Saliency Ranking using Deep Learning

by

Mahmoud Kalash

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

Master of Science

Department of Computer Science
University of Manitoba
Winnipeg, MB, Canada

Thesis advisor

Dr. Neil Bruce

Author

Mahmoud Kalash

Saliency Ranking using Deep Learning

Abstract

Salient object detection is a problem that has been considered in detail and many solutions proposed. In this thesis, we argue that work to date has addressed a problem that is relatively ill-posed. Specifically, there is not universal agreement about what constitutes a salient object when multiple observers are queried which implies a relative rank exists on salient objects. In this thesis, we solve this more general problem that considers relative rank. A novel deep learning solution is proposed based on a hierarchical representation of relative saliency and stage-wise refinement to address both of the saliency ranking and subitizing tasks. We also present methods for deriving suitable ranked salient object instances to generate a large scale dataset for saliency ranking, along with metrics suitable to measuring success in a relative object saliency landscape. Our approach exceeds performance of any prior work across all metrics considered (both traditional and newly proposed).

Contents

List of Figures

List of Tables

Acknowledgments

First and foremost, I praise God, the Almighty, merciful and passionate for providing me this opportunity, granting me the capability to proceed successfully, and honoring me in having so many amazing people leading and pushing me through in the accomplishment of this thesis.

I would like to take this unique opportunity to express my heartfelt thanks and sincere gratitude to my supervisor Dr. Neil Bruce for giving me the opportunity to work under his supervision and for his continuous support and guidance throughout my Master's program. This accomplishment would not have been possible without him. Dr. Bruce has truly been an inspiration for me and I have learned quite a lot from him. His thoughtful reviews and suggestions during my research have made my expedition outstanding. Words are insufficient to appreciate him. I especially thank him for the exemplary guidance, monitoring, constant encouragement, understanding, kindness, care and empathy.

I would also choose this moment to thank my committee members, Dr. Yang Wang, Dr. Noman Mohammed and Dr. Mohammad Jafari Jozani, who have provided valuable suggestions and constructive feedback in perfecting my thesis.

I wish to express my gratitude for the financial support provided by Dr. Bruce the Department of Computer Science at the University of Manitoba, and the Government of Manitoba which was essential for the completion of this degree.

I would like to take this opportunity to express my gratitude to my wonderful lab mates in Computer Vision Lab for their time, support, suggestions and genuine kindness that helped sustain a positive atmosphere in the lab. I would also like to thank all the support staff in the department for their assistance.

Also, this report, and indeed the completion of this Master of Science, would not have been possible without the love, eternal patience and support of the most important people in my life my wonderful parents and family members.

Finally, but by no means the least, I would like to recognize the continuous motivation and support from my acquaintances in Winnipeg and at the University of Manitoba – an incredible group of people who made my journey an enjoyable one.

This thesis is dedicated to my parents
for their love, endless support
and encouragement.

Publications

Some of the ideas, materials and figures in this thesis have appeared previously in the following publication:

1. M. A. Islam*, **M. Kalash*** and N. Bruce. Revisiting Salient Object Detection: Simultaneous Detection, Ranking, and Subitizing of Multiple Salient Objects. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, USA, June 2018.

* indicates equal contribution

Glossary

Convolution Layer Convolution layers apply a convolution operations (linear filtering) to extract edge, color, and shape information from the input image. The convolution layer operates on a subregion of the input image and produces a single value for each subregion. 14

Convolutional Neural Networks (CNNs) A feed-forward neural network that extracts features from local regions of an input image. CNNs are currently considered the state-of-the-art neural network architecture for many computer vision tasks. Most CNNs contain a combination of convolutional layers, pooling layers and fully connected layers. 7

Fixation map This term is mostly used when talking about eye tracking. If an observer is wearing an eye tracker with a sampling rate of 60 Hz and freely viewing an image, 60 individual gaze locations can be collected every second. If gaze points are spatially and temporally close, they form a fixation which signifies a period where the observer's gaze was locked at an object. Fixations are usually visualized in the form of a binary image or as a heatmap after applying a gaussian blur, which is usually referred to as a fixation map. v, 4

Instance-wise map Provides a labelling for all pixels that belong to the same object in the image. Instances that belongs to the same semantic category have distinct labels. v, 4

Object Region Proposals Object Region Proposals are regions of image that are

presumed to contain significant objects based on a determination made by some algorithm [2]. 7

Principal component analysis (PCA) PCA is a dimensionality-reduction technique used to reduce a set of statistics (variables) to a smaller set while keeping as much variance present in the original data as possible. The first principal component captures the direction of highest variability in the data, and the successive components capture as much of the remaining variability as possible and so forth. PCA is usually used to make data easy to explore and visualize as it emphasizes variation and reveals strong patterns in a dataset. 31

Salient object detection/segmentation A problem domain where the objective is to select an object or objects in an image that are important, striking, or draw attention. 1

Salient object subitizing Research indicates that people are capable of identifying up to 4 items by a simple glance at a given image [3]. This *fast counting* ability is referred to as Salient Object Subitizing [4]. Therefore, the goal in this problem domain is to count the number of salient objects disregarding their relative importance. 2

Semantic Segmentation The objective in this problem domain is to assign a semantic category (e.g person, dog, ... etc) to each pixel in an image. 14

Spatial Pooling Spatial pooling reduces the spatial dimensions of the data flowing through the network to decrease the amount of parameters and computational

cost. Different pooling techniques are usually adopted in convolutional neural architectures. *Average pooling* takes the average of several values within a region (e.g. 2x2 non-overlapping regions of input data) whereas *max pooling* keeps the maximum value and discards the remaining values. Pooling layers are typically inserted between successive convolutional layers and max pooling is the most common type of downsampling used in convolutional networks. 14

Stochastic Gradient Descent (SGD) SGD is a gradient-based optimization technique that updates network parameters during the training phase. Mini-batch stochastic gradient descent is a very common approach to improve computational efficiency. In this approach, the gradients of the neural network are calculated using backpropagation and network parameter updates are performed based on a batch instead of a single examplar leading to faster convergence. 22

Superpixel A Superpixel is a group of connected pixels that have similar colors and brightness. This concept was first introduced in [5]. 7

Chapter 1

Introduction

The problem of Salient object detection/segmentation [6; 7; 8; 9] has been well studied, and much progress has been made. The objective in this problem domain is to select an object or objects in an image that are important, striking, stand-out or draw attention. The majority of work in salient object detection considers either a single salient object [10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21] or multiple salient objects [22; 23; 24], but does not consider that what is salient may vary from one person to another, and certain objects may be met with more universal agreement concerning their importance. In this work, we bring to light a consideration that has long been neglected in this domain i.e individual observers may have differences in opinion about what is salient, and moreover, the definition of a salient object is relatively equivocal. This implies that while one or more object(s) may be salient, there may be more agreement for certain object than others. With respect to a problem definition, salient object detection can be extended to a problem of ranking salient objects.

There is a paucity of data that includes salient objects that are hand-segmented by multiple observers. It is important to note that any labels provided by a small number of observers (including one) does not allow for discerning the relative importance of objects. Implicit assignment of relative salience based on gaze data [25] also presents difficulties, given a different cognitive process than a calculated decision that involves manual labeling [26]. Moreover, gaze data is relatively challenging to interpret given factors such as centre bias, visuomotor constraints, and other latent factors [27; 28]. To overcome some of these shortcomings, we have re-purposed the PASCAL-S dataset [9] via further processing to provide a set of data that overcomes some of the limitations of traditional efforts.

Therefore, in this thesis we consider the problem of salient object detection more broadly. This includes detection of all salient regions in an image, and accounting for inter-observer variability by assigning confidence to different salient regions. We augment the PASCAL-S dataset via further processing to provide ground truth in a form that accounts for relative salience. Success is measured against other algorithms based on the rank order of salient objects relative to ground truth orderings in addition to traditional metrics. Recent efforts also consider the problem of *Salient object subitizing*. It is our contention that this determination should be possible by a model that provides detection of salient objects (see Fig. 1.1). Therefore, we also allow our network to subitize.

Although the PASCAL-S dataset [9] may be made suitable for addressing the problem of ranking salient objects; however, in order to train deep learning models a significant number of examples are needed; this is also true of evaluation of algorithms

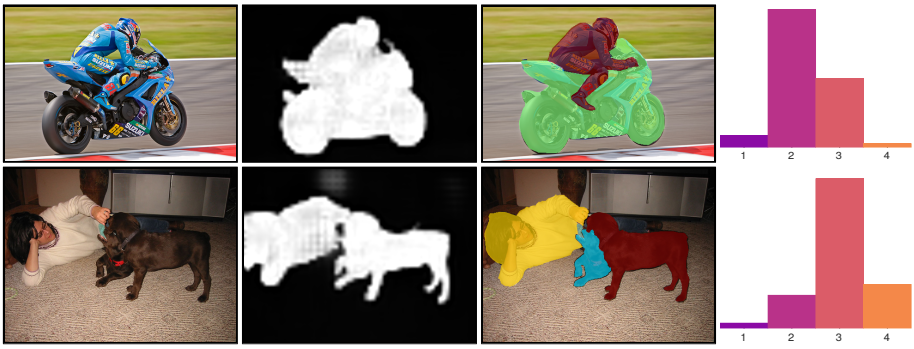


Figure 1.1: We present a solution in the form of a deep neural network to detect salient objects, consider the relative ranking of salience of these objects, and predict the total number of salient objects. Left to right: input image, detected salient regions, rank order of salient objects, confidence score for salient object count (subitizing). Colors indicate the rank order of different salient object instances.

that seek to assign a relative rank to salient objects.

Therefore, we extend our work in this thesis to present a dataset for saliency ranking, and associated benchmarks, based on images from the MS-COCO dataset [29]. This is accomplished by combining existing measurements diagnostic of human attention [30] with existing object annotations. This process is more challenging and nuanced than one might initially expect; MS-COCO labels only cover specific object categories, vary significantly in the precision with which objects are segmented, and include many examples that are over-segmented or under-segmented.

With that said, we propose a set of methods that prune an initial set of 15k labeled images based on a careful choice of formal criteria for inclusion/rejection. Images or labels that remain are assigned a relative ranking based on a simulated gaze tracking process [30] (see Fig. 1.2). While there are significant differences between manual choice of salient objects and simulated gaze data, we demonstrate how rank values based on the latter can be treated to produce rankings that approximate the former.

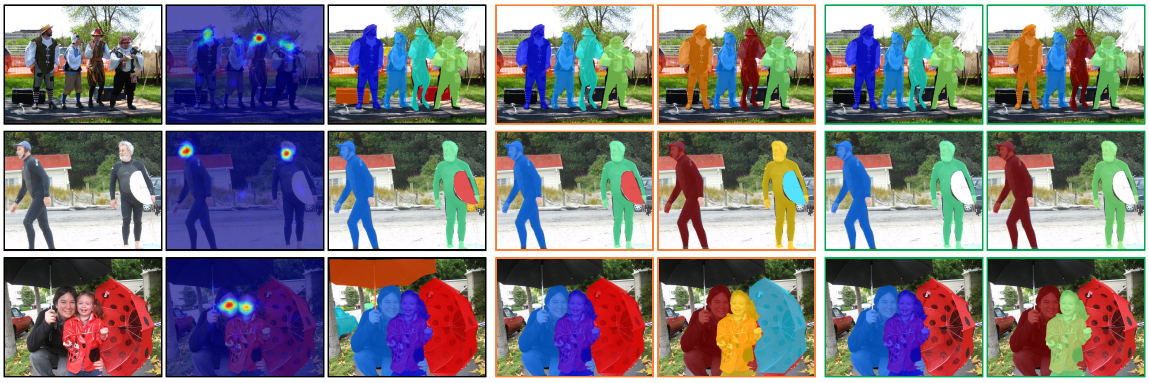


Figure 1.2: An illustration of the COCO-SalRank dataset. Our dataset provides salient object instances and their relative rank order (relative salience). Due to the large number of instances present in some images, an instance pruning process assigns a rank only to instances that receive a sufficiently high degree of attention. We provide two different versions of proposed ranking dataset in the form of a noisy and cleaned version. Left to right: input image, Fixation map, original Instance-wise map, pruned map for version I, rank order of salient objects, pruned map for version II, rank order of salient objects. Colors indicate the rank order of salient object instances with chroma corresponding to a numeric scale.

This process is validated in comparing rank-order assignments based on manual selection on an alternative dataset using the same criteria. Moreover, we demonstrate that training of models on the dataset we provide produces more capable models than those trained on PASCAL-S [9], even in the case that training uses no images from PASCAL-S. It has been noted that in training deep learning models, there is considerable robustness to even large amounts of label noise [31]. With that said, our experimentation shows only a very small deviation in algorithmic assignment of rank when compared against click-based ground truth, and also presents a general end-to-end approach for generating saliency ranking data suitable for crowdsourcing.

1.1 Contributions

We summarize our main contributions as follows:

- We generalize the problem of salient object detection to salient object ranking which includes inter-observer variability of saliency and considers relative rank of salient objects.
- We present a novel deep learning solution that is based on a hierarchical representation of relative saliency and stage-wise stack refinement mechanism. The proposed model provides predictions of salient objects according to the traditional form of this problem, and multiple salient object detection and relative ranking. We also show that the problem of salient object subitizing can be addressed with the same network.
- We introduce a novel set of methods that make use of existing gaze or related data paired with object annotations to generate a large scale benchmark dataset for saliency ranking. We also propose metrics suitable to measuring success in a relative object saliency landscape.
- We provide new state-of-the-art baseline scores for the saliency ranking problem on PASCAL-S and the proposed dataset, while also providing a corpus of data and code to the community that provides significant value for training and evaluation for a relatively nascent problem domain.
- We perform extensive experiments and analysis that demonstrate the effectiveness of our proposed model and dataset. Our results show state-of-the-art performance for all metrics considered (both traditional and newly proposed).

1.2 Thesis Organization

The remainder of the thesis is organized as follows. In Chapter 2, we briefly discuss previous work related to this thesis. In Chapter 3, we present our novel deep learning model to solve the more general problem that considers relative rank using the augmented ground truth of the PASCAL-S dataset, and we propose suitable metrics for this realm of study. We also re-purpose the proposed model to tackle the problem of salient object subitizing. Finally, we examine the performance of our approach across both traditional an newly proposed metrics. In Chapter 4, we introduce a novel set of methods that utilizes existing gaze data along with object annotations to generate large rank ordered data for salient objects and we present new metrics suitable to measuring algorithm performance. We re-orientate some sate-of-the-art algorithms to address salient object ranking and we provide a performance comparison on the newly proposed data to establish initial baselines. In Chapter 5, we conclude this thesis and discuss possible future directions.

Chapter 2

Related Work

2.1 Salient Object Detection:

Convolutional Neural Networks (CNNs) have raised the bar in performance for many problems in computer vision including salient object detection. CNN based models are able to extract more representative and complex features than hand crafted features used in less contemporary work [32; 6; 33] which has promoted widespread adoption.

Some CNN based methods [16; 17; 18; 34; 20; 21] exploit Superpixel and Object Region Proposals to achieve accurate salient object detection . Such methods follow a multi-branch architecture where a CNN is used to extract semantic information across different levels of abstraction to generate an initial saliency prediction. Subsequently, new branches are added to obtain superpixels or object region proposals, which are used to improve precision of the predictions.

As an alternative to superpixels and object region proposals, other methods [35;

13; 10] predict saliency per-pixel by aggregating multi-level features. Luo et al. [35] integrate local and global features through a CNN that is structured as a multi-resolution grid. Hou et al. [13] implement stage-wise short connections between shallow and deeper feature maps for more precise detection and inferred the final saliency map considering only middle layer features. Zhang et al. [10] combine multi-level features as cues to generate and recursively fine-tune multi-resolution saliency maps which are refined by boundary preserving refinement blocks and then fused to produce final predictions.

Other methods [19; 14; 11] use an end-to-end encoder-decoder architecture that produces an initial coarse saliency map and then refines it stage-by-stage to provide better localization of salient objects. Liu and Han [19] propose a network that combines local contextual information step-by-step with a coarse saliency map. Wang et al. [14] propose a recurrent fully convolutional network for saliency detection that includes priors to correct initial saliency detection errors. Zhang et al. [11] incorporate a reformulated dropout after specific convolutional layers to quantify uncertainty in the convolutional features, and a new upsampling method to reduce artifacts of deconvolution which results in a better boundary for salient object detection.

In contrast to the above described approaches, we achieve spatial precision through stage-wise refinement by applying novel mechanisms to control information flow through the network while also importantly including a *stacking* strategy that implicitly carries the information necessary to determine relative saliency.

2.2 Salient Object Subitizing:

Recent work [36; 12] has also addressed the problem of subitizing salient objects in images. This task involves counting the number of salient objects, regardless of their importance or semantic category. The first salient object subitizing network proposed in [36] applies a feed-forward CNN to treat the problem as a classification task. He et al. [12] combine the subitizing task with detection by exploring the interaction between numeric and spatial representations. Our proposal provides a specific determination of the number of salient objects, recognizes variability in this number, and also provides output as a distribution that reflects this variability.

2.3 Universal Saliency Detection Benchmarks

There has long been growing interest in cognitive science disciplines to understand where people look and direct their gaze while interacting with complex indoor or outdoor scenes. This can be examined by direct measurement of gaze, or by manual selection of important regions, with the latter presenting a situation less prone to low-level biases [27]. Predictive models that address both of these processes have been defined as visual saliency detection/segmentation. However, in the selection task (as with free-viewing gaze), there is no universal agreement on what constitutes a salient object when opinions are elicited from multiple observers. Until very recently [1], the literature has failed to acknowledge this nuance of salient object detection. A large number of proposals have been made on methods for predicting salient targets; these studies focus on a binary prediction of a *universal saliency* label that considers only

the most salient object [10; 11; 12; 13; 14; 15; 16; 17; 18; 19; 20; 21]. Other studies consider multiple salient objects [22; 23; 24] but do not consider the variability that may exist across humans in deciding what is salient. In deference to the apparently deeper problem definition that may be attached to salient object detection, recent work [1] extended the traditional problem to the salient object ranking. However, given that this implies a distinct problem domain, there are limitations on what data is currently available for training and/or testing models.

Several saliency detection [6; 7; 8; 9; 30] or eye tracking [30] datasets have been curated and shared with the community to promote saliency research. The majority of these datasets share common features in providing ground-truth annotation as a binary mask (background vs salient object) assigned to each image that is based on one observer, or subject to a threshold when a few opinions are present. Currently, only the Pascal-S is widely available for addressing the saliency ranking problem (with suitable post-processing) as it provides ground-truth corresponding to multiple observers' agreement across 12 observers. In order to progress further in addressing this problem, there is a dire need for larger-scale datasets to provide a stronger based set of data for training and evaluating models. Pascal-S provides ground-truth that implicitly captures relative salience, but only for ≈ 1000 examples. The current emphasis on, and success of deep learning architectures generally requires large-scale datasets that implies an even greater need for alternative datasets and suitable benchmark results and metrics. Moreover, the array of current saliency detection datasets is heavily entrenched in the traditional problem definition and associated experimental analysis. We therefore seek to allow research efforts on this problem to rapidly

progress, in ensuring availability of large-scale data that serves to improve saliency ranking performance among models, and to allow for immediate adoption of solutions that explore new research directions in assessing *relative saliency* as well as complementing existing datasets and the historical legacy of work in this area.

Obtaining ground-truth by manual labeling is crucial for computer vision applications. Amazon Mechanical Turk (AMT) has been used extensively for labeling large-scale datasets in distributing the labeling task among many human annotators. The crowdsourcing approach has been directed to different tasks that derive desired output labels from human annotation (e.g. assigning a category label, segmenting salient objects, providing bounding boxes). In this thesis, we propose a novel approach to provide a benchmark dataset which involves a hybrid of algorithmic coalescence of multiple disparate common labels from common datasets (MSCOCO [29]), with optional refinement by a human as a secondary stage. This approach is validated by comparing rank-order assignments with the smaller extant alternative dataset that has exact labels.

Very recently, the saliency ranking problem was proposed along with an effective deep learning solution [1]. This presents a deeper problem than traditional salient object detection. In this work, we specifically emphasize promoting rapid progress in this domain by providing new state-of-the-art baseline scores as well as a new benchmark dataset. This is also accompanied by in depth analysis of datasets, nuances of saliency ranking, and considerations going forward. We expect that this presents another significant contribution that will provide stronger capabilities for models and guidance on model success advancing progress in the field related to this problem.

Chapter 3

Revisiting Salient Object Detection: Simultaneous Detection, Ranking, and Subitizing of Multiple Salient Objects

In this chapter, we present our proposed network architecture for simultaneously detecting, ranking, and subitizing multiple salient objects which introduces a stack refinement mechanism to account for relative saliency. We also discuss the proposed processing method of the PASCAL-S dataset to represent relative saliency in the ground truth. Note that we are interested in this dataset because it is unique in having multiple, explicitly tagged salient regions provided by a reasonable sample size of observers. Finally, we demonstrate the effectiveness of our architecture by performing comprehensive experiments and comparison with state-of-the-art methods.

3.1 Proposed Network Architecture

We propose a new end-to-end framework for solving the problem of detecting multiple salient objects and ranking the objects according to their degree of salience. Our proposed salient object detection network is inspired by the success of convolution-deconvolution pipelines [37; 19; 38] that include a feed-forward network for initial coarse-level prediction. Then, we provide a stage-wise refinement mechanism over which predictions of finer structures are gradually restored. Fig. 3.1 shows the overall architecture of our proposed network. The encoder stage serves as a feature extractor that transforms the input image to a rich feature representation, while the refinement stages attempt to recover lost contextual information to yield accurate predictions and ranking.

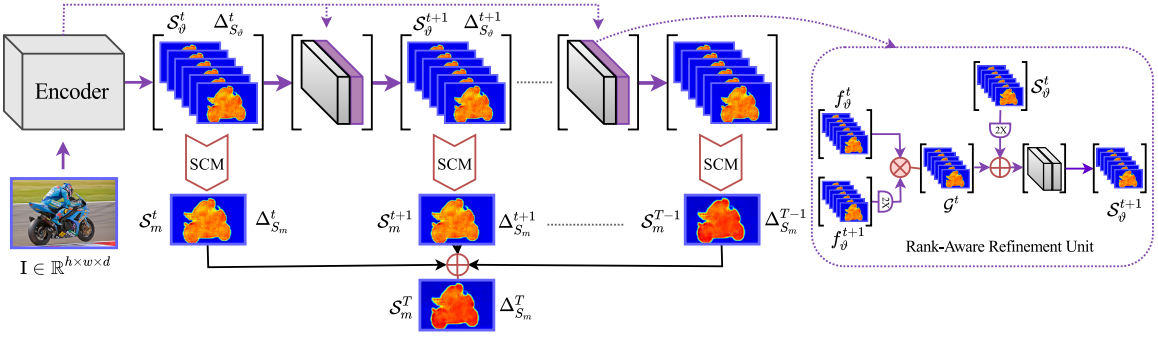


Figure 3.1: Illustration of our proposed network architecture. In the encoder network, the input image is processed with a feed-forward encoder to generate a coarse nested relative salience stack ($\mathcal{S}_{\vartheta}^{t}$). We append a Stacked Convolutional Module (SCM) on top of $\mathcal{S}_{\vartheta}^{t}$ to obtain a coarse saliency map $\mathcal{S}_{m}^{t}$. Then, a stage-wise refinement network, comprised of rank-aware refinement units (dotted box in the figure), successively refines each preceding NRSS ($\mathcal{S}_{\vartheta}^{t}$) and produces a refined NRSS ($\mathcal{S}_{\vartheta}^{t+1}$). A fusion layer combines predictions from all stages to generate the final saliency map ($\mathcal{S}_{m}^{T}$). We provide supervision ($\Delta_{S_{\vartheta}}^{t}$, $\Delta_{S_{m}}^{t}$) at the outputs ($\mathcal{S}_{\vartheta}^{t}$, $\mathcal{S}_{m}^{t}$) of each refinement stage. The architecture based on iterative refinement of a stacked representation is capable of effectively detecting multiple salient objects.

We begin by describing how the initial coarse saliency map is generated in section 3.1.1. This is followed by a detailed description of the stage-wise refinement network, and multi-stage saliency map fusion in sections 3.1.2 and section 3.1.3 respectively.

3.1.1 Feed-forward Network for Coarse Prediction

Recent feed-forward deep learning models applied to high-level vision tasks (e.g. image classification [39; 40], object detection [41]) employ a cascade comprised of repeated convolution stages followed by Spatial Pooling. Down-sampling by pooling allows the model to achieve a highly detailed semantic feature representation with relatively poor spatial resolution at the deepest stage of encoding, also marked by spatial coverage of filters that is much larger in extent. The loss of spatial resolution is not problematic for recognition problems; however, pixel-wise labeling tasks (e.g. Semantic Segmentation, salient object detection) require pixel-precise information to produce accurate predictions. Thus, we choose Resnet-101 [39] as our encoder network (fundamental building block) due to its superior performance in classification and segmentation tasks. Following prior works on pixel-wise labeling [42; 38], we use the dilated ResNet-101 [42] to balance the semantic context and fine details, resulting in an output feature map reduced by a factor of 8. More specifically, given an input image $I \in \mathbb{R}^{h \times w \times d}$, our encoder network produces a feature map of size $\lfloor \frac{h}{8}, \frac{w}{8} \rfloor$. To augment the backbone of the encoder network with a top-down refinement network, we first attach one extra Convolution Layer with 3×3 kernel and 12 channels to obtain a *Nested Relative Salience Stack* (NRSS). Then, we append a *Stacked Convolutional*

Module (SCM) to compute the coarse level saliency score for each pixel. It is worth noting that our encoder network is flexible enough to be replaced with any other baseline network e.g. VGG-16 [40], DenseNet-101 [43]. Moreover, we utilize atrous pyramid pooling [42] to gather more global contextual information. The described operations can be expressed as

$$\mathcal{S}_{\vartheta}^{t} = \mathcal{C}_{3\times3}(\mathcal{F}_{s}(I;\mathcal{W});\Theta), \;\; \mathcal{S}_{m}^{t} = \xi(\mathcal{S}_{\vartheta}^{t}) \tag{3.1}$$

where I is the input image and $(\mathcal{W}, \Theta)$ denote the parameters of the convolution $\mathcal{C}$. $\mathcal{S}_{\vartheta}^{t}$ is the coarse level NRSS for stage t that encapsulates different degrees of saliency for each pixel (akin to a prediction of the proportion of observers that might agree an object is salient), $\mathcal{S}_{m}^{t}$ refers to the coarse level saliency map, and ξ refers to SCM. $\mathcal{F}_{s}(.)$ denotes the output feature map generated by the encoder network. The SCM consists of three convolutional layers for generating the desired saliency map. The initial convolutional layer has 6 channels with a 3×3 kernel, followed by two convolutional layers having 3 channels with 3×3 kernel and one channel with 1×1 kernel respectively. Each of the channels in the SCM learns a soft weight for each spatial location of the nested relative salience stack in order to label pixels based on confidence that they belong to a salient object.

3.1.2 Stage-wise Refinement Network

Most existing works [19; 15; 10; 13] that have shown success for salient object detection typically share a common structure of stage-wise decoding to recover per-pixel categorization. Although the deepest stage of an encoder has the richest possible feature representation, relying only on convolution and unpooling at the decoding

stages to recover lost information may degrade the quality of predictions [38]. So, the spatial resolution that is lost at the deepest layer may be gradually recovered from earlier representations. This intuition appears in proposed refinement based models that include skip connections [44; 38; 10; 13] between encoder and decoder layers. However, how to effectively combine local and global contextual information remains an area deserving further analysis. Inspired by the success of refinement based approaches [44; 45; 38], we propose a multi-stage fusion based refinement network to recover lost contextual information in the decoding stage by combining an initial coarse representation with finer features represented at earlier layers. The refinement network is comprised of successive stages of rank-aware refinement units that attempt to recover missing spatial details in each stage of refinement and also preserve the relative rank order of salient objects. Each stage refinement unit takes the preceding NRSS with earlier finer scale representations as inputs and carries out a sequence of operations to generate a refined NRSS that contributes to obtain a refined saliency map. Note that refining the hierarchical NRSS implies that the refinement unit is leveraging the degree of agreement at different levels of SCMs to iteratively improve confidence in relative rank and overall saliency. As a final stage, refined saliency maps generated by the SCMs are fused to obtain the overall saliency map.

Rank-Aware Refinement Unit

Previous saliency detection networks [15; 19] proposed refinement across different levels by directly integrating representations from earlier features. Following [38], we integrate gate units in our rank-aware refinement unit that control the information

passed forward to filter out the ambiguity relating to figure-ground and salient objects. The initial NRSS ($\mathcal{S}_\vartheta^t$) generated by the feed-forward encoder provides input for the first refinement unit. Note that one can interpret $\mathcal{S}_\vartheta^t$ as the predicted saliency map in the decoding process, but our model forces the channel dimension to be the same as the number of participants involved in labeling salient objects. The refinement unit takes the gated feature map $\mathcal{G}^t$ generated by the gate unit [38] as a second input. As suggested by [38], we obtain $\mathcal{G}^t$ by combining two consecutive feature maps (f_ϑ^t and f_ϑ^{t+1}) from the encoder network (see dotted box in Fig. 3.1). We first upsample the preceding $\mathcal{S}_\vartheta^t$ to double its size. A transformation function $\mathcal{T}_f$ comprised of a sequence of operations is applied on upsampled $\mathcal{S}_\vartheta^t$ and $\mathcal{G}^t$ to obtain the refined NRSS ($\mathcal{S}_\vartheta^{t+1}$). We then append the *SCM* module on top of $\mathcal{S}_\vartheta^{t+1}$ to generate the refined saliency map $\mathcal{S}_m^{t+1}$. Finally, the predicted $\mathcal{S}_\vartheta^{t+1}$ is fed to the next stage rank-aware refinement unit. Note that, we only forward the NRSS to the next stage, allowing the network to learn contrast between different levels of confidence for salient objects. Unlike other approaches, we apply supervision for both of the refined NRSS and the refined saliency map. The procedure for obtaining the refined NRSS and the refined saliency map for all stages is identical. The described operations may be summarized as follows:

$$\mathcal{S}_\vartheta^{t+1} = w^b * \mathcal{T}_f(\mathcal{G}^t, u(\mathcal{S}_\vartheta^t)),\ S_m^{t+1} = w_s^b * \xi(\mathcal{S}_\vartheta^{t+1}) \tag{3.2}$$

where u represents the upsample operation; w^b and w_s^b denotes the parameter for the transformation function $\mathcal{T}_f$ and SCM (ξ in the equation) respectively. Note that t refers to particular stage of the refinement process.

3.1.3 Multi-Stage Saliency Map Fusion

Predicted saliency maps at different stages of the refinement units are capable of finding the location of salient regions with increasingly sharper boundaries. Since all the rank-aware refinement units are stacked together on top of each other, the network allows each stage to learn specific features that are of value in the refinement process. These phenomena motivate us to combine different level SCMs predictions, since the internal connection between them is not explicitly present in the network structure. To facilitate interaction, we add a fusion layer at the end of network that concatenates the predicted saliency maps of different stages, resulting in a fused feature map $\mathcal{S}_m^{\hat{f}}$. Then, we apply a 1×1 convolution layer Υ to produce the final predicted saliency map $\mathcal{S}_m^T$ of our network. Note that our network has T predictions, including one fused prediction and T-1 stage-wise predictions. We can write the operations as follows:

$$\mathcal{S}_m^{\hat{f}} = \eth(\mathcal{S}_m^t, \mathcal{S}_m^{t+1}, ..., \mathcal{S}_m^{T-1}),\ \mathcal{S}_m^T = w^f * \Upsilon(\mathcal{S}_m^{\hat{f}}) \tag{3.3}$$

where $\eth$ denotes the cross channel concatenation; w^f is the resultant parameter for obtaining the final prediction.

3.1.4 Stacked Representation of Ground-truth

The ground-truth for salient object detection or segmentation contains a set of numbers defining the degree of saliency for each pixel. The traditional way of generating binary masks is by thresholding which implies that there is no notion of relative salience. Since we aim to explicitly model observer agreement, using traditional binary ground-truth masks is unsuitable. To address this problem, we propose

to generate a set of stacked ground-truth maps that corresponds to different levels of saliency (defined by inter-observer agreement). Given a ground-truth saliency map $\mathcal{G}_m$, we obtain a stack $\mathcal{G}_\vartheta$ of N ground-truth maps $(\mathcal{G}_i, \mathcal{G}_{i+1}, ..., \mathcal{G}_N)$ where each map $\mathcal{G}_i$ includes a binary indication that at least i observers judged an object to be salient (represented at a per-pixel level). N is the number of different participants involved in labeling the salient objects. The stacked ground-truth saliency maps $\mathcal{G}_\vartheta$ provides better separation for multiple salient objects (see Eq. (3.4) for illustration) and also naturally acts as the relative rank order that allows the network to learn to focus on degree of salience. It is important to note the nested nature of the stacked ground truth wherein $\mathcal{G}_{i+1} \subseteq \mathcal{G}_i$. This is important conceptually as a representation wherein $\mathcal{G}_i = 1 \iff$ exactly i observers agree, results in zeroed layers in the ground truth stack, and large changes to ground truth based on small differences in degree of agreement.

$$\mathcal{G}_\vartheta = \left[\begin{array}{c}\mathcal{G}_i\\ \end{array}\right] \left[\begin{array}{c}\mathcal{G}_{i+1}\\ \end{array}\right] \left[\begin{array}{c}\mathcal{G}_{i+2}\\ \end{array}\right] \left[\begin{array}{c}...\\ \end{array}\right] \left[\begin{array}{c}\mathcal{G}_{i+6}\\ \end{array}\right] \left[\begin{array}{c}\mathcal{G}_{i+7}\\ \end{array}\right] \left[\begin{array}{c}...\\ \end{array}\right] \left[\begin{array}{c}\mathcal{G}_N\\ \end{array}\right] \tag{3.4}$$

3.1.5 Salient Object Subitizing Network

Previous works [36; 12] treat subitizing as a straight-forward classification task. Similar to our multiple salient object detection network, the subitizing network is also based on ResNet-101 [39] except we remove the last block. We append a fully connected layer at the end to generate confidence scores for each of 0, 1, 2, 3, and 4+ salient objects existing in the input image followed by another fully connected layer leads to generate final confidence scores for each category. The reasoning behind this

is that a single layer allows for accumulation of confidence tied to salience while two layers allows for reasoning about relative salience. We use our pre-trained detection model to train the subitizing network. As a classifier, the subitizing network reduces two cross entropy losses $\ell^1_{sub}(c, n)$ and $\ell^f_{sub}(c_f, n)$ between the number of salient objects n in ground-truth, and the total predicted objects.

A New Dataset for Salient Object Subitizing:

Since salient object subitizing is not a widely addressed problem, a limited number of datasets [36] were created. In order to facilitate the study of this problem in more complex scenarios, we create the subitizing ground-truth for the Pascal-S dataset [9] that provides instance-wise counting as labels. The distribution of the images in Pascal-S dataset with respect to different categories is shown in Table 3.1. It is an evident from the table that, there is a considerable number of images with more than two salient objects but only few images with more than 7. We initially include all instances of salient objects in the labeling process. To reduce imbalance between different categories, we create another ground-truth set where we only categorize the images as 1, 2, 3, and 4+ salient objects.

# Salient Object	1	2	3	4	5	6	7	8+	Total
#Images	300	227	136	72	43	28	18	26	**850**
Distribution (%)	0.35	0.27	0.16	0.08	0.05	0.03	0.02	0.03	**1**

Table 3.1: Count and percentage of images corresponding to different numbers of salient objects in the Pascal-S dataset.

3.1.6 Training the Network

Our proposed network produces a sequence of nested relative salience stacks (NRSS) and saliency maps at each stage of refinement; however, we are principally interested in the final fused saliency map. Each stage of the network is encouraged to repeatedly produce NRSS and a saliency map with increasingly finer details by leveraging preceding NRSS representations. We apply an auxiliary loss at the output of each refinement stage along with an overall master loss at the end of the network. Both of the losses help the optimization process. In more specific terms, let $I \in \mathbb{R}^{h\times w\times 3}$ be a training image with ground-truth saliency map $\mathcal{G}_m \in \mathbb{R}^{h\times w}$. As described in section 3.1.4, we generate a stack of ground-truth saliency maps $\mathcal{G}_\vartheta \in \mathbb{R}^{h\times w\times 12}$. To apply supervision on the NRSS (S^t_ϑ) and saliency map S^t_m, we first down-sample $\mathcal{G}_\vartheta$ and $\mathcal{G}_m$ to the size of S^t_ϑ generated at each stage resulting in $\mathcal{G}^t_\vartheta$ and $\mathcal{G}^t_m$. Then, at each refinement stage we define pixel-wise euclidean loss $\Delta^t_{S_\vartheta}$ and $\Delta^t_{S_m}$ to measure the difference between $(S^t_\vartheta, \mathcal{G}^t_\vartheta)$ and $(S^t_m, \mathcal{G}^t_m)$ respectively. We can summarize these operations as:

$$\begin{aligned}
\Delta^t_{S_\vartheta}(W) &= \frac{1}{2dN}\sum_{i=1}^{d}\sum_{z=1}^{N}(x_i(z)-y_i(z))^2 \\
\Delta^t_{S_m}(W) &= \frac{1}{2d}\sum_{i=1}^{d}(x_i-y_i)^2 \\
L^t_{aux}(W) &= \Delta^t_{S_\vartheta}+\Delta^t_{S_m}
\end{aligned} \tag{3.5}$$

where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^d$ (d denotes the spatial resolution) are the vectorized ground-truth and predicted saliency map. x_i and y_i refer to a particular pixel of S^t_ϑ and G^t_ϑ respectively. W denotes the parameters of whole network and N refers to

total number of ground-truth slices (N =12 in our case). The final loss function of the network combining master and auxiliary losses can be written as:

$$L_{final}(W) = L_{mas}(W) + \sum_{t=1}^{T-1} \lambda_t L_{aux}^{t}(W) \tag{3.6}$$

where $L_{mas}(W)$ refers to the euclidean loss function computed on the final predicted saliency map $\mathcal{S}_m^T$. We set λ_t to 1 for all stages to balance the loss, which remains continuously differentiable. Each stage of prediction contains information related to two predictions, allowing our network to propagate supervised information from deep layers. This also begins with aligning the weights with the initial coarse representation, leading to a coarse-to-fine learning process. The fused prediction generally appears much better than other stage-wise predictions since it contains the aggregated information from all the refinement stages. For saliency inference, we can simply feed an image of arbitrary size to the network and use the fused prediction as our final saliency map.

3.2 Experiments

The core of our model follows a structure based on ResNet-101 [39] with pre-trained weights to initialize the encoder portion and random initialization of any newly added layers that follow the encoder drawn from a standard gaussian distribution. A few variants of the basic architecture are proposed, and the network is implemented using Caffe [46] with both training and testing done on Nvidia Titan X GPU. The network is trained using Stochastic Gradient Descent (SGD) for 20k iter-

ations with momentum of 0.9, weight decay of 0.0005 and the "poly" learning rate policy. Testing uses the full resolution image while training relies on random crops for memory savings. We report numbers for the following variants that are described in what follows:

- **RSDNet:** This network includes dilatedResNet-101 [42] + NRSS + SCM.
- **RSDNet-A:** This network is the same as RSDNet except the ground-truth is scaled by a factor of 1000, encouraging the network to explicitly learn deeper contrast.
- **RSDNet-B:** The structure follows RSDNet except that an atrous pyramid pooling module is added.
- **RSDNet-C:** RSDNet-B + the ground-truth scaling.
- **RSDNet-R:** RSDNet with stage-wise rank-aware refinement units + multi-stage saliency map fusion to generate the final prediction map.

3.2.1 Datasets and Evaluation Metrics

Datasets: The Pascal-S dataset includes 850 natural images with multiple complex objects derived from the PASCAL VOC 2012 validation set [47]. We randomly split the Pascal-S dataset into two subsets (425 for training and 425 for testing). In this dataset, salient object labels are based on an experiment using 12 participants to label salient objects. Virtually all existing approaches for salient object segmentation or detection threshold the ground-truth saliency map to obtain a binary saliency map. This operation seems somewhat arbitrary since the threshold can require consensus among k observers, and the value of k varies from one study to another. This is

one of the most highly used salient object segmentation datasets, but is unique in having multiple explicitly tagged salient regions provided by a reasonable sample size of observers. Since a key objective of this work is to rank salient objects in an image, we use the original ground-truth maps (each pixel having a value corresponding to the number of observers that deemed it to be a salient object) rather than trying to predict a binary output based on an arguably contentious thresholding process.

Evaluation Metrics: For the multiple salient object detection task, we use four different standard metrics to measure performance including precision-recall (PR) curves, F-measure (maximal along the curve), Area under ROC curve (AUC), and mean absolute error (MAE). Since some of these rely on binary decisions, we threshold the ground-truth saliency map based on the number of participants that deem an object salient, resulting in 12 binary ground truth maps. For each binary ground truth map, multiple thresholds of a predicted saliency map allow for calculation of the true positive rate (TPR), false positive rate (FPR), precision and recall, and corresponding ROC and PR curves. Given that methods that predate this work are trained based on varying thresholds and consider a binary ground truth map, scores are reported based on the binary ground truth map that produces the best AUC or F-measure score (and the corresponding curves are shown). Max F-measure, average F-measure and median F-measure are also reported to provide a sense of how performance varies as a function of the threshold chosen. We also report the MAE score i.e. the average pixel-wise difference between the predicted saliency map and the binary ground-truth map that produces the minimum score.

In ordered to evaluate the rank order of salient objects, we introduce the *Salient*

Object Ranking (SOR) metric which is defined as the Spearman's Rank-Order Correlation between the ground truth rank order and the predicted rank order of salient objects. SOR score is normalized to [0 1] for ease of interpretation. Scores are reported based on the average SOR score for each method considering the whole dataset.

3.2.2 Performance Comparison with State-of-the-art

The problem of evaluating salient detection models is challenging in itself which has contributed to differences among benchmarks that are used. In light of these considerations, the specific evaluation we have applied to all the methods aims to remove any advantages of one algorithm over another. We compare our proposed method with recent state-of-the-art approaches, including Amulet [10], UCF [11], DSS [13], NLDF [35], DHSNet [19], MDF [21], ELD [18], MTDS [34], MC [20], HS [6], HDCT [33], DSR [32], and DRFI [48]. For fair comparison, we build the evaluation code based on the publicly available code provided in [49] and we use saliency maps provided by authors of models compared against, or by running their pre-trained models with recommended parameter settings.

Quantitative Evaluation: Table 3.2 shows the performance score of all the variants of our model, and other recent methods on salient object detection. It is evident that, RSDNet-R outperforms other recent approaches for all evaluation metrics, which establishes the effectiveness of our proposed hierarchical nested relative salience stack. From the results we have few fundamental observations: (1) Our network improves the max F-measure by a considerable margin on the Pascal-S dataset which indicates that our model is general enough that it achieves higher precision with higher recall

*	AUC	max-F_m	med-F_m	avg-F_m	MAE	**SOR**
DRFI [48]	0.887	0.716	0.583	0.504	0.216	0.726
DSR [32]	0.871	0.696	0.628	0.583	0.186	0.728
HDCT [33]	0.809	0.654	0.567	0.523	0.214	0.645
HS [6]	0.837	0.702	0.634	0.596	0.263	0.714
MC [20]	0.870	0.717	0.616	0.573	0.216	0.732
MTDS [34]	0.941	0.805	0.731	0.664	0.176	0.782
ELD [18]	0.916	0.789	0.784	0.774	0.123	0.792
MDF [21]	0.892	0.787	0.746	0.730	0.138	0.768
DHSNet [19]	0.927	0.837	0.833	0.822	**0.092**	0.781
NLDF [35]	0.933	0.846	0.843	**0.836**	0.099	0.783
DSS [13]	0.918	0.841	0.838	0.830	0.099	0.770
AMULET [10]	0.957	0.865	**0.854**	**0.841**	0.097	0.788
UCF [11]	0.959	0.858	0.840	0.813	0.123	0.792
RSDNet	**0.972**	0.873	**0.854**	0.834	**0.091**	0.825
RSDNet-**A**	**0.973**	**0.874**	0.851	0.796	0.103	0.838
RSDNet-**B**	0.969	**0.877**	**0.857**	0.831	0.100	**0.840**
RSDNet-**C**	**0.972**	**0.874**	0.850	0.795	0.110	**0.848**
RSDNet-**R**	**0.971**	**0.880**	**0.861**	**0.837**	**0.090**	**0.852**

Table 3.2: Quantitative comparison of methods including AUC, max F-measure (higher is better), median F-measure, average F-measure, MAE (lower is better), and SOR (higher is better). The best three results are shown in red, violet and blue respectively.

(see Fig. 3.2). (2) Our model decreases the overall MAE on the Pascal-S dataset and achieves higher area under the ROC curve (AUC) score compared to the baselines shown in Fig. 3.2. (3) Although our model is only trained on a subset of Pascal-S, it significantly outperforms other algorithms that also leverage large-scale saliency datasets. Overall, this analysis hints at strengths of the proposed hierarchical stacked refinement strategy to provide a more accurate saliency map. In addition, it is worth mentioning that RDSNet-R outperforms all the recent deep learning based methods intended for salient object detection/segmentation without any post-processing techniques such as CRF that are typically used to boost scores.

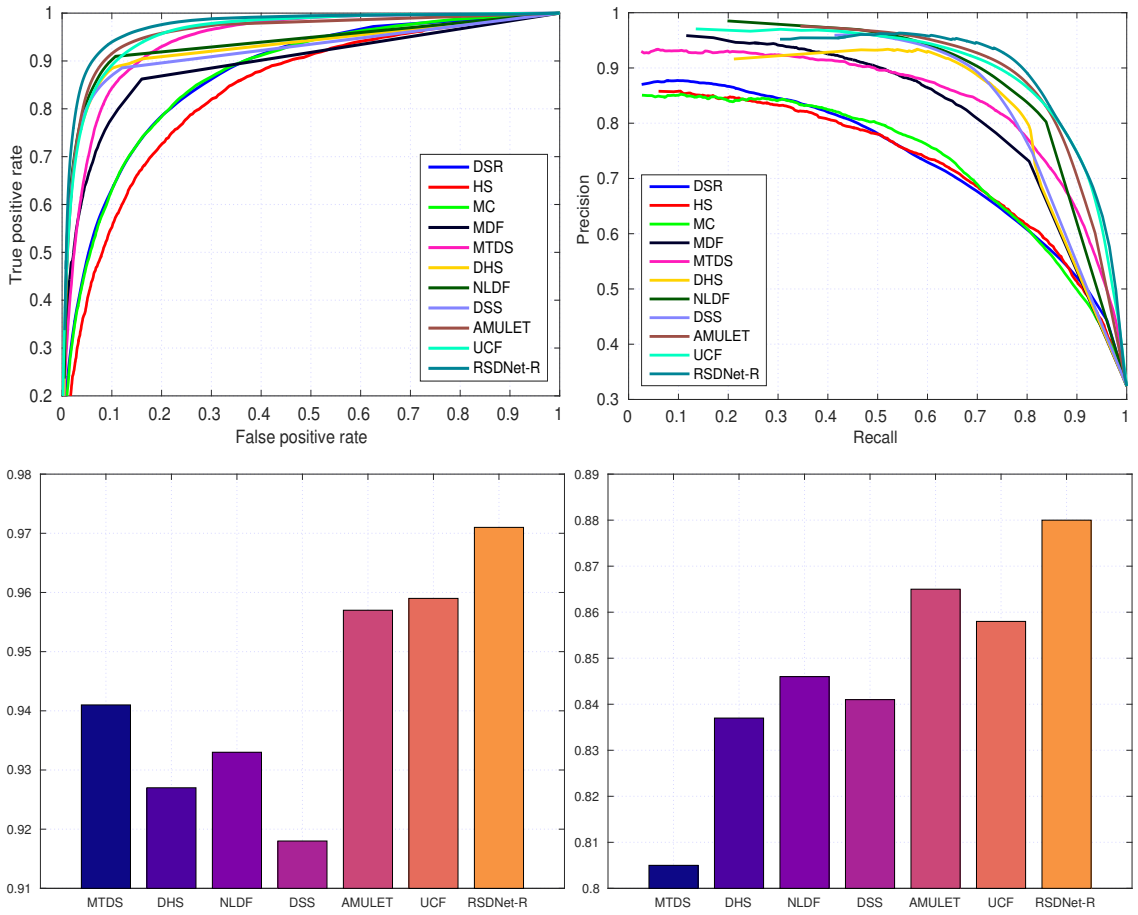


Figure 3.2: Top left: ROC curves corresponding to different state-of-the-art methods. Bottom left: AUC score on the Pascal-S dataset for different approaches. Top right: Precision-Recall curves for salient region prediction corresponding to a variety of algorithms. Bottom right: max F-Score on the PASCAL-S dataset. Our method consistently outperforms other baselines.

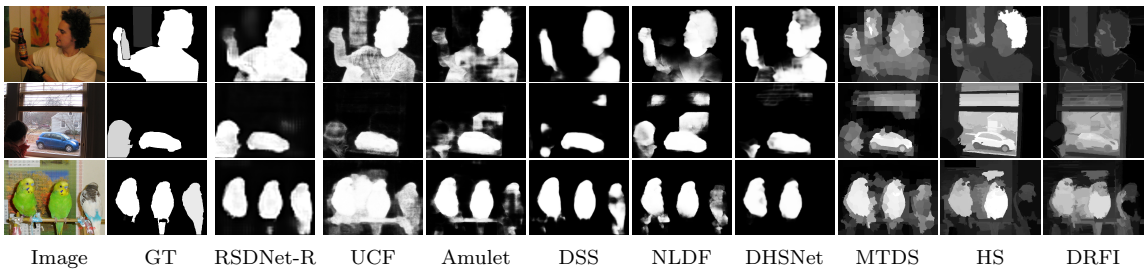


Figure 3.3: Predicted salient object regions for the Pascal-S dataset. Each row shows outputs corresponding to different algorithms designed for the salient object detection/segmentation task.

Qualitative Evaluation: Fig. 3.3 depicts a visual comparison of RSDNet-R with respect to other state-of-the-art methods. We can see that our method can predict salient regions accurately and produces output closer to ground-truth maps in various challenging cases e.g., instances touching the image boundary (1^{st} & 2^{nd} rows), multiple instances of same object (3^{rd} row). The nested relative salience stack at each stage provides distinct representations to differentiate between multiple salient objects and allows for reasoning about their relative salience to take place.

Application: Ranking by Detection

As salient instance ranking is a completely new problem, there is not existing benchmark. In order to promote this direction of studying this problem, we are interested in finding the ranking of salient objects from the predicted saliency map. Rank order of a salient instance is obtained by averaging the degree of saliency within that instance mask. We can write the operation as follows:

$$\text{Rank}(\mathcal{S}_m^T(\delta)) = \frac{\sum_{i=1}^{\rho_\delta} \delta(x_i, y_i)}{\rho_\delta} \tag{3.7}$$

where δ represents a particular instance of the predicted saliency map (S_m^T), ρ_δ denotes total numbers of pixels δ contains, and $\delta(x_i, y_i)$ refers to saliency score for the pixel (x_i, y_i). While there may exist alternatives for defining rank order, this is an intuitive way of assigning this score. With that said, we expect that this is another interesting nuance of the problem to explore further; specifically salience vs. scale, and part-whole relationships. Note that we do not need to change the network architecture to obtain the desired ranking. Instead we use the provided instance-wise segmentation and saliency map to calculate the ranking for each image.

To demonstrate the effectiveness of our approach, we compare the overall ranking score with recent state-of-the-art approaches. It is worth noting that no prior methods report results for salient instance ranking. In an effort to provide a fair comparison, we use saliency maps provided by authors of models compared against, or by running their pre-trained models with recommended parameter settings. We apply the proposed SOR evaluation metric to report how different models gauge relative salience. The last column in Table 3.2 shows the SOR score of our approach and comparisons with other state-of-the-art methods. We achieve 85.2% correlation score for the best variant of our model. The proposed method significantly outperforms other approaches in ranking multiple salient objects and our analysis shows that learning salient object detection implicitly learns rank to some extent, but explicitly learning rank can also improve salient object detection irrespective of how the ground truth is defined. Fig. 3.4 shows a qualitative comparison of the state-of-the-art approaches designed for salient object detection. Note that the role of ranking for more than three objects is particularly pronounced.

Application: Salient Object Subitizing

As mentioned prior, salient object detection, ranking, and subitizing are interrelated. It is therefore natural to consider whether salient region prediction and ranking provide guidance to subitize. A copy of the detection network is further trained to perform subitizing on Pascal-S. For simplicity (and in line with prior work [36; 12]), we train our system only for predicting objects either for 1, 2, 3, or 4+ and report the Average Precision (AP) [50] in Table 3.3.

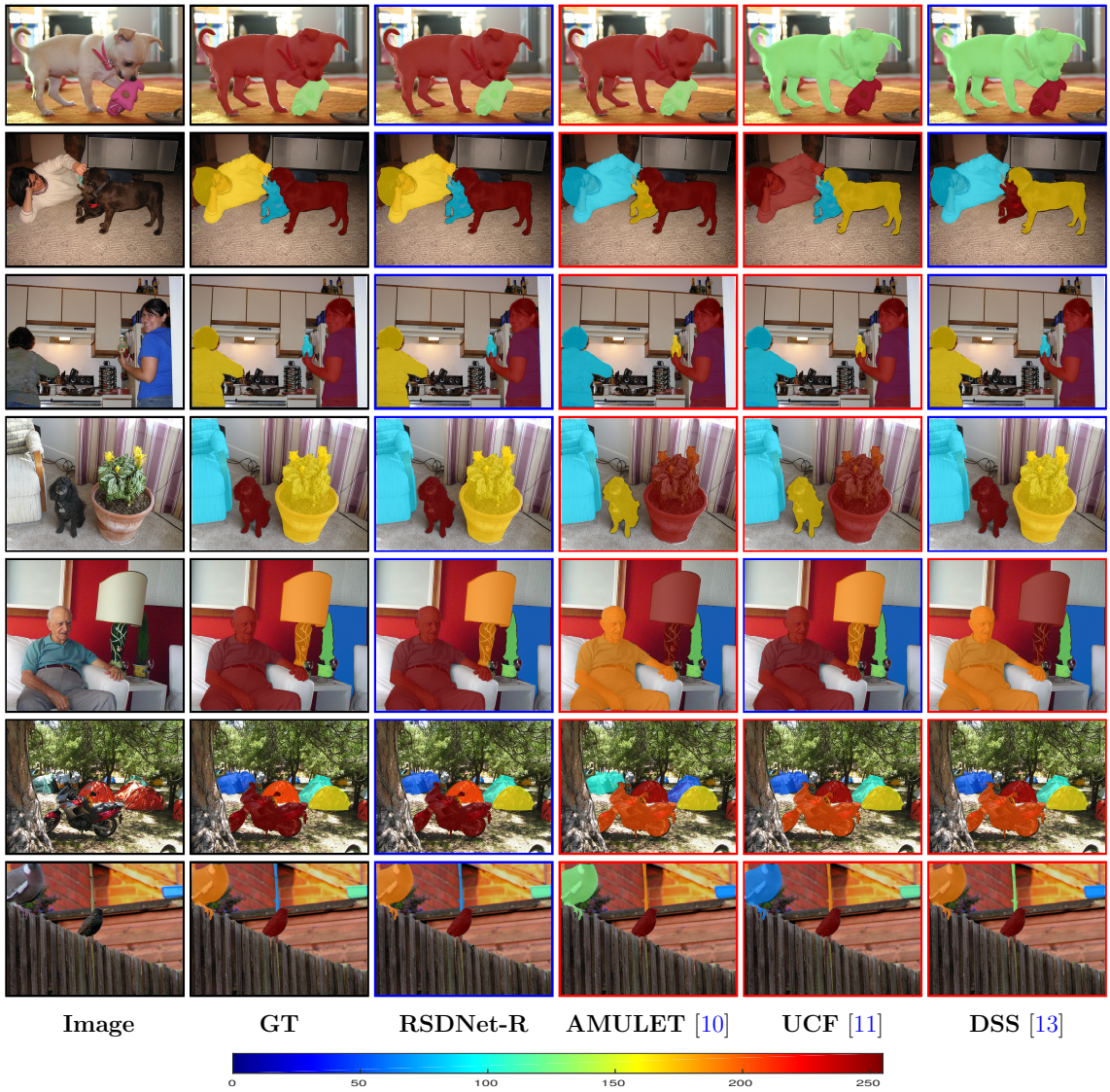


Figure 3.4: Qualitative depiction of rank order of salient objects. Relative rank is indicated by the assigned color. Blue and red image borders indicate correct and incorrect ranking respectively.

*	1	2	3	4+	mean
RSDNet	0.62	0.42	0.20	0.55	0.45

Table 3.3: Average Precision (AP) on Pascal-S dataset.

Since this is the first work to perform subitizing on the Pascal-S dataset, we do not have any baselines to compare with. To make comparison possible, we fine-tune and evaluate our model on the SOS dataset [36], and report the AP and weighted AP (overall) scores in Table 3.4. Our proposed model achieves state-of-the-art results on this dataset compared to baselines.

*	0	1	2	3	4+	mean	overall
count	338	617	219	137	69	-	-
%	0.24	0.45	0.16	0.10	0.05	-	-
HOG [36]	0.65	0.62	0.32	0.29	0.14	0.40	0.52
GIST [36]	0.69	0.66	0.32	0.23	0.22	0.42	0.55
IFV [36]	0.84	0.69	0.32	0.24	0.44	0.50	0.61
CNN [36]	0.92	0.82	0.34	0.31	0.56	0.59	0.70
SOS [36]	0.93	0.90	0.51	0.48	0.65	0.69	0.79
RSDNet	**0.95**	**0.92**	**0.61**	**0.59**	**0.67**	**0.75**	**0.83**

Table 3.4: Overall and Average Precision (AP) on the SOS dataset.

3.2.3 Examining the Nested Relative Salience Stack

Comparison of slices of the nested relative salience stack can be challenging as differences between some layer pairs may be subtle, and contrast can differ across layers. We therefore examine variability among NRSS layers through Principal component analysis (PCA) to determine regions where greatest variability (and signal) exists. Fig. 3.5 shows the top three principal components as an RGB image where the first principal component (which captures the most variance across layers) is mapped to the R-channel, the second principal component is mapped to the G-channel and so forth. Salient areas in the ground truth are captured in the variability across layers demonstrating the value of our stacking mechanism for saliency ranking. Moreover,

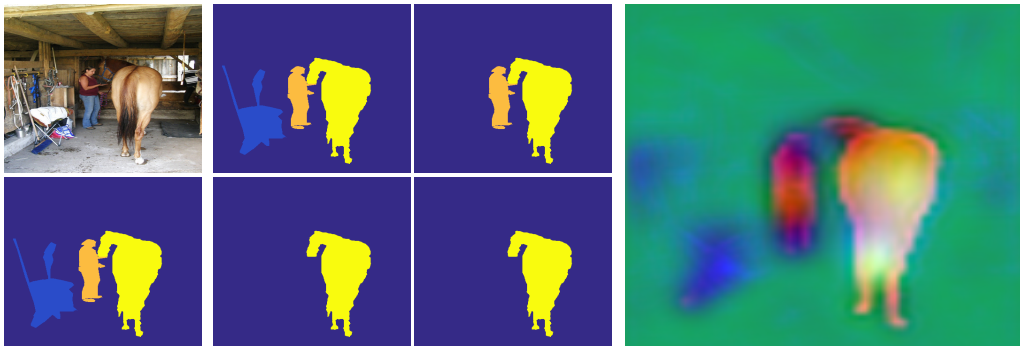


Figure 3.5: Visualization of Principal component analysis (PCA) for the final prediction stack (NRSS) of our model. The first column shows the image and its ground truth. Second and third columns show a selection of ground truth stack slices. The final column provides a visualization of the top three principal components for our predicted stack as an RGB image. Note that the contribution of the top three components itself is diagnostic with respect to relative salience.

it is nearly possible to read a relative ranking directly from this visualization wherein high values for the first 2 eigenvectors result in yellow, the first only red, etc.

We also report the AUC score and max F-measure for each slice (denoted as S) in Table 3.5. Compared to baselines, our proposed method achieves better scores across all ground truth thresholds, that correspond to the different numbers of participants showing agreement that an object is salient. This further shows the effectiveness of the stacking mechanism and predicting relative salience, which results in improvements no matter how the ground truth is determined (if considered as a binary quantity).

3.2.4 Failure Cases

Despite good performance for the majority of cases; there are instances that are more challenging to predict (see Fig. 3.6). Sometimes, the ground truth has multiple objects with the same degree of saliency (ties in participants agreeing) (see 1st row

*	S-1		S-2		S-3		S-4		S-5		S-6		S-7		S-8		S-9		S-10		S-11		S-12	
	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm	AUC	Fm
DRFI [48]	0.872	0.716	0.884	0.701	0.886	0.691	0.887	0.686	0.883	0.68	0.879	0.677	0.882	0.675	0.869	0.664	0.854	0.658	0.825	0.64	0.775	0.613	0.628	0.556
DSR [32]	0.859	0.696	0.866	0.678	0.87	0.671	0.871	0.666	0.867	0.661	0.861	0.658	0.864	0.656	0.851	0.646	0.837	0.64	0.809	0.625	0.759	0.595	0.616	0.54
HDCT [33]	0.793	0.654	0.803	0.636	0.809	0.624	0.808	0.616	0.805	0.611	0.799	0.607	0.802	0.604	0.789	0.59	0.776	0.584	0.75	0.569	0.705	0.545	0.571	0.495
HS [6]	0.828	0.702	0.834	0.678	0.835	0.665	0.837	0.659	0.834	0.655	0.828	0.65	0.829	0.647	0.815	0.636	0.801	0.629	0.773	0.611	0.728	0.585	0.591	0.529
MC [20]	0.855	0.717	0.863	0.694	0.867	0.683	0.87	0.678	0.866	0.672	0.861	0.668	0.864	0.666	0.851	0.654	0.836	0.647	0.808	0.631	0.761	0.607	0.618	0.55
MTDS [34]	0.922	0.805	0.938	0.791	0.941	0.779	0.941	0.772	0.937	0.766	0.93	0.763	0.931	0.76	0.918	0.748	0.901	0.739	0.869	0.718	0.816	0.69	0.661	0.627
ELD [18]	0.885	0.789	0.903	0.779	0.912	0.777	0.916	0.776	0.915	0.774	0.910	0.770	0.913	0.768	0.900	0.755	0.884	0.749	0.855	0.733	0.805	0.708	0.653	0.642
MDF [21]	0.866	0.787	0.882	0.775	0.888	0.767	0.892	0.761	0.888	0.756	0.883	0.753	0.885	0.75	0.876	0.743	0.861	0.736	0.83	0.713	0.78	0.688	0.633	0.627
DHSNet [19]	0.888	0.837	0.911	0.832	0.922	0.831	0.927	0.829	0.926	0.824	0.919	0.820	0.923	0.818	0.912	0.807	0.897	0.800	0.865	0.774	0.815	0.748	0.660	0.680
NLDF [35]	0.900	0.846	0.922	0.840	0.931	0.836	0.933	0.831	0.930	0.827	0.922	0.821	0.925	0.818	0.913	0.809	0.897	0.802	0.865	0.782	0.812	0.751	0.660	0.680
DSS [13]	0.883	0.841	0.906	0.839	0.916	0.832	0.918	0.825	0.915	0.821	0.910	0.819	0.912	0.816	0.901	0.805	0.886	0.799	0.855	0.779	0.802	0.745	0.651	0.675
AMULET [10]	0.932	0.865	0.949	0.856	0.954	0.850	0.957	0.847	0.952	0.840	0.944	0.834	0.946	0.829	0.933	0.819	0.918	0.813	0.884	0.791	0.827	0.760	0.671	0.688
UCF [11]	0.940	0.858	0.955	0.845	0.959	0.838	0.959	0.831	0.956	0.829	0.947	0.825	0.949	0.823	0.935	0.813	0.918	0.806	0.885	0.785	0.827	0.754	0.672	0.689
RSDNet	0.950	0.872	0.966	0.873	0.970	0.870	**0.972**	0.868	**0.967**	0.860	**0.957**	0.854	**0.957**	0.850	**0.945**	0.842	**0.926**	0.834	**0.893**	0.812	0.836	0.774	0.676	**0.705**
RSDNet-**A**	**0.952**	0.874	**0.967**	0.874	**0.972**	0.871	**0.973**	**0.869**	**0.968**	0.860	**0.958**	**0.856**	**0.958**	**0.853**	**0.946**	**0.846**	**0.928**	**0.836**	**0.895**	**0.815**	**0.837**	**0.778**	**0.677**	**0.707**
RSDNet-**B**	0.948	**0.877**	0.963	**0.877**	0.968	**0.873**	0.969	**0.871**	0.964	**0.862**	0.954	**0.856**	0.954	0.852	0.942	0.844	0.923	0.833	0.889	0.810	0.831	0.774	0.672	0.702
RSDNet-**C**	**0.955**	0.874	**0.968**	0.872	**0.971**	0.869	**0.972**	0.867	**0.967**	0.859	**0.958**	0.854	**0.958**	0.851	**0.946**	0.843	**0.928**	0.835	**0.895**	**0.813**	**0.838**	0.775	**0.678**	0.699
RSDNet-**R**	0.951	**0.880**	0.965	**0.879**	0.969	**0.874**	0.971	**0.871**	0.966	**0.866**	0.956	**0.859**	0.956	**0.854**	0.944	**0.849**	0.925	**0.838**	0.892	**0.815**	0.833	**0.776**	0.674	0.701

Table 3.5: Quantitative comparison (AUC & Fm) with state-of-the-art methods across all ground truth thresholds, each corresponding to agreement among a specific number participants. Best and second best scores are shown in red and blue respectively.

in Fig. 3.6). Other failures of ranking happen when there is considerable diversity in agreement on what is salient in an image (as shown in the second row) or when there is occlusion among two objects which have a relatively close degree of saliency as shown in the last row.

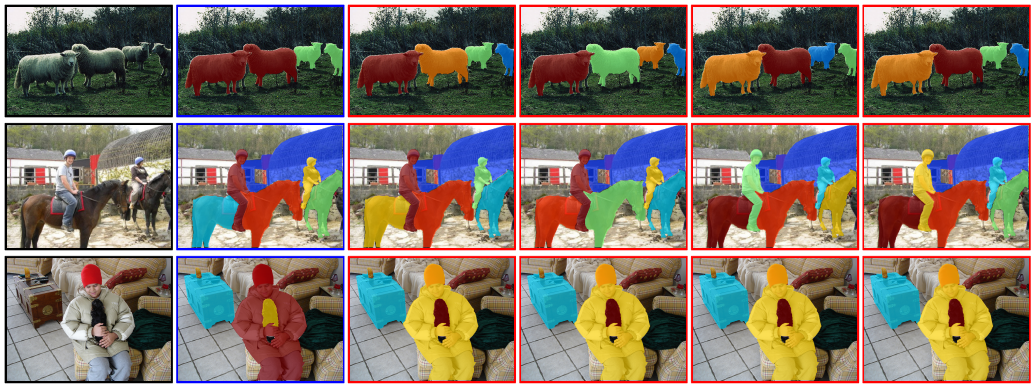


Figure 3.6: Shown are some illustrative examples of disagreements in rank between model and ground truth. These are most common for ties in the ground truth, and for scenes with many salient objects.

Chapter 4

Saliency Ranking: Moving Forward in Relative Saliency Prediction with a Large-scale Dataset

In this chapter, we present our novel approach to generate a large scale dataset for the problem of salient object ranking based on publicly available annotations. Towards this endeavor, we introduce a data refinement pipeline that utilizes instance-wise segmentations and fixation maps to assign a rank order to salient objects in an image. We also show how different parameters of the proposed algorithm can alter the quality of the generated dataset. Finally, we experiment using different state-of-the-art methods to establish initial baselines for both salient object ranking and salient object detection tasks.

4.1 COCO-SalRank Dataset

We start with constructing a dataset suitable for saliency ranking analysis, which will be released to the computer vision community to promote saliency ranking research. Designing a large-scale dataset requires a large number of decisions including data collection, processing, and the annotation protocol. Our choices were driven by the end goal of enabling immediate significant progress in the field of saliency ranking and allowing deeper exploration of relative saliency.

4.1.1 Description of the COCO-SalRank dataset

One rule of thumb for constructing a dataset is to first analyze existing datasets for the same task. There exist a significant number of saliency detection datasets but the majority provide ground-truth in binary notion. Since saliency ranking requires ground-truth in the form of multiple observers' agreement, a primary objective is to arrive at a dataset that provides a faithful rank order of the salient objects. To do so, we at least require images with several distinct measures of what may be salient. Given the relatedness to human allocation of gaze, it is natural to consider what value may be derived from considering fixation maps across observers in conjunction with cases where instance-wise label maps are also present. Specifically, there is the opportunity to leverage fixation maps to assign rankings among category items present within instance-wise label maps. We therefore make the careful choice of MS COCO images that include multiple simulated fixation annotations from the training and validation data of the SALICON dataset [30] to construct our proposed dataset. We obtain instance-wise mask annotations from the MS-COCO dataset [29] as SALICON

images are chosen from MS-COCO.

Directly combining these two data sources to achieve the desired objective is a more significant challenge than one might expect on the surface for reasons that are highlighted throughout this chapter. Moreover, we have already discussed significant differences between gaze data and click-based assessments of salient object for ranking. For this reason, we also include careful evaluation on data where both gaze and click based selections are present to validate that the end result comes very close in approximating click-based ranking data. Given the set of images with instance-wise labels and fixation maps, we propose a novel approach to provide saliency ranking ground-truth. Initially we create a *noisy* version (version I) of the proposed dataset (COCO-SalRank) that includes 7047 training images and 3363 test images. Subsequently, we produce a refined *clean* version (version II) of annotations with 3052 training and 1381 test images. Note that in deriving the clean dataset, the cleaned data (version II) is also manually checked to remove obvious outliers, and to ensure labeling fidelity. The total number of images removed manually based on visual inspection is 605. In addition to the salient ranking data, we also generate a new set of ground-truth images for salient object subitizing given the parallel interest in this problem within the research community. The following presents details for deriving these datasets, including data refinement that have been considered in creating the COCO-SalRank dataset (and rationale):

- **Number of Instances:** We restrict the total number of instances in an image to maximum five.

- **Ranking Ties:** No instances are assigned tied values due to differences in

relative saliency within labeled instance regions.

Considering the above-mentioned factors, we provide two different sets of ground-truth data that may each carry value in model training or performance evaluation. The justification for this latter statement is borne out in analysis of results.

4.1.2 Ground Truth Annotation

To obtain the saliency ranking ground-truth, as alluded to earlier, we propose a novel approach that assigns a rank order to salient objects in an image given instance-wise segmentations and associated stimulated fixation maps. The instance-wise labeling in MS-COCO [29] is not consistent with the fixation maps provided by SALICON, so we can not directly use these two sets of annotations in order to generate the saliency ranking annotation. There are many instances which are labeled in the instance map but are not a point of focus in the fixation map. Similarly, some instances have reasonable fixation density but are not labeled as an instance in the mask. Other challenges include over or under-segmented images, unlabeled non-class objects, and that explicit ranking from manual selection are unavailable for this data. To overcome these limitations, we have arrived through experimentation at a data refinement pipeline that is shown to produce faithful rankings when measured against alternate smaller-scale data where gaze, instance labels, and explicit selection are all present. The methods that address this challenge are described in what follows.

We first apply Gaussian blurring on the fixation locations to obtain the new fixation map $\mathcal{F}_i'$. This is a crucial step in our approach since fixation locations provided by SALICON [30] are generated by mouse tracking instead of a traditional eye

tracker which implies a less diffuse distribution on the focus of attention such that fixation points may not overlap with the corresponding instances (even if they are proximal). In addition, as the labeling in MS-COCO does not capture the border of the instances accurately, blurring the fixation locations is especially important to allow density from border fixations to diffuse into the defined mask area. Another important step in our annotation protocol is pruning the original instance mask. As mentioned earlier, regions labeled in instance-wise label maps may not include all salient objects. Instance masks may therefore be pruned based on few carefully chosen criteria. Given the set of new fixation maps $\mathcal{F}'$ and the provided instance-wise maps $\mathcal{I}$, we propose a *Saliency Ranking* algorithm (Algorithm 1) that generates the ranking ground-truth. Algorithm 1 describes the general set of steps to obtain a rank order of salient instances in an image.

First, we calculate overlap ϑ between the new fixation map $\mathcal{F}'$ and the instance-wise map $\mathcal{I}_i$ to remove non-salient instances. Then, we generate a ranking score $\mathbb{R}_\chi$ for each salient instance χ by dividing the total saliency captured by χ in ϑ to the size of the instance raised to the power ϵ. The *prune* function takes the newly generated fixation map $\mathcal{F}'$, instance map $\mathcal{I}_i$, and two parameters (α_1&α_2) as input, resulting in a pruned instance mask, $\mathcal{I}'_i$. The prune function focuses on removing instances with the following conditions: (1) If the size of a particular instance is greater than a certain threshold α_1 (generally due to under-segmentation) or (2) The rank score of an instance $\mathbb{R}_\chi$ is less than α_2 (generally due to receiving very little attention). Given the pruned instance map $\mathcal{I}'_i$, we again calculate the overlap ϑ' between $\mathcal{F}'$ and $\mathcal{I}'_i$ in order to disqualify the pruned instance maps $\mathcal{I}'$. We apply the following conditions to filter

Algorithm 1 Saliency Ranking

function SALRANK($\mathcal{I}, \mathcal{F}, \sigma, \mu, \xi, \ell, \gamma$)

▷ instance maps $\mathcal{I}$, fixation maps $\mathcal{F}$

for `each instance map` $\mathcal{I}_i \in \mathcal{I}$ **do**

Rank list, $\mathbb{R}$

$\mathcal{F}'_i = Gaussian\ (\sigma, \mu, \mathcal{F}_i)$

overlap, $\vartheta = \mathcal{F}'_i \times \mathcal{I}_i$

for `each instance` $\chi \in \mathcal{I}_i$ **do**

score, $\mathbb{R}_\chi = \frac{\sum \vartheta(\chi)}{\sqrt[\epsilon]{\text{size}(\chi)}}$

end for

$\mathcal{I}'_i =$ **Prune**$(\mathcal{I}_i, \mathcal{F}'_i, \mathbb{R}, \alpha_1, \alpha_2)$

overlap, $\vartheta' = \mathcal{I}'_i \times \mathcal{F}_i$

total instances, $\rho = unique(\mathcal{I}'_i)$

if $\rho > \xi$ **or** $\frac{\sum(\vartheta')}{\sum(\mathcal{F}'_i)} < \ell$ **then**

ignore instance map $\mathcal{I}_i$

end if

if $\sum(\mathcal{I}'_i \neq 0) > \gamma$ **then**

ignore instance map $\mathcal{I}_i$

end if

Ranked gt = **Rank** $(\mathcal{I}'_i, \mathbb{R})$

end for

end function

Algorithm 2 Pruned Instance Mask

function Prune$(\mathcal{I}_i, \mathcal{F}'_i, \mathbb{R}, \alpha_1, \alpha_2)$

▷ instance map $\mathcal{I}_i$, fixation map $\mathcal{F}'_i$

for `each instance` $\chi \in \mathcal{I}_i$ **do**

if size$(\chi) > \alpha_1$ or $\mathbb{R}_\chi < \alpha_2$ **then**

remove instance χ

end if

end for

return pruned instance map $\mathcal{I}'_i$

end function

function Rank$(\mathcal{I}'_i, \mathbb{R})$

▷ pruned instance $\mathcal{I}'_i$, Rank list $\mathbb{R}$

for `each instance` $\chi \in \mathcal{I}_i$ **do**

if *Relative* **then**

Assign relative salience

end if

if *Absolute* **then**

Assign absolute salience

end if

end for

end function

instance maps: (a) The total number of instances is less than ξ (b) The total saliency captured by the pruned instance mask is greater than ℓ compared to the fixation map (c) The ratio of background vs salient instances satisfies a certain threshold γ. Table 4.1 demonstrate the set of parameters used to generate both versions of the proposed ground-truth labels.

COCO-SalRank

version I (noisy)							version II (clean)						
σ	μ	ξ	ℓ	γ	α_1	α_2	σ	μ	ξ	ℓ	γ	α_1	α_2
10.5	80	5	0.4	0.65	0.4	0.7	10.5	80	5	0.7	0.65	0.4	0.9

Table 4.1: The set of parameters that are used in our labeling process. Note that in version II, we specifically tighten ℓ and α_2 to obtain more clean and reliable annotations.

If an instance map meets all of the three conditions, we employ the *Rank* function that assign relative salience to each instance based on their rank score. Note that we propose two different strategies (relative and absolute) to assign numeric rank in a range of [0, 255] and this is briefly described in Sec 4.1.4.

4.1.3 Dataset Analysis

We have proposed in the early part of this chapter that our proposed dataset is more suitable for saliency ranking than existing saliency detection datasets. It is evident that it is mandatory to have multi-user agreement to assign the rank score for individual salient instances. In the following sections, we discuss a few aspects of the proposed ranking algorithm and their justification.

Effect of Blurring and instance size in ranking

As mentioned prior, we use instance maps from MS-COCO and the fixation locations from SALICON in order to generate ranking labels. For each instance in an instance map, we calculate a score that represents the degree of saliency for that instance. A natural approach to calculate the rank score is to compute the amount of saliency that an instance captures. However, larger objects tend to capture more fixations than smaller objects which implies that larger objects may be implicitly biased towards having a higher rank order compared to smaller ones. Therefore, the size of the object is a key deciding factor in the process of calculating the ranking score for each object. Furthermore, the degree of Gaussian burring of the fixation locations also influences the amount of saliency that an instance may capture and as noted, has a proximity effect in relation to object instances and is also influenced by the undefined boundary region.

In order to examine how the size of each instance and the blurring of fixation locations might sway the rank for each instance, and to investigate the possible ways to integrate these variables in the calculation of the ranking score, we measure the impact of the ranking parameters against the PASCAL-S dataset [9] since it provides both fixation locations and the manually chosen rank of salient objects annotated by multiple observers. In other words, we use the fixation location and the segmentation masks provided by the PASCAL-S dataset to generate our ranking label which is compared against the ranking labels provided by the PASCAL-S dataset. We compute each instance ranking score according to equation 4.3.

We firstly examine the effect of changing α by fixing the Gaussian burring filter

size to $\mu = 80$ and the standard deviation $\sigma = 10.5$. We change the value of α and calculate the Salient Object Ranking (SOR) [1] score between the predicted and the ground truth ranking label provided by PASCAL-S as shown in Table 4.2.

Similarly, we examine the Gaussian burring effect by fixing the power $\alpha = 0.3$ and changing the standard deviation of the Gaussian blurring σ. Note that the filter size corresponding to a specific σ in Table 4.2 is $\mu = \sigma * 7$. We found empirically that a Gaussian burring filter size $\mu = 80$, the standard deviation $\sigma = 10.5$, and $\alpha = 0.3$ tends to correspond to the set of conditions for which fixations best determine manually chosen rankings (based on SOR score for PASCAL-S). Thus, we apply these values as our threshold while generating ground truth for our proposed COCO-SalRank dataset. Fig. 4.1 further demonstrates both of those effects.

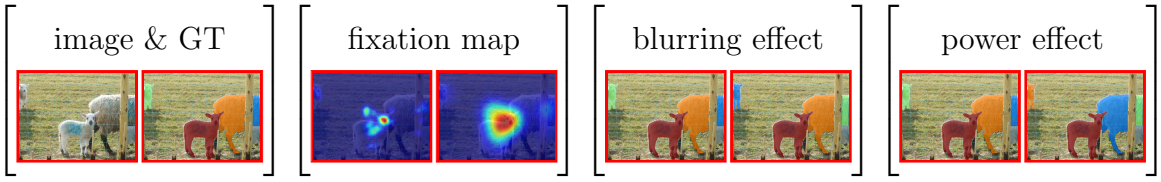


Figure 4.1: Sets left to right are: input image and ground truth rank, fixation maps blurred with different Gaussian filters, predicted rank that correspond to fixation maps in the previous set ($\alpha = 0.3$), predicted rank that corresponds to two different α ($\sigma = 10.5, \mu = 80$).

α	1	0.8	0.6	0.4	0.3	0.2	0.1
SOR	0.76	0.83	0.86	0.89	0.90	0.88	0.87

sigma	5	13	21	29	37	41
SOR	0.89	0.89	0.88	0.87	0.86	0.85

Table 4.2: The impact of applying different power α, filter size (μ), standard deviation σ on ranking performance. Note that in the right set, we fix $\alpha = 0.3$ and use the filter size equivalent to $\mu = \sigma \times 7$ whereas in left set we fix both $\mu = 80, \sigma = 10.5$.

Why Not Using Existing Saliency Detection datasets?

The primary objective is to obtain the rank order of the salient objects, so it is crucial for us to provide assignment of relative salience based on fixation data that ensures reliable ground-truth annotation. Existing datasets such as ECSSD [6], MSRA-10K [7], DUT-OMRON [8] provide ground-truth in a traditional way by thresholding which implies that there is no notion of relative salience. Since saliency ranking aims to explicitly model observer agreement, one can not directly use the existing saliency detection datasets for retrieving the rank order.

4.1.4 Relative and Absolute Salience Assignment

Existing datasets (Pascal-S [9]) contain information that allows for implicit assignment of relative salience based on agreement among multiple observers while in our case we propose to assign rank value under two different settings (*Relative* and *Absolute*). In the relative setting, we assign rank value based on total number of instances in the mask and their rank score, $\mathbb{R}_\chi$ where χ is a particular instance. For example, if we have total τ number of instances in the mask, we divide the range [0, 255] by τ to obtain the numeric rank value. While in the absolute case, rank values are assigned based on the percentile of the rank score set and then rescaled to the range [50 255] which corresponds to [20% - 100%] of the gray-scale levels. Similar to [1], we generate a stack representation of the ground truth where each stack consists of 5 slices since a cap of 5 objects is set in filtering to avoid over-segmentation. For the relative case, The first slice of the stack contains the most salient object, the second slice contains the top 2 salient objects and so on. However, for the absolute case, the

first slice accounts for less than 20% of the fixations, the second one accounts for less than 40% of the fixations, and so on. The stacked representation of the ground-truth in relative and absolute settings are shown in Eq. 4.1 and Eq. 4.2 respectively. As shown in this example (in the case of 5 ranked instances), a new instance is added to each slice in the relative case, whereas, multiple instances may be added at once in the absolute case if they reside within the same percentile rank.

$$\mathcal{R}_\vartheta = \left[\begin{array}{c}\mathcal{R}_i\\ \\ \end{array}\right]\left[\begin{array}{c}\mathcal{R}_{i+1}\\ \\ \end{array}\right]\left[\begin{array}{c}\mathcal{R}_{i+2}\\ \\ \end{array}\right]\left[\begin{array}{c}\mathcal{R}_{i+3}\\ \\ \end{array}\right]\left[\begin{array}{c}\mathcal{R}_{i+4}\\ \\ \end{array}\right] \tag{4.1}$$

$$\mathcal{A}_\vartheta = \left[\begin{array}{c}A_i\\ \\ \end{array}\right]\left[\begin{array}{c}A_{i+1}\\ \\ \end{array}\right]\left[\begin{array}{c}A_{i+2}\\ \\ \end{array}\right]\left[\begin{array}{c}A_{i+3}\\ \\ \end{array}\right]\left[\begin{array}{c}A_{i+4}\\ \\ \end{array}\right] \tag{4.2}$$

4.1.5 Ranking Mechanism

As saliency ranking is a newly proposed problem, there is no universally agreed upon metric to obtain the rank order of salient instances. We start with the metric proposed in [1] and explore a few alternatives ranking mechanisms. Firstly, rank order of a salient instance is obtained by averaging the degree of saliency within that instance mask. Then we propose to assign the rank order from an output saliency map by dividing the total degree of saliency within an instance to its size raised to a certain power. Finally, we obtain the rank order by taking the max value of the saliency within the instance region. These considerations relate to the earlier observations concerning the size-saliency tradeoff and its relation to ranking. We can

write these metrics as follows:

$$\text{Rank} = \begin{cases} \text{SOR}_{\text{avg}}(\mathcal{S}(\delta)) = & \frac{\sum_{i=1}^{\rho_\delta} \delta(x_i, y_i)}{\rho_\delta} \\ \text{SOR}_{\text{pow}}(\mathcal{S}(\delta); \alpha) = & \frac{\sum_{i=1}^{\rho_\delta} \delta(x_i, y_i)}{\rho_\delta^\alpha} \quad \alpha = 0.3 \\ \text{SOR}_{\text{max}}(\mathcal{S}(\delta)) = & \max(\delta(x_i, y_i)) \end{cases} \tag{4.3}$$

where δ represents a particular instance of the predicted saliency map (S), ρ_δ denotes total numbers of pixels δ contains, and $\delta(x_i, y_i)$ refers to saliency score for the pixel (x_i, y_i).

4.1.6 A New Set of Ground-truth for Salient Object Subitizing

To promote the study of salient object subitizing [36; 12; 23] in more complex scenarios and with a greater amount of data, we provide a new set of ground truth that provides an instance-wise count. Similar to saliency ranking, to remove the ambiguity in subitizing results, we restrict the images to have at most five objects. The distribution of images in both the versions of proposed dataset with respect to several categories is shown in Table 4.3. We can see that there are a considerable number of images with more than 5 objects. The images are categorized as 1, 2, 3, 4, and 5 salient objects. It is worth mentioning that we calculate the count for each image after obtaining the ranking ground-truth.

count	1	2	3	4	5
train	734	2062	2133	1441	677
test	346	996	1029	656	336
total	1080	3058	3162	2097	1013

count	1	2	3	4	5
train	489	1165	911	398	89
test	219	567	393	158	44
total	708	1732	1304	556	133

Table 4.3: Count of images corresponding to different numbers of salient objects in the proposed COCO-SalRank dataset. Left : version I, Right: version II

4.2 Qualitative Examples of generated ground-truth for the COCO-SalRank Dataset

Fig. 4.2 depicts visual examples of generated saliency ranking ground-truth on the COCO-SalRank dataset with respect to fixation maps. It is an evident that our algorithm produces ground-truth maps that have an intuitive correspondence with rank and that are consistent with fixation maps in various challenging cases.

4.3 Challenging Saliency Ranking Cases

Despite the consistent ground-truth quality for the majority of cases, there are samples that are especially challenging to assign rank order by algorithmic means (see Fig. 4.3). Sometimes, fixations are distributed among multiple objects which makes the ranking task challenging. Other failures of ranking happen when a higher ranked object overlaps with or exists close to a less salient object. In that case, the lower ranked object receives some attention attributed to fixations from the higher ranked instance. An additional challenge is the lack of consistency in instance-wise labeling in MS-COCO.

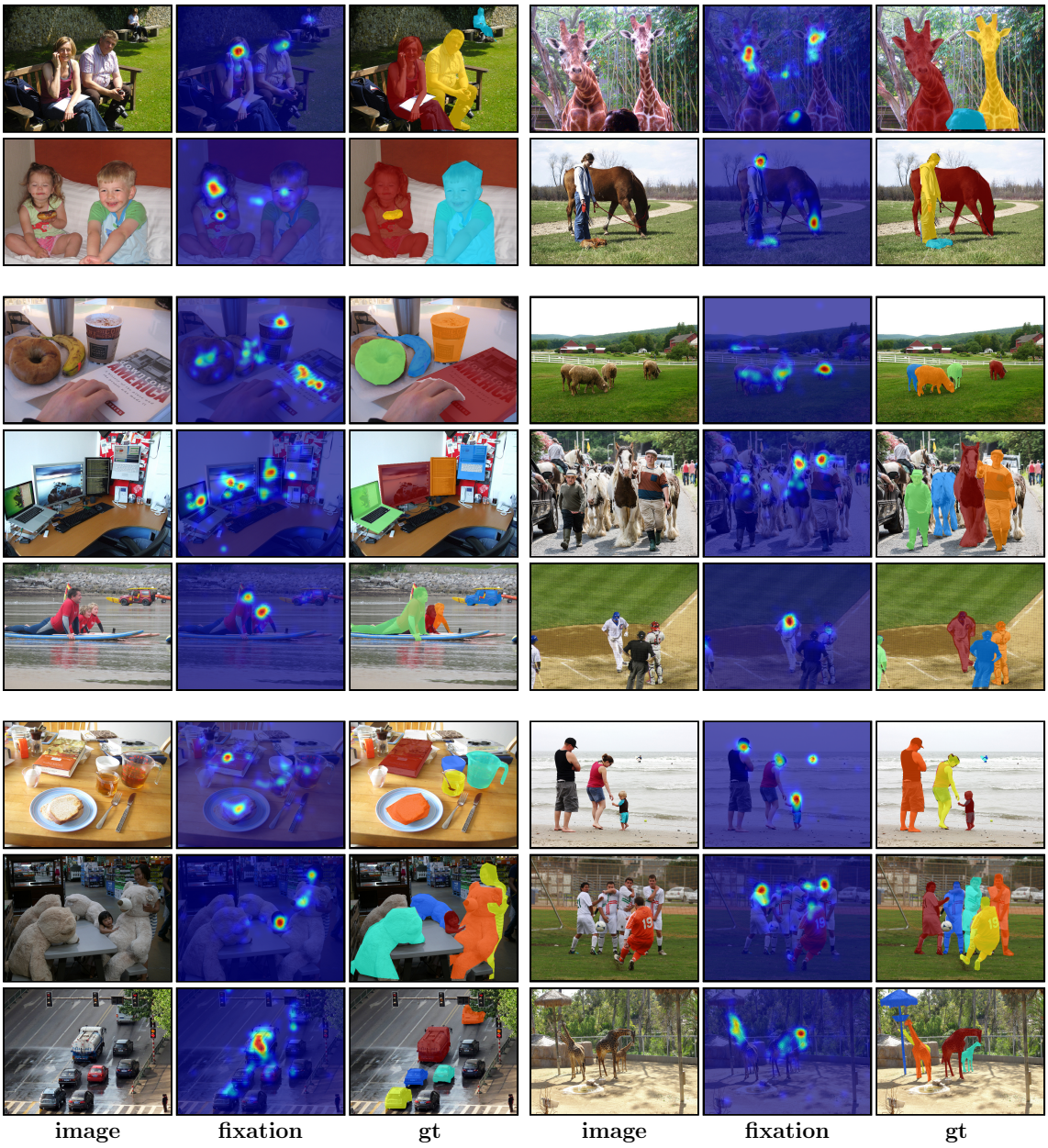


Figure 4.2: Qualitative illustration of obtained ground-truth samples on COCO-SalRank dataset. Relative rank is indicated by the assigned color. The consistency among fixation maps and ground-truth ranking shows good agreement and an intuitive ranking for our proposed dataset. Note that all the ground-truth samples are chosen from *noisy* version of the dataset.

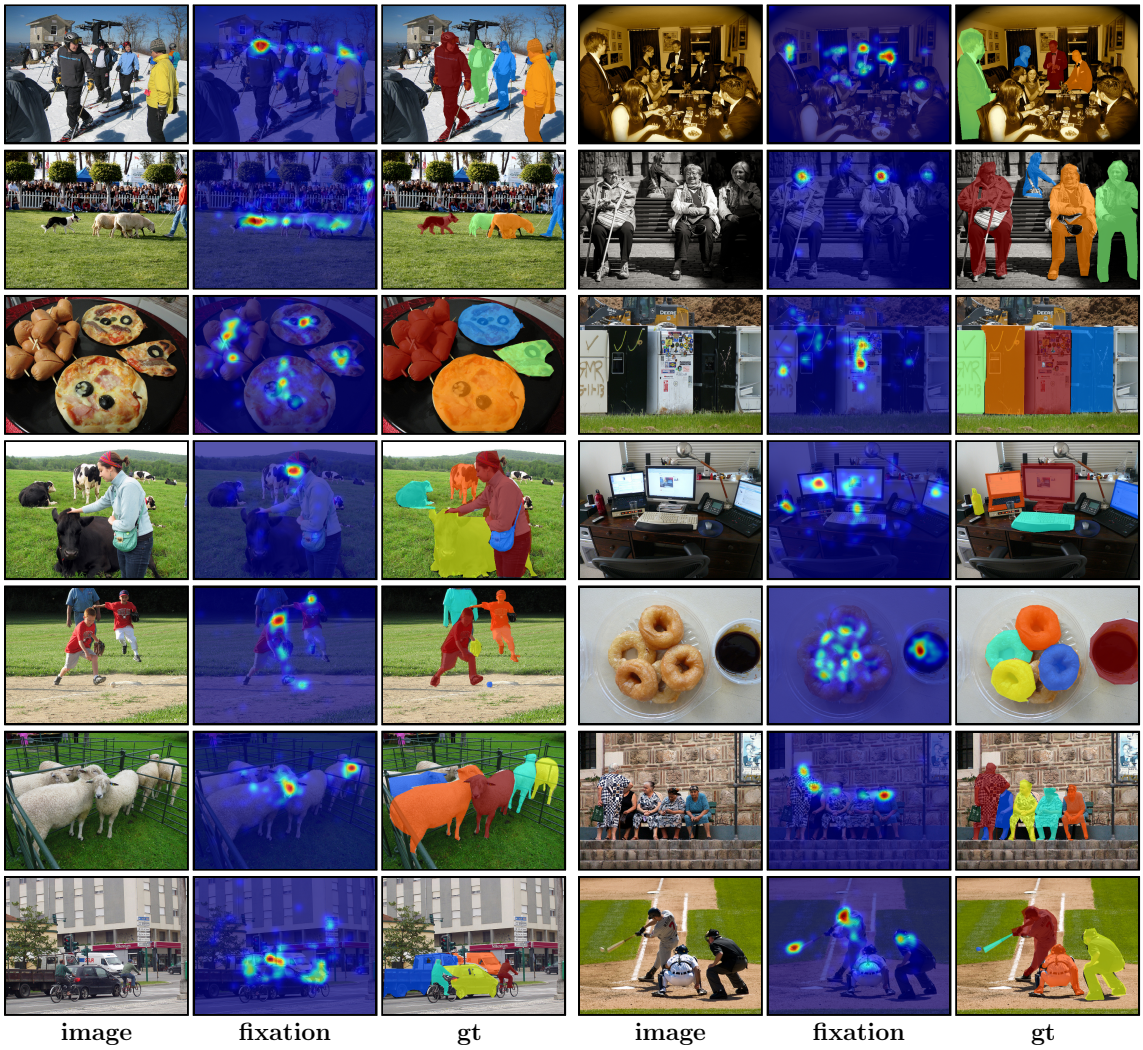


Figure 4.3: Shown are some illustrative examples of inconsistency among fixation maps and the generated ground-truth ranking. These cases are most common for overlapping instances, and for scenes with fixations spread over multiple salient objects.

4.4 Experiments

In this section, we conduct several experiments on the proposed COCO-SalRank dataset to establish baseline performance metrics. In Sec. 4.4.2, we evaluate the saliency ranking performance of several state-of-the-art saliency detection architectures on our dataset, revealing that the proposed dataset provides the possibility to

achieve state-of-the-art accuracy, even when tested without any training on PASCAL-S. For further analysis on our dataset, we carry out multiple salient object detection tasks in Section Sec. 4.4.5.

4.4.1 Experimental Settings

To report performance numbers, we train a few baseline models with the recommended parameter settings provided in publicly available code. For fair comparison, all the experiments only use the training set of our proposed dataset without any data augmentation. In saliency ranking networks, we train with the following hyper-parameters settings: batch size (4), poly learning rate ($2.5e^{-11}$), momentum (0.9), and number of iterations (10k). Testing uses the full resolution image while training relies on random crops for memory savings.

Evaluation Protocols: We evaluate all the baseline models under two different settings: *Relative Ranking* and *Absolute Ranking*. Note that relative and absolute ranking settings differ in the way of assigning a relative salience score when generating rankings.

Evaluation Metrics: For saliency ranking, we use the Salient Object Ranking (SOR) metric proposed in [1] and few proposed alternatives. Note that all variants of SOR are based on Spearman's Correlation between the ground truth ranking and the predicted ranking of salient objects. For salient region detection tasks, we use three different universally accepted metrics (F-measure, Area under ROC curve (AUC), and mean absolute error (MAE)) to report performance numbers. Given that there are multiple possible binary ground-truth maps that may be associated with

a representation of salient objects that includes rank, in considering the traditional problem of salient object detection and associated metrics it is important to consider different binary interpretations that could result depending on the chosen threshold. The most natural means of doing this is to apply the aforementioned metrics on a per-slice basis and we also present corresponding results.

4.4.2 Saliency Ranking Baselines

In our experimental settings, any networks intended for classification, segmentation, or saliency detection paired with a Nested Relative Salience Stack (NRSS) [1] can be used to generate a rank-aware saliency map. We obtain the rank order of predicted salient regions using existing and our proposed ranking metrics. Based on the performance and reproducibility of existing methods, we choose RSDNet [1], DeepLabv2-VGG [42], and PSPNet [51] networks to predict saliency for our dataset. It is worth noting that the inclusion of NRSS at the end of the network architectures allows any model to be trained for the saliency ranking task. For fair comparison, we have employed the same strategies for ground-truth representation, and layer supervision for all these baselines.

4.4.3 Ranking Evaluation on Pascal-S Dataset

Since Pascal-S [9] is the only dataset that provides ranking ground-truth from multiple observers, to justify the correctness of our proposed dataset we perform several experiments on the Pascal-S dataset under different settings shown in Table 4.4. First, we train several baseline models using our proposed dataset and evaluate on

$*$	RSDNet*	RSDNet‡	RSDNet†	RSDNet [1]	UCF [35]	Amulet [35]	DSS [35]	NLDF [35]	DHS [19]
SOR_{avg}	0.848	0.862	0.832	0.825	0.792	0.788	0.770	0.783	0.781
SOR_{pow}	0.848	0.843	0.867	0.839	0.820	0.823	0.834	0.850	0.826
SOR_{max}	0.831	0.855	0.857	0.824	0.837	0.840	0.810	0.851	0.824

Table 4.4: Performance Comparison of Saliency ranking score of several networks on Pascal-S dataset. Note that all the baseline numbers are reported from [1].

the test subset of the Pascal-S dataset. It is a evident from Table 4.4 that RSDNet† (*train* : COCO-SalRank, *test*: test set of Pascal-S) outperforms RSDNet in term of saliency ranking. When we fine-tune the trained model with the train set of Pascal-S, RSDNet‡ further improves the ranking performance by a considerable margin. Note that we perform the same experiments for both versions of our dataset under relative ranking scenario. When we train RSDNet with COCO-SalRank (version II) and test on Pascal-S we achieve a significant boost (+2%) with RSDNet* compared to RSDNet. One could argue that the reason behind the improvement of ranking performance is the large training set compared to the small one of Pascal-S, which further underscores the fidelity and effectiveness of our proposed dataset. Fig. 4.4 presents a qualitative comparison of the state-of-the-art approaches designed for salient instance ranking or detection. Note that the role of ranking for more than three objects (2^{nd} row) is especially pronounced in revealing the significance of our dataset.

4.4.4 Ranking Evaluation Under Relative and Absolute Rank Setting

The saliency ranking performance of all baseline models under relative and absolute settings are listed in Table 4.5 (version I & version II). The performance in terms of all metrics reveals the generality and reliability of the proposed dataset. Given

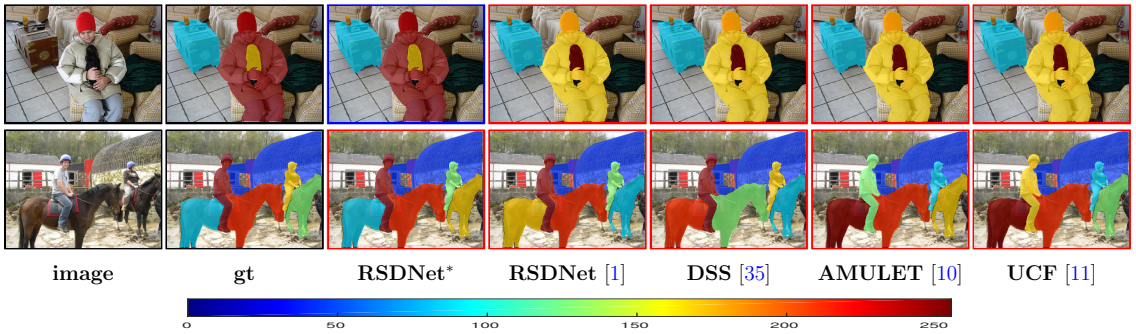


Figure 4.4: Qualitative illustration of rank order of salient instances on Pascal-S dataset. Relative rank is indicated by the assigned color.

the more limited number of training samples, baseline networks achieve a slightly lower ranking score from version II compared to version I, which implies and interesting trade-off between signal-to-noise in the data (version I) and sheer volume of data. Also the two different settings (relative vs absolute) maintain a comparable performance across models and metrics. Further, as shown in Table 4.5, the SOR_{max} ranking strategy shows higher scores for all baselines under different settings, demonstrating value in considering different vantage points in how assignments of rank are derived from an underlying saliency map.

*	Methods	Relative			Absolute		
		SOR_{avg}	SOR_{pow}	SOR_{max}	SOR_{avg}	SOR_{pow}	SOR_{max}
ver I	RSDNet [1]	0.736	0.727	0.767	0.732	0.726	0.761
	PSPNet + NRSS [51]	0.720	0.709	0.764	0.730	0.725	0.770
	DeepLabv2-VGG + NRSS [42]	0.708	0.710	0.753	0.716	0.713	0.750
ver II	RSDNet [1]	0.715	0.693	0.758	0.709	0.696	0.745
	PSPNet + NRSS [51]	0.689	0.682	0.754	0.695	0.692	0.771
	DeepLabv2-VGG + NRSS [42]	0.680	0.682	0.742	0.676	0.688	0.727

Table 4.5: Saliency ranking performance comparison for different methods subject to relative and absolute ranking settings on our COCO-SalRank dataset.

Fig. 4.5 shows a visual comparison of saliency ranking on COCO-SalRank dataset with respect to different baselines. We can see that the baselines can predict rank or-

der of salient objects quite accurately and this produces output closer to ground-truth maps in various challenging cases. Recall that each model is paired with the nested relative salience stack (NRSS) [1] at each slice and provides distinct representations to differentiate between multiple salient objects and allows for reasoning about their relative salience to take place.

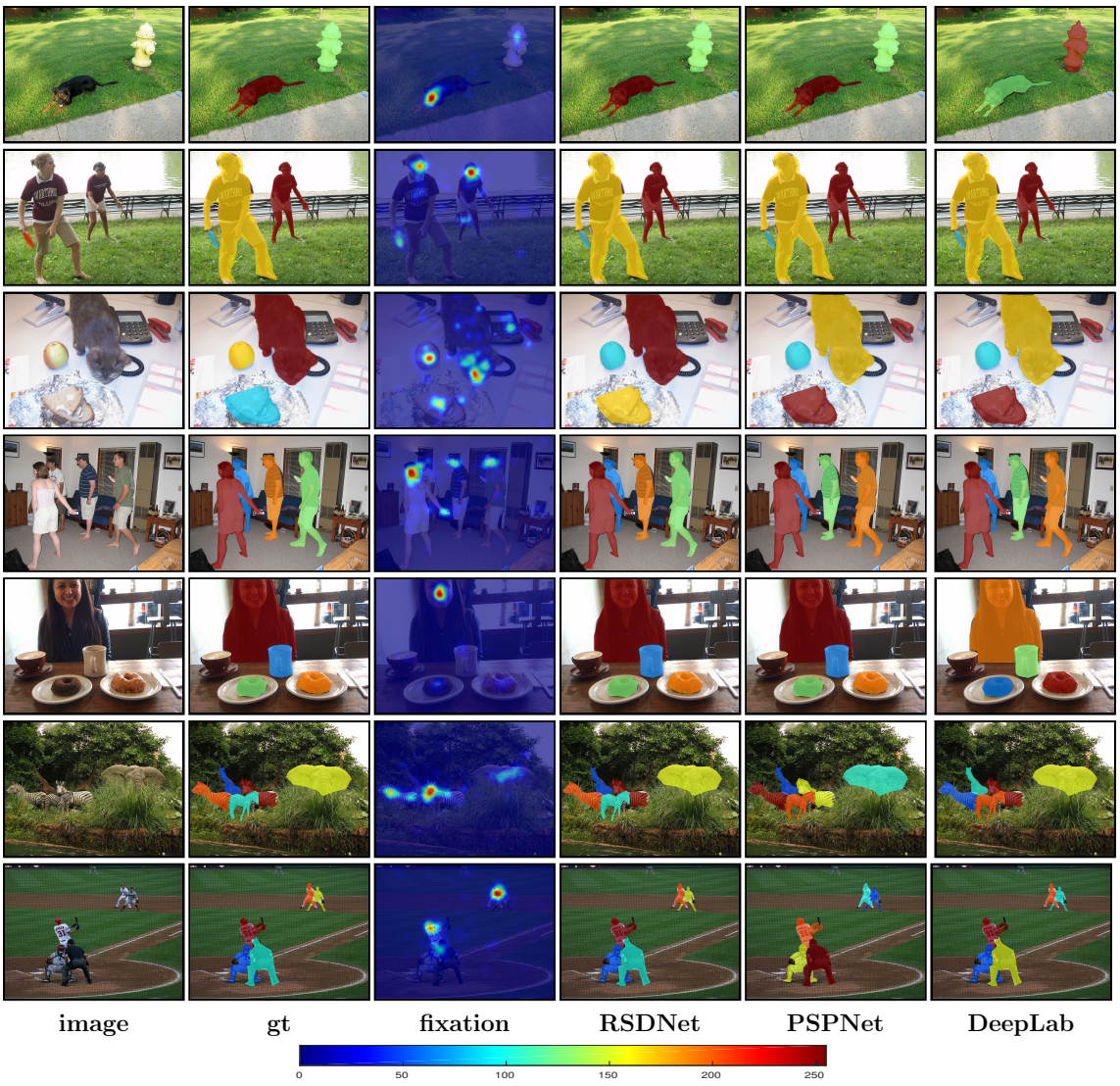


Figure 4.5: Qualitative illustration of rank order of salient objects on COCO-SalRank dataset. Relative rank is indicated by the assigned color.

4.4.5 Detection Evaluation Under Relative and Absolute Rank

To further evaluate the generalization capability and reliability of our proposed dataset, we use the predicted saliency maps for the traditional problem of salient object detection. Note that we do not explicitly train the networks for detection; instead we use the same predicted saliency map produced by networks trained for the ranking problem. We start with performing comprehensive analysis on a subset of the Pascal-S dataset to demonstrate the value of our dataset on multiple salient object detection with performance values shown in Table 4.6. Similar to (Sec. 4.4.4), we first train a baseline network (RSDNet) with our dataset (version I) under the relative rank setting and test on Pascal-S. Further, in RSDNet‡ we fine-tune the trained model with the training subset of Pascal-S and evaluate on the test set. From Table 4.6, we can see that RSDNet‡ outperforms RSDNet in terms of all metrics with a reasonable margin. Fig. 4.6 depicts a visual comparison of baseline methods along with RSDNet‡. Considering the different strategy for assigning relative salience on the Pascal-S dataset, the baseline models achieve comparable numbers only when trained on our dataset. RSDNet‡ achieves better performance on all the metrics compared to the RSDNet trained on the subset of Pascal-S.

$*$	RSDNet*	RSDNet‡	RSDNet†	RSDNet [1]	UCF [35]	Amulet [35]	DSS [35]	NLDF [35]	DHS [19]
max-F_m	0.849	0.889	0.855	0.873	0.858	0.865	0.841	0.846	0.837
med-F_m	0.817	0.855	0.819	0.854	0.840	0.854	0.838	0.843	0.833
avg-F_m	0.787	0.834	0.789	0.834	0.813	0.841	0.830	0.836	0.822
AUC	0.963	0.977	0.962	0.972	0.959	0.957	0.918	0.933	0.927
MAE	0.092	0.082	0.095	0.091	0.123	0.097	0.099	0.099	0.092

Table 4.6: Quantitative comparison of baseline methods including max F-measure (higher is better), average F-measure, AUC, and MAE (lower is better) on Pascal-S dataset.

Figure 4.6: Predicted salient object regions for the Pascal-S dataset. Each row shows outputs corresponding to different algorithms designed for the salient object detection/segmentation task.

Furthermore, we use precision-recall (PR) curves, and Area under ROC curve (AUC) metrics to measure detection performance as shown in Fig. 4.7. Our proposed dataset improves the max F-measure by a considerable margin on the Pascal-S dataset (RSDNet‡) which indicates that our dataset is general enough that it helps to achieve higher precision with higher recall. Also the baseline trained with our dataset achieves higher area under the ROC curve (AUC) score compared to the baselines shown.

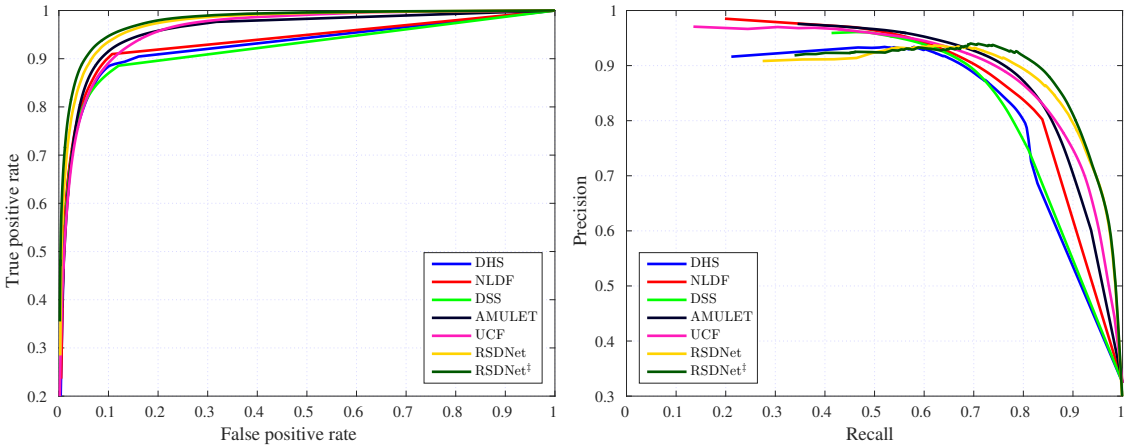


Figure 4.7: Left: ROC curves corresponding to different methods on the PASCAL-S dataset. Right: Precision-Recall curves for saliency ranking corresponding to different baselines.

For further analysis, we report detection performance for baseline models on our dataset (version I & version II). Table 4.7 shows the comparison of detection score for

baseline methods in terms of F-measure, AUC, and MAE. Similar to saliency ranking performance, we see consistent performance across different versions, metrics, and ranking settings.

*	Methods	Relative					Absolute				
		max-F_m	med-F_m	avg-F_m	AUC	MAE	max-F_m	med-F_m	avg-F_m	AUC	MAE
ver I	RSDNet [1]	0.780	0.740	0.705	0.936	0.135	0.783	0.718	0.663	0.944	0.158
	PSPNet + NRSS [51]	0.796	0.754	0.728	0.944	0.136	0.794	0.764	0.737	0.944	0.132
	DeepLab + NRSS [42]	0.743	0.688	0.649	0.917	0.162	0.735	0.682	0.631	0.911	0.166
ver II	RSDNet [1]	0.843	0.803	0.777	0.954	0.138	0.832	0.801	0.762	0.953	0.149
	PSPNet + NRSS [51]	0.850	0.817	0.792	0.957	0.13	0.845	0.824	0.79	0.956	0.133
	DeepLab + NRSS [42]	0.805	0.760	0.729	0.934	0.165	0.800	0.758	0.692	0.932	0.175

Table 4.7: Quantitative comparison of baseline methods including max, median and average F-measure (higher is better), AUC, and MAE (lower is better) on our proposed dataset (ver I & ver II) under relative and absolute ranking settings.

4.4.6 Effect of Nested Relative Salience Stack on COCO-SalRank

Inspired by RSDNet [1], we show comparison of slices of the nested relative salience stack for relative and absolute cases through principle component analysis (PCA). This tends to reveal regions where greatest variability exists in NRSS and has some diagnosticity for relative salience. Fig. 4.8 depicts the top three principle component which are most likely to capture highest variance across slices.

4.4.7 Cross-Dataset Evaluation

To further investigate the role and value of the proposed saliency ranking dataset in detail, we conduct cross-dataset evaluation on two different respective versions of the dataset (*noisy* & *clean*) under relative ranking setting. First, we report saliency ranking performance of baselines for all the alternative metrics in Table 4.8. As shown

Figure 4.8: Visualization of Principal Component Analysis (PCA) for the final prediction stack (NRSS) for both the relative (left set) and the absolute (right set) cases. For each set, the first column shows the image and its ground truth. A selection of ground truth stack slices is shown in second and third column. The top three principal components for our predicted stack is visualized as an RGB column.

in Table 4.8, the set of baseline methods achieves better performance in the scenario (train:(vI), test: (vII)) for relative ranking settings when compared with the numbers reported in Table 4.5. However, when we use the clean version (vII) for training, and evaluate the model on the noisy version (vI), all the baselines achieve a lower saliency ranking performance compared to the reported number in Table 4.5. Fig. 4.9 shows the qualitative comparison of cross-dataset evaluation for the two different scenarios. This analysis hints at strengths of the proposed saliency ranking dataset in it's value for boosting performance in training algorithms for saliency ranking, and also for validation and testing. We further show the detection performance of

Methods	train:(vI), test: (vII)			train:(vII), test: (vI)		
	SOR_{avg}	SOR_{pow}	SOR_{max}	SOR_{avg}	SOR_{pow}	SOR_{max}
RSDNet [1]	0.721	0.696	0.780	0.727	0.724	0.753
PSPNet + NRSS [51]	0.697	0.684	0.781	0.700	0.706	0.729
DeepLabv2-VGG + NRSS [42]	0.680	0.680	0.752	0.709	0.715	0.748

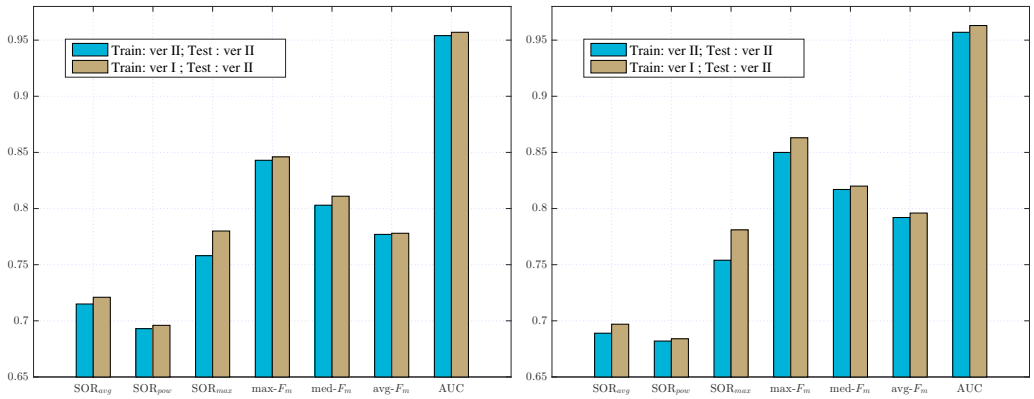
Table 4.8: Saliency ranking performance comparison for different methods with respect to cross-dataset evaluation under the relative ranking setting.

different baseline architectures (using NRSS) under similar settings on our proposed dataset, and report the results in Table 4.9. In analyzing the two tables, we see a

clear demonstration of performance up/down which reveals the effectiveness of the *noisy* version of the dataset.

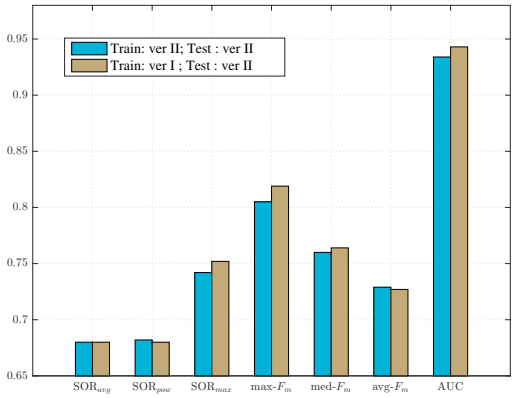
Methods	train:(vI), test: (vII)					train:(vII), test: (vI)				
	max-F_m	med-F_m	avg-F_m	AUC	MAE	max-F_m	med-F_m	avg-F_m	AUC	MAE
RSDNet [1]	0.846	0.811	0.778	0.957	0.14	0.754	0.712	0.681	0.921	0.141
PSPNet + NRSS [51]	0.863	0.82	0.796	0.963	0.126	0.762	0.726	0.702	0.923	0.144
DeepLab + NRSS [42]	0.819	0.764	0.727	0.943	0.167	0.708	0.662	0.629	0.894	0.169

Table 4.9: Quantitative comparison of baselines including max, median, and average F-measure, AUC, and MAE with respect to cross dataset evaluation under relative ranking setting.



(a) RSDNet

(b) PSPNet



(c) DeepLabv2-VGG16

Figure 4.9: Comparison of two different training and testing scenarios with respect to ranking and detection metrics.

Chapter 5

Conclusion and Future Work

In this thesis, we have presented a neural framework for detecting, ranking, and subitizing multiple salient objects that introduces a stack refinement mechanism to achieve better performance. Central to the success of this approach, is how to represent relative saliency both in terms of ground truth, and *in network* in a manner that produces stable performance. We highlight the fact that to date, salient object detection has assumed a relatively limited, and sometimes inconsistent problem definition. Comprehensive experiments demonstrate that the proposed architecture outperforms state-of-the-art approaches across a broad gamut of metrics. Moreover, we have considered the problem of saliency ranking from a data driven standpoint and in doing so have presented a large scale benchmark dataset for saliency ranking. We have also introduced a novel set of methods that make use of existing gaze or related data paired with object annotations to generate large rank ordered data for salient objects. We validate the proposed benchmark by applying state-of-the-art saliency models, and demonstrate the value of a nested representation in defining a relative

ranking. We also show that using our dataset, all models can achieve higher performance across a set of metrics for salient object ranking, in addition to traditional detection tasks.

Many interesting research questions arise from the approach and results presented in this thesis. One avenue for further investigation is to inspect different representations of the ground truth stack and their impact on the overall performance of both saliency ranking and detection tasks. Another interesting avenue to examine is the way to assign the rank order based on the predicted saliency maps. Although some solutions were proposed in this thesis, our main goal is not to find the best way to assign rank order. Thus, further exploration of different methods to assign rank order will also be a fruitful path for further research.

Bibliography

[1] M. A. Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *CVPR*, 2018. viii, 9, 10, 11, 42, 43, 44, 49, 50, 51, 52, 53, 54, 56, 57, 58

[2] S. Manen, M. Guillaumin, and L. Van Gool, "Prime object proposals with randomized prim's algorithm," in *ICCV*, 2013. xv

[3] W. S. Jevons, "The power of numerical discrimination," 1871. xv

[4] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann, "The discrimination of visual number," *The American journal of psychology*, vol. 62, no. 4, pp. 498–525, 1949. xv

[5] X. Ren and J. Malik, "Learning a classification model for segmentation," in *ICCV*, 2003. xvi

[6] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013. 1, 7, 10, 25, 26, 33, 43

[7] K. Shi, K. Wang, J. Lu, and L. Lin, "Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *CVPR*, 2013. 1, 10, 43

[8] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013. 1, 10, 43

[9] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014. 1, 2, 4, 10, 20, 41, 43, 50

[10] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *ICCV*, 2017. 1, 8, 10, 15, 16, 25, 26, 30, 33, 52

[11] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, 2017. 1, 8, 10, 25, 26, 30, 33, 52

[12] S. He, J. Jiao, X. Zhang, G. Han, and R. W. Lau, "Delving into salient object subitizing and detection," in *ICCV*, 2017. 1, 9, 10, 19, 29, 45

[13] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017. 1, 8, 10, 15, 16, 25, 26, 30, 33

[14] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016. 1, 8, 10

[15] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *CVPR*, 2017. 1, 10, 15, 16

[16] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *CVPR*, 2017. 1, 7, 10

[17] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016. 1, 7, 10

[18] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016. 1, 7, 10, 25, 26, 33

[19] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016. 1, 8, 10, 13, 15, 16, 25, 26, 33, 51, 54

[20] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015. 1, 7, 10, 25, 26, 33

[21] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015. 1, 7, 10, 25, 26, 33

[22] S. Jia, Y. Liang, X. Chen, Y. Gu, J. Yang, N. Kasabov, and Y. Qiao, "Adaptive location for multiple salient objects detection," in *NIPS*, 2015. 1, 10

[23] M. Najibi, F. Yang, Q. Wang, and R. Piramuthu, "Towards the success rate of one: Real-time unconstrained salient object detection," *arXiv:1708.00079*, 2017. 1, 10, 45

[24] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Unconstrained salient object detection via proposal subset optimization," in *CVPR*, 2016. 1, 10

[25] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *CVPR*, 2017. 2

[26] K. Koehler, F. Guo, S. Zhang, and M. P. Eckstein, "What do saliency models predict?" *Journal of vision*, vol. 14, no. 3, pp. 14–14, 2014. 2

[27] N. D. Bruce, C. Wloka, N. Frosst, S. Rahman, and J. K. Tsotsos, "On computational modeling of visual saliency: Examining what's right, and what's left," *Vision research*, vol. 116, pp. 95–112, 2015. 2, 9

[28] A. Açık, A. Bartel, and P. Koenig, "Real and implied motion at the center of gaze," *Journal of Vision*, vol. 14, no. 1, pp. 2–2, 2014. 2

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, 2014. 3, 11, 35, 37

[30] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *CVPR*, 2015. 3, 10, 35, 37

[31] D. Rolnick, A. Veit, S. Belongie, and N. Shavit, "Deep learning is robust to massive label noise," *arXiv preprint:1705.10694*, 2017. 4

[32] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *ICCV*, 2013. 7, 25, 26, 33

[33] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *CVPR*, 2014. 7, 25, 26, 33

[34] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *TIP*, 2016. 7, 25, 26, 33

[35] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *CVPR*, 2017. 7, 8, 25, 26, 33, 51, 52, 54

[36] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech, "Salient object subitizing," in *CVPR*, 2015. 9, 19, 20, 29, 31, 45

[37] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *ICCV*, 2015. 13

[38] M. A. Islam, M. Rochan, N. D. Bruce, and Y. Wang, "Gated feedback refinement network for dense image labeling," in *CVPR*, 2017. 13, 14, 16, 17

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016. 14, 19, 22

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv:1409.1556*, 2014. 14, 15

[41] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *NIPS*, 2015. 14

[42] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *TPAMI*, vol. 40, no. 4, pp. 834–848, 2018. 14, 15, 23, 50, 52, 56, 57, 58

[43] G. Huang, Z. Liu, L. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017. 15

[44] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015. 16

[45] M. A. Islam, S. Naha, M. Rochan, N. Bruce, and Y. Wang, "Label refinement network for coarse-to-fine semantic segmentation," *arXiv preprint arXiv:1703.00551*, 2017. 16

[46] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv:1408.5093*, 2014. 22

[47] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *IJCV*, vol. 88, no. 2, pp. 303–338, 2010. 23

[48] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013. 25, 26, 33

[49] X. Li, Y. Li, C. Shen, A. Dick, and A. Van Den Hengel, "Contextual hypergraph modeling for salient object detection," in *ICCV*, 2013. 25

[50] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results," http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html. 29

[51] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *CVPR*, 2017. 50, 52, 56, 57, 58