IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 17, 2020, accepted May 29, 2020, date of publication June 3, 2020, date of current version June 15, 2020.

*Digital Object Identifier 10.1109/ACCESS.2020.2999627*

# SdBAN: Salient Object Detection Using Bilateral Attention Network With Dice Coefficient Loss

**DONGGOO KANG, SANGWOO PARK, (Student Member, IEEE), AND JOONKI PAIK, (Senior Member, IEEE)**
Department of Image, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Joonki Paik (paikj@cau.ac.kr)

This work was supported in part by the Institute for Information and Communications Technology Planning and Evaluation (IITP) through the Korea Government (MSIT) Development of Global Multi-Target Tracking and Event Prediction Techniques-Based On Real-Time Large-Scale Video Analysis under Grant 2014-0-00077, and in part by the Institute for Information and Communications Technology Promotion (IITP) through the Korea Government (MSIT) Intelligent Defense Boundary Surveillance Technology Using Collaborative Reinforced Learning of Embedded Edge Camera and Image Analysis under Grant 2017-0-00250.

**ABSTRACT** Visual attention plays an important role in saliency detection by highlighting meaningful context regions. In this paper, we present a novel saliency detection method using a bilateral attention network. The proposed network consists of two branches: i) a spatial path using an encoder-decoder structure to learn spatial cues and ii) a context path using an attention mechanism to learn contextual cues. The feature aggregation module is finally used to predict salient objects by concatenating the cues. To optimize the weights of the network in the sense of minimizing the class imbalance problem, we minimize the dice coefficient loss together with the classical cross-entropy loss. The proposed network can predict salient regions in an end-to-end manner without post-processing. Experimental results show that the proposed network achieved better performance than existing state-of-the-art methods in most cases. Furthermore, the proposed network takes only 0.03 seconds to process a $224 \times 224$ image. The code for the proposed method can be found at the following URL: https://github.com/tiruss/SdBAN

**INDEX TERMS** Salient object detection, deep learning, dice coefficient, attention mechanism.

## I. INTRODUCTION

Saliency detection aims at extracting the most visually noticeable region in an image. Unlike other segmentation approaches such as semantic segmentation and boundary detection, saliency detection only distinguishes the most visually attractive and interesting object from the background. It can be applied to various computer vision fields such as image segmentation [1], object recognition [2], action recognition [3], weakly supervised semantic segmentation [4]–[7], visual tracking [8], video compression [9], [10] and video summarization [11].

Existing hand-crafted feature-based saliency detection methods commonly measure the contrast. Itti and Koch proposed contrast difference between the center pixel and its neighborhood [12]. Klein and Frintrop used Kullback-Leibler Divergence (KLD) to measure the difference [13]. However, these difference measurement-based saliency detection methods commonly fails when there is no significant difference between the object and background, or the background has a complex pattern or clutters. Wang *et al.* applied a learning-based discriminative model to guarantee high performance in various types of domains [14]. To provide a pre-specified prior, they need additional pre- and post-processing steps. Kong *et al.* proposed an exemplar-aided method that complement heuristic saliency assumptions by leveraging only a few exemplar images [15]. Zeng *et al.* proposed a game-theoretic method that does not require labeled training data [16]. Zhou *et al.* proposed a superpixel-based two-layer diffusion process [17].

In recent years, convolutional neural networks (CNNs) have demonstrated unparalleled performance in the salient object detection and segmentation fields. Specifically, fully convolutional networks (FCNs) greatly improve the ability to preserve spatial information [18]. Mnih *et al.* proposed U-shape structure to reduce the loss of details of an object [19]. By fusing the hierarchical features of the backbone network, the U-shape structure gradually increases the

The associate editor coordinating the review of this manuscript and approving it for publication was Long Wang.

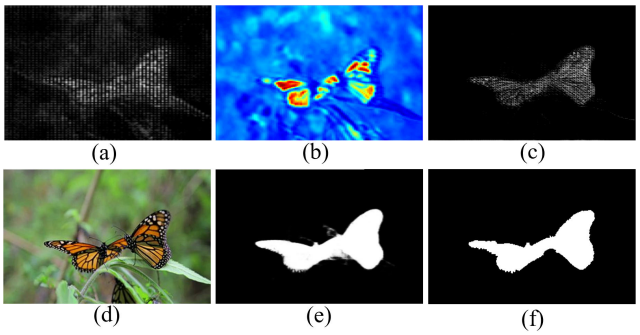



**FIGURE 1.** Example of learned spatial and context feature maps: (a) feature map computed by the spatial path using the U-net architecture, (b) feature map of the context path using the pixel-wise attention mechanism, (c) fusion of (a) and (b) using the feature fusion block, (d) the input image, (e) the prediction result using the proposed method, and (f) the ground truth mask.

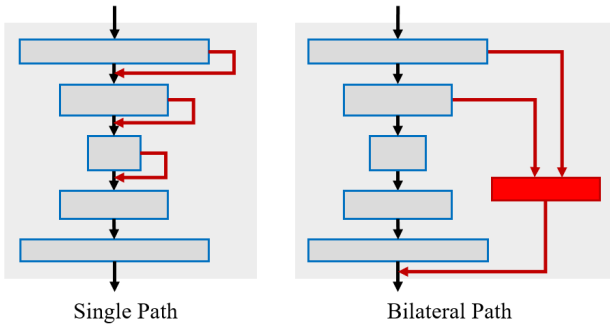



**FIGURE 2.** Conventional attention network with a single path(left) and the proposed network with an additional bilateral path. Red arrow represent the attention path, and the red box represents the concatenation operation of attention maps in each layer.

spatial resolution and fills some missing details. For that reason, recent saliency detection studies are based on FCNs and U-shape structures. He *et al*. proposed a super-pixel-wise convolutional neural network using hierarchical contrast features [20]. For each scale of super-pixel, two contrast sequences were fed into the convolutional network for more detailed features. Li and Yu proposed a deep contrast network to emphasize the contrast information [21]. It concatenates a pixel-level FCN stream and a segment-wise spatial pooling stream. A fully connected conditional random field (CRF) is also used for refining the output from the contrast network. Liu *et al*. used a hierarchical recurrent convolutional neural network for saliency detection [22]. This network consists of two stages: i) generating a coarse output map using a deep CNN and ii) hierarchical refinement of the details using a recurrent CNN. Both Li's and Liu's works commonly used multiscale features that are extracted by convolutional layers. Hou *et al*. proposed skip-connections between layers to find a salient object in a deep neural network without loss of information [23]. Hu *et al*. tried to find the salient object by minimizing the loss of each pooling layer and refinement using guided super-pixel filtering [24]. Fu *et al*. proposed a deep framework for salient object detection that effectively fuses multi-scale outputs [25]. To fuse differently scaled outputs, they proposed: i) a linear model using a fully connected layer, ii) a nonlinear model using the FCN for concatenation, and iii) a joint fusion of the two models. Edge-based salient object detection approaches were recently proposed in [26], [27]. Zhao *et al*. proposed an edge guidance network using an explicit edge modeling method [26], which estimates deterministic object boundaries by adding a complementary salient edge to multi-scale information. Wu *et al* proposed a stacked cross refinement network for an edge-aware network [27], which simultaneously leans both saliency map and salient object boundaries using consecutively stacked cross refinement units (CRUs). Compared with existing hand-crafted feature-based methods, CNN-based methods can produce generalized results in various domains, and give a significantly imporved performance without using pre-specified priors. However, since these methods learn the entire image, background may be detected as a salient object when the size of the salient object is smaller than the background. This is called class imbalance problem, and we will discuss about related experimental results in section IV-E2.

To solve this problem, we present a novel saliency detection method using an attention mechanism to assign a higher weight to informative regions. The proposed network consists of two branches: spatial path and context path. The spatial path has an encoder-decoder structure with skip-connections to learn the spatial information. On the other hand, the context path has an attention mechanism to learn the context information. We also propose a feature aggregation block to effectively concatenate two branches without loss of information as shown in Fig. 2. To train the proposed network, we minimize a harmonic loss function that combines the dice coefficient and cross entropy losses. The cross entropy loss cannot solve the class imbalance problem by itself since it tends to decrease when the object size is small. The dice coefficient is devised as an index to measure the similarity of two images. By minimizing the dice coefficient loss, background is ignored and only the object region is considered. As a result, minimization of the dice coefficient loss can solve the class imbalance problem. Since learning with only dice coefficient loss becomes unstable, we added the classical cross-entropy loss for stable learning without class imbalance problem.

The main contributions of this work are summarized as follows:

1) We present a *bilateral attention network* to learn both spatial and context information. The spatial path is an encoder-decoder structure with skip-connection, which is robust to object size variations. The context path assigns a higher weight to the informative region of the image through the attention mechanism. This process will be proved to be robust even when the background is complex and the difference between object and background is not significant.
2) We propose *a harmonic loss function* that combines the dice coefficient loss and cross-entropy loss for stