Bi-directional Object-Context Prioritization Learning for Saliency Ranking

Xin Tian[1,2,3]   Ke Xu[2,†]   Xin Yang[1,†]   Lin Du[3]   Baocai Yin[1]   Rynson W.H. Lau[2,‡]

[1]Dalian University of Technology   [2]City University of Hong Kong

[3]AI Application Research Center (AARC), Huawei Technologies Co., Ltd.

{xin.tian.831, kkangwing}@gmail.com, {xinyang, ybc}@dlut.edu.cn,

dulin09@huawei.com, Rynson.Lau@cityu.edu.hk

Abstract

*The saliency ranking task is recently proposed to study the visual behavior that humans would typically shift their attention over different objects of a scene based on their degrees of saliency. Existing approaches focus on learning either object-object or object-scene relations. Such a strategy follows the idea of **object-based attention** in Psychology, but it tends to favor objects with strong semantics (e.g., humans), resulting in unrealistic saliency ranking. We observe that **spatial attention** works concurrently with object-based attention in the human visual recognition system. During the recognition process, the human spatial attention mechanism would move, engage, and disengage from region to region (i.e., context to context). This inspires us to model region-level interactions, in addition to object-level reasoning, for saliency ranking. Hence, we propose a novel bi-directional method to unify spatial attention and object-based attention for saliency ranking. Our model has two novel modules: (1) a selective object saliency (SOS) module to model object-based attention via inferring the semantic representation of salient objects, and (2) an object-context-object relation (OCOR) module to allocate saliency ranks to objects by jointly modeling object-context and context-object interactions of salient objects. Extensive experiments show that our approach outperforms existing state-of-the-art methods. Code and pretrained model are available at* https://github.com/GrassBro/OCOR.

1. Introduction

Saliency detection is a fundamental task in computer vision. Previous works mainly focus on the salient object / instance detection tasks [9, 17, 38], and the gaze prediction task [33]. Recently, Siris *et al*. [34] propose to study a novel task called saliency ranking, which aims to detect the salient instances and infer their saliency ranks simultaneously. By

---

† Ke Xu and Xin Yang are joint corresponding authors.
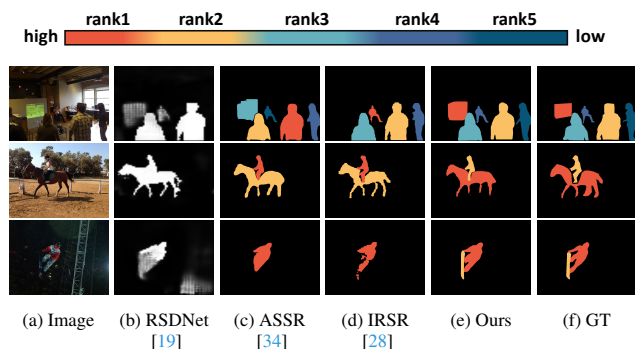
‡ Rynson Lau leads this project.

Figure 1. Existing salient ranking approaches [19, 28, 34] produce unrealistic saliency ranks. (b) RSDNet [19] is a pixel-level method and does not predict object-level saliency ranks well. (c) ASSR [34] and (d) IRSR [28] explore object-object and object-scene relations for inferring saliency ranks. However, they favor objects with strong semantics and tend to assign people with high saliency ranks. (e) Our method explores both spatial and object-based attentions with a bi-directional object-context prioritization learning formulation, yielding faithful saliency ranking results.

mimicking how humans change their attention across the scenes depending on the saliency ranks, saliency ranking can benefit many down-stream visual tasks, *e.g*., image manipulation [6, 49], scene understanding [37], important person identification [24] and their interaction reasoning [10].

Islam *et al*. [19] propose the first saliency ranking work, which directly predicts a relative saliency map with different pixel values indicating different saliency degrees, as shown in Figure 1(b). However, this method only studies the relative saliency of pixels. Later, Siris *et al*. [34] propose to study the salient object ranking as humans shift attention from object to object. They propose to model the relations between objects and global context for reasoning their ranks. Liu *et al*. [28] further propose a neural graph-based method to learn relations between objects and local contexts as well as relations between objects. However, these two methods [28, 34] tend to favor objects with strong semantics (*e.g*., people) as shown in Figure 1, resulting in incorrect saliency ranks. For example, in the first and sec-

ond rows of Figure 1(c,d), they assign people with highest saliency ranks, despite the visual distinctiveness of the green screen in the first example and the relatively larger horse in the second example. In the last row, the two methods do not even consider the skateboard as a salient object.

In this paper, we tackle the saliency ranking problem based on the observation that as revealed by psychological studies [2,8], **spatial attention** and **object-based attention** work concurrently in the human visual system. While object-based attention directs our views to the candidate objects or perceptual groups via a preattentive segmentation of the scene [8], the human spatial attention mechanism allows us to process the scene sequentially through prioritizing the regions (where the objects belong to) based on low-level visual stimuli (*e.g.*, rich colors), functionalities of the objects, and interactions among the objects. This inspires us to jointly exploit spatial attention and object-based attention for saliency ranking. Based on this observation, we formulate a bi-directional object-context prioritization learning approach to model both region-level and object-level relations. We first propose a Selective Object Saliency (SOS) module to model object-based attention via inferring and enriching the semantic representations of the salient objects based on their local contexts. We then propose an object-context-object relation (OCOR) module to exploit the spatial attention mechanism by reasoning in the object-context and context-object bi-directional way. We formulate a multi-head attention mechanism to model the way that an object with its context would interact with other objects with their contexts. As shown in Figure 1(e), our approach produces more faithful saliency ranks over the state-of-the-art methods. For example, our method can detect the screen (first row) and the horse (second row) as the most salient objects according to their visual distinctiveness. In the third row, our method can detect and rank the skateboard by modeling its interaction with the person.

To summarize, this work has three main contributions: *1)* Inspired by the psychological studies, we propose a novel bi-directional object-context prioritization learning approach for saliency ranking, by jointly exploiting spatial and object-based attention mechanisms. *2)* We propose a novel selective object saliency (SOS) module for modeling object-based attention, and a novel object-context-object relation (OCOR) module for modeling spatial attention via inferring the relations of objects in a bi-directional object-context and context-object manner. *3)* We conduct extensive experiments to analyze our approach and verify its superior performance over the state-of-the-art methods.

2. Related Work

2.1. Saliency Ranking

Saliency ranking is a new task. It studies the visual phenomenon that objects in a daily scene generally have different saliency degrees, which draws the observer's attention sequentially across the objects. Islam *et al*. [19] make an initial attempt on this problem, but they only study the pixel-level relative saliency. This is evident from their data collection step, where multiple objects are sometimes annotated with the same saliency rank. Siris *et al*. [34] propose a new dataset and a model to facilitate research on this task. Specifically, they exploit the statistics of fixations on objects to build a large-scale dataset, and design a network with an object-context relation module to learn saliency ranks. Liu *et al*. [28] further propose another dataset with less annotation noise, and investigate object-object relations for the task. Similarly, Fang *et al*. [12] also propose to model object-object relations but embed the spatial coordinates of objects as spatial cues in the learning step.

By modeling object-based attention, the above methods can take advantages of the learned semantic representations of objects. However, they tend to produce unreasonable saliency ranks, *e.g.*, they tend to rank humans higher than stuff with visually distinguishing hues. In contrast, in this paper, we propose a bi-directional object-context prioritization method to combine spatial attention with object-based attention for saliency ranking.

2.2. Salient Object Detection (SOD)

SOD is a long-standing problem with a lot of methods proposed. It aims to detect conspicuous objects in a scene. Earlier methods [1, 7, 31, 48] mainly rely on hand-crafted features (*e.g.*, color, brightness, and textures) to detect salient objects. These methods often fail in complex scenarios due to the limited representation capacities of low-level hand-crafted features. Recent deep learning-based methods [42] have achieved superior performances. They mainly incorporate two kinds of deep techniques, *deep feature fusion* [44,51] and *feature attention* [27,35,52,53]. Deep feature fusion aims to aggregate multi-level context information including low-level stimulus and high-level semantics for SOD, while feature attention reweights multi-scale features and enhances the context learning to help the model focus on the salient regions and suppress noise from the background regions.

Unlike SOD, the saliency ranking task needs to detect salient instances and then determine their saliency ranks. Hence, existing salient object detection methods cannot be directly applied to the saliency ranking task.

2.3. Salient Instance Detection (SID)

A few methods are proposed to detect salient objects at the instance level. Zhang *et al*. [50] propose a Maximum a Posteriori optimization based method to detect salient instances with bounding boxes. Li *et al*. [22] propose to leverage the instance-aware saliency contours for detecting instance-level objects. Fan *et al*. [11] propose to combine
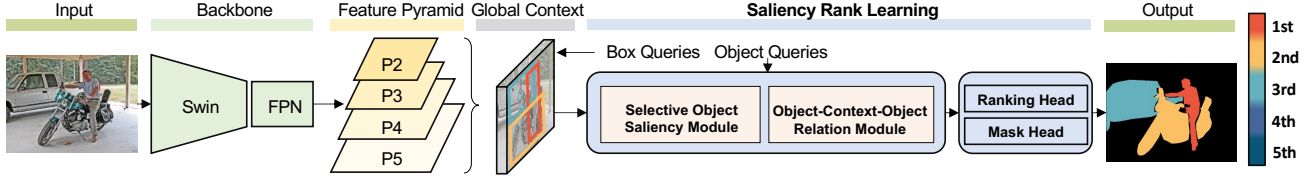
Figure 2. Overview of our proposed network. Given an input image, we first apply the query based object detection method [13, 36] to extract global context features and leverage a set of learnable salient object proposals (*i.e.*, box and object queries that encode object locations and rich object characteristics), to help predict the final saliency ranks. The Saliency Rank Learning process has (1) a SOS module to capture and enrich object-level semantic representation, (2) an OCOR module to model the interactions from one object with its context to other objects with their contexts, *i.e.*, bi-directional object-context and context-object relations learning, and (3) ranking and mask heads to infer object-wise saliency ranks on top of the improved features from the SOS and OCOR modules.

the object detection model FPN [26] with a segmentation branch to detect salient instances. Tian *et al.* [38, 39] propose a weakly-supervised method that exploits class labels and subitizing labels for SID.

SID methods can provide instance-level information of salient objects, but they do not attempt to rank the saliency degrees of the detected salient instances.

3. Methdology

Psychological studies [2, 8] reveal that **spatial attention** and **object-based attention** work cooperatively in the human visual systems in order to process multiple visual inputs in a scene sequentially. On the one hand, object-based attention tends to fix humans' views to the candidate objects by preattentive segmentation of the scene, as humans are more likely to be attracted to objects that they are most familiar with. On the other hand, spatial attention directs humans' views from region to region via a joint process of multiple factors, such as regional visual stimuli, functionalities of objects, and their contextual interactions. This motivates us to exploit object-based attention and spatial attention to design our saliency ranking model.

Figure 2 shows the overview of our bi-directional object-context prioritization model. Given an input image, we first use the query-based object detection method [13, 36] to extract global features and generate a set of object features based on object proposals, *i.e.*, box and object queries that encode object locations and rich object characteristics. We then feed them into two novel modules, Selective Object Saliency (SOS) module and Object-Context-Object Relation (OCOR) module, which model the two attention mechanisms for reasoning the saliency ranks. Finally, our model learns prioritization information of objects for saliency ranking.

3.1. Selective Object Saliency Module

Following the spirit of the object-based attention mechanism, the goal of our SOS module is to capture and enhance the semantic representation of salient objects. It has been shown that channels of deep features response to diverse semantic components [4, 54]. Previous channel-wise
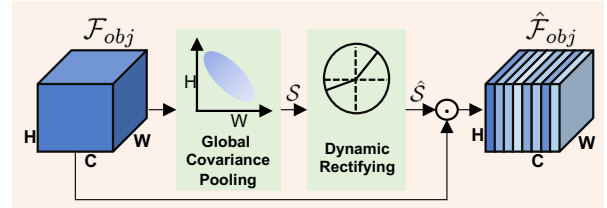

Figure 3. Structure of our SOS module.

attentions [18, 45] mainly aim to highlight the discriminative channels according to the ground truth categories while suppressing the responses from other channels. However, simply suppressing the low responses from the other channels may not be suitable for saliency ranking, as these less discriminative channels may also be informative. They may serve as the contexts to correlate objects to each other or as the global context. Hence, we propose to extend existing channel-wise attentions in two aspects. First, we leverage the global covariance pooling [23, 41] to learn object representations as well as their correlations to both local and global contexts. Second, we learn a group of dynamic rectifying functions to reallocate attentions to the channels based on the high-order feature statistics computed by the global covariance pooling. Consequently, they jointly capture fine-grained object information for learning object-based representations. Figure 3 shows our SOS module.

Global Covariance Pooling investigates cross-channel correlations for modeling high-order feature statistics. Given the context features $\mathcal{F}_{context}$ and object proposals (*i.e.*, box and object queries), we first extract object features $\mathcal{F}_{obj} \in R^{H \times W \times C}$ using ROIAlign [15], where H, W, and C indicate the height, width, and channel dimension of the features. We then split $\mathcal{F}_{obj}$ along the channel dimension to generate channel-wise features $\mathcal{F}_{obj} = [\mathcal{F}_{obj}^1, \mathcal{F}_{obj}^2, ..., \mathcal{F}_{obj}^K] \in R^{C \times K}$, where $K = W \times H$ and $k \in [1, K]$, for computing the covariance normalization matrix $\mathcal{M}$ as:

$$\mathcal{M} = \left(\frac{1}{K} \sum_{k=1}^{K} \left(\mathcal{F}_{obj}^k - \mu \right) \left(\mathcal{F}_{obj}^k - \mu \right)^T \right)^{\alpha} \in R^{C \times C},$$

(1)

where $\mu = \left(\frac{1}{K}\sum_{k=1}^{K}\mathcal{F}_{obj}^k\right) \in R^{1\times K}$ is the channel-wise mean vector. α is a normalization hyper parameter. We then compute the global covariance pooling $\mathcal{GCP}$ on $\mathcal{M}$ as:

$$s_c = \mathcal{GCP}(\mathcal{M}_c) = \frac{1}{C}\sum_{c=1}^{C}\mathcal{M}_c, \qquad (2)$$

where $\mathcal{S} = [s_1, s_2, ..., s_C] \in R^{C\times 1}$ is the high-order channel-wise feature statistics, which encodes objects and their relations to both local and global contexts.

Dynamic Rectifying Functions further dynamically reallocate attentions to each feature channel based on the channel-wise statistics $\mathcal{S}$. It first learns to reweight each s_c as:

$$\hat{s}_c = \max\{a_c^1(\mathcal{S})s_c + b_c^1(\mathcal{S}), a_c^2(\mathcal{S})s_c + b_c^2(\mathcal{S})\}, \qquad (3)$$

where a_c^1, b_c^1, a_c^2, and b_c^2 are learnable parameters, which form two sets of coefficients (*i.e.*, (a_c^1, b_c^1) and (a_c^2, b_c^2)) to formulate two piece-wise functions for updating s_c to $\hat{s}_c$. Since the max operation in Eq. 3 is not differentiable, it is reformulated as:

$$[\Delta a_{1:C}^1, \Delta a_{1:C}^2, \Delta b_{1:C}^1, \Delta b_{1:C}^2] \in R^{4\times C} =$$
$$2\sigma(FC_{R/C\to 4\times C}(ReLU(FC_{C\to R/C}(\mathcal{S})))) - 1, \qquad (4)$$
$$and, a_{1:C}^1 = 1 + \lambda_a\Delta a_{1:C}^1,\ a_{1:C}^2 = \lambda_a\Delta a_{1:C}^2,$$
$$b_{1:C}^1 = \lambda_b\Delta b_{1:C}^1,\ b_{1:C}^2 = \lambda_b\Delta b_{1:C}^2, \qquad (5)$$

where $FC_{C\to R/C}$ is a fully connected layer, which changes the feature dimension from C to R/C. σ is a Sigmoid function. λ_a, λ_b are hyper-parameters that are set to 1 and 0.5. $a_{1:C}^1 = [a_1^1, a_2^1, ..., a_C^1] \in R^{1\times C}$. $a_{1:C}^2$, $b_{1:C}^1$, and $b_{1:C}^2$ are defined similar to $a_{1:C}^1$.

Note that both a and b are related not only to the input s_c, but also to others $s_{i\neq c} \in \mathcal{S}$ via two FC layers. Hence, we establish the cross-channel correlations and compute the enhanced $\hat{\mathcal{F}}_{obj}$ as:

$$\hat{\mathcal{F}}_{obj} = \mathcal{F}_{obj} \diamond \hat{S}, \qquad (6)$$

where $\hat{S} = [\hat{s_1}, \hat{s_2}, ..., \hat{s_c}] \in R^{C\times 1}$, and $\diamond$ denotes channel-wise multiplication.

Given the globally pooled high-order channel-wise feature statistics, the dynamic rectifying function refines it via linear functions described in Eq. 3, where slopes and intercepts (*e.g.*, (a_c^1, b_c^1)) of the linear functions are adaptively learned via Eq. 4 and 5. We leverage two sets of slopes and intercepts to help adjust negative/positive channel responses separately. In this way, fine-grained attentions can be adjusted via the slopes and intercepts before they are assigned to the feature channels, to accommodate for the saliency ranking task.
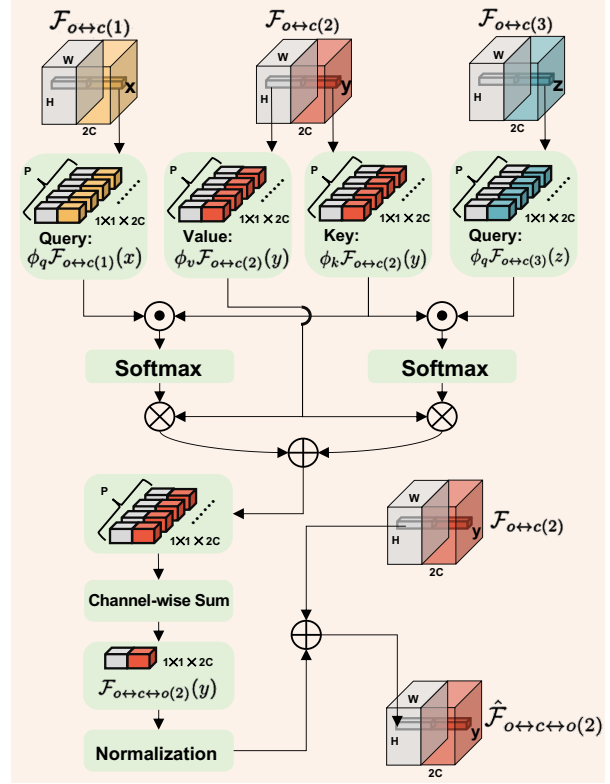

Figure 4. Structure of our OCOR module.

3.2. Object-Context-Object Relation Module

The OCOR module aims to model the spatial attention of the human visual system for learning how to prioritize regions. To this end, we first encode the object-context relation based on the enhanced object representation of the SOS module, and then establish the bi-directional object-context-object relation, to mimic the spatial scanning process of humans.

Formulation of object-context relation. After the SOS module, we obtain N object features $\hat{\mathcal{F}}_{obj(n)} \in R^{H\times W\times C}$. We then rescale $\mathcal{F}_{context}$ to have the same spatial size ($H \times W$) as $\hat{\mathcal{F}}_{obj(n)}$, and establish the object-context relation by concatenating each $\hat{\mathcal{F}}_{obj(n)}$ with $\mathcal{F}_{context}$ to form a series of object-context relation features $\{\mathcal{F}_{o\leftrightarrow c(n)}\}_{i=1}^{N} \in R^{H\times W\times 2C}$ for the N objects.

Formulation of object-context-object relation. Based on the object-context relations, we build bi-directional object-context-object relations $\{\mathcal{F}_{o\leftrightarrow c\leftrightarrow o(n)}\}_{n=1}^{N}$ to model how attentions shift from region to region, accompanied with objects interacting with objects through contexts. Specifically, we exploit groups of linear projections to compute long-range interactions between different object-context relation features, inspired by the multi-head attention mechanism [40]. We denote $i, j \in N$ as two different objects, $x, y \in K$ as spatial locations, and ϕ as a linear projec-

tion function. We utilize P distinct projection functions, ϕ^p, $p \in [1, P]$, to obtain Key k, Query q, and Value v embedding of the object-context relation features. For example, $\phi_v^p \mathcal{F}_{o \leftrightarrow c(i)}(x)$ denotes the value embedding v of object-context relation of object i at location x using the p-th linear projection function. The interaction can be modeled as:

$$Att_{\langle i,j \rangle}^p(x, y) = \varrho(A_{\langle i,j \rangle}^p(x, y)) \times \phi_v^p(\mathcal{F}_{o \leftrightarrow c(i)}(x)), \quad (7)$$

where $Att_{\langle i,j \rangle}^p(x, y)$ measures how object i with its spatial context x interacts with object j with its spatial context y. ϱ is the weighting function (*i.e.*, Softmax) to compute the importance of the embedding of the object-context representation, and $A_{\langle i,j \rangle}^p(x, y)$ encodes the context information of $\mathcal{F}_{o \leftrightarrow c(i)}(x)$. They are calculated as:

$$\varrho(A_{\langle i,j \rangle}^p(x, y)) = \frac{e^{A_{\langle i,j \rangle}^p(x,y)}}{\sum_{p' \in P} e^{A_{\langle i,j \rangle}^{p'}(x,y)}}, \quad (8)$$

$$A_{\langle i,j \rangle}^p(x, y) = \phi_k^p(\mathcal{F}_{o \leftrightarrow c(i)}(x))^T \cdot \phi_q^p(\mathcal{F}_{o \leftrightarrow c(j)}(y)), \quad (9)$$

where $\cdot$ is the dot production between two feature vectors. We compute the object-context-object relation features as:

$$\mathcal{F}_{o \leftrightarrow c \leftrightarrow o(i)}(x) = \sum_{j=1, j \neq i}^{N} \sum_{y=1}^{K} \sum_{p=1}^{P} Att_{\langle i,j \rangle}^p(x, y). \quad (10)$$

Finally, we normalize $\mathcal{F}_{o \leftrightarrow c \leftrightarrow o}$ and fuse it with the input $\mathcal{F}_{o \leftrightarrow c}$ to produce the final output $\hat{\mathcal{F}}_{o \leftrightarrow c \leftrightarrow o}$. Figure 4 shows the structure of our OCOR module.

3.3. Learning of Saliency Rank

To learn saliency ranking based on our SOS and OCOR modules, we formulate the ranking step as a multi-stage query-based detection process, and follow [14] to initialize with T query stages ($1 \leq t \leq T$). The advantage of this multi-stage strategy is that the box and object queries (q_{box}^t, and q_{obj}^t), and the corresponding object features $\mathcal{F}_{obj}^t$ can be improved stage by stage. At each stage, we perform the following three sub-tasks. First, we extract the object features of the current stage t (*i.e.*, $\mathcal{F}_{obj}^t$) from the box query of stage $t-1$ (*i.e.*, q_{box}^{t-1}) as:

$$\mathcal{F}_{obj}^t = \text{ROIAlign}(\mathcal{F}_{context}, q_{box}^{t-1}). \quad (11)$$

Second, we obtain the improved object features $\tilde{\mathcal{F}}_{obj}^t$ and object query q_{obj}^t as:

$$\tilde{\mathcal{F}}_{obj}^t, q_{obj}^t = H_{rank}^t \left(\underbrace{f_{OCOR}(f_{SOS}(\mathcal{F}_{obj}^t))}_{\hat{\mathcal{F}}_{o \leftrightarrow c \leftrightarrow o}^t}, f_{SA}(q_{obj}^{t-1}) \right), \quad (12)$$

where f_{SOS} and f_{OCOR} are our SOS and OCOR modules, f_{SA} is the multi-head self-attention [40], and H_{rank}^t is the

ranking head at stage t. Third, we update the box query using the box prediction branch $\mathcal{B}$:

$$q_{box}^t = \mathcal{B}_t(\tilde{\mathcal{F}}_{obj}^t). \quad (13)$$

At the final stage, the object features $\tilde{\mathcal{F}}_{obj}^T$ are fed into the ranking head for predicting the final saliency ranks, and the mask head for predicting the salient object masks.

4. Experiments

4.1. Experimental Setups

Datasets and Metrics. Our experiments are conducted on the publicly available ASSR [34] and IRSR [28] datasets. ASSR [34] ranks 5 salient objects per image based on the sequential eye gaze information. It provides 7,464, 1,436, and 2,418 images for training, validation, and testing, respectively. IRSR [28] considers both eye gaze sequences and the duration of eye gaze for labeling the saliency ranks. It also manually filters out inappropriate annotations. In IRSR, each image involves at most 8 salient objects with ranks. It contains 8,988 images, which are divided into 6,059 images for training and 2,929 images for test.

We adopt three metrics, *i.e.*, Salient Object Ranking (SOR) [19, 34], Segmentation-Aware SOR (SA-SOR) [28], and Mean Absolute Error (MAE), to evaluate our method. SOR computes the Spaearman's rank-order correlation between the prediction and ground truth of the saliency ranking order. This metric indicates the prediction quality of relative saliency among objects rather than a particular saliency ranking number for an object. SA-SOR calculates the Pearson correlation between the prediction and ground truth of saliency ranks. It also penalizes the errors of detecting non-salient objects and false ranking. MAE measures averaged per-pixel differences between the prediction and ground truth of saliency map. Hence, it also helps measure the ranking quality of an object depending on its overlapping degree with the ground truth.

Implementation Details. Our network is built upon the multi-stage query-based detector [14, 36]. Following their settings, the stage number T is set to 6, and the query number N is set to 100. We utilize the Swin transformer [29] pretrained on ImageNet [20] as our backbone. H and W are initially set to 7 in the SOS module, and then reduced to 2 in the OCOR module. We set the α to $\frac{1}{2}$ and C to 256. We experimentally set $R = 4$ and $P = 8$ for our SOS and OCOR modules, respectively.

We train our model for 60 epochs on each dataset. The learning rate starts from 2.5×10^{-5}, and is divided by 10 at epochs 25 and 45. To optimize the model, we adopt AdamW optimizer with a 1×10^{-4} weight decay. Images are resized to 800×800 resolution. We use random flip for data augmentation. The batch size is set to 16 for training

on 4 Tesla GPU cards. We train the ranking head with set prediction loss [14, 36] and saliency ranking loss [28], and the mask head with dice loss [30]. We test our model on a single Tesla GPU card with images of 800×800 resolution.

4.2. Baselines

Since saliency ranking is a relatively new task with only four methods proposed, to evaluate our method comprehensively, we design baseline methods for comparisons. We find the Semantic Instance Segmentation task related to our task in that they have a similar output formulation (*i.e.*, object instance maps + corresponding semantic categories) to our task (*i.e.*, salient instance maps + saliency ranks). Hence, we include eight representative semantic instance segmentation methods as baselines for comparisons.

- Mask R-CNN [15] is a popular two-stage detector. All existing salient object ranking methods [12, 28, 34] are built on it. Here, we have two versions: a ResNet-based Mask R-CNN, and a Swin-based Mask R-CNN.

- BlendMask [5] is a single-stage anchor-free method that integrates high-level task semantics (from top-down) and low-level finer details (from bottom-up). It provides a simpler baseline compared to the anchor-based detectors [15].

- CenterMask [21] is also a single-stage anchor-free network consisting of a channel attention-based backbone (VoVNet) and a spatial attention-based mask head for extracting informative semantics and suppressing background noise, respectively.

- SOLO [43] proposes to classify and detect an instance at each location of the feature grid. They directly combine coordinate information with candidate object features so that the detection can be more sensitive to the object centroids and boundaries.

- Cascade R-CNN [3] and QueryInst [14] detect instances in a progressive manner. They utilize cascaded detectors in a network to refine the detection results stage-by-stage.

- CBNetV2 [25] proposes to connect parallel backbones in a top-down pathway. The higher-level features, with coarser space information and finer semantic information, of the former backbone can be used to enrich lower-level features of the later backbone.

To adapt these methods for saliency ranking, we modify their output layers according to the amounts of saliency ranks in the saliency ranking datasets. We train these methods using the saliency ranking loss [28].

4.3. Main Results

We compare our method to **<u>four</u>** existing saliency ranking methods: RSDNet [19], ASSR [34], IRSR [28], and SOR [12]; **<u>four</u>** salient object detection methods used by [34] for comparison: S4Net [11], BASNet [32], CPD-R [46], and SCRN [47]; and our **eight** baselines from semantic instance segmentation: CenterMask [21], SOLO [43], BlendMask [5], ResNet-based Mask R-CNN [15], Swin-based Mask R-CNN [15], Cascade R-CNN [3], QueryInst [14], and CBNetV2 [25].

Quantitative Comparisons. Table 1 shows the quantitative results. From the comparison on the ASSR testing set, we can see that S4Net performs the second best on SOR, but the second worst based on MAE. This is because missing objects are not penalized by SOR but are considered by MAE. SOD-based methods perform relatively bad on SOR, as they do not have ability to rank object saliency. Several SIS baselines perform well on the SR task. Overall, the proposed method outperforms all methods compared by a large margin on all three metrics and on both testing sets.

Qualitative Comparisons. We further qualitatively evaluate our method, as shown in Figure 5. Due to limited space, we only compare with the best performing methods in Table 1. We can see that existing methods commonly suffer from detecting non-salient objects, inferring incorrect saliency ranks among objects, and under-/over-detection of objects. The visual comparison show that our method is able to individualize salient objects and infer their saliency ranks accurately for diverse scenes.

4.4. Module Analysis

Ablation Study of Our Model. We begin by studying the effectiveness of the proposed SOS and OCOR modules. Table 2 shows the results. By removing each or both of them from the network, we can see that the performance drops significantly. This shows both SOS and OCOR modules are important to the saliency ranking performance.

Internal Analysis of the SOS Module. We study how Global Covariance Pooling (GCP) and Dynamic Rectifying (DR) functions in the SOS module influence the saliency ranking performance. Table 3 shows the results, from which we can see that GCP and DR perform much more better when they work together than individually.

Alternatives of the OCOR Module. Finally, we exploit four alternative strategies to build bi-directional object-context relations. **S1**: Given two object features and the context features, we concatenate them together and feed them to two conv. layers to learn their context relation; **S2**: Based on S1, we apply channel-wise attention [45] on the fused features; **S3**: Based on S1, we apply spatial-wise attention [45] on the fused features; **S4**: Based on S1, we apply both channel-wise and spatial-wise attentions [45] on

Table 1. Quantitative comparison with 4 state-of-the-art saliency ranking methods, 4 salient object detection methods used by [34] for comparison, and our 8 baselines in Section 4.2. We show their original tasks and backbones used (*i.e.*, ResNet [16], VoVNet [21], and Swin [29]) in the 2*nd* and 3*rd* columns. SID, SOD, SIS, and SR represent salient instance detection, salient object detection, semantic instance segmentation, and saliency ranking, respectively. Methods with † denotes that their results are copied from their original papers. - denotes missing results due to the lacking of publicly available implementations/results. Best performances are marked in **bold**.

Method	Original Task	Backbone	Benchmark Datasets and Evaluation Metrics					
			ASSR Test Set [34]			IRSR Test Set [28]		
			SA-SOR↑	SOR↑	MAE↓	SA-SOR↑	SOR↑	MAE↓
S4Net† [11]	SID	ResNet-50	-	0.891	0.150	-	-	-
BASNet† [32]	SOD	ResNet-34	-	0.707	0.115	-	-	-
CPD-R† [46]	SOD	ResNet-50	-	0.766	0.100	-	-	-
SCRN† [47]	SOD	ResNet-50	-	0.756	0.116	-	-	-
RSDNet [19]	SR	ResNet-101	0.499	0.717	0.158	0.460	0.735	0.129
ASSR [34]	SR	ResNet-101	0.667	0.792	0.101	0.388	0.714	0.125
IRSR [28]	SR	ResNet-50	0.709	0.811	0.105	0.565	0.806	0.085
SOR† [12]	SR and SIS (joint learning)	VoVNet-39	-	0.841	0.081	-	-	-
CenterMask [21]	SIS	VoVNet-99	0.672	0.813	0.099	0.509	0.771	0.113
SOLO [43]	SIS	ResNet-101	0.655	0.805	0.112	0.499	0.745	0.126
BlendMask [5]	SIS	ResNet-101	0.694	0.822	0.094	0.531	0.785	0.098
Mask R-CNN [15]	SIS	ResNet-101	0.632	0.739	0.123	0.480	0.699	0.137
Mask R-CNN [15]	SIS	Swin-L	0.643	0.750	0.118	0.489	0.712	0.128
Cascade R-CNN [3]	SIS	Swin-L	0.699	0.816	0.100	0.520	0.766	0.105
QueryInst [14]	SIS	Swin-L	0.715	0.837	0.095	0.542	0.799	0.087
CBNetV2 [25]	SIS	Cascaded Swin-L	0.704	0.827	0.101	0.514	0.780	0.091
Ours	SR	Swin-L	**0.738**	**0.904**	**0.078**	**0.578**	**0.834**	**0.079**

Table 2. Ablation study of the SOS and OCOR modules.

Method	SA-SOR ↑	SOR↑	MAE↓
w/o SOS, OCOR	0.715	0.837	0.095
w/o SOS	0.729	0.885	0.084
w/o OCOR	0.722	0.870	0.090
Ours	**0.738**	**0.904**	**0.078**

Table 3. Internal analysis of the SOS module.

Method	SA-SOR ↑	SOR↑	MAE↓
Baseline	0.715	0.837	0.095
Baseline + GCP	0.715	0.840	0.092
Baseline + DR	0.719	0.857	0.094
Baseline + SOS	**0.722**	**0.870**	**0.090**

Table 4. Comparison of different strategies for modeling object-context-object relation.

Method	SA-SOR ↑	SOR↑	MAE↓
Baseline	0.715	0.837	0.095
Baseline + S1	0.718	0.842	0.092
Baseline + S2	0.719	0.845	0.093
Baseline + S3	0.722	0.853	0.090
Baseline + S4	0.722	0.857	0.088
Baseline + OCOR	**0.729**	**0.885**	**0.084**

the fused features. Table 4 shows that our OCOR module outperforms all these baselines, verifying that our OCOR module can capture region-level context relating to an object and build bidirectional object-context relations well.

5. Conclusion

In this paper, we have studied the saliency ranking problem. We observe that the human visual system leverage both spatial and object-based attentions for handling the vi-

sual inputs. We therefore propose a bi-directional Object-Context Prioritization Learning approach for saliency ranking. A selective object saliency module is proposed to model object-based attention via capturing and enhancing object semantic representation. An object-context-object relation module is proposed to model spatial attention through studying how an object with its regional context would interact with other objects with their regional contexts. Extensive experiments have verified the effectiveness of our method against state-of-the-art methods.

Our work does have a limitation. When objects with the same functionality appear in a scene with a narrow context, our model may fail to infer correct saliency ranking. Figure 6 shows two examples. In the first row, the left and right
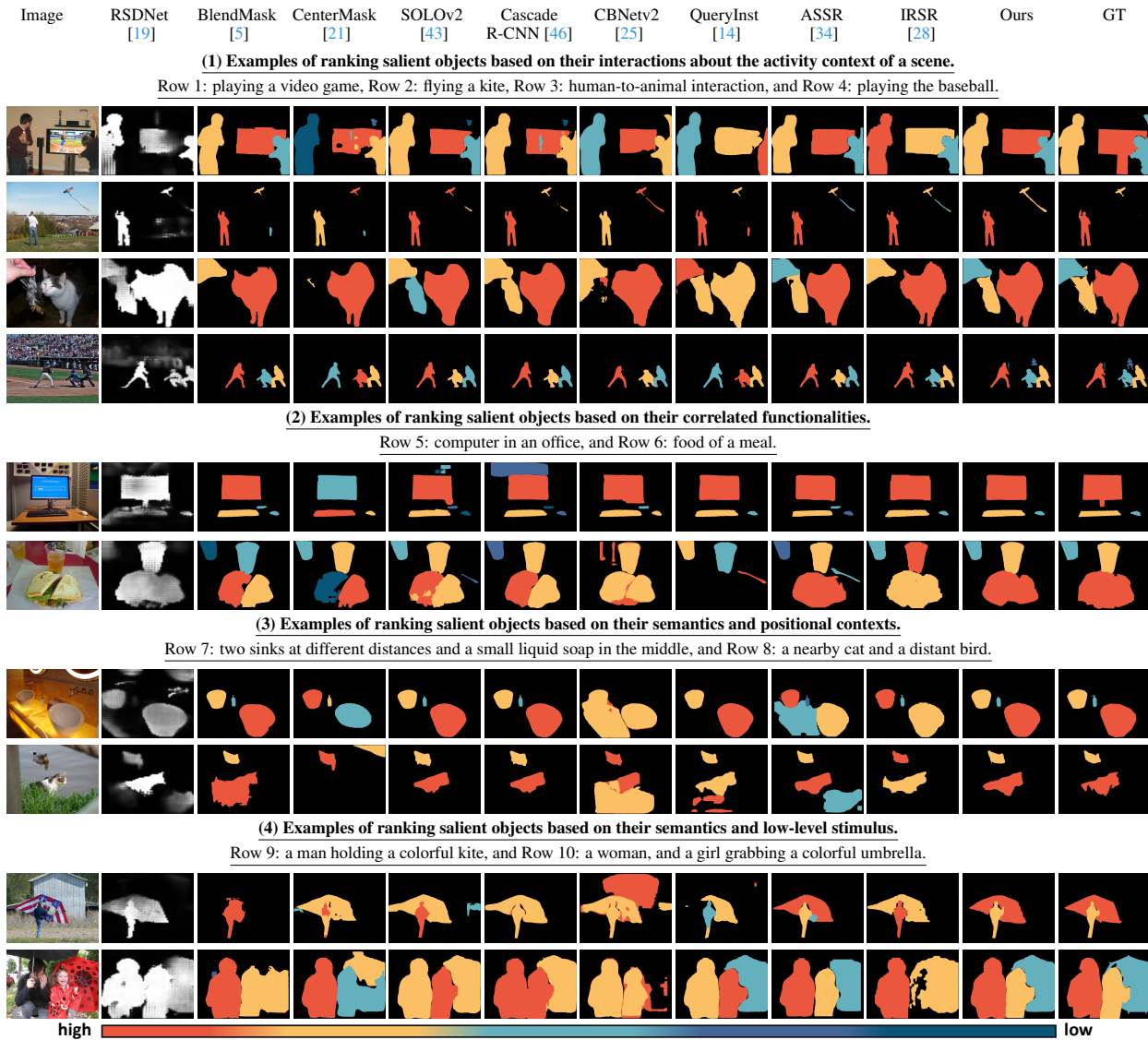
Figure 5. Qualitative comparison of our method with existing saliency ranking approaches and baseline methods. Above each image group, we summarize the scenarios of the images used for the test.
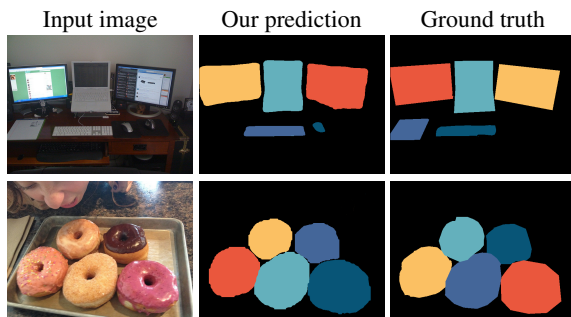

Figure 6. Our method may fail to infer accurate saliency ranks for objects with a similar functionality and interaction with the global contexts.

screens interact with the global context symmetrically. In the second row, these donuts are all visually attractive. We suspect that in these scenarios, even humans may not have a consistent ranking. In the future, we would like to delve deeper into the human visual system for more cues to help with the saliency ranking task.

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009. 2

[2] John R Anderson, Daniel Bothell, Michael D Byrne, Scott Douglass, Christian Lebiere, and Yulin Qin. An integrated theory of the mind. *Psychological review*, 2004. 2, 3

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *CVPR*, 2018. 6, 7

[4] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. *arXiv:2103.14899*, 2021. 3

[5] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yong-ming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *CVPR*, 2020. 6, 7, 8

[6] Jiansheng Chen, Gaocheng Bai, Shaoheng Liang, and Zhengqin Li. Automatic image cropping: A computational complexity study. In *CVPR*, 2016. 1

[7] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *TPAMI*, 2014. 2

[8] Robert Egly, Jon Driver, and Robert D Rafal. Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology: General*, 1994. 2, 3

[9] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, 2019. 1

[10] Lifeng Fan, Wenguan Wang, Siyuan Huang, Xinyu Tang, and Song-Chun Zhu. Understanding human gaze communication by spatio-temporal graph reasoning. In *ICCV*, 2019. 1

[11] Ruochen Fan, Ming-Ming Cheng, Qibin Hou, Tai-Jiang Mu, Jingdong Wang, and Shi-Min Hu. S4net: Single stage salient-instance segmentation. In *CVPR*, 2019. 2, 6, 7

[12] Hao Fang, Daoxin Zhang, Yi Zhang, Minghao Chen, Jiawei Li, Yao Hu, Deng Cai, and Xiaofei He. Salient object ranking with position-preserved attention. In *ICCV*, 2021. 2, 6, 7

[13] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *ICCV*, 2021. 3

[14] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Queryinst: Parallelly supervised mask query for instance segmentation. In *ICCV*, 2021. 5, 6, 7, 8

[15] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 3, 6, 7

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

[17] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 2015. 1

[18] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 3

[19] Md Amirul Islam, Mahmoud Kalash, and Neil DB Bruce. Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects. In *CVPR*, 2018. 1, 2, 5, 6, 7, 8

[20] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 5

[21] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *CVPR*, 2020. 6, 7, 8

[22] Guanbin Li, Yuan Xie, Liang Lin, and Yizhou Yu. Instance-level salient object segmentation. In *CVPR*, 2017. 2

[23] Peihua Li, Jiangtao Xie, Qilong Wang, and Wangmeng Zuo. Is second-order information helpful for large-scale visual recognition? In *CVPR*, 2017. 3

[24] Wei-Hong Li, Fa-Ting Hong, and Wei-Shi Zheng. Learning to learn relation for important people detection in still images. In *CVPR*, 2019. 1

[25] Tingting Liang, Xiaojie Chu, Yudong Liu, Yongtao Wang, Zhi Tang, Wei Chu, Jingdong Chen, and Haibin Ling. Cb-netv2: A composite backbone network architecture for object detection. *arXiv:2107.00420*, 2021. 6, 7, 8

[26] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 3

[27] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018. 2

[28] Nian Liu, Long Li, Wangbo Zhao, Junwei Han, and Ling Shao. Instance-level relative saliency ranking with graph reasoning. *TPAMI*, 2021. 1, 2, 5, 6, 7, 8

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 5, 7

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 6

[31] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012. 2

[32] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 6, 7

[33] Adria Recasens, Carl Vondrick, Aditya Khosla, and Antonio Torralba. Following gaze in video. In *ICCV*, 2017. 1

[34] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Inferring attention shift ranks of objects for image saliency. In *CVPR*, 2020. 1, 2, 5, 6, 7, 8

[35] Avishek Siris, Jianbo Jiao, Gary KL Tam, Xianghua Xie, and Rynson WH Lau. Scene context-aware salient object detection. In *ICCV*, 2021. 2

[36] Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *CVPR*, 2021. 3, 5, 6

[37] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson W. H. Lau. Night-time scene parsing with a large real dataset. *IEEE TIP*, 2021. 1

[38] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Weakly-supervised salient instance detection. In *BMVC*, 2020. 1, 3

[39] Xin Tian, Ke Xu, Xin Yang, Baocai Yin, and Rynson WH Lau. Learning to detect instance-level salient objects using complementary image labels. *IJCV*, 2022. 3

[40] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszko-reit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4, 5

[41] Qilong Wang, Jiangtao Xie, Wangmeng Zuo, Lei Zhang, and Peihua Li. Deep cnns meet global covariance pooling: Better representation and generalization. *TPAMI*, 2020. 3

[42] Wenguan Wang, Qiuxia Lai, Huazhu Fu, Jianbing Shen, Haibin Ling, and Ruigang Yang. Salient object detection in the deep learning era: An in-depth survey. *TPAMI*, 2021. 2

[43] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *ECCV*, 2020. 6, 7, 8

[44] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, 2020. 2

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018. 3, 6

[46] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019. 6, 7, 8

[47] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, 2019. 6, 7

[48] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via graph-based manifold ranking. In *CVPR*, 2013. 2

[49] Xin Yang, Ke Xu, Shaozhe Chen, Shengfeng He, Baocai Yin Yin, and Rynson Lau. Active matting. In *NeurIPS*, 2018. 1

[50] Jianming Zhang, Stan Sclaroff, Zhe Lin, Xiaohui Shen, Brian Price, and Radomir Mech. Unconstrained salient object detection via proposal subset optimization. In *CVPR*, 2016. 2

[51] Lu Zhang, Ju Dai, Huchuan Lu, You He, and Gang Wang. A bi-directional message passing model for salient object detection. In *CVPR*, 2018. 2

[52] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018. 2

[53] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019. 2

[54] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, 2016. 3