Scene Context-Aware Salient Object Detection

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

Abstract

Salient object detection identifies objects in an image that grab visual attention. Although contextual features are considered in recent literature, they often fail in real-world complex scenarios. We observe that this is mainly due to two issues: First, most existing datasets consist of simple foregrounds and backgrounds that hardly represent real-life scenarios. Second, current methods only learn contextual features of salient objects, which are insufficient to model high-level semantics for saliency reasoning in complex scenes. To address these problems, we first construct a new large-scale dataset with complex scenes in this paper. We then propose a context-aware learning approach to explicitly exploit the semantic scene contexts. Specifically, two modules are proposed to achieve the goal: 1) a Semantic Scene Context Refinement module to enhance contextual features learned from salient objects with scene context, and 2) a Contextual Instance Transformer to learn contextual relations between objects and scene context. To our knowledge, such high-level semantic contextual information of image scenes is under-explored for saliency detection in the literature. Extensive experiments demonstrate that the proposed approach outperforms state-of-the-art techniques in complex scenarios for saliency detection, and transfers well to other existing datasets. The code and dataset are available at https://github.com/SirisAvishek/Scene_Context_Aware_Saliency.

1. Introduction

Salient object detection explores the problem of identifying objects that “pop-out” and grab visual attention in an image or video. The task has been widely used as a preprocessing step for many vision applications such as image/video compression [59, 11], video object segmentation [50], image captioning [56], and image parsing [21]. All these vision tasks are proposed for real-world images with complex scenes.

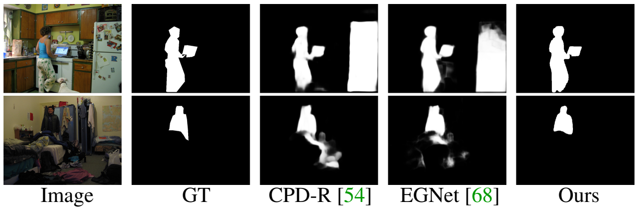

Figure 1: Examples of real-world complex scenarios where existing methods (e.g. [54, 68]) may not capture semantic scene contexts well, leading to incorrect detection of distractors. Whereas our model is able to capture semantic contexts of the scenes.

Recently, saliency research has grown rapidly through the success of CNNs, which are able to capture better feature representations compared to hand-crafted features [32, 5, 57]. State-of-the-art saliency models mainly extract and aggregate contextual information from spatial relations, including multi-scale and local-global features in various manners [67, 30, 37]. Although good performance has been demonstrated, these models are mainly trained on binary saliency labels that are class-agnostic. Training on such labels only can limit the ability of networks to learn semantic contextual features (higher-level understanding) that would otherwise help model various relationship of objects within complex image scenes [29, 65]. Fig. 1 shows two examples of real-world complex scenarios where existing models perform poorly. The top row shows a kitchen scene with a salient person and a distractor (*e.g.*, fridge with similar texture). Existing models can not capture the semantic knowledge of the distractor and are unable to differentiate it from the person’s attire, resulting in incorrect saliency for that distractor. It is a similar case for the bottom row which involves a bedroom scene with a salient person surrounded by many distractors (*e.g.*, clothes and objects with similar texture). The above observations motivate us to ask the question: *Can we learn and use discriminative semantic context to improve saliency modeling in challenging complex scenes with rich context?*

Further to our observations, we find that most of the ex-

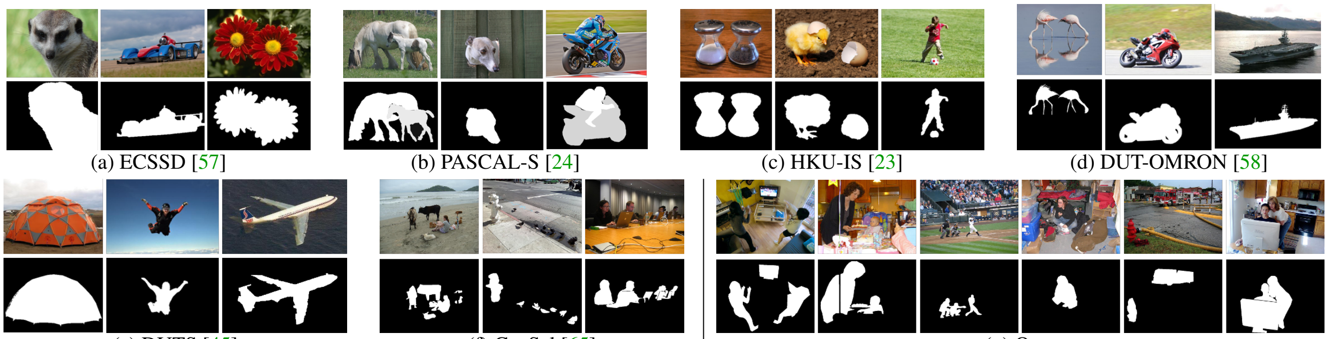
(a) ECSSD [57]
(b) PASCAL-S [24]
(c) HKU-IS [23]
(d) DUT-OMRON [58]
(e) DUTS [45]
(f) CapSal [65]
(g) Ours

Figure 2: Comparison between existing datasets and the proposed new challenging dataset. Existing popular salient object datasets (a – f) are not very challenging. In contrast, our proposed dataset (g) contains more complex scenes due to the increase in the number of objects in the foreground/background as well as non-salient distractors.

isting salient object detection datasets [36, 4, 57, 24, 23, 58, 45] consist of images with few objects and simple backgrounds. Example images from these datasets are illustrated in Fig. 2(a-f). These images are relatively simple for salient object detection in the wild, where images are typically complex with lots of objects and complex backgrounds, as shown in Fig. 2(g).

