Salient Object Detection by Contextual Refinement

Sayanti Bardhan
Indian Institute of Technology Madras, Chennai 600036, India
National Institute of Ocean Technology, Chennai 600100, India
sayantibardhan@gmail.com

Abstract

Context plays an important role in the saliency prediction task. In this work, we propose a saliency detection framework that not only extracts visual features but also models two kinds of context including object-object relationships within a single image and scene contextual information. Specifically, we develop a novel saliency detection framework with a Contextual Refinement Module (CRM) which consists of two sub-networks, Object Relation Unit (ORU) and Scene Context Unit (SCU). ORU encodes the object-object relationship based on object relative position and object co-occurrence pattern in an image, by graphical approach, while SCU incorporates the scene contextual information of an image. Object Relation Unit (ORU) and Scene Context Unit (SCU) captures complementary contextual information to give a holistic estimation of salient regions. Extensive experiments show the effectiveness of modelling object relations and scene context in boosting the performance of saliency prediction. In particular, our framework outperforms the state-of-the-art models on challenging benchmark datasets.

1. Introduction

The goal of salient object detection is to identify the most visually distinctive object or region in an image, that attracts human attention. Recently, with the rapid development of deep learning techniques, progress in the performance of salient object detection models have been substantial. Effective deep network architectures for saliency prediction include adaptive aggregation [40], short connections [9] among others. Till date, despite significant improvement in accuracy, saliency detection models have performance constraints. In this work, we observe that such performance constraints are largely mitigated if saliency detection model makes use of context.

Context refers to circumstances that form the setting of an event or environment [2]. Image contains rich contextual information including object relationships and scene context [5]. In an image, contextual information determines the relative importance of objects in the image, which in turn determines the saliency of an object [25]. Here, we propose to explore two types of context to mitigate the performance constraints of saliency models: (i) Context in terms of object-object relationship within a single image. Here, the object-object relationship refers to the visually observable interaction between a pair of objects or a pair of subject and an object [44, 34]. (ii) Scene-level contextual information in an image. Scene context refers to the surrounding information around the object of interest [21].

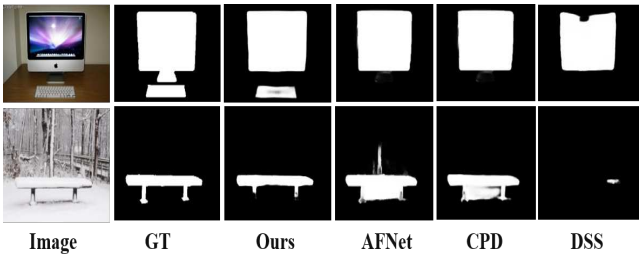

Figure 1. State-of-the-art methods, AFNet [7], CPD [32], DSS [9] fail to segment computer keyboard (first row) and bench against its background (second row). Our method overcomes such limitations.

To further support the importance of contextual information in saliency prediction, we take the help of two examples in Fig. 1. In the first row of Fig. 1, the image demonstrates a computer monitor with its keyboard. It is observed that competitive methods like AFNet [7], CPD [32] and DSS [9] detects only the computer screen as a salient foreground and fails to segment the keyboard. However, it is quite common for computer and its keyboard to co-occur in a single image. Therefore, if the object-object relationship in an image is modelled in terms of object relative position and object co-occurrence pattern, such failures can be easily avoided. Proposed model overcomes such limitations to segment both computer monitor and keyboard by modelling the relationship between objects in an image. In the second row of Fig. 1, because of similar physical

appearance (colour) of foreground and background, state-of-the-art models like AFNet [7], CPD [32] and DSS [9] fail to detect the bench (whole or in part) against the background. This is primarily because deep neural network based saliency frameworks ignore the important scene-level contextual information and extract features only based on physical attributes. Our method models the scene contextual information and segments the bench as salient against its background. Hence, Fig. 1 demonstrates that modelling two types of context enables our framework to imitate human vision system better and approach saliency detection as a reasoning problem (*e.g*. computer and its keyboard).

In this work, we propose a novel saliency prediction model with Context Refinement Module (CRM) that refines saliency maps based on contextual information. Our novel Context Refinement Module has two units: (i) Object Relation Unit (ORU), which models the relationship between objects in an image (ii) Scene Context Unit (SCU), which captures scene contextual information. Our proposed framework utilizes VGGNet [45] as its backbone and is built upon DSS [9] architecture. Obtained side outputs [9] of the proposed network are rich in features from shallower and deeper layers [9], but lack contextual information. To get context rich features, proposed Context Refinement Module (CRM) refines these side outputs. Within CRM, the Object Relation Unit models object-object relationship by a graphical approach based on object relative position and object co-occurrence pattern in an image. In Object Relation Unit, objects in an image are treated as nodes in a graph and object-object relationship in an image are modelled as edges in the graph. On the other hand, Scene Context Unit within CRM, models scene contextual information from an input image to refine network generated saliency map. SCU utilizes Convolutional Gated Recurrent Unit (Conv-GRU) as memory module for scene context modelling. Object Relation Unit (ORU) and Scene Context Unit (SCU) in CRM, capture complementary contextual information. Context rich saliency maps from ORU and SCU are fused to generate a final estimation of saliency regions. It may be noted here that this paper builds on our previous work [2], that proposes the utilization of only scene context for saliency modelling.

To summarize, our key contributions are as follows:
(1) We propose a novel *Contextual Refinement Module* (CRM) to model image context for accurate saliency detection. Contextual Refinement Module models context in terms of: (i) Object-object relationship in a single image with the help of *Object Relation Unit* (ORU). (ii) Scene contextual information with the help of *Scene Context Unit* (SCU). We further introduce a *fusion* of complementary contextual information accumulated from ORU and SCU to give the final saliency map.
(2) We propose the Object Relation Unit that models the relationship between objects based on *object relative position and object co-occurrence pattern* in an image, by a graphical approach.
(3) To the best of our knowledge, the proposed saliency framework is the *first work reported so far to explore context in terms of object-object relationship*.

2. Related Works

2.1. Deep Models for Saliency Detection

Compared to traditional saliency prediction methods [19, 23], deep neural network based saliency detection models [28, 41, 10, 36, 38, 13, 4, 31, 27, 15] have achieved considerable improvement in performance. It may also be noted here, that this paper focuses solely on saliency detection and not visual attention in an image. Hou *et al*. [9] introduce short connection to skip-layer structures in HED architecture [33] to take advantage of multi-scale and multi-level features. Feng *et al*. [7] propose Attentive Feedback Modules to explore object structures and Boundary-Enhanced Loss to further learn the boundaries of salient objects. Qin *et al*. [20] introduces a framework composed of densely supervised Encoder-Decoder network with residual refinement module and hybrid loss for Boundary-Aware salient object detection. Wu *et al*. [32] utilize cascaded partial decoder architecture for saliency prediction. Wang *et al*. [30] propose a framework that integrates bottom-up and top-down saliency inference in an iterative and cooperative manner. Zhuge *et al*. [45] introduce Convolutional Guided Filter and embedding learning architecture to embed initial saliency map into feature vectors and recursively narrow the gap between stage-wise prediction and ground truth. While these deep learning based models achieve considerable accuracy, they ignore context of an image. Ignoring context constraints performance of saliency prediction models.

2.2. Saliency Detection utilizing Contextual Information

Contextual information is an important aspect for saliency detection. Context modelling is also utilized in various tasks like scene graph generation [34], object detection [16] among others. For the detection of salient regions, Zhao *et al*. [43], Zhang *et al*. [40] and Luo *et al*. [18] utilize multi-level contexts. Zhang *et al*. [39] use a symmetrical network to learn complementary visual features under the guidance of lossless feature reflection. It also utilizes weighted structural loss that integrates location, semantic and contextual information of salient objects. Guan *et al*. [8] propose edge-aware saliency detection method based on multi-scale pyramid pooling layers and extra boundary information that preserve sharp boundaries of salient objects and extract rich global context information. Sayanti *et al*.

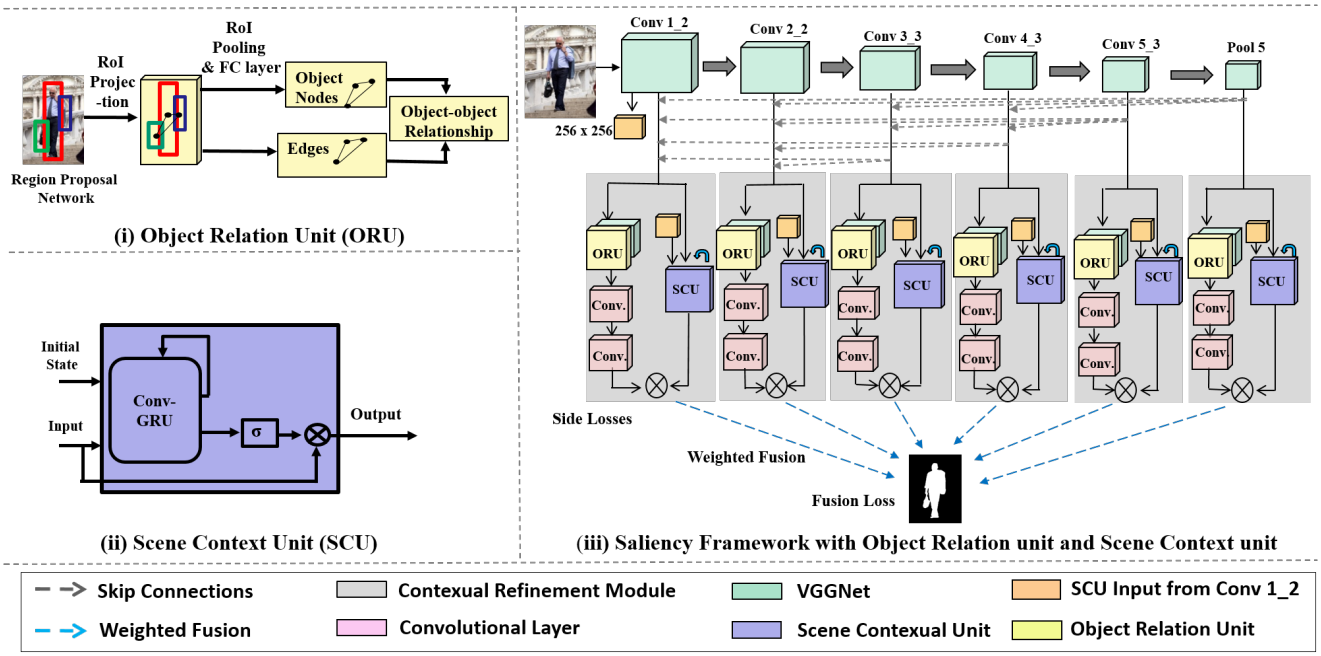

Figure 2. Proposed network overview. (i) Details of Object Relation Unit (ORU). (ii) Details of Scene Context Unit (SCU). (iii) Overview of proposed saliency prediction framework. Short Connection (grey dashed arrows) are introduced to six side outputs of VGGNet. Context Refinement Module (CRM) (grey boxes) refines saliency maps by modelling context with the help of ORU and SCU. ORU (yellow box) is fused with network generated visual features (green box). This is followed by two convolutional layers (pink box) to capture both visual and contextual features. The output from SCU (blue box) is multiplied with output from convolutional layers (pink boxes). This output from CRM undergoes weighted fusion. Fusion loss and side loss give accurate saliency predictions.

[2] model scene context with Convolutional GRU. Zhang *et al*. [37], Wang *et al*. [29] and Zhang *et al*. [42] also exploit contextual information for image saliency detection. While these models mostly model contextual information based on dilated convolutions and multi-scale feature concatenation, our proposed model explores context in terms of object-object relationship and scene contextual information.

Difference from saliency deep networks utilizing contextual information: While our previous work [2] propose the utilization of only scene context, a crucial distinction from [2] is that we propose a saliency prediction method to explicitly take both object-object relationship and scene context in an image into consideration. To effectively leverage the contextual information, our framework proposes a novel Contextual Refinement Module, that models object-object relationship with a graphical approach based on the object relative position and object co-occurrence pattern in an image. Contextual Refinement Module exploits scene contextual information with Conv-GRU. Further, fusion of this complementary contextual information gives a holistic estimation of saliency regions.

3. Our Approach

Image contains rich contextual information including object relationships and scene context [5]. Our objective is to improve saliency detection by utilizing image contextual information. We propose a saliency prediction architecture comprising novel Context Refinement Module (CRM). CRM consists of (i) Object Relation Unit (ORU) to explore context based on object-object relationship and (ii) Scene Context Unit (SCU) to model scene contextual information. Following sub-sections provide more details.

3.1. Object Relation Unit

We leverage context in terms of object relative position and object co-occurrence pattern in an image for accurate saliency detection. However, to capture such the relationship between objects in an image, we need visual and spatial features of objects in an image. This is challenging because objects may appear at different locations in an image, with arbitrary size and scale. Addressing these constraints, we propose the utilization of Faster R-CNN [22] to obtain features of objects in an image. Faster R-CNN is a state-of-the-art object detector, that is also computationally less expensive [22].

Object Relation Unit utilizes Faster R-CNN to produce features of objects in an image, as shown in Fig. 2(i). Region Proposal Network [22] of Faster R-CNN gives region proposals that might contain objects. Further, we use Non-Maximum Suppression [6] to choose a fixed number of Re-

gion of Interests (ROIs). Here, the fixed number of Region of Interests for an image is 256 [17]. For each obtained Region of Interest, v_i, ROI pooling layer followed by a fully connected layer, gives a fixed size visual feature map, f_i^v. Further, to model the object-object relationship in an image, we propose a graphical approach.

Graphical Modelling: At this juncture, we construct a graph $G = (V, E)$, such that every $v \in V$ represent an object node and edge $e \in E$ represents the relationship between each pair of the object nodes. For computing the directed edge $e_{j \to i}$, from node v_j to v_i, we utilize visual and spatial features of v_j and v_i. Obtained $e_{j \to i}$ is scalar and represents the influence of v_j on v_i. As in [17], we compute $e_{j \to i}$ as:

$$e_{j \to i} = tanh(W_v[f_i^v, f_j^v]) * relu(W_p R_{j \to i}^p) \quad (1)$$

where, W_p and W_v are learnable weight matrices. f_i^v and f_j^v represent visual features of nodes i and j respectively. $R_{j \to i}^p$ denotes the spatial position relationship between node i and j and is calculated as [17]:

$$R_{j \to i}^p = [w_i, h_i, s_i, w_j, h_j, s_j, \frac{(x_i - x_j)}{w_j}, \frac{(y_i - y_j)}{h_j}, \frac{(x_i - x_j)^2}{w_j^2}, \frac{(y_i - y_j)^2}{h_j^2}, log(\frac{w_i}{w_j}), log(\frac{h_i}{h_j})] \quad (2)$$

where w_i, h_i are the width and height of the i-th ROI, (x_i, y_i) is the centre of i-th ROI and s_i is the area of i-th ROI. Here $e_{j \to i}$ represents the object-object relationship. As seen in equation 1, it is reasonable that the object-object relationship, $e_{j \to i}$ is determined by the visual cues, f_i^v & f_j^v, and relative object position, $R_{j \to i}^p$ of the object nodes, v_i and v_j. Thus, $e_{j \to i}$ forms a matrix of dimension, (number of ROI $\times$ number of ROI), for a given image. This forms the output of Object Relation Unit. We further combine this matrix representing the relationship between the object nodes with network generated features to capture the fine details of salient objects in an image, as detailed in Section 3.3.

3.2. Scene Context Unit

Scene Context Unit (SCU) models context in terms of contextual information of scene for accurate salient object detection. SCU utilizes Convolutional Gated Recurrent Unit (Conv-GRU) [1] as a memory module to accumulate scene contextual information of an image. Here, we use Conv-GRU instead of Gated Recurrent Unit because [1]: (i) GRU directly applied on images, ignores strong local correlation among pixels of image feature maps. (ii) GRU requires more parameters to be processed than Conv-GRU.

Figure 2(ii) shows the layout of the Scene Context Unit. Here, unlike other recurrent network based models, SCU doesn't initialize hidden state as a random or empty vector [2], instead, SCU utilizes network generated side outputs as the initial hidden state. As in [2], image features (i.e., Conv 1_2 layer features from the proposed network) is fed as SCU input. This enables SCU that comprises of a Convolutional GRU, to refine the network generated saliency map based on scene context learnt from input image features. Conv-GRU output is fed as the updated initial state, at the next time step. However, for all time steps, the SCU input remains the same [2]. After two such time steps, SCU output obtained is fed to sigmoid and then fused with SCU input by multiplication to generate the scene-context rich saliency map, as shown in Fig. 2(ii). The number of time steps is chosen empirically [2]. The output of SCU is further fused with Object Relation Unit output, to give the final saliency estimation, as described in Section 3.3.

3.3. Saliency Prediction Framework

Our proposed saliency prediction network uses VGG-16 [45] as its backbone. Our network is built upon DSS [9] architecture, that takes full advantage of the multi-scale and multi-level features from both the deeper and shallower layers of the network. Our proposed architecture is illustrated in Figure 2(iii). Short connections [9] to side outputs of VGGNet are introduced (indicated by the grey dashed arrows in Fig. 2(iii)). This allows concatenation of multi-scale features and gives output maps of dimension 256×256 [9].

Further to refine the maps and incorporate contextual information, we utilize our proposed Contextual Refinement Module (CRM) on each of the six side outputs, as indicated in Fig. 2(iii). Proposed Contextual Refinement Module has two parts: (i) Object Relation Unit (ORU), which captures the object-object relationship in an image, as detailed in Section 3.1. (ii) Scene Context Unit (SCU), which incorporates the scene contextual information as detailed in Section 3.2.

Object Relation Unit outputs a matrix of dimension 256×256, given the number of ROIs or object nodes for any given image is 256. This matrix represents the relationship between object nodes in an image. We concatenate this matrix (illustrated as yellow boxes in Fig. 2(iii)) with the network generated side outputs (green boxes shown in Fig. 2(iii)) to capture both visual features and contextual information in terms of object relationship. This is followed by two Convolutional layers (pink boxes as shown in Fig. 2(iii)) and a sigmoid layer to give a holistic estimation of salient objects in an image. Further, this saliency map is fused with the output map from SCU, both of dimension 256×256, by multiplication operation, to give context rich saliency maps. These context rich saliency maps of dimension 256×256 forms the output of the Contextual Refine-

Table 1. Quantitative comparison with state-of-the arts in terms of F_β and MAE on standard saliency datasets. Three best performing models are shown in RED, GREEN and BLUE respectively for each dataset. MAE: lower the better and F_β: higher the better

Methods		HKU-IS		DUTS		ECSSD		DUT-OMORON		PASCAL-S	
Name	Year	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
OURS	-	**0.933**	**0.030**	**0.862**	**0.042**	**0.966**	**0.030**	**0.798**	**0.053**	**0.866**	**0.070**
AFNet[7]	CVPR-19	0.923	0.036	**0.862**	0.046	0.935	0.042	**0.797**	0.057	**0.868**	**0.071**
BASNet[20]	CVPR-19	**0.928**	**0.032**	**0.860**	0.050	**0.942**	**0.037**	**0.805**	**0.056**	0.854	0.076
CPD[32]	CVPR-19	**0.924**	**0.033**	**0.864**	**0.043**	**0.936**	0.040	0.794	0.057	**0.866**	0.074
ICTB[30]	CVPR-19	0.920	**0.030**	0.832	**0.045**	0.923	0.041	0.772	**0.055**	0.849	**0.072**
DEFNet[45]	AAAI-19	0.907	**0.033**	0.821	**0.045**	0.915	**0.036**	0.769	0.062	0.826	**0.070**
CTUNet[2]	ICONIP-19	0.932	0.034	0.833	0.047	0.938	0.042	0.788	0.054	0.859	0.073
DSS[9]	PAMI-19	0.913	0.039	0.791	0.056	0.915	0.052	0.729	0.066	0.830	0.080
SFCN[39]	TIP-19	0.906	0.035	0.742	0.062	0.911	0.042	0.718	0.064	0.813	0.732
ECN[8]	SPL-19	0.879	0.036	0.785	0.047	0.901	0.044	-	-	0.812	0.075
ASNet[31]	CVPR-18	0.920	0.035	0.831	0.060	0.928	0.043	-	-	0.857	**0.072**
DGRL[29]	CVPR-18	0.882	0.037	0.768	0.051	0.903	0.045	0.709	0.063	**0.868**	0.079
BDMP[37]	CVPR-18	0.920	0.038	0.850	0.049	0.928	0.044	0.692	0.064	**0.862**	0.074
PAGRN[42]	CVPR-18	0.886	0.048	0.788	0.055	0.891	0.064	0.711	0.072	0.803	0.092
CKT[13]	ECCV-18	0.896	0.048	0.807	0.062	0.910	0.054	0.757	0.071	0.846	0.081
RA[4]	ECCV-18	0.913	0.045	0.831	0.058	0.918	0.059	0.786	0.062	0.834	0.104
LPSD[36]	CVPR-18	0.899	0.039	0.787	0.059	0.908	0.049	0.780	0.060	0.811	0.091
LFR[38]	IJCAI-18	0.875	0.039	0.716	0.083	0.880	0.052	0.696	0.086	0.772	0.105
SRM[28]	ICCV-17	0.873	0.046	0.757	0.058	0.892	0.054	0.707	0.069	0.801	0.085

ment Module. Contextual Refinement Module output, as illustrated by the blue dashed arrows in Fig. 2(iii), undergoes weighted fusion [9] to give saliency estimation. To further preserve boundary information and improve spatial coherence, we utilize fully connected Conditional Random Field (CRF) [11] to obtain the final saliency map output.

Training with deep supervision: As in [33] and [9], we use Cross entropy loss at side output layers after Context Refinement Module and fusion layer of the proposed framework, as illustrated in Fig. 2(iii). The overall loss function, $\tilde{L}_{final}$ is defined as [9]:

$$\tilde{L}_{final}\left(f_w, s, W, \tilde{w}\right) = \tilde{L}_{side}\left(s, W, \tilde{w}\right) + \tilde{L}_{fuse}\left(f_w, s, W, \tilde{w},\right) \quad (3)$$

where, $\tilde{L}_{side}$ is side loss and $\tilde{L}_{fuse}$ is fusion loss as defined in [9]. $\tilde{w}$ are side output weights, s is short connection weights within side outputs, W are the collection of network layer parameters and f_w is the fusion weight.

4. Experimental Results and Analyses

4.1. Implementation Details

Our proposed architecture is trained on DUT-OMRON dataset [35]. We randomly partition DUT-OMORON dataset for train and test. We augment the DUT-OMORON training set of 3500 images by horizontal flipping. Our framework is optimized with Stochastic Gradient Descend. Learning Rate is initially set to 0.01 and is reduced every 10 epochs by 10%. Other hyperparameters used in this paper are: weight decay ($1e^{-4}$), momentum (0.9) and batch size (8). We use Pytorch library[1] and deploy our network on NVIDIA GTX1080ti GPU with 11 GB RAM.

We use the standard metrics [9]: Mean Absolute Error (MAE), F measure (β=0.3) and Precision-Recall (PR) Curve for quantitative evaluation of our proposed framework. Further details of these standard evaluation metrics can be found in [3]. We evaluate our framework on PASCAL-S [14], ECSSD [24], HKU-IS [12], DUTS-TE (DUTS test set) [26] and DUT-OMORON (partitioned for testing) [35] saliency datasets.

4.2. Comparison with the state-of-the-art frameworks

We compare the performance of our model with 18 state-of-the-art deep-learning based saliency frameworks, AFNet[7], BASNet[20], CPD[32], ICTB[30], CTUNet[2], DSS[9], DEFNet[45], SFCN[39], ECN[8], BDMP[37], DGRL[29], PAGRN[42], LPSD[36], ASNet[31], RA[4], CKT[13], LFR[38] and SRM[28]. Saliency maps of these methods are produced by either result published by authors or by executing source codes provided by them.

Quantitative Comparison: We adopt two metrics, F_β

[1] https://pytorch.org

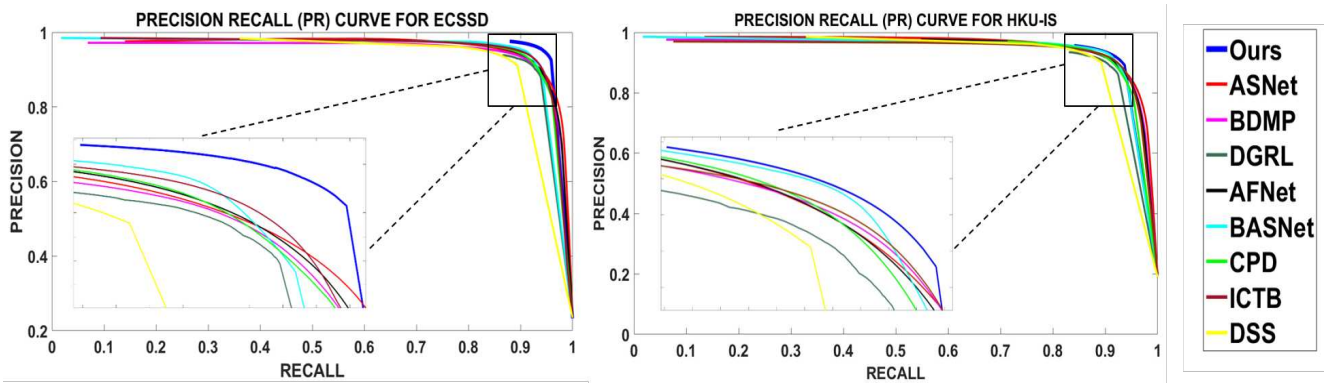

Figure 3. Precision-Recall Curve on ECSSD and HKU-IS datasets

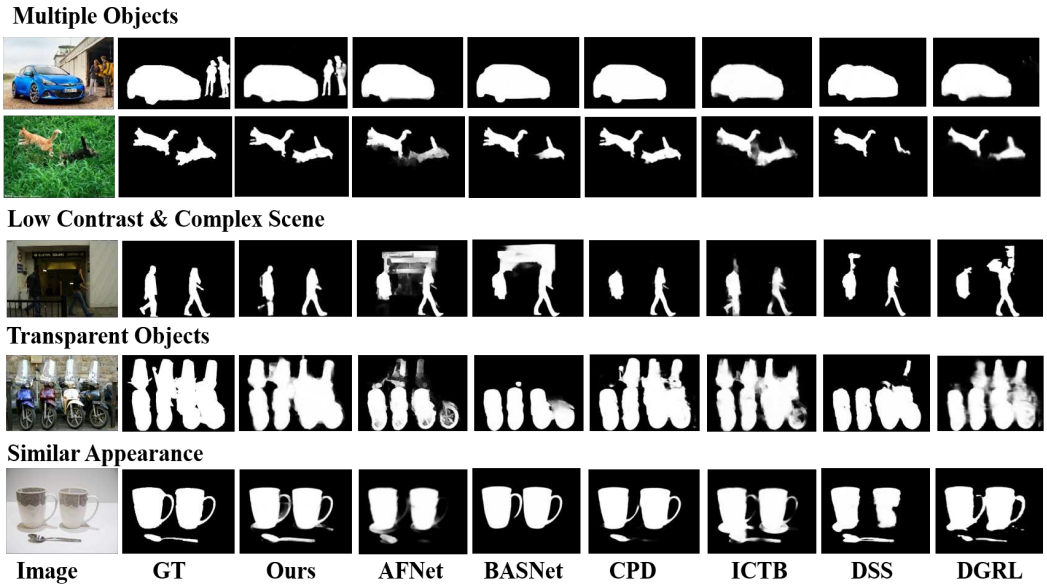

Figure 4. Visual comparison with best performing models. Complexity in images are mentioned.

and MAE to quantitatively compare our approach to the state-of-the-art frameworks. The quantitative results are shown in Table 1. As seen, our framework outperforms all the other state-of-the-art models. The observations from Table 1 are as follows: (i) In HKU-IS and ECSSD dataset, our framework beats other competitive models by a large margin. It improves the F_β measure by 2.5% & 0.5% and reduces the lowest MAE score by 16% & 6% on ECSSD and HKU-IS respectively. (ii) In DUTS, DUT-OMORON and PASCAL-S dataset our framework beats the other existing methods in MAE metric by 1.4%, 2.3% and 3.6% respectively. (iii) In F_β measure, our method is beaten by a small margin of 0.2%, 0.23% and 0.8% by the highest F_β measure in DUTS, DUT-OMORON and PASCAL-S datasets respectively.

We also quantitatively compare our framework with other state-of-the-art approaches in terms of Precision-Recall (PR) curve. Figure 3 illustrates the PR curve for HKU-IS and ECSSD, the two challenging saliency datasets. It may be noted here that our Precision-Recall curve terminates earlier in both datasets due to high confidence (contrast) expressed in our saliency maps. Figure 3 shows that for high recall values (i.e., greater than 0.8), the proposed method beats the existing saliency models.

Visual Comparison: The visual comparison of our approach with the competitive models is demonstrated in Fig. 4. Here, we select images which incorporate a variety of difficult circumstances, like multiple objects, complex scenes, low contrast, transparent objects and images with similar appearance between foreground and background, from standard saliency datasets. We compare our results only with the six best performing methods (i.e., AFNet[7], BASNet[20], CPD[32], ICTB[30], DSS[9] and DGRL[29]). Results of [45] could not be included here because its source codes or saliency maps are not publicly available.

Table 2. Computation time comparison with competitive models

Models	**Ours**	AFNet[7]	BASNet[20]	CPD[32]	DSS[9]	DGRL[29]
Time(sec)	**0.08**	0.03	0.04	0.02	0.05	0.20

Table 3. Ablation study of proposed framework. We change one component at a time, to assess the individual contributions

Methods	ECSSD		HKU-IS		DUTS	
	F_β	MAE	F_β	MAE	F_β	MAE
With Contextual Refinement Module	**0.966**	**0.030**	**0.933**	**0.030**	**0.862**	**0.042**
w/o Object Relation Unit	0.938	0.042	0.932	0.034	0.833	0.047
w/o Scene Context Unit	0.951	0.040	0.931	0.031	0.849	0.045
w/o Contextual Refinement Module	0.915	0.052	0.913	0.039	0.791	0.056

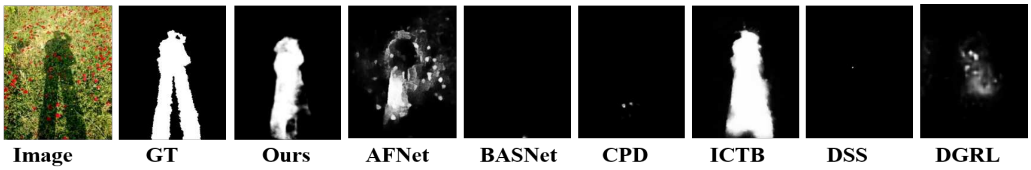

Figure 5. A case of failure of our framework

For multiple objects (first two rows of Fig. 4), our method gives result closest to the ground truth (GT). In the first row, all competing models detect only the blue car, because of its physical appearance (colour) and fail to segment people standing close to this car. In contrast, our method detects both the blue car and people standing close to it. In images with complex scenes and low contrast between foreground and background (third row of Fig. 4), state-of-the-art models detect part of the background as foreground and vice-versa. Our model best segments the salient objects in the image. Our method performs well in detecting transparent objects (i.e., front windshield of the bikes) in the fourth row of Fig. 4 compared to the other competing models. In situations like similar appearance between foreground and background, our model gives output closest to the ground truth, as indicated in the fifth row of Fig. 4 and earlier in the second row of Fig. 1. Our model on average consistently performs the best in all difficult circumstances in Fig. 4. Our framework, in contrast to the competing models, is capable of modelling context in terms of object-object relationship and scene contextual information, that helps to locate the salient regions in an image. *Computation Time:* We compare our average computation time with best performing models (i.e., AFNet[7], BASNet[20], CPD[32], DSS[9] and DGRL[29]). Table 2 shows that our computation time is comparable with other state-of-the-art frameworks.

4.3. Analysis of the proposed architecture

Contextual Refinement Module effectiveness: We analyze the contribution of each component of the Contextual Refinement Module, i.e., Scene Context Unit and Object Relation Unit, by ablation experiments. From Table 3, we observe that our proposed Contextual Refinement Module increases F_β by 5.5% and reduces MAE by 30% on average for the three datasets. We observe that the Scene Context Unit boosts the performance by 3.2% & 17.3% in F_β and MAE respectively and Object Relation Unit enhances the performance by 4.3% & 20.6% in F_β and MAE on average in three datasets. Similar results are also observed in DUT-OMORON and PASCAL-S datasets. The accuracy and efficiency of Context Refinement Module outperform the other cases considered in Table 3, which validates the effectiveness of the proposed framework.

Failure Case: In this section, we analyse failure cases for our framework. One such failure case where models (our proposed framework, as well as state-of-the-art models published earlier) fail is shown in Fig. 5. It is seen that our model fails to detect the shadow in an image that demonstrates low contrast between foreground and background. Training our model with more complex scenes and low contrast images can substantially improve the performance of our model in such cases.

5. Conclusion

We propose a novel architecture that utilizes two kinds of context including object-object relationship within an image and scene contextual information. We design Context Refinement Module comprising of: (i) Object Relation Unit, to model the relationship between objects in an image based on object relative position and object co-occurrence pattern, by graphical approach (ii) Scene Context Unit, to explore scene contextual information in an image with Convolutional Gated Recurrent Unit. This is the first-ever work that utilizes context in terms of both the relationship between objects and scene context for efficient saliency detection. Experiments demonstrate that our model outperforms

state-of-the-art saliency prediction frameworks.

References

[1] Nicolas Ballas, Li Yao, Christopher Joseph Pal, and Aaron C. Courville. Delving deeper into convolutional networks for learning video representations. *CoRR*, abs/1511.06432, 2016. 4

[2] Sayanti Bardhan, Sukhendu Das, and Shibu Jacob. Visual saliency detection via convolutional gated recurrent units. In Tom Gedeon, Kok Wai Wong, and Minho Lee, editors, *Neural Information Processing. ICONIP*, pages 162–174, Cham, 2019. Springer International Publishing. 1, 2, 3, 4, 5

[3] A. Borji, M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, Dec 2015. 5

[4] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 5

[5] S. K. Divvala, D. Hoiem, J. H. Hays, A. A. Efros, and M. Hebert. An empirical study of context in object detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1271–1278, June 2009. 1, 3

[6] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, Sep. 2010. 3

[7] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 5, 6, 7

[8] W. Guan, T. Wang, J. Qi, L. Zhang, and H. Lu. Edge-aware convolution neural network based salient object detection. *IEEE Signal Processing Letters*, 26(1):114–118, Jan 2019. 2, 5

[9] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, April 2019. 1, 2, 4, 5, 6, 7

[10] P. Hu, B. Shuai, J. Liu, and G. Wang. Deep level sets for salient object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 540–549, July 2017. 2

[11] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 109–117. Curran Associates, Inc., 2011. 5

[12] G. Li and Y. Yu. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing*, 25(11):5012–5024, Nov 2016. 5

[13] Xin Li, Fan Yang, Hong Cheng, Wei Liu, and Dinggang Shen. Contour knowledge transfer for salient object detection. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2, 5

[14] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, June 2014. 5

[15] N. Liu, J. Han, and M. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3089–3098, June 2018. 2

[16] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6985–6994, 2018. 2

[17] Yong Liu, Ruiping Wang, Shiguang Shan, and Xilin Chen. Structure inference net: Object detection using scene-level context and instance-level relationships. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 4

[18] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P. Jodoin. Non-local deep features for salient object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6593–6601, July 2017. 2

[19] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank. Salient object detection via structured matrix decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):818–832, April 2017. 2

[20] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2, 5, 6, 7

[21] A. Rabinovich and S. Belongie. Scenes vs. objects: A comparative study of two approaches to context based recognition. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 92–99, June 2009. 1

[22] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 91–99. Curran Associates, Inc., 2015. 3

[23] Sudeshna Roy and Sukhendu Das. Multi-criteria Energy Minimization with Boundedness, Edge-density and Rarity, for Object Saliency in Natural Images . In *The Ninth Indian Conference on Computer Vision, Graphics, Image Processing (ICVGIP)*, 2014. Bangalore, India, 14-17 December 2014. 2

[24] J. Shi, Q. Yan, L. Xu, and J. Jia. Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4):717–729, April 2016. 5

[25] Antonio Torralba, Monica S. Castelhano, Aude Oliva, 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*, 113:2006, 2006. 1

[26] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level su-

pervision. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3796–3805, July 2017. 5

[27] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Salient object detection with recurrent fully convolutional networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018. 2

[28] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stagewise refinement model for detecting salient objects in images. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4039–4048, Oct 2017. 2, 5

[29] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji. Detect globally, refine locally: A novel approach to saliency detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3127–3135, June 2018. 3, 5, 6, 7

[30] W. Wang, J. Shen, M. Cheng, and L. Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5961–5970, June 2019. 2, 5, 6

[31] Wenguan Wang, Jianbing Shen, Xingping Dong, and Ali Borji. Salient object detection driven by fixation prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2, 5

[32] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 2, 5, 6, 7

[33] S. Xie and Z. Tu. Holistically-nested edge detection. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1395–1403, Dec 2015. 2, 5

[34] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei. Scene graph generation by iterative message passing. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3097–3106, July 2017. 1, 2

[35] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang. Saliency detection via graph-based manifold ranking. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3166–3173, June 2013. 5

[36] Y. Zeng, H. Lu, L. Zhang, M. Feng, and A. Borji. Learning to promote saliency detectors. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1644–1653, June 2018. 2, 5

[37] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang. A bi-directional message passing model for salient object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1741–1750, June 2018. 3, 5

[38] Pingping Zhang, Wei Liu, Huchuan Lu, and Chunhua Shen. Salient object detection by lossless feature reflection. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, IJCAI'18, pages 1149–1155. AAAI Press, 2018. 2, 5

[39] P. Zhang, W. Liu, H. Lu, and C. Shen. Salient object detection with lossless feature reflection and weighted structural loss. *IEEE Transactions on Image Processing*, 28(6):3048–3060, June 2019. 2, 5

[40] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 202–211, Oct 2017. 1, 2

[41] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 212–221, Oct 2017. 2

[42] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 5

[43] R. Zhao, W. Ouyang, H. Li, and X. Wang. Saliency detection by multi-context deep learning. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1265–1274, June 2015. 2

[44] Langming Zhou, Jian Zhao, Jianshu Li, Li Yuan, and Jiashi Feng. Object relation detection based on one-shot learning. *ArXiv*, abs/1807.05857, 2018. 1

[45] Yunzhi Zhuge, Yu Zeng, and Huchuan Lu. Deep embedding features for salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9340–9347, 07 2019. 2, 4, 5, 6