SeqRank: Sequential Ranking of Salient Objects

Huankang Guan and Rynson W.H. Lau*

Department of Computer Science, City University of Hong Kong
Huankang.Guan@my.cityu.edu.hk, Rynson.Lau@cityu.edu.hk

Abstract

Salient Object Ranking (SOR) is the process of predicting the order of an observer's attention to objects when viewing a complex scene. Existing SOR methods primarily focus on ranking various scene objects simultaneously by exploring their spatial and semantic properties. However, their solutions of **simultaneously** ranking all salient objects do not align with human viewing behavior, and may result in incorrect attention shift predictions. We observe that humans view a scene through a **sequential** and **continuous** process involving a cycle of foveating to objects of interest with our foveal vision while using peripheral vision to prepare for the next fixation location. For instance, when we see a flying kite, our foveal vision captures the kite itself, while our peripheral vision can help us locate the person controlling it such that we can smoothly divert our attention to it next. By repeatedly carrying out this cycle, we can gain a thorough understanding of the entire scene. Based on this observation, we propose to model the dynamic interplay between foveal and peripheral vision to predict human attention shifts sequentially. To this end, we propose a novel SOR model, SeqRank, which reproduces foveal vision to extract high-acuity visual features for accurate salient instance segmentation while also modeling peripheral vision to select the object that is likely to grab the viewer's attention next. By incorporating both types of vision, our model can mimic human viewing behavior better and provide a more faithful ranking among various scene objects. Most notably, our model improves the SA-SOR/MAE scores by +6.1%/-13.0% on IRSR, compared with the state-of-the-art. Extensive experiments show the superior performance of our model on the SOR benchmarks. Code is available at https://github.com/guanhuankang/SeqRank.

Introduction

Salient object detection (He et al. 2017b; Qin et al. 2019; Zhao et al. 2019; Liu et al. 2021b; Wu et al. 2022; Wang et al. 2023) aims to identify objects that naturally attract human attention in a cluttered visual world. Although SOD can tell which objects are more likely to grab human attention, it fails to reveal how human attention shifts among them. This issue has recently led to the development of Salient Object Ranking (SOR) (Siris et al. 2020), which is to predict the visiting order of an observer's attention to various scene objects based on their degrees of salience, allowing the most salient object to be attended to first, followed by the less salient objects, resulting in a sequential visiting of different parts of a complex scene. SOR can help understand how humans interpret images and benefit many downstream tasks, such as image editing (Aberman et al. 2022; Miangoleh et al. 2023), scene understanding (Du et al. 2019; Li et al. 2023), and human-robot interaction (Schillaci, Bodiroža, and Hafner 2013).

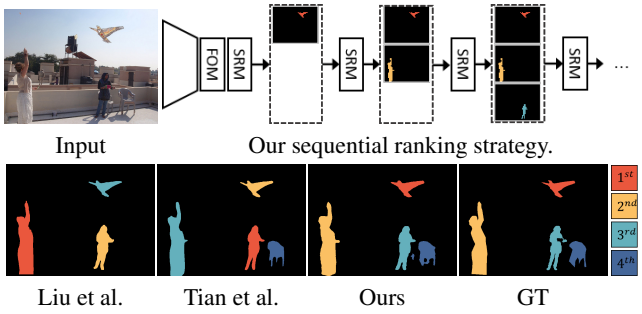

Figure 1: We present a novel salient object ranking approach, in which the objects are detected in a sequential order, allowing the most salient object to be detected first, followed by the less salient one and so on. Our sequential ranking strategy is consistent with human viewing behavior and gives a more accurate prediction than the state-of-the-art methods (Liu et al. 2021a; Tian et al. 2022a), which rank objects simultaneously.

Salient object ranking is a rather new research field that has witnessed some recent progress. For instance, (Siris et al. 2020) study the object-scene context to predict attention shifts. (Fang et al. 2021a) suggest adding position information explicitly to enhance ranking performance. (Tian et al. 2022a) propose combining both spatial and object-based attention in the ranking model for more realistic attention shift predictions. These methods have demonstrated some advantages in salient object ranking. However, they are limited to exploring semantics and spatial attributes to learn how objects compete for saliency, raising concerns about spatial and semantic biases. Besides, they operate under an unrealistic assumption that the saliency ranking of all potential salient objects in the scene could be inferred **simultane-**

*Rynson W.H. Lau is the corresponding author.

ously, which is different from how humans view the scene and may lead to incorrect attention shift predictions.

We observe that humans view the visual world in a **sequential** and **continuous** manner, with an ongoing interaction between visual stimuli and our visual system (Hooge and Erkelens 1999; Ludwig, Davies, and Eckstein 2014; Wolf, Belopolsky, and Lappe 2022). Specifically, we use the central part of our visual field (*i.e.,* foveal vision) to fixate on objects for perceiving object details and the outer region of our visual field (*i.e.,* peripheral vision) to search for the next fixation location so that we can smoothly divert our attention to it next. By repeatedly engaging in this cycle, we can prioritize our attention on various scene objects and ultimately gain a comprehensive understanding of the entire scene. Based on this observation, we propose to model the dynamic interplay between foveal and peripheral vision to infer saliency ranking in a sequential manner. Our sequential ranking strategy enables us to explicitly explore the temporal relationships between previously visited objects and subsequently attended locations (Wolf, Belopolsky, and Lappe 2022), which is neglected by previous works.

To this end, we propose **SeqRank**, a query-based salient object ranking method driven by the human sequential viewing behavior. SeqRank includes two novel modules: a *Fovea Module (FOM)* to extract high-acuity visual features, and a *Sequential Ranking Module (SRM)* to search for the subsequent fixation location. Our FOM is inspired by the human fovea, which is a small, central area of the retina that is responsible for sharp, detailed vision and color perception. FOM works by fixating on a specific region of an image and extracting detailed visual features from that region to enable the model to accurately represent the high-acuity visual information there. The proposed SRM aims to mimic human peripheral vision by selecting the next fixation location from the periphery region. This region is usually not as sharp as the central visual field but contains highly compressive visual information, which can be used by SRM to predict where the viewer's attention is likely to shift to next. By repeatedly invoking SRM, as shown in Figure 1, we can sequentially detect all salient instances in an order that reflects the sequence of human attention shifts. By incorporating both types of vision, our model can better mimic human viewing behavior and provide a more faithful ranking among various salient objects. We conduct extensive experiments to show the superior performance of our model.

In summary, our main contributions of this work include:

1. We propose a novel approach for salient object ranking. It learns to sequentially infer the saliency ranks of various scene objects by modeling both foveal and peripheral vision, and we make the first attempt to explore temporal relationships between objects for SOR.
2. We propose our SOR model, SeqRank, with two innovative modules, *i.e.,* Fovea Module and Sequential Ranking Module. These two modules work coherently for a natural and realistic attention shift prediction.
3. Extensive experiments are conducted to confirm the effectiveness of our approach, and our model achieves new state-of-the-art results on the existing SOR benchmarks.

Related Work

Salient Object Ranking (SOR) is first studied by (Islam, Kalash, and Bruce 2018), who suggest ranking scene objects by the level of agreement among multiple observers who consider the objects to be salient. However, their rank-aware network can only output pixel-level saliency contrasts. Later, (Siris et al. 2020) complement the concept of SOR by incorporating psychological and neuroscientific evidence (Neisser 2014; Desimone and Duncan 1995), and consider SOR as a task of predicting the visiting order of human attention to distinct objects in the scene. They explore the object-scene context for attention shifts prediction, and introduce a large-scale salient object ranking benchmark for evaluation, which has become widely used by subsequent works. Then, (Fang et al. 2021a) study the impact of position embedding and feature interaction between objects and show their benefits in the SOR task. (Liu et al. 2021a) propose another dataset with fewer annotation errors and introduce graph convolution for object-level reasoning, further enhancing the ranking performance. (Tian et al. 2022a) emphasize the interaction between objects and context, proposing to model both spatial and object-based attention.

While these works have made significant progress in salient object ranking, their exploration is limited to semantics and spatial properties for learning how objects compete for saliency simultaneously. In this work, we propose to model the dynamic interplay between foveal and peripheral vision, inspired by the human sequential viewing behavior, to predict attention shifts sequentially. This enables an explicit exploration of the temporal relationship between objects, which is greatly neglected by previous works.

Salient Object Detection (SOD) is a topic closely related to SOR and has been widely studied. It aims to identify and locate objects that naturally capture human attention in a scene. Early methods in SOD primarily used low-level cues, such as background priors (Li et al. 2013; Jiang et al. 2013; Zhu et al. 2014) and center priors (Cheng et al. 2013; Yan et al. 2013). Yet, these traditional approaches often struggle with a lack of high-level semantics. Later on, deep learning-based methods (Qin et al. 2019; Zhao et al. 2019; Liu, Han, and Yang 2020; Wu et al. 2022; Wang et al. 2023) become popular and show impressive results. Most of these works rely on multi-level or multi-scale feature fusion strategies to achieve their success. Recently, transformer-based architecture is also introduced to salient object detection, *e.g.,* visual saliency transfomer (Liu et al. 2021b).

Despite the success, SOD approaches can only disentangle salient objects from the background. They cannot recognize distinct object instances. To tackle this issue, **Salient Instance Detection (SID)** has recently been proposed. It aims to detect multiple salient instances in an image. Unlike SOD approaches, SID methods (Li et al. 2017; Fan et al. 2019; Tian et al. 2022b) usually include an additional object proposal stage for instance discovery and then learn to identify the salient ones from the background.

In contrast to SOD and SID, salient object ranking is more challenging as it not only requires detecting distinct salient instances but also assigning a rank to each instance to indi-

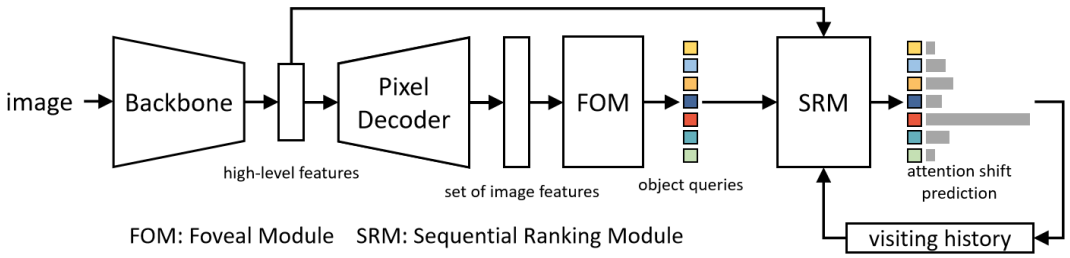

Figure 2: SeqRank is composed of a backbone network, a pixel decoder, a Fovea Module (FOM), and a Sequential Ranking Module (SRM). The FOM learns to progressively refine the learnable object queries from image features, while the SRM predicts the next object that is likely to be visited, conditioning on the previous visiting history. By continuously updating the visiting history and invoking the SRM, all salient objects can be detected in a sequential order that reflects how human attention shifts among them.

cate how human attention shifts from one to another based on their salience degrees.

Our Approach

The design of SeqRank is inspired by the human sequential viewing behavior. It first learns to detect salient instances in the scene, and then ranks them in a sequential order such that the most salient object is visited first, followed by the less salient ones. Two novel modules are proposed for this purpose: *Fovea Module (FOM)* and *Sequential Ranking Module (SRM)*, which aim to simulate human foveal and peripheral vision, respectively.

The overall architecture of SeqRank is illustrated in Figure 2. SeqRank receives an RGB image as input and employs a bottom-up backbone network, such as ResNet (He et al. 2016) or Swin Transformer (Liu et al. 2021c), for features extraction. A pixel decoder, *e.g.,* FPN (Lin et al. 2017), is included to restore the spatial information from low-level features and produces a set of image features at different resolutions, denoted as $feat_5$, $feat_4$, $feat_3$ and $feat_2$ from low to high. Note that the resolutions of $feat_5$, $feat_4$, $feat_3$ and $feat_2$ are $1/32$, $1/16$, $1/8$ and $1/4$ of the input resolution, respectively. The set of image features is then fed to FOM, which learns to progressively refine the learnable object queries for high-acuity salient instance segmentation. After that, SRM attempts to predict which object is likely to grab human attention, conditioning on the previous visiting history. Through iteratively invoking SRM with the increasing visiting history, we obtain a list of salient instances that reflects the sequence of human attention shifts.

Fovea Module (FOM)

FOM is proposed to extract high-acuity visual features and facilitate an accurate salient instance segmentation by taking inspiration from the human fovea, which is a small region in the retina of the eye containing many cones packed closely to allow it to fixate at a small region of the scene for sharp details. FOM works by progressively refining the learnable object queries from image features $feat_i$, where

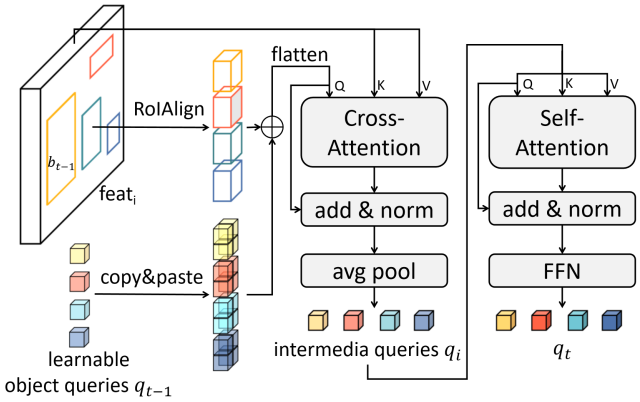

Figure 3: The Fovea Layer is a variant of the transformer decoder. It first applies RoIAlign to obtain object-part-aware target features, which are then added to object-level queries. The cross-attention layer comes before the self-attention layer, so that the image features can be involved earlier. An average pooling is inserted after the cross-attention layer for aggregating the object-part-aware queries back to object-level queries, which are then sent to the self-attention layer for inter-object relationships modeling. Note that positional embeddings are omitted in this figure for readability.

$i \in \{3, 4, 5\}$. Each refining stage is formulated as:

$$q_t = f(q_{t-1}, b_{t-1}, feat_i), \quad (1)$$

$$b_t = mlp(q_t), \quad (2)$$

where $q_t \in R^{N \times d}$ and $b_t \in R^{N \times 4}$ are the object queries and bounding box prediction at stage t. N is the number of object queries. We formulate f as a fovea layer shown in Figure 3. To help fixate on the targets and learn high-acuity visual information, it first performs RoIAlign (He et al. 2017a) on $feat_i$ to extract object-part-aware features, which are then attached to q_{t-1} forming a new set of object-part-aware queries, denoted as $q_{fovea} \in R^{Nhw \times d}$, where h, w are the height and width of the RoI window. We further apply a cross-attention layer (Vaswani et al. 2017) to update q_{foveal} from $feat_i$. After that, we use an average pooling operation

to recover the updated q_{fovea} back to object-level queries $q_i \in R^{N\times d}$. A self-attention layer for modeling inter-object relationships is appended at the end and outputs q_t, which can be used for predicting bounding boxes b_t with a 3-layer MLP.

FOM is composed of six fovea layers. Each layer receives image features at a single scale from the set of $\{feat_3,\ feat_4,\ feat_5\}$. Specifically, inspired by Mask2Former (Cheng et al. 2022), the six layers are assigned with $feat_5,\ feat_4,\ feat_3,\ feat_5,\ feat_4,\ feat_3$. Noteworthy, FOM works differently from Mask2Former, which learns localized features by constraining cross-attention to within predicted instance regions. In contrast, FOM fixates on objects through the guidance of RoI features while enabling object queries to attend to every location within the image, such that it could potentially be more robust to early prediction errors.

Finally, we apply a dot product between $feat_2$ and the object queries q from the last fovea layer, followed by a sigmoid function to output salient instance masks. A classification head is used to indicate which object queries are activated for outputting salient instance masks.

Sequential Ranking Module (SRM)

SRM is proposed to predict the next object that is likely to catch human attention conditioning on the previous visiting history (Wolf, Belopolsky, and Lappe 2022). It is motivated by human peripheral vision, in which humans prepare the next fixation location by scanning the peripheral region (a region with low resolution but high information compression) and then shift their gaze smoothly to it afterwards. Figure 4 shows the structure of SRM. We encode the visiting history with one-hot encoding and expand it with a learnable memory embedding to match the dimensionality d of the object queries $q \in R^{N\times d}$. Mathematically, the encoding process of the visiting history can be written as:

$$v = onehot(visiting\,history), \tag{3}$$

$$v_e = v \cdot memo, \tag{4}$$

where $onehot$ is the one-hot encoding function. $v \in \{0,1\}^{N\times 1}$ is the one-hot vector of the visiting history, where 1 means the object was visited and 0 means it was not. $memo \in R^{1\times d}$ is a learnable memory embedding for expanding v to d-dimension space. We then inject the visiting history into object queries with an addition, followed by a self-attention layer to allow the queries to exchange visiting information. This can be written as:

$$q_v = SelfAttn(Q{=}q{+}v_e,\ K{=}q{+}v_e,\ V{=}q{+}v_e), \tag{5}$$

$$q_V = FFN(q_v), \tag{6}$$

where $SelfAttn$ is the self-attention layer. FFN is the feed-forward neural network.

We use the high-level features from the last layer of the backbone to mimic the visual features of the peripheral region, since high-level features are in low-resolution and contain rich semantics. We first perform a cross-attention from high-level features (as $queries$) to the object queries (as

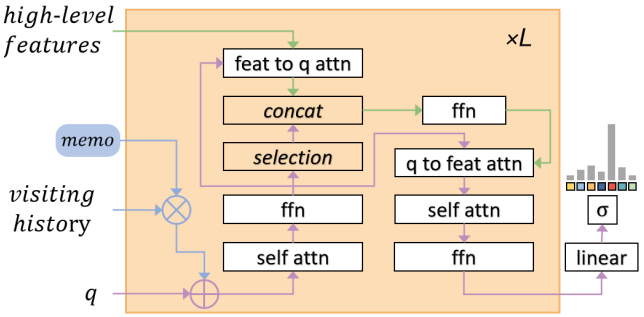

Figure 4: Sequential Ranking Module (SRM) learns to predict the next salient object conditioning on the visiting history. memo: learnable memory embedding. self attn: self-attention layer. ffn: feed-forward neural network. A to B attn: cross-attention layer with A as $query$ and B as $key, value$. linear: linear projection. σ: sigmoid function. Note that shortcuts, LayerNorm and positional embeddings are omitted in this figure for readability.

$key, value$) for the alignment between them, as:

$$feat_{align} = CrossAttn(Q{=}feat_{high},\ K{=}q_V,\ V{=}q_V), \tag{7}$$

where $feat_{high} \in R^{P\times d}$ are the high-level features. P is the number pixels in $feat_{high}$. $CrossAttn$ is the cross-attention layer. To model the saliency competition among distinct locations, we select the current fixated object query $q_c \in R^d$ from $q_V \in R^{N\times d}$ and concatenate q_c with each pixel of $feat_{align}$, resulting in $feat_{concat} \in R^{P\times 2d}$, such that each pixel contains both local and currently fixated object information:

$$q_c = q_V[c], \tag{8}$$

$$feat_{concat}[i] = concat(feat_{align}[i], q_c), \tag{9}$$

$$feat_{Concat} = FFN(feat_{concat}), \tag{10}$$

where c is the index of the currently fixated object. i is the index to all pixels of $feat_{align}$, and $concat$ is a concatenate operation. After that, object queries q_V is updated from $feat_{Concat}$ via a cross-attention, followed by a self-attention and a feed-forward network allowing an exploration of the temporal relationships between objects. Finally, we use a linear projection followed by a sigmoid function to produce the likelihood of an object to be visited next:

$$likelihood = \sigma(linear(q_{out})), \tag{11}$$

where $q_{out} \in R^{N\times d}$ is the final object queries by SRM, and $likelihood \in [0,1]^N$ indicates the likelihood of an object being visited next.

During inference, we use SRM to iteratively find all salient instances in the scene. The first run of SRM detects the most salient object. After adding it to the visiting history, we run SRM again to find the next object to be fixated on. This process is repeated until there are no more salient instances or the maximum number of runs, *i.e.,* N, is reached.

Method	Task	ASSR Test Set (2418)			IRSR Test Set (2929)		
		SA-SOR↑	SOR↑	MAE↓	SA-SOR↑	SOR↑	MAE↓
VST (Liu et al. 2021b)	SOD	0.422	0.643	9.99	0.183	0.571	8.75
MENet (Wang et al. 2023)	SOD	0.369	0.627	9.60	0.162	0.558	8.25
S4Net (Fan et al. 2019)	SID	0.451	0.649	14.4	0.224	0.611	12.1
QueryInst (Fang et al. 2021b)	IS	0.596	0.865	8.52	0.538	0.816	<u>7.13</u>
Mask2Former (Cheng et al. 2022)	IS	0.635	0.867	<u>7.31</u>	0.521	0.799	7.14
RSDNet (Islam, Kalash, and Bruce 2018)	SOR	0.386	0.692	18.2	0.326	0.663	18.5
ASRNet (Siris et al. 2020)	SOR	0.590	0.770	9.39	0.346	0.681	9.44
PPA (Fang et al. 2021a)	SOR	0.635	0.863	8.52	0.521	0.797	8.08
IRSR (Liu et al. 2021a)	SOR	<u>0.650</u>	0.854	9.73	<u>0.543</u>	0.815	7.79
OCOR (Tian et al. 2022a)	SOR	0.541	**0.873**	10.2	0.504	<u>0.820</u>	8.45
Ours	SOR	**0.685**	<u>0.870</u>	**7.22**	**0.576**	**0.822**	**6.20**

Table 1: Quantitative Comparison. SOD: Salient Object Detection. SID: Salient Instance Detection. IS: Instance Segmentation. SOR: Salient Object Ranking. The best is marked in bold and the second-best is marked with an underline.

Training Strategy

We use the binary cross-entropy loss and the dice loss (Milletari, Navab, and Ahmadi 2016) for our salient instance masks. The bounding box is supervised with ℓ_1 loss and GIoU loss (Rezatofighi et al. 2019), while the classification head adopts binary cross-entropy loss. Therefore, the final loss for FOM is:

$$\ell_{FOM} = \lambda_m \ell_{mask} + \lambda_b \ell_{bbox} + \lambda_c \ell_{cls}, \tag{12}$$

where λ_m is set to 5.0, λ_b is set to 2.0 for GIoU loss and 5.0 for ℓ_1 loss, and λ_c is set to 10.0 for predictions matched with a ground truth and 0.1 for the "no object", *i.e.,* predictions have not been matched with any ground truth. The matching mechanism used is the Hungarian algorithm following (Carion et al. 2020; Fang et al. 2021b; Cheng et al. 2022).

We use the binary cross-entropy loss for SRM:

$$\ell_{SRM} = \lambda_s \sum_{j=1}^{n} \ell_{bce}(likelihood_j, GT_j), \tag{13}$$

where λ_s is set to 5.0, n is the number of salient objects in the image, ℓ_{bce} is a binary cross-entropy loss function and $likelihood_j$ is the likelihood of an object being visited in the j^{th} round prediction by SRM. It is noteworthy that the corresponding ground truth, $GT_j \in \{0,1\}^N$, either has only one salient object or no salient object, since our SRM predicts salient objects one by one and predicts no object when there are no more salient instances in the scene. Thus, the final loss for SeqRank can be written as:

$$\ell_{SeqRank} = \ell_{FOM} + \ell_{SRM}. \tag{14}$$

Experiments

Datasets and Evaluation Metrics

We conduct experiments on the public SOR benchmarks, ASSR (Siris et al. 2020) and IRSR (Liu et al. 2021a). ASSR is constructed from MS-COCO (Lin et al. 2014) and SALICON (Jiang et al. 2015) and comprises 7646 images for training, 1436 images for validation and 2418 images for testing. Each image is labeled with at most five salient instances with ranks. IRSR consists of 6059 training images and 2929 testing images, with fewer annotation errors but more challenging as each image contains up to eight ranked salient instances. The images in both SOR benchmarks cover a wide range of scenes and common objects, presenting a great challenge for the SOR task.

We evaluate our results using three metrics: Salient Object Ranking (SOR) (Islam, Kalash, and Bruce 2018), Segmentation-Aware SOR (SA-SOR) (Liu et al. 2021a), and Mean Absolute Error (MAE). SOR is formulated as the Spearman's rank correlation between the predicted rankings and the ground truth, with values normalized to a range of 0.0 to 1.0 and higher values indicating better performance. SA-SOR calculates the Pearson correlation between the predicted rankings and the ground truth, and it also penalizes missed salient instances or false negative predictions. SA-SOR measures the consistency between the predicted saliency ranking and the ground truth, with values ranging from -1.0 to 1.0, where a positive/negative value indicates a positive/negative correlation. MAE reflects the quality of salient object segmentation, and smaller means better.

Implementation Details

We use Swin Transformer (Liu et al. 2021c) pretrained on ImageNet (Krizhevsky, Sutskever, and Hinton 2012) as our backbone network and FPN (Lin et al. 2017) as our pixel decoder. The height and width of the RoI window for our FOM are set dynamically according to the scales of the $feat_i$, with $h = w = 1/2/3$ when $i = 5/4/3$. We set $N = 100$, $d = 256$, $L = 2$ and $P = \frac{H}{32} \times \frac{W}{32}$, where H, W are the resolution of the input image, which is set to 800×800. During training, suggested by (Cheng et al. 2022), we calculate ℓ_{mask} on K randomly sampled points instead of the whole mask to improve training efficiency and reduce training memory. K is set to 12544, *i.e.,* 112×112 points. We trained SeqRank for 30k iterations with a batch size of 32. We adopt AdamW optimizer with a weight decay of 1e-4. The initial learning rate is set to 1e-4 and is decayed to 1e-5 after 20k iterations. We use random flip for data augmentation. We develop SeqRank using Detectron2 (Wu et al. 2019) and train it with 4 NVIDIA A100-SXM4-80GB.

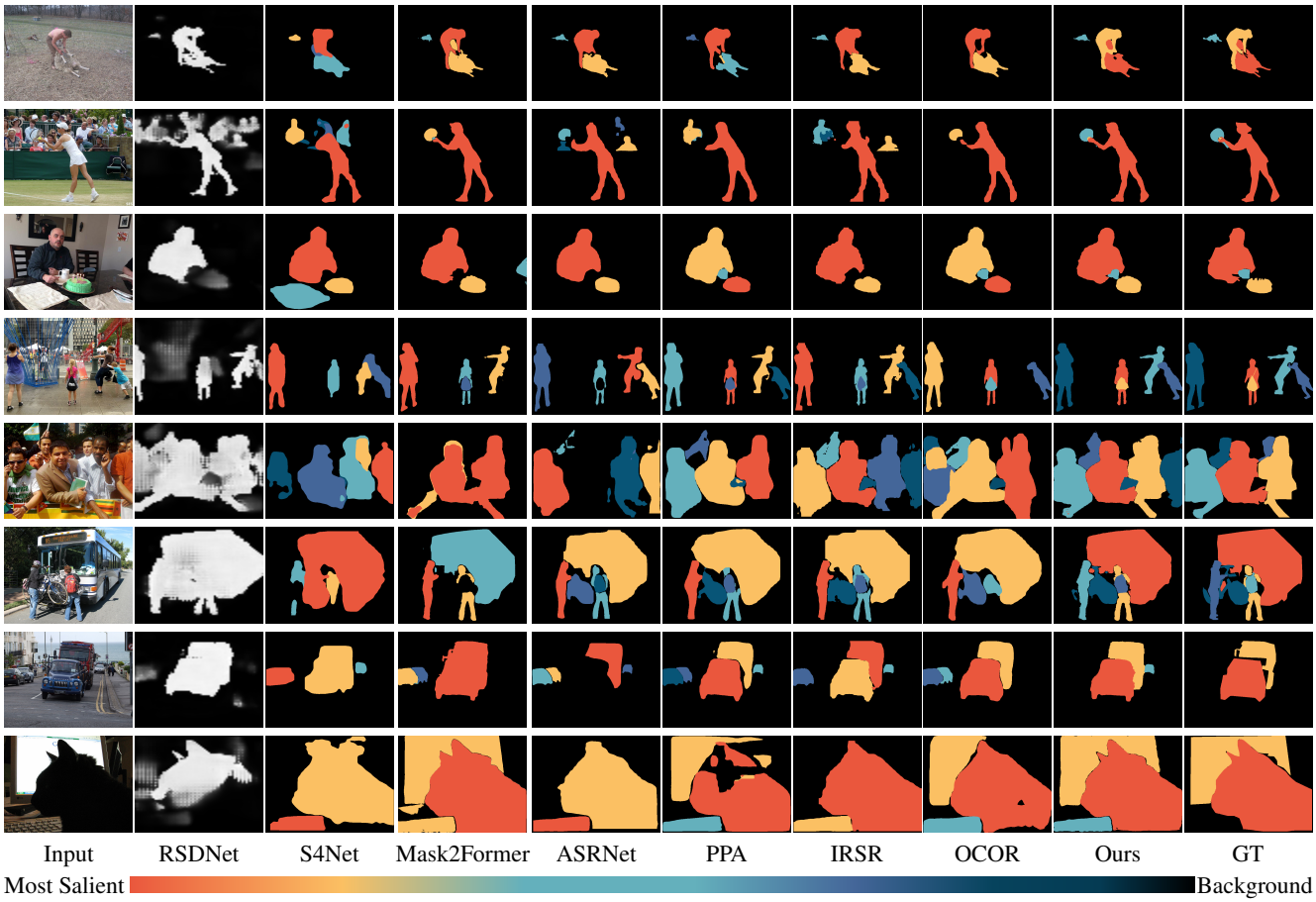

Figure 5: Visual Comparison. Salient instances are colorized using varying color temperatures, ranging from warm to cool, indicating the order in which they are visited. In general, our model produces more favorable results compared to other methods.

Quantitative Results

To thoroughly evaluate our approach, we compare it with 10 other related methods, including two SOD methods (Liu et al. 2021b; Wang et al. 2023), one SID method (Fan et al. 2019), two instance segmentation methods (Fang et al. 2021b; Cheng et al. 2022) and five SOR methods (Islam, Kalash, and Bruce 2018; Siris et al. 2020; Fang et al. 2021a; Liu et al. 2021a; Tian et al. 2022a). For a fair comparison, we retrain all these methods on both the ASSR and IRSR benchmarks. For SOD and SID methods, we compute the saliency ranks based on the average saliency intensity following (2018). For instance segmentation methods, we consider the rank labels to be the category labels. Table 1 shows the quantitative results. Our method consistently outperforms all other compared methods in nearly every metric on both benchmarks. Notably, our SA-SOR scores surpass the second-best by a clear margin of 5.4% and 6.1% on ASSR and IRSR, respectively, demonstrating our approach's superior ability to predict human attention shifts. Additionally, our method significantly improves the MAE scores in the more challenging IRSR benchmark, which contains images with more salient instances. These results clearly highlight the advantages of the proposed SeqRank.

Visual Comparison

We also present the visual comparison in Figure 5. In general, our approach can produce more favorable results compared to the other methods. For example, in the second row, our method accurately identifies the tennis player as the most salient object and then shifts focus to the tennis ball that is small in size. In contrast, other approaches incorrectly highlight either another person or the tennis racket as the second most salient object. In the third row, our method predicts a natural ranking by identifying the man as the most salient object, followed by the cake and then the cup near it. Additionally, our approach is capable of detecting and ranking multiple salient instances in complex scenes. For instance, in the fifth row, despite the presence of a crowd of people and many small distractors, our method still produces pleasing segmentation and ranking results.

Ablation Study

We now conduct experiments to understand the effectiveness of the proposed modules and strategies. For simplicity, all experiments are conducted on the ASSR dataset.

Fovea Layer is the basic component of our FOM. To validate its effectiveness, we replace it with a standard trans-

Method	SA-SOR	SOR	MAE
transformer decoder layer	0.672	0.849	7.82
fovea layer (h=w=1/1/1)	0.680	0.864	7.35
fovea layer (h=w=1/2/3)	**0.685**	**0.870**	**7.22**

Table 2: Analysis on the Fovea Layer. h, w are the height and the width of the RoI window. We set h and w according to the scales of $feat_i$, where $i \in \{5, 4, 3\}$.

Method	#FL	SA-SOR	SOR	MAE	FLOPs
$feat_5$ only	3	0.676	0.859	7.57	512.4G
$feat_3$ only	3	0.678	0.870	7.44	516.1G
multi-scale	3	0.682	**0.872**	7.59	513.9G
multi-scale	6	**0.685**	0.870	**7.22**	516.4G

Table 3: Multi-scale v.s. single-scale. Multi-scale strategy uses $feat_5$, $feat_4$ and $feat_3$ in turn, while single-scale strategy only uses image features at a single scale for all fovea layers. #FL: the number of fovea layers.

former decoder layer (Vaswani et al. 2017). We further analyze the impact of varying the size of the RoI window. As shown in Table 2, the fovea layer consistently achieves better results than the transformer decoder layer, regardless of the size of the RoI window. In particular, our aggressive settings (h=w=1/2/3) lead to a 7.7% decrease in the MAE scores.

Multi-scale strategy is adopted by our FOM, which updates object queries by using $feat_5$, $feat_4$ and $feat_3$ in sequence. We compare it with the single-scale strategy in Table 3. It shows that the high-resolution features ($feat_3$) are more important for accurate segmentation but at the expense of increased FLOPs. We adopt the multi-scale strategy since it strikes a balance between performance and FLOPs.

Sequential ranking of salient objects is the core design of SRM. This is in contrast to previous SOR models, which rank salient objects in parallel by directly predicting their saliency scores or ranks. Table 4 compares the two ranking strategies and shows that the sequential ranking strategy performs better, regardless of whether the baseline or SRM is used. Besides, SRM with the sequential strategy can lead to the best performance, demonstrating the effectiveness of our SRM and sequential ranking strategy, which agrees with human viewing behaviors.

Sequential Ranking Module. We study the importance of each component by removing them one at a time. As shown in Table 5, we can see that the "feat to q attn" layer is more important for our SRM design. We think this is because the subsequent saliency competition modeling involves the concatenation between the pixel embedding and the object query, which demands the alignment between image features and object queries. When we further remove the selection and concatenate operations, *i.e.,* baseline, we find that both SA-SOR and SOR improve moderately. Moreover, we conduct ablation experiments to understand the importance of the first and second self-attention layers. Our results show that the second self-attention layer, which is behind the "q to feat attn" layer, is more useful for this task.

Generalization Ability. We further evaluate SeqRank's generalization ability with new internet-sourced images. The results, shown in Figure 6, demonstrate its robust performance and generalization across various scenarios.

Method	Mode	SA-SOR	SOR	MAE
w/o SRM	Parallel	0.669	0.862	7.55
Baseline	Parallel	0.665	0.861	7.41
Baseline	Sequential	0.670	0.870	7.47
SRM	Parallel	0.676	0.857	7.68
SRM	Sequential	**0.685**	**0.870**	**7.22**

Table 4: Sequential ranking v.s. Parallel ranking. w/o SRM: remove SRM and rank salient objects in parallel. Baseline: a cross-attention layer followed by a self-attention layer and a feed-forward neural network.

Method	SA-SOR	SOR	MAE
w/o feat to q attn	0.665	0.861	7.31
w/o first self-attn	0.679	**0.873**	7.39
w/o second self-attn	0.672	0.864	7.25
Baseline	0.670	0.870	7.47
SRM	**0.685**	0.870	**7.22**

Table 5: Analysis on the SRM. To validate the importance of each component, we remove them one at a time. Baseline: only keep the last two attention layers and the last feed-forward neural network.

Figure 6: Examples from Internet. The input images are in the top row and the predicted results are in the bottom row.

Conclusions

In this work, we propose SeqRank, a novel method for salient object ranking (SOR). It predicts the order of human attention on various scene objects in a sequential manner, allowing us to model the temporal relationships between objects. Moreover, we propose *Fovea Module (FOM)* and *Sequential Ranking Module (SRM)* to facilitate an accurate and realistic prediction. The FOM stimulates human foveal vision to learn high-acuity visual features, while the SRM aims to mimic human peripheral vision to predict where the viewer's attention is likely to shift to next. We conduct extensive experiments to demonstrate the superior performance of SeqRank. We hope our work can inspire future research in this direction and facilitate various applications that require understanding human attention.

Acknowledgments

This work was in part supported by an Adobe Research Gift (No.: 9229152).

References

Aberman, K.; He, J.; Gandelsman, Y.; Mosseri, I.; Jacobs, D. E.; Kohlhoff, K.; Pritch, Y.; and Rubinstein, M. 2022. Deep saliency prior for reducing visual distraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19851–19860.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J., eds., *Computer Vision - ECCV 2020 - 16th European Conference*, volume 12346, 213–229. Springer.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1290–1299.

Cheng, M.-M.; Warrell, J.; Lin, W.-Y.; Zheng, S.; Vineet, V.; and Crook, N. 2013. Efficient salient region detection with soft image abstraction. In *Proceedings of the IEEE International Conference on Computer vision*, 1529–1536.

Desimone, R.; and Duncan, J. 1995. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1): 193–222.

Du, L.; Li, L.; Wei, D.; and Mao, J. 2019. Saliency-guided single shot multibox detector for target detection in SAR images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(5): 3366–3376.

Fan, R.; Cheng, M.-M.; Hou, Q.; Mu, T.-J.; Wang, J.; and Hu, S.-M. 2019. S4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6103–6112.

Fang, H.; Zhang, D.; Zhang, Y.; Chen, M.; Li, J.; Hu, Y.; Cai, D.; and He, X. 2021a. Salient object ranking with position-preserved attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16331–16341.

Fang, Y.; Yang, S.; Wang, X.; Li, Y.; Fang, C.; Shan, Y.; Feng, B.; and Liu, W. 2021b. Instances as queries. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6910–6919.

He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017a. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, 2961–2969.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, S.; Jiao, J.; Zhang, X.; Han, G.; and Lau, R. W. 2017b. Delving into salient object subitizing and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, 1059–1067.

Hooge, I. T. C.; and Erkelens, C. J. 1999. Peripheral vision and oculomotor control during visual search. *Vision research*, 39(8): 1567–1575.

Islam, M. A.; Kalash, M.; and Bruce, N. D. 2018. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7142–7150.

Jiang, B.; Zhang, L.; Lu, H.; Yang, C.; and Yang, M.-H. 2013. Saliency detection via absorbing markov chain. In *Proceedings of the IEEE international conference on computer vision*, 1665–1672.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1072–1080.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.

Li, G.; Xie, Y.; Lin, L.; and Yu, Y. 2017. Instance-level salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2386–2395.

Li, H.; Zhang, D.; Liu, N.; Cheng, L.; Dai, Y.; Zhang, C.; Wang, X.; and Han, J. 2023. Boosting Low-Data Instance Segmentation by Unsupervised Pre-training with Saliency Prompt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15485–15494.

Li, X.; Lu, H.; Zhang, L.; Ruan, X.; and Yang, M.-H. 2013. Saliency detection via dense and sparse reconstruction. In *Proceedings of the IEEE international conference on computer vision*, 2976–2983.

Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; and Belongie, S. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2117–2125.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, 740–755. Springer.

Liu, N.; Han, J.; and Yang, M.-H. 2020. PiCANet: Pixel-wise contextual attention learning for accurate saliency detection. *IEEE Transactions on Image Processing*, 29: 6438–6451.

Liu, N.; Li, L.; Zhao, W.; Han, J.; and Shao, L. 2021a. Instance-level relative saliency ranking with graph reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11): 8321–8337.

Liu, N.; Zhang, N.; Wan, K.; Shao, L.; and Han, J. 2021b. Visual saliency transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4722–4732.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021c. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Ludwig, C. J.; Davies, J. R.; and Eckstein, M. P. 2014. Foveal analysis and peripheral selection during active visual sampling. *Proceedings of the National Academy of Sciences*, 111(2): E291–E299.

Miangoleh, S. M. H.; Bylinskii, Z.; Kee, E.; Shechtman, E.; and Aksoy, Y. 2023. Realistic Saliency Guided Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 186–194.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571. Ieee.

Neisser, U. 2014. *Cognitive psychology: Classic edition*. Psychology press.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 7479–7489.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I. D.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 658–666.

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

Siris, A.; Jiao, J.; Tam, G. K.; Xie, X.; and Lau, R. W. 2020. Inferring attention shift ranks of objects for image saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12133–12143.

Tian, X.; Xu, K.; Yang, X.; Du, L.; Yin, B.; and Lau, R. W. 2022a. Bi-directional object-context prioritization learning for saliency ranking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5882–5891.

Tian, X.; Xu, K.; Yang, X.; Yin, B.; and Lau, R. W. 2022b. Learning to detect instance-level salient objects using complementary image labels. *International Journal of Computer Vision*, 130(3): 729–746.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Wang, R.; Fan, X.; Wang, T.; and He, X. 2023. Pixels, Regions, and Objects: Multiple Enhancement for Salient Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10031–10040.

Wolf, C.; Belopolsky, A. V.; and Lappe, M. 2022. Current foveal inspection and previous peripheral preview influence subsequent eye movement decisions. *Iscience*, 25(9).

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Wu, Y.-H.; Liu, Y.; Zhang, L.; Cheng, M.-M.; and Ren, B. 2022. EDN: Salient object detection via extremely-downsampled network. *IEEE Transactions on Image Processing*, 31: 3125–3136.

Yan, Q.; Xu, L.; Shi, J.; and Jia, J. 2013. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1155–1162.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8779–8788.

Zhu, W.; Liang, S.; Wei, Y.; and Sun, J. 2014. Saliency optimization from robust background detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2814–2821.