Psychological studies suggest that semantic scene context influences eye movements and attention [43], revealing the relationship between salient objects and the surrounding image scenes. To the best of our knowledge, saliency detection with high-level scene context and spatial context is under-explored, with only two related works [65, 29] addressing a similar problem. Zhang *et al*. [65] propose to leverage captions as the semantic scene context for improving salient object prediction. However, reliance on generated captions can be detrimental to saliency prediction, especially if they are incorrect. On the other hand, DSCLRCN [29] derives their scene context features from an image-level scene classification model, whereas the extracted features are too abstract, containing only an overall representation without capturing object relationships within the scene.

The above-mentioned limitations further motivate us to explore the use of semantic scene and spatial context for salient object detection in real-world scenarios with complex scenes. To this end, we first construct a novel dataset comprising of images with rich context (more details in Sec. 3). We then propose a context-aware saliency modeling framework to leverage semantic scene context features. Specifically, we introduce Instance Context Segmentation and Stuff Context Segmentation to semantically segment *Things* and *Stuff*. These two components perform panoptic segmentation on the whole scene, providing detailed semantics of a given image. However, we find that not all the semantic information play an effective role in defining the semantic scene context of an image. As a result, we propose a novel *Semantic Scene Context Refinement* (SSCR) module to fuse and augment information of salient object features with surrounding semantic scene context for improving saliency reasoning. To further exploit semantic scene context, we propose a *Contextual Instance Transformer* (CIT) to capture the relationship between objects and the scene context.

In summary, our main contributions include:

- We propose a semantic scene context-aware framework for salient object detection, which explores the semantic relationship between salient objects and the scene context.
- We propose a *Semantic Scene Context Refinement* module to extract and enhance semantic scene context features that are highly related to the image scene. We further propose a new *Contextual Instance Transformer* to learn the contextual relations between objects and scene context for saliency reasoning.
- We build (and will make available) a new salient object detection dataset with real-world complex scenes to consider semantic scene contexts.
- Extensive experiments demonstrate that the proposed approach outperforms the state-of-the-art methods on our dataset and also generalizes well to existing datasets.

2. Related work

2.1. Salient Object Detection

Early traditional saliency work were mainly based on computational methods that combined low-level features [16]. However, in this paper we focus on deep learning based methods. Previous deep learning salient object detection methods used multi-layer perception (MLP) to predict a saliency score for each pixel in an image [13, 23, 69, 22]. Though these MLP-based models outperform traditional hand-crafted saliency methods [52, 5], they were unable to capture spatial information effectively due to the use of fixed fully connected layers. Later methods tackled this issue by utilizing fully convolutional networks (FCNs) [34], building their success on semantic segmentation.

Many of the recent deep learning based saliency models were built on the FCNs with various strategies to combine multi-scale contextual information. They mostly embedded modules for extracting and aggregating context features from different layers in the network [64, 55, 9, 37, 71, 63]. Typically, they employed side outputs from different layers in their encoders, and aggregated those side outputs with the layers in their decoders [66, 14, 46, 49, 68, 51, 72, 62]. Su *et al*. [41] further extracted multi-scale contextual features using varying dilated convolutions, while recurrent blocks were applied in [28] and [48]. Subsequent works proposed to explicitly combine local and global contextual features through (a) the use of separate networks [60, 3], (b) additional convolutions after the final convolutional layer in an encoder [35, 33] and (c) adopting a Pyramid Pooling Module [47, 27]. Attention mechanisms have also been exploited to enhance multi-scale contextual features by capturing the interaction between pixels in local [20] and global contexts [30, 39, 70, 15].

Recently, Zeng *et al*. [61] proposed to unify the task of salient object detection and weakly-supervised semantic segmentation. They introduced a saliency aggregation module that used saliency scores to weight corresponding semantic segmentation, in order to generate the final saliency map. Aydemir *et al*. [1] used object detection to produce dissimilarity scores based on visual appearance and relative size, so as to enhance the saliency contrast among objects. Liu *et al*. [31] employed Transformers [44] to propagate context among image patches.

Although the above-mentioned saliency methods have shown significant improvements, they still struggle with complex scene images that are rich in semantic context. These methods mainly learn a limited scope of discriminative spatial context features in multiple scales. Networks are generally trained with binary saliency annotations only, and fail to effectively learn high-level semantics. We address this problem by including a joint task of segmentation to identify *Things* and *Stuff* in the scene explicitly. This allows our saliency network to explore semantic scene context, and to enhance the saliency reasoning of multiple objects and their relations in complex scenarios.

2.2. Semantic Scene Context in Saliency

Gist features are considered as an abstract low-level scene representation. Torralba *et al*. [42] combined scene representation from holistic low-dimensional encoding with low-level saliency in a statistical framework for modelling attention. Peter *et al*. [38] proposed a technique to learn the mapping between low-level gist features and recorded eye movements during video gameplay. Judd *et al*. [18] combined low to high level features to model attention. They use horizontal lines detector trained from mid-level gist features as their mid-level features.

High-level semantic scene context is mostly under-explored for saliency detection. Liu and Han [29] proposed to use an existing scene classification network for extracting scene context features. Zhang *et al*. [65] encoded scene context by using a captioning network to capture the "major" objects in a scene. In our work, we employ semantic segmentation for capturing high-level semantic scene context features. While [29] mainly captured an overall representation of a scene, we obtain much more detailed semantic information from a scene through our segmentation. Their target task was for eye fixation points prediction, whereas, we focus on salient object detection. Additionally, we explore the semantic relationships between all the objects in a given scene, while [65] is limited to those objects mentioned in the captions only.

Goferman *et al*. [10] define a new interpretation for saliency, by introducing GT background regions (context) with a GT salient object based on image description. Unlike [10], we consider *Things* and *Stuff* context segmentation as an auxiliary task to obtain our scene context features.

3. Proposed Dataset

As aforementioned, existing salient object datasets mostly contain images that do not well represent real-world scenes. The CapSal [65] dataset contains real-world images, however the ground-truth salient objects are often heavily biased towards the caption data. The consequence is that all objects that relate to the caption are often considered salient, regardless of whether each of those objects are individually visually salient or not (see Fig. 2(f)).

We propose a new dataset to support the modeling of saliency in real-world scenes containing rich semantic context. Our dataset is based on MS-COCO [26] and SALICON [17]. MS-COCO provides images of challenging scenarios and annotations of semantic segmentation of object instances (*Things*) and regions (*Stuff*). SALICON provides the mouse-based fixation sequence of respective images. Our dataset is constructed in two phases: 1) automatic ground-truth saliency generation and image filtering, and (2) manual image filtering.

(1) Automatic phase. We automate the ground-truth salient objects generation based on the observations in [8], where Fosco *et al.* found that humans generally gaze at people during [0,0.5] seconds, and then move towards other objects during [0.5,3] seconds. After the first 3 seconds, there are more fixations on *Stuff* regions. Based on these observations, we collect salient objects if the SALICON fixation points, in the range [0,3] seconds, fall on the MS-COCO annotations of an object segmentation in an image. An object is further labelled as ground-truth salient if more than

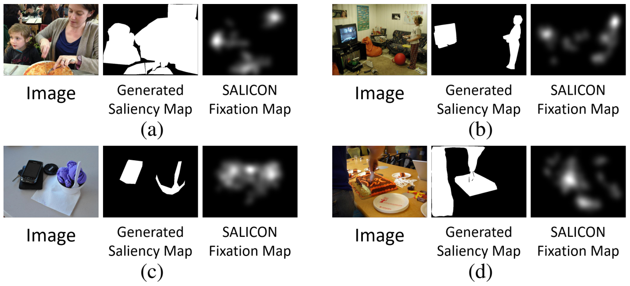

Figure 3: Examples of visual comparison during the manual image filtering (Phase – 2) when we construct the proposed dataset. In (a) and (b), the generated saliency maps are comparable to the corresponding SALICON fixation maps, which are kept in our final dataset. Whereas those images with larger discrepancies, as shown in (c) and (d), are removed from the dataset.

half of the observers fixate on this object. Once ground-truth saliency is generated for all available 15,000 images, an automated filtering step is applied to ensure the images are complex and contain rich context. It is a trade-off between complexity of images and the number of resultant images. We find that a minimum of 4 objects and at least 2 object categories per image produce a good set of complex images, whilst retaining a higher number of images in the constructed dataset.

The above automated step however may run into issues when the annotations of foreground and background objects overlap. First, in some of the images in the MS-COCO dataset (*e.g*., food on bench / dining table), we observe that the background object (*e.g*., bench, dining table) are often large objects. They are incorrectly considered salient simply because the fixations fall on both the foreground and background objects. Second, some objects (*e.g*., car, train) are easy to collect fixation as they cover a large portion of the background in some images (*e.g*., a person in a car), but they are clearly not salient (*e.g*., compared to the foreground person). In the former cases, since MS-COCO does not provide depth information, we manually go through the dataset, identify those background categories and exclude them for saliency ground-truth generation. In the latter cases, we also omit all objects where its area is larger than 60% (a threshold we empirically decided) of the image. These steps are carried out as a pre-filtering step before the above automated process. They ensure that large objects (typically background) are not given saliency scores.

(2) Manual phase. To ensure the quality of the constructed dataset, we manually inspect if the generated salient object map from phase-(1) is consistent with the corresponding SALICON fixation map, following a similar procedure in [65]. Specifically, we would like to ensure that the peak fixations in SALICON also land on objects that are identified as salient in our generated saliency maps. For example, Fig. 3(c) and 3(d) show two images that are removed, because there are large discrepancies between the peaks of fixation maps and the chosen salient objects in the generated saliency maps. This step removes inconsistent annotations that may arise from the automatic process.

Table 1: Comparison of the average number of objects and object categories per image among existing datasets and our dataset.

Dataset	#Avg. Obj.	#Avg. Obj. Cat.
ECSSD [57]	1.32	1.28
PASCAL-S [24]	2.08	1.80
HKU-IS [23]	2.12	1.68
DUT-OMRON [58]	1.44	1.24
DUTS [45]	1.56	1.36
Ours	12.79	4.62

After the two phases, our final dataset consists of 5,534 training and 2,554 testing images. Table 1 compares the average number of objects and categories per image in existing datasets and our dataset. Existing datasets do not provide object segmentation or category data. Therefore, we report the statistics in the table by sampling 25 images randomly from each dataset and manually counting the objects and categories. It shows that our dataset contains images with a higher count of objects and categories that is much closer to real-world scenes. More details of the above-mentioned dataset creation steps, and further statistics of our dataset are provided in the supplementary material.

4. Proposed Method

In this section, we first introduce the backbone (Sec. 4.1) of our network and discuss how contextual features are extracted and utilized. Then we specify how the proposed modules (Sec. 4.2 and Sec. 4.3) take advantage of the contextual features, in order to refine and augment features for saliency. Finally, we detail the Salient Instance Network (Sec. 4.4) for the task of salient object detection. An overview of the proposed framework is illustrated in Fig. 4.

4.1. Backbone

Our network is built on the Mask-RCNN [12] architecture and we extract multi-scale features from the FPN [25]. We utilize the multi-scale features as input for 3 operations, namely, (1) object proposal, (2) context segmentation and (3) context feature refinement.

(1) Object proposal. We apply RPN and RoIAlign [12] on the multi-scale features to generate object instance proposals and corresponding object features.

(2) Context segmentation. We include a Shared Context Segmentation Decoder for Instance and Stuff Context Segmentation, in order to extract semantic context features for a given scene. The decoder take the multi-scale features as

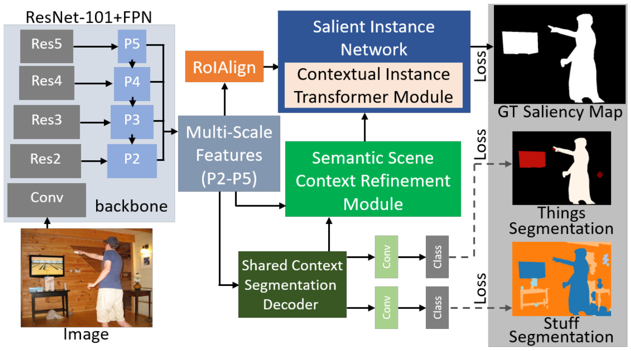

Figure 4: An overview of the proposed network. Our model extracts semantic features from the Shared Context Segmentation Decoder. The decoder is trained to reconstruct features for generating *Things* and *Stuff* categories. Our Semantic Scene Context Refinement (SSCR) module (Sec. 4.2) then utilizes the semantic features and multi-scale features to build the augmented scene context features, correlating the semantics of an image. Our Contextual Instance Transformer (CIT) module (Sec. 4.3), inside the Salient Instance Network, learns relationships between objects and scene context, and enhance saliency reasoning.

input and reconstructs features for segmentation of *Things* and *Stuff* categories. From the decoder we extract semantic features $f^C \in \mathbb{R}^{\frac{W}{4} \times \frac{H}{4} \times 128}$, where $W \times H$ is the spatial dimensions for image I. The decoder follows similar design as in [19] to combine the multi-scale features for segmentation. The output features of the decoder are then passed through two separate convolution layers for generating *Things* and then *Stuff* context segmentation.

(3) Context feature refinement. Third, we combine the multi-scale features with the features extracted from context segmentation **(2)**, producing the refined scene context features for boosting saliency reasoning.

These context features are used by the proposed SSCR (Sec. 4.2) and CIT (Sec. 4.3). SSCR builds the final scene context features by aggregating only useful context information. CIT learns the relationships between the scene context features and object features. The final salient object classification is detailed in Sec. 4.4.

4.2. Semantic Scene Context Refinement (SSCR)

Previous works suggest that not all context information (*e.g*., distractors) is relevant and useful to the final prediction task [30, 48, 67]. To address this problem we design this module to enhance the semantic information that has strong correlation to the scene context. This allows the network to augment contextual information learned only from saliency annotations with strong semantic scene context.

We build our semantic scene context features by refining the context features f^C obtained from the context segmentation decoder and multi-scale features (Fig. 5). We only use feature levels [P3, P4, P5] from the multi-scale features, as these levels contain higher-level contextual features [70].

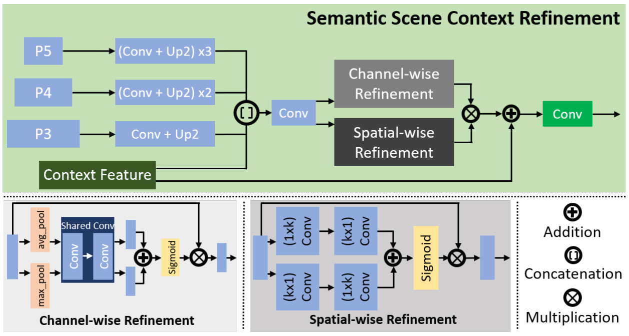

Figure 5: Illustration of the Semantic Scene Context Refinement (SSCR) module.

The three levels of multi-scale features are applied with operations similar to those in context segmentation, resulting in features $p^3, p^4, p^5 \in \mathbb{R}^{\frac{W}{4} \times \frac{H}{4} \times 128}$. We fuse these features into $f^F \in \mathbb{R}^{\frac{W}{4} \times \frac{H}{4} \times 128}$ by concatenating the multi-scale features with context features and applying a (1×1) convolution layer. The concatenation helps suppress saliency distractors by utilising the scene context information. Next, we refine f^F in a channel-wise and spatial-wise manner.

Channel-wise refinement. In CNN design, typically different semantic information are activated by distinct channel features [2, 67]. We include a channel-wise attention mechanism to weight channel features that are more relevant to semantic information. Given a set of features, we apply average pooling, max pooling and two convolution layers with a ReLU and sigmoid activation. We then multiply the processed features (i.e., C_a, C_m) with the original feature x as:

$$\begin{aligned} CR(x) &= x \times Sigmoid(C_a(x) + C_m(x)) \\ C_a(x) &= conv_2(ReLU(conv_1(avgpool(x), W_1)), W_2) \\ C_m(x) &= conv_2(ReLU(conv_1(maxpool(x), W_1)), W_2) \end{aligned} \tag{1}$$

where $x = f^F$. W_1 and W_2 represent the parameters of the two convolution layers.

Spatial-wise refinement. Similarly, we include spatial-wise attention that leverages useful spatial information. Given a set of features, we employ two sets of double convolution layers with alternating kernels, where one set contains kernels {1×k, k×1} and the other contains {k×1, 1×k}. The resulting features from the two sets of convolution layers are added and sigmoid activation is applied to generate a spatial attention map. We weight the original feature x with the attention maps (i.e., S_1, S_2) through multiplication:

$$\begin{aligned} SR(x) &= x \times Sigmoid(S_1(x) + S_2(x)) \\ S_1(x) &= x \times ReLU(conv_2(ReLU(conv_1(s, W_1)), W_2)) \\ S_2(x) &= x \times ReLU(conv_4(ReLU(conv_3(s, W_3)), W_4)) \end{aligned} \tag{2}$$

where (W_1, W_2) and (W_3, W_4) are the parameters of the two sets of convolution layers, with respect to $\{1\times$k, k$\times1\}$ and $\{$k$\times1$, $1\times$k$\}$ kernels. After performing channel- and spatial-wise refinement on the fused features f^F, we combine the two outputs with Hadamard Multiplication. The product is further fused with the original context features (f^C) by addition and a final convolution is applied. This generates our final semantic scene context features f_{sc}.

$$f_{sc} = ReLU(conv((CR(f^F) \times SR(f^F)) + f^C, W_{sc})) \quad (3)$$

where W_{sc} are the parameters of the final convolution layer. The process enables the enhancement of context from saliency features and scene context features, which are learned from salient object detection and context segmentation, respectively.

4.3. Contextual Instance Transformer (CIT)

It is observed in the literature that scene context influences eye movements [43]. However, most existing salient object detection methods do not model such high level understanding and relationships, not to mention, guiding saliency prediction in complex real-world scenes. As shown in Fig. 1, saliency of individual objects requires semantic information about other objects and scene context to infer the high-level relationships and to differentiate objects from distractors. This module aims to learn relationships between objects and scene context for saliency reasoning.

We adapt transformers [44] to learn the dependencies between individual object features and scene context features in object-to-object and object-to-context relationships. We divide the module into two parts (see Fig. 6). The first part is designed to learn relationships among the objects only, whereas, the second part learns relationships between individual objects and scene context. We use a scaled dot product attention layer with a single head on both types of relationship:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \quad (4)$$

where $\sqrt{d}$ refers to normalization based on the feature size. Q, K and V are matrices corresponding to Queries, Keys and Values. Specifically, Q is projected from object features, while K and V are generated from either object features or scene context features. Multiplication of Q and K, followed by a softmax, produces an output that represents the degree of correlation between the feature vectors in Q and K. This is then used to weight the information of objects represented by the latent features V:

$$\begin{aligned} T_{OO} &= Attention(W_{q1}F^{o'}, W_{k1}F^{o'}, W_{v1}F^{o'}) \\ T_{OC} &= Attention(W_{q2}F^{o'}, W_{k2}f_{sc}, W_{v2}f_{sc}) \end{aligned} \quad (5)$$

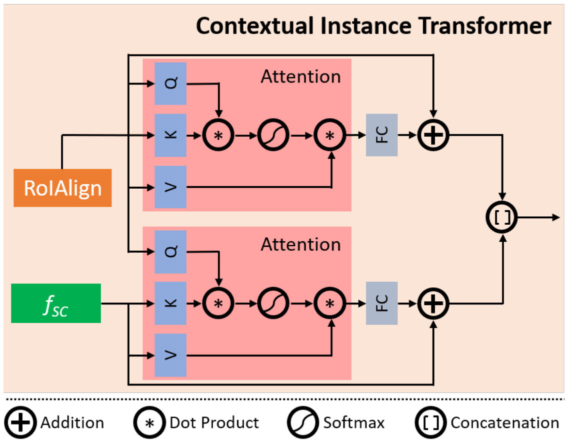

Figure 6: Illustration of the Contextual Instance Transformer (CIT) module.

where T_{OO} and T_{OC} are attention features modeling object-to-object and object-to-context relationships. $W_{\{q1,q2\}}$, $W_{\{k1,k2\}}$ and $W_{\{v1,v2\}}$ are parameters of fully connected and convolution layers for linear projection. $F^{o'}$ refers to object features from RoIAlign and one fully connected layer (Sec. 4.4). During the attention in T_{OC}, K and V are flattened to become 1-D vectors (same as the object features). We then apply a fully connected layer and residual connection to both attention features (T_{OO} and T_{OC}). For the residual connection applied to T_{OC}, we first average pool the scene context features (f_{sc}) to transform the features into a 1-D vector. Finally, the two object-to-object and object-to-context relationship features are concatenated for our subsequent saliency classification (Sec. 4.4).

4.4. Salient Instance Network

Salient Instance Network performs the main salient object classification task from input object features, allowing our method to perform saliency reasoning on the object-level. It adapts from the second stage of Mask-RCNN, which consists of networks for predicting object class, bounding box and mask segmentation. We modify the network for salient object detection and enhance saliency prediction with scene context, visualized in Fig. 7.

Our backbone (Sec. 4.1) generates candidate object features and predict their saliency. RPN and RoIAlign generate 2D features of individual object candidates. It is followed by a flatten and a fully connected layer. A feature vector, $f^o_i \in \mathbb{R}^{1024}$, is produced for each object, leading to a set of object features $F^o = \{f^o_1, f^o_2, \ldots, f^o_N\}$, where N=512 is the maximum number of object proposals. We obtain our final object features by fusing with object-to-object and object-to-context relationship information after employing CIT and two fully connected layers. A modified classification layer then determines the saliency of each object.

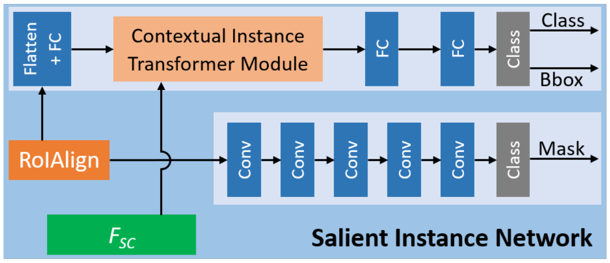

Figure 7: Illustration of the Salient Instance Network.

5. Experiments and Results

5.1. Dataset and Evaluation Metrics

Dataset. Our evaluation is mainly carried out on the proposed dataset. We use a training set of 5,534 images for training and 2,554 images for testing. Popular salient object datasets are not suitable for training our model (*e.g.*, no object instance and semantic segmentation annotations), we do not report evaluation on those datasets here. Instead, we refer readers to the supplementary materials, where we provide comparison results of our model with state-of-the-arts on existing datasets.

Evaluation metrics. We use three metrics namely, F-measure, Mean Absolute Error (MAE) and E-measure [6], to evaluate the performance of our model and state-of-the-arts. The F-measure provides a score of the overall performance in regards to the quality of the predicted saliency map. It is formulated by a weighted combination of Precision and Recall:

$$F_\beta = \frac{(1+\beta^2) Precision \times Recall}{\beta^2 Precision + Recall} \tag{6}$$

where β^2=0.3. MAE calculates the average per-pixel disparity between predicted and ground-truth saliency maps. E-measure computes the pixel-level and image-level errors simultaneously to measure binary foreground similarities.

5.2. Implementation Details

We use ResNet-101 pre-trained on MS-COCO [26] as part of our backbone. Our network share similar architecture and parameters with Mask-RCNN [12] and use the same loss functions for saliency prediction. We use cross entropy loss for both instance and stuff context segmentation networks. Our model is based on the detectron2 framework [53] and is trained on a single NVIDIA GTX 1080 Ti GPU, for 30 epochs. A SGD optimizer with initial learning rate 0.001 is used, along with weight decay (10^{-4}) and momentum (0.9). We apply random cropping, flipping and multi-scale image training for data augmentation.

Table 2: Quantitative comparison with state-of-the-art methods on our dataset. avgF refers to the average F-measure taken and E_m refers to E-measure. Red and blue indicate best and second best performances, respectively.

Method	avgF ↑	E_m ↑	MAE ↓
BASNet [40]	0.706	0.823	0.087
CapSal [65]	0.797	0.853	0.082
CPD-R [54]	0.803	0.854	0.074
PFANet [70]	0.676	0.772	0.131
S4Net [7]	0.625	0.720	0.149
EGNet [68]	0.815	0.863	0.067
SCRN [55]	0.786	0.842	0.076
ITSD [72]	0.776	0.854	0.070
LDF [37]	0.808	0.852	0.070
MINet [51]	0.810	0.861	0.067
Ours	0.849	0.872	0.062

5.3. Comparison with State-of-the-Arts

We compare against 9 state-of-the-art methods in salient objects detection, including BASNet [40], CapSal [65], CPD-R [54], PFANet [70], EGNet [68], SCRN [55], ITSD [72], LDF [51] and MINet [37]. Furthermore, we compare with S4Net [7] (salient instance segmentation), which also builds on the Mask-RCNN architecture like CapSal and our model. Note that in the comparison, CapSal is the only method not trained on our dataset (direct testing only with their pre-trained weights). CapSal requires GT captions data corresponding to GT saliency annotations. We also run into issues running their provided source code[1].

Quantitative evaluation. We report the experimental results comparing the proposed model with state-of-the-arts in Table 2. It shows that our model quite significantly outperforms existing state-of-the-arts across all metrics. In particular, our model show substantial improvement in the average F-measure, with a performance increase of 4.17% over the second best method.

Qualitative evaluation. We further showcase the performance of our model in Fig. 8, which displays visual comparisons between our model and 10 state-of-the-art methods. Our model is able to correctly pick out unique and interesting salient objects among multiple distractors by utilizing the context of image scenes. This is often not the case for the other methods as they are unable to effectively distinguish between salient objects and distractors. The bottom row images further illustrate our model exploiting seman-

[1] We tried the CapSal source code for pre-processing captions data (https://github.com/zhangludl/code-and-dataset-for-CapSal), but were unable to adapt their code for our dataset.

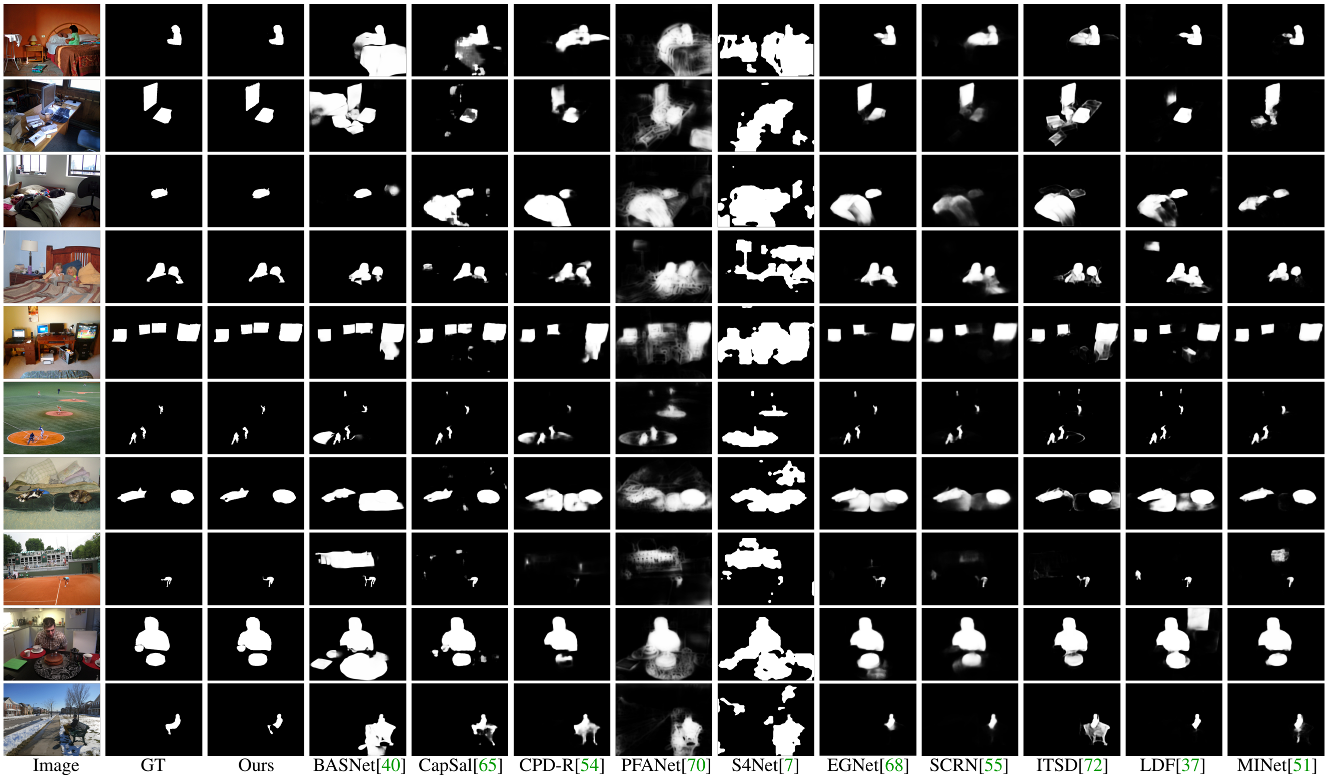

Figure 8: Qualitative comparison of the proposed method with ten other state-of-the-art saliency methods.

Table 3: Ablation study of the proposed model on our dataset. Base: Mask-RCNN architecture, ISCG: Instance/Stuff Context Segmentation, SSCR: Semantic Scene Context Refinement, CIT: Contextual Instance Transformer.

Method	avgF ↑	E_m ↑	MAE ↓
Base	0.826	0.851	0.069
Base+ISCG	0.841	0.866	0.063
Base+ISCG+SSCR	0.845	0.869	0.062
Base+ISCG+CIT	0.849	0.871	0.062
Base+ISCG+SSCR+CIT	0.849	0.872	0.062

tic information in order to fully segment the salient person from the bench. The other methods do not capture such semantic information. They suffer from additional false saliency on part of the bench or unable to segment salient object correctly.

5.4. Ablation Study

We perform additional experiments to evaluate the effectiveness of our proposed modules. These results are shown in Table 3. It shows that the proposed modules produce improvements to the baseline saliency network. Our full model achieves the best overall performance and state-of-the-art results. This suggests that the proposed modules are able to effectively extract and enhance scene context information, then integrate them for saliency reasoning.

6. Conclusion

In this paper, we observe that existing salient object detection methods do not fully capture the semantic context of complex image scenes, leading them to produce false saliency of distractors and missing prediction of salient objects with relations to the scene context. We have also found that popular saliency benchmark datasets mostly contain images of simple scene structure, and do not provide real-world scenarios involving complex scenes with rich context. We have tackled these problems by proposing a new challenging dataset with complex scenes and a saliency model that exploits semantic scene context for improving saliency reasoning. Experimental results show that the proposed model outperforms state-of-the-art methods on the proposed dataset.

Acknowledgements: This work was partially supported by a GRF grant from RGC of Hong Kong (Ref.: 11205620). Avishek Siris is supported by the Swansea Science DTC Postgraduate Research Scholarship. Jianbo Jiao is supported by the EPSRC Programme Grant Visual AI EP/T028572/1.

References

[1] Bahar Aydemir, Deblina Bhattacharjee, Seungryong Kim, Tong Zhang, Mathieu Salzmann, and Sabine Süsstrunk. Modeling object dissimilarity for deep saliency prediction. *arXiv:2104.03864*, 2021.

[2] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, pages 5659–5667, 2017.

[3] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, volume 34, pages 10599–10606, 2020.

[4] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, and Shi-Min Hu. Salientshape: group saliency in image collections. *The Visual Computer*, 30(4):443–453, 2014.

[5] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 37(3):569–582, 2014.

[6] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, 2018.

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

[8] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How much time do you have? modeling multi-duration saliency. In *CVPR*, pages 4473–4482, 2020.

[9] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. *arXiv:2003.05643*, 2020.

[10] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE TPAMI*, 34(10):1915–1926, 2011.

[11] Hadi Hadizadeh and Ivan V Bajić. Saliency-aware video compression. *IEEE TIP*, 23(1):19–33, 2013.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017.

[13] Shengfeng He, Rynson WH Lau, Wenxi Liu, Zhe Huang, and Qingxiong Yang. Supercnn: A superpixelwise convolutional neural network for salient object detection. *IJCV*, 115(3):330–344, 2015.

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

[15] Xiaowei Hu, Chi-Wing Fu, Lei Zhu, Tianyu Wang, and Pheng-Ann Heng. Sac-net: spatial attenuation context for salient object detection. *IEEE TCSVT*, 2020.

[16] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI*, 20(11):1254–1259, 1998.

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

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

[19] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. In *CVPR*, pages 6399–6408, 2019.

[20] Jason Kuen, Zhenhua Wang, and Gang Wang. Recurrent attentional networks for saliency detection. In *CVPR*, pages 3668–3677, 2016.

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

[22] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, pages 660–668, 2016.

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

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

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

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755, 2014.

[27] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019.

[28] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016.

[29] Nian Liu and Junwei Han. A deep spatial contextual long-term recurrent convolutional network for saliency detection. *IEEE TIP*, 27(7):3264–3274, 2018.

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

[31] Nian Liu, Ni Zhang, Kaiyuan Wan, Junwei Han, and Ling Shao. Visual saliency transformer. *arXiv:2104.12099*, 2021.

[32] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE TPAMI*, 33(2):353–367, 2010.

[33] Yi Liu, Jungong Han, Qiang Zhang, and Caifeng Shan. Deep salient object detection with contextual information guidance. *IEEE TIP*, 29:360–374, 2019.

[34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

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

[36] Vida Movahedi and James H Elder. Design and perceptual validation of performance measures for salient object segmentation. In *CVPR Workshops*, pages 49–56, 2010.

[37] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, pages 9413–9422, 2020.

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

[39] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019.

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

[41] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *ICCV*, pages 3799–3808, 2019.

[42] Antonio Torralba. Modeling global scene factors in attention. *JOSA A*, 20(7):1407–1418, 2003.

[43] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 5998–6008, 2017.

[45] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017.

[46] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Salient object detection with recurrent fully convolutional networks. *IEEE TPAMI*, 41(7):1734–1746, 2018.

[47] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. A stagewise refinement model for detecting salient objects in images. In *ICCV*, pages 4019–4028, 2017.

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

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

[50] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2017.

[51] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13025–13034, 2020.

[52] Yichen Wei, Fang Wen, Wangjiang Zhu, and Jian Sun. Geodesic saliency using background priors. In *ECCV*, pages 29–42, 2012.

[53] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2, 2019.

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

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

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

[57] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *CVPR*, pages 1155–1162, 2013.

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

[59] Stella Yu and Dimitri Lisin. Image compression based on visual saliency at individual scales. In *International Symposium on Visual Computing*, pages 157–166, 2009.

[60] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, pages 7234–7243, 2019.

[61] Yu Zeng, Yunzhi Zhuge, Huchuan Lu, and Lihe Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *ICCV*, pages 7223–7233, 2019.

[62] Jing Zhang, Jianwen Xie, and Nick Barnes. Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. *arXiv:2007.12211*, 2020.

[63] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, pages 12546–12555, 2020.

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

[65] Lu Zhang, Jianming Zhang, Zhe Lin, Huchuan Lu, and You He. Capsal: Leveraging captioning to boost semantics for salient object detection. In *CVPR*, pages 6024–6033, 2019.

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

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

[68] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019.

[69] Rui Zhao, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. Saliency detection by multi-context deep learning. In *CVPR*, pages 1265–1274, 2015.

[70] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, pages 3085–3094, 2019.

[71] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Suppress and balance: A simple gated network for salient object detection. *arXiv:2007.08074*, 2020.

[72] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, pages 9141–9150, 2020.