]. We consider this module as a process of combining bottom-up and semantic attributes of objects [61].

4.5. Salient Attention Shift Order (SASO)

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 is 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 the detected salient objects 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. Then we add and multiply the initial score with a value relative to the predicted rank. Objects that are supposedly of higher rank will accumulate higher scores. This is inspired from [2] which determines object saliency rank by the descending average pixel saliency value of each object. By doing so, we can ensure distinct saliency rank be predicted for each object. Finally, we consider the top5 saliency rank order of objects from their descending score values.

5. Experiments

We discuss the experiment setup, dataset and evaluation metrics in Sec. 5.1. Then we compare with state-of-the-art methods in Sec. 5.2 and offer an ablation study in Sec. 5.3.

5.1. Experimental Setup

Implementation Details: Our implementation is built on Keras and Tensorflow as we adopt the Mask-RCNN code implementation from [1]. We fine-tune our backbone components of Mask-RCNN on salient objects before proceeding to train our final model on salient object ranking. A pre-trained ResNet-101 [18] initialize the convolutional layers of the Mask-RCNN. All images during training and testing are resized to 1024 $\times$ 1024 before feeding into the network, due to the adopted source code. During inference we resize output saliency map back to original size of 640 $\times$ 480. Our network is trained on a single Nvidia GTX 1080 Ti GPU. We set 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. Standard Gradient Descent optimization is used with gradient norm clipping set at 5.0. Learning rate is set to 0.001, with learning momentum and weight decay configured to 0.9 and 0.0001 respectively.

Datasets: Our dataset employs the same set of images and fixation sequence from SALICON, and use object segmentation masks from MS-COCO. The SALICON dataset consists of 10K training, 5K validation and testing images. There is no annotation for the test set. As a result, we use the training and validation image sets to build our dataset. As mentioned in Sec. 3 we consider saliecy ranking based on the fixation sequence of the first 5 distinct objects visited without repetition (*DistFixSeq*). The choice of the method is supported by our user study. We discard images with no object annotation, and those images containing smaller object that is completely enclosed by larger one. Finally, we use images containing at least two salient objects (*i.e.* at least two ranks) for our salient object ranking task. Our resulting dataset is split into 7646 training, 1436 validation and 2418 test images randomly.

Evaluation Metrics: We employ the Salient Object Ranking (SOR) metric introduced in [2] for evaluation. It is formulated as the Spearman's Rank-Order correlation between the rank order of predicted salient objects and ground-truth. The correlation metric measures the strength and direction of the monotonic relationship between two rank variables with [-1, 1] indicating negative to positive correlation. However it does not cater the case when there are no common objects between the two rank variables. For example, when one technique completely predicts the wrong set of objects from ground-truth, SOR is not defined.

Method	MAE ↓	SOR ↑	#Images used in SOR out of 2418 ↑
RSDNet [2]	0.139	0.728	2418
S4Net [13]	0.150	0.891	1507
BASNet [44]	0.115	0.707	2402
CPD-R [59]	0.100	0.766	2417
SCRN [60]	0.116	0.756	2418
Ours	0.105	0.792	2367

Table 1: Comparison to state-of-the-art techniques 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 green, respectively.

Therefore, we further report how many images were used to calculate the average SOR for the whole test set. The reported SOR measurement is all normalized to [0,1].

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

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

Quantitative Evaluation: We compare against five state-of-the-art techniques, namely RSDNet [2], S4Net [13], BASNet [44], CPD-R [59] and SCRN [60]. We compare with RSDNet as it first introduces saliency rank problem. All these salient object detection networks do not predict object segmentation but provide a single binary map only.

We also compare with S4Net as their network have similar structure to our backbone and output object instance segmentation. We modify their source codes to predict up to 6 classes (5 ranks and 1 background) for each object, instead of the binary (1 saliency, 1 background) prediction in [13]. Then, we apply our method of inference to obtain distinct saliency ranks. We also compare against BASNet, CPD-R and SCRN, which are the current state-of-the-art salient object detection techniques. For these models and RSDNet, we obtain the predicted saliency ranks of ground-truth objects by first averaging the pixel saliency values. Object rank is determined by descending order of such averages.

The experimental results are shown in Table. 1. It shows that our technique outperforms state-of-the-art techniques on the proposed dataset. We have the best overall performance with better scores among all measurements (MAE, SOR and Images used). Note that RSDNet is able to use all images during the calculation for SOR. It is because the single binary saliency maps generated by RSDNet often contain many false saliency. Noise or very weak saliency are often propagated throughout the image and reach parts of

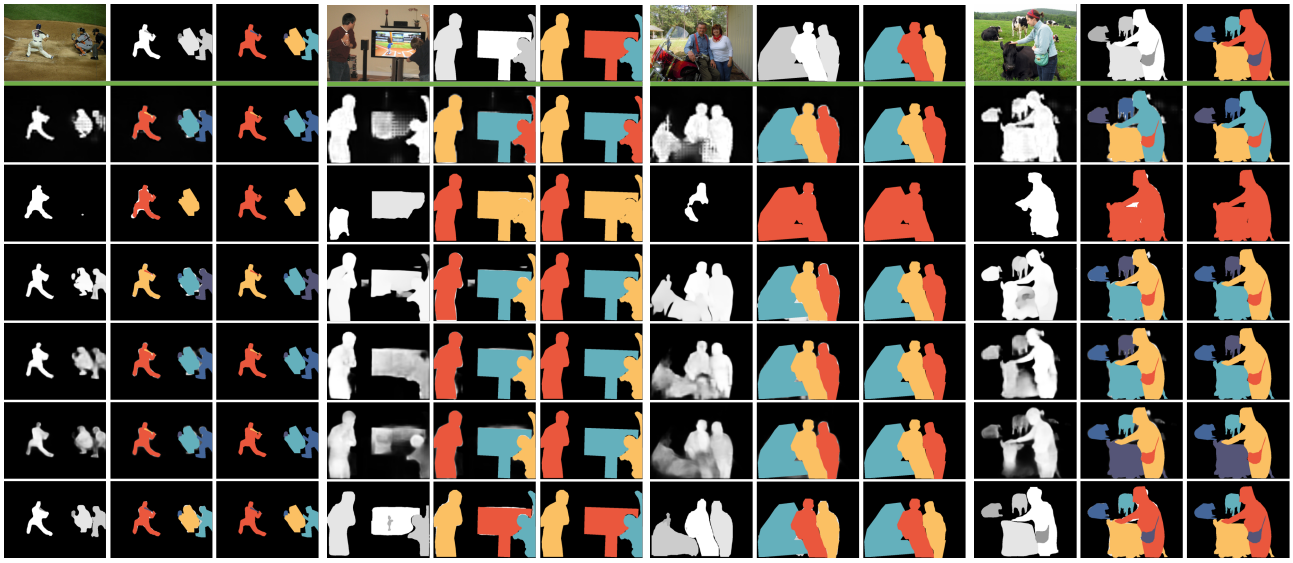

Figure 7: Qualitative comparison of the proposed method with five state of the arts. Top row (image, ground truth saliency map, ground truth ranks). Second row: RSDNet. Third row: S4Net. Fourth row: BASNet. Fifth row: CPD-R. Sixth row: SCRN. Last row: our results.

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 images in the test set. The rest is not used because 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 higher SOR and lower MAE simultaneously. Though S4Net has the highest score, it also has the highest MAE. It means that S4Net only perform well to predict a small subset but not all salient objects and their ranks. SOR excludes any missing objects and does not penalize such missing prediction. The high MAE of S4Net indicates both incorrect prediction of saliency maps and object ranks.

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

Overall, our technique performs best with highest SOR score that uses most images whilst maintains a good MAE.

Qualitative Evaluation: We showcase results in Fig. 7 for qualitative comparisons. Our proposed network directly generate a saliency rank map that segment each object instance and predict their respective ranks simultaneously. The saliency maps obtained from RSDNet often contain many false saliency and do not always capture the whole object cleanly. S4Net often predicts wrong and fewer object proposals than ours. Proposal of fewer objects lead to less available objects for SOR calculation and thus unreliable SOR score. BASNet and our saliency maps are cleaner. However, BASNet, RSDNet, CPD-R and SCRN often mix up the respective object ranks. This shows the usefulness of our salieny rank technique that infers attention shift order.

Method	MAE ↓	SOR ↑	#Images used in SOR from 2418 ↑
Bb+SASO	0.109	0.773	2353
Bb+SASO+SAM	**0.104**	0.775	2366
Bb+SASO+SMM	0.111	0.769	2361
Bb+SASO+SAM+SMM	0.105	**0.792**	**2367**

Table 2: Ablation study of different variants of model architecture. Bb is the backbone network and SASO is the small salient attention shift order classification network.

5.3. Ablation Study

We compare some basic variants to our full model in Table. 2. The full model provides the best overall performance. It provides the highest SOR scores using the largest number of images. The MAE is also comparable to the best case. These show the usefulness of each proposed component when combined together.

6. Conclusion

In this paper, we propose the first salieny 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. Next we propose a novel salieny rank prediction technique that infers attention shift order. Our technique shows good results that outperforms five state-of-the-art techniques on the proposed saliency rank dataset.

References

[1] Waleed Abdulla. Mask r-cnn for object detection and instance segmentation on keras and tensorflow. https://github.com/matterport/Mask_RCNN, 2017. 4327

[2] Md Amirul Islam, Mahmoud Kalash, and Neil D. B. Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4321, 4322, 4323, 4326, 4327, 4332, 4334, 4335

[3] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018. 4323, 4325

[4] David F Baldwin and Michael C Mozer. Controlling attention with noise: The cue-combination model of visual search. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 28, 2006. 4323

[5] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 438–445. IEEE, 2012. 4322

[6] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 4323

[7] Ali Borji and Laurent Itti. Exploiting local and global patch rarities for saliency detection. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 478–485. IEEE, 2012. 4324

[8] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 470–477. IEEE, 2012. 4322

[9] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2956–2964, 2015. 4323, 4326

[10] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 234–250, 2018. 4322

[11] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 4322, 4323

[12] Wolfgang Einhäuser, Merrielle Spain, and Pietro Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14):18–18, 2008. 4324, 4325, 4326

[13] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6103–6112, 2019. 4325, 4327, 4332, 4334, 4335

[14] Jie Feng, Yichen Wei, Litian Tao, Chao Zhang, and Jian Sun. Salient object detection by composition. In *2011 International Conference on Computer Vision*, pages 1028–1035. IEEE, 2011. 4325

[15] Dashan Gao, Sunhyoung Han, and Nuno Vasconcelos. Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(6):989–1005, 2009. 4324

[16] Siavash Gorji and James J Clark. Attentional push: A deep convolutional network for augmenting image salience with shared attention modeling in social scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2510–2519, 2017. 4321, 4323, 4326

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 2980–2988. IEEE, 2017. 4324, 4325

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4327

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3203–3212, 2017. 4323

[20] Ping Hu, Bing Shuai, Jun Liu, and Gang Wang. Deep level sets for salient object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2300–2309, 2017. 4322

[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. 4321

[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. 4321, 4324

[23] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259, 1998. 4321, 4324

[24] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, 2015. 4323, 4332

[25] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 4322, 4326

[26] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987. 4321

[27] Baisheng Lai and Xiaojin Gong. Saliency guided dictionary learning for weakly-supervised image parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3630–3639, 2016. 4321

[28] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015. 4322

[29] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016. 4322

[30] Jia Li, Yonghong Tian, Tiejun Huang, and Wen Gao. Multi-task rank learning for visual saliency estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(5):623–636, 2011. 4323

[31] Jia Li, Dong Xu, and Wen Gao. Removing label ambiguity in learning-based visual saliency estimation. *IEEE Transactions on image processing*, 21(4):1513–1525, 2012. 4323

[32] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 355–370, 2018. 4322

[33] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014. 4321, 4322, 4323

[34] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4325

[35] 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*, 2014. 4323, 4332

[36] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, 2018. 4323

[37] 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*, volume 2, page 7, 2017. 4322

[38] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018. 4323, 4326

[39] Ulric Neisser. *Cognitive psychology: Classic edition*. Psychology Press, 2014. 4321, 4323

[40] Andrea Palazzi, Davide Abati, Francesco Solera, and Rita Cucchiara. Predicting the driver's focus of attention: the dr (eye) ve project. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1720–1733, 2018. 4321

[41] Junting Pan, Elisa Sayrol, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Shallow and deep convolutional networks for saliency prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 598–606, 2016. 4322

[42] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*, 2016. 4323

[43] Robert J Peters and Laurent Itti. Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 4322, 4326

[44] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7479–7489, 2019. 4327, 4332, 4334, 4335

[45] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? In *Advances in Neural Information Processing Systems*, pages 199–207, 2015. 4321, 4322, 4323

[46] Alexander M Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. *arXiv preprint arXiv:1509.00685*, 2015. 4323

[47] 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. 4321

[48] Parthipan Siva, Chris Russell, Tao Xiang, and Lourdes Agapito. Looking beyond the image: Unsupervised learning for object saliency and detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3238–3245, 2013. 4325

[49] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 715–731, 2018. 4322

[50] Antonio Torralba. Modeling global scene factors in attention. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 20 7:1407–18, 2003. 4326

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4323, 4326

[52] Fei Wang, Mengqing Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. Residual attention network for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2017. 4323, 4326

[53] Qiaosong Wang, Wen Zheng, and Robinson Piramuthu. Grab: Visual saliency via novel graph model and background priors. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 535–543, 2016. 4323

[54] 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 *Proceed-*

ings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3127–3135, 2018. 4323

[55] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017. 4322

[56] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5968–5977, 2019. 4323

[57] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1711–1720, 2018. 4323

[58] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 4323, 4326

[59] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3907–3916, 2019. 4327, 4332, 4334, 4335

[60] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7264–7273, 2019. 4327, 4332, 4334, 4335

[61] 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. 4325, 4326

[62] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015. 4321

[63] Mai Xu, Yun Ren, and Zulin Wang. Learning to predict saliency on face images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3907–3915, 2015. 4322

[64] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, 2013. 4323

[65] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 4326

[66] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29, 2016. 4323, 4326

[67] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5733–5742, 2016. 4325

[68] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1741–1750, 2018. 4322, 4323

[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. 4324

[70] Lihe Zhang, Chuan Yang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Ranking saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1892–1904, 2016. 4323

[71] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017. 4322

[72] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 212–221, 2017. 4322

[73] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018. 4323

[74] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Person re-identification by salience matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2528–2535, 2013. 4321

[75] Rui Zhao, Wanli Ouyang, and Xiaogang Wang. Unsupervised salience learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3586–3593, 2013. 4321

Supplementary Material

In this supplementary material, we provide more details and comparisons of our implementation. These include:

- A more detailed description of the data collection process: the ground-truth generation methods (Sec. 7.1) and the user study and participants (Sec. 7.2),
- Some additional details of our implementation and design rationale (Sec. 8), and
- Some additional details when evaluating against state-of-the-arts [2, 13, 44, 59, 60], (Sec. 9) and further comparison with [13] (Sec. 10)

7. Saliency Rank Dataset from Attention Shift

First, we provide further details of the three main approaches that we propose to generate our ground-truth saliency rank annotations, and our user study.

7.1. Data Collection

To our knowledge, there are no large 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 [35] with the SALICON dataset [24]. MS-COCO contains complex images with ground-truth object segmentation, whilst 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) fixation maps for each image. We exploit these two sources and consider three main approaches to generate our ground-truth saliency rank annotations.

Approach 1: For a given image, we follow each of 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 the object contains (*i.e.*, the higher the aggregated scores, the more salient the object and the higher its rank). The number of fixation points varies among observers and leads to a large difference in scores.

We first assign scores to pixel values using fixation points from the SALICON [24] dataset. Then we get the score for objects based on the values of pixels that belong to those objects. More specifically, for every image $I \in \mathbb{R}^{W \times H}$ of dimension $W \times H$, there are N number of 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 i^{th} fixation in the sequence F^j of length t. We then assign a score to image pixel p if the fixation f_i^j falls on p using:

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

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

where v_p denotes the score at a pixel $p \in I$ aggregating from all N observers' fixation data. The function g takes the temporal order i^{th} of a fixation point in the sequence into account, and assigns lower values to fixation points if they are latter in the sequence.

Note that we are interested in the importance of the order of fixation points. We thus do not take into account the duration of fixation points in our formulation. There are large variances in the duration of fixations among different observers. Considering the durations of fixation points would cause the scoring to fluctuate greatly. Further, it is difficult (if not impossible) to obtain the exact duration of each fixation point whilst the fixations are obtained from a re-sampling process [24]. 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, and lead to 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. We do not know which methods would reflect how humans rank multiple objects in term of saliency. We try four methods to aggregate scores for subsequent saliency ranks of objects, namely: *FixSeq-avg* (average score), *FixSeq-max* (maximum score), FixSeq-avgPmax (average + maximum score) and *FixSeq-avgMmax* (average $\times$ 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 inside an object $p \in o$. We define:

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

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

$$\begin{aligned} \textit{FixSeq-avgPmax}(o, I) = \textit{FixSeq-avg}(o, I) \\ + \textit{FixSeq-max}(o, I), \end{aligned} \tag{5}$$

$$\begin{aligned} \textit{FixSeq-avgMmax}(o, I) = \textit{FixSeq-avg}(o, I) \\ \times \textit{FixSeq-max}(o, I). \end{aligned} \tag{6}$$

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 concern the object sizes. 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 Eq. 3 and Eq. 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*.

In our user study, we use $T = 10$ as the number of top salient objects for ground-truth rank. Note that we only use top-5 during our prediction task. We then sort all objects in descending order of the saliency score, and each object is given a distinct rank.

Approach 2: This approach also considers temporal order. However, we only focus on the first T distinct objects and ignore repeated fixations on already visited objects. Moreover, we directly assign a score to the whole object if a fixation point resides in its segmentation. We calls this method *DistFixSeq*.

Specifically, we define a new sequence $\hat{f}_i^n$ by removing fixations that fall on objects that are already visited by earlier fixations in f_i^n. We then define *DistFixSeq*, for each object o in an image I:

$$\textit{DistFixSeq}(o, I) = \frac{1}{N}\sum_{j}^{N}\sum_{i}^{T} h(\hat{f}_i^j), \quad \text{if } \hat{f}_i^n \in o \tag{7}$$

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

where $T = 10$. h assigns higher scores to objects if they are observed earlier. Eq. 7 takes into account only the first T objects, then average it across all N observers. We then obtain the ranks of objects in the order of descending scores.

Approach 3: We use fixation maps in this approach as the source for saliency score. We directly take intensity values from the fixation map as pixel scores v_p. Similar to **Approach 1**, we expand this approach into four methods to generate the final scores for each object. Accordingly, we have *FixMap-avg* (average score), *FixMap-max* (maximum score), *FixMap-avgPmax* (average + maximum score) and *FixMap-avgMmax* (average $\times$ maximum score). These four methods compute the final scores of objects in the same way to their counterparts in **Approach 1** (as in Eq. 3-6). Again, we consider the first distinct T objects, and assign the saliency rank in the order of descending scores.

Saliency Map: Apart from assigning a distinct rank to each object, we also produce a saliency map for each image. Objects are given an initial saliency value according to their rank (*i.e.*, Rank 1 = 1, Rank 2 = 0.9, Rank 3 = 0.8, ..., Rank 10 = 0.1). These saliency values are further multiplied by 255 and the results are assigned to the corresponding object pixels to generate the final saliency map consequently.

Figure 8: 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 pick the map that best respects their "order of attractiveness". The green box indicates the map picked by one of the participants.

7.2. User Study

We conduct a user study with 11 participants (8 males, 3 females), in order to find out which of the 9 methods produces the best attention shift order that respects human judgement. We take the best method as our technique to generate the final ground-truth saliency rank in our dataset.

For each image, the participants were presented with the image and the nine corresponding saliency rank maps arranged in a grid. Fig. 8 show a screenshot example of the annotation tool used in the user study. After a briefing session on how to use the annotation tool, every participant is told to observe the image first, then pick the maps that show objects with "order of decreasing attractiveness". 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. After all the annotations, participants were rewarded with a £25 Amazon gift voucher for their time. The annotation result is shown in Figure 3 in the main paper. It shows that human judgement of saliency rank (decreasing attractiveness) correlates very well to the maps generated by human attention shift.

8. Implementation Details

Pre-processing and Training: In the main paper, we report our network results based on the training from a pre-processing strategy. Our pre-processing step outputs fea-

tures from the backbone (Sec. 4.2 in the main paper) to save computation and training time. Consideration of this strategy also stemmed from the issue that our earlier network designs cannot fit into the memory of a single GPU card (NVIDIA GTX 1080 Ti 11GB) for training.

Our pre-processing strategy first generates object proposals for each image. We take the top M object proposals, whose probability scores were greater than 0.5. We chose $M = 30$, as it covers all objects for majority of our dataset containing an average of around 11 objects per each image. Next, we generate the corresponding object features and segmentation output for each object proposal. During the pre-processing step, we also generate the "$P5$" pyramid features from the backbone network, which we later use in the Selective Attention Module (main paper, Sec. 4.3). Finally, we train the rest of our network for saliency ranking using these pre-generated features as input.

Inference: In our current implementation, the object proposals come from the backbone network pre-trained for binary saliency prediction only. That is, it does not consider multiple saliency ranks. As a consequence, we do not use the confidence score of the object proposals (from binary classification) during our inference stage for rank prediction. Instead, we choose to use the softmax rank classification probabilities as our initial scores for distinct ranking (the last step in Sec. 4.5 of the main paper).

9. Comparison with State-of-the-arts

As noted by the caption of Table 2 in the main paper, we directly evaluate RSDNet [2] on our dataset using their pre-trained weights, for two reasons. First, the idea and model of RSDNet are based on the agreement of twelve observers on binary saliency prediction. Our training dataset, however, is based on attention shift order of the most five salient objects. Their training strategy does not fit well to the nature of our dataset. Second, practically, when we try to train their model on their dataset, or to adapt and train their model on our dataset (using their available source codes), both cases do not converge. We thus use their model with pre-trained weights to evaluate on our dataset.

For S4Net [13], we modify the prediction layer in the salient object detection and segmentation heads from binary prediction (salient, background) to multiple saliency rank prediction (5 ranks, 1 background), and train on our dataset. We find that S4Net mostly predicts the same saliency rank (rank 1) during inference with standard classification. We apply the same inference method involved in our network (main paper, Sec. 4.5) to S4Net. This allows S4Net to produce distinct saliency rank predictions and enable fair comparison with our network.

We provide more qualitative comparisons between RSDNet [2], S4Net [13], BASNet [44], CPD-R [59], SCRN [60] and ours in Fig. 9 and 10.

Table 3: Quantitative comparison with S4Net for the task of salient instance detection on our dataset. Note that we do not include comparison with RSDNet and BASNet since they are unable to perform this task.

Method	mAP^r@0.5 ↑	mAP^r@0.7 ↑
S4Net [13]	16.9 %	10.7 %
Ours	**59.4 %**	**49.9 %**

10. Further Comparison with S4Net

Like S4Net, our network is able to generate individual segmentation for each salient object instance. We further compare our network to S4Net on the task of salient instance detection. We do not include comparison with RSDNet and BASNet as they are unable to produce ouput of salient object instances. We use the mean Average Precision mAP^r ($r = 0.5/0.7$) to measure performance similarly as in [13]. Table 3 reports the results between S4Net and our network for salient instance detection on our dataset. The table shows that our network outperforms S4Net by a large margin. This shows that S4Net is not able to handle the primary task of salient object ranking, which is the focus of this paper. S4Net predicts very few salient objects when compared to our network (see Fig. 7) and misses the prediction of saliency towards corresponding ground-truth objects in over one third of the test set (indicated by #Images used in Table. 1, main paper).

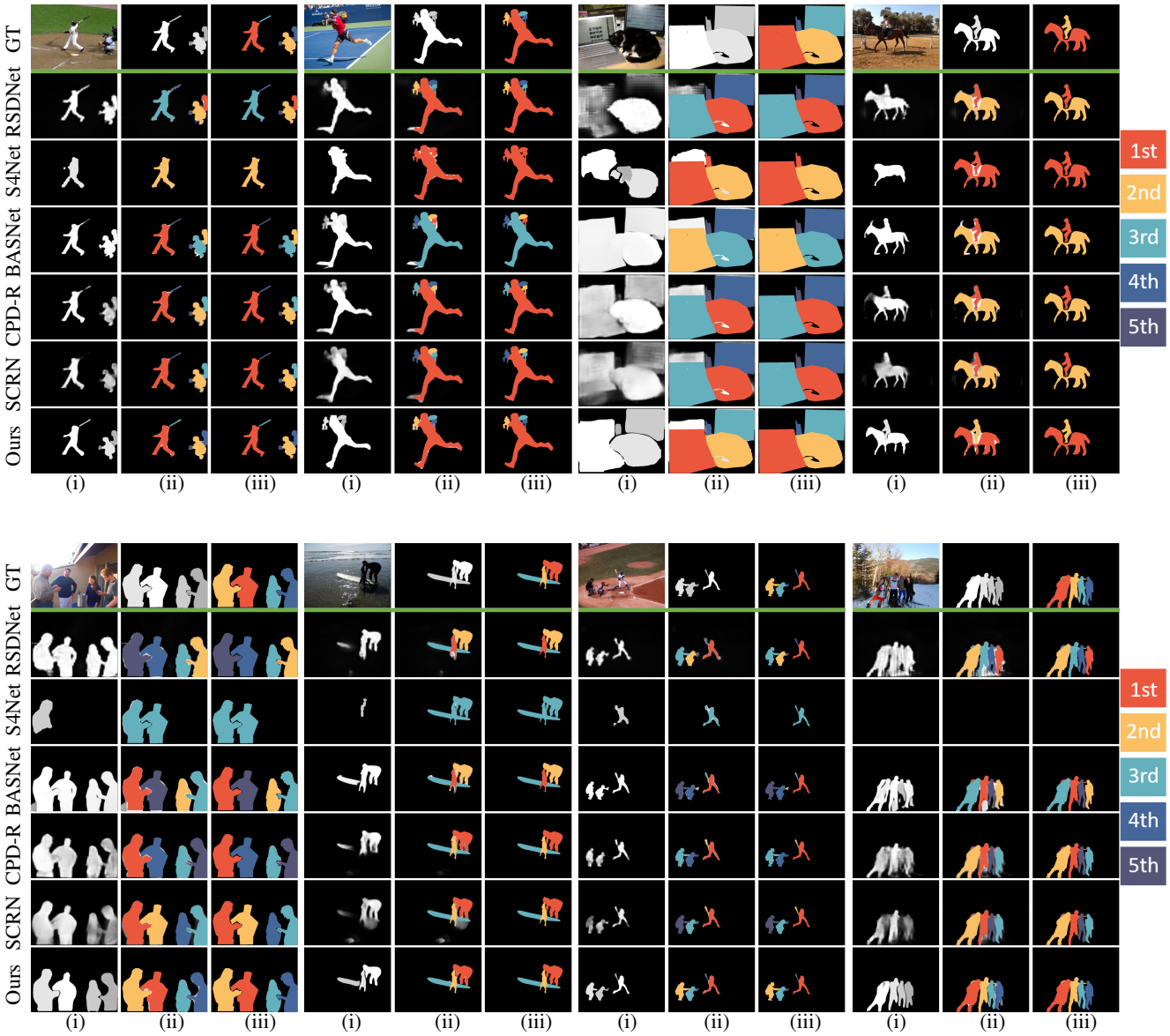

Figure 9: Further qualitative comparisons between RSDNet [2], S4Net [13], BASNet [44], CPD-R [59], SCRN [60] and our network. The top row (GT) in each of the 3 sub-figures shows 4 sets of examples. In each of the examples, we show respectively the image, the ground truth saliency map and the ground truth ranks. Each row of the 4 networks shows their respective results: (i) saliency prediction map, (ii) saliency prediction map with predicted rank of ground-truth object segments colorized on top, and (iii) corresponding map that contains only the predicted rank of ground-truth objects. Specifically, in each example, (i) provides a direct comparison of the predicted saliency maps (greyscales) against the ground-truth map. The column (ii) visualizes the false saliency and rank prediction from each methods. The column (iii) compares predicted saliency rank of ground-truth objects and their corresponding ground-truth rank. We use (iii) ground-truth object segmentation to obtain their predicted saliency ranks for numerical evaluation.

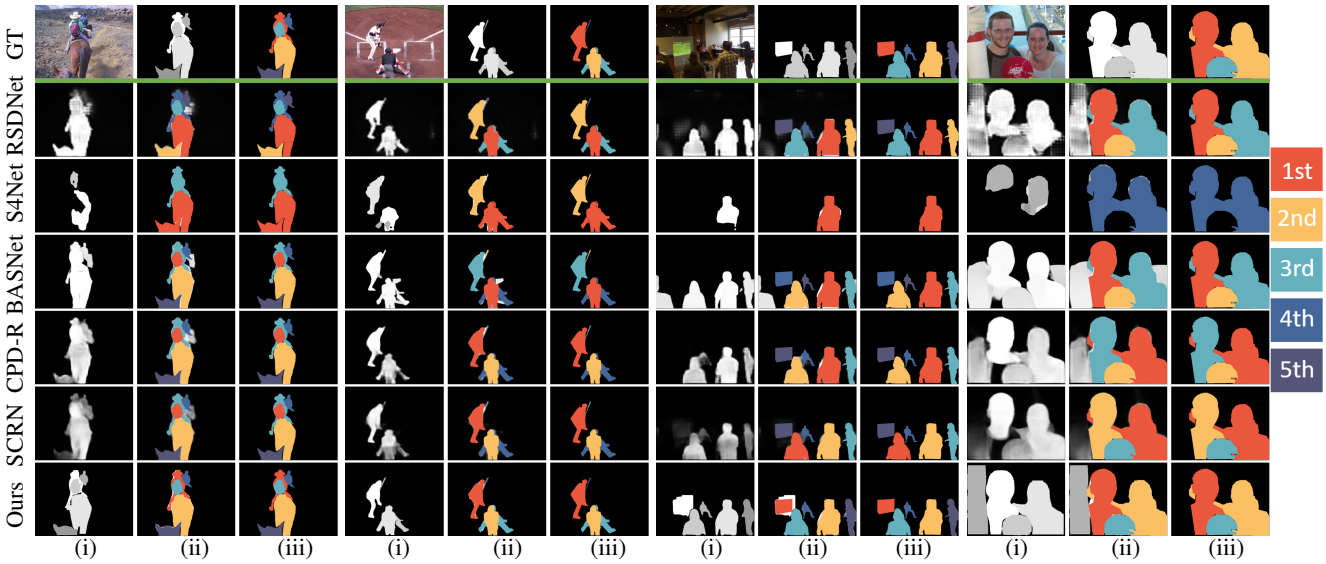

Figure 10: Further qualitative comparisons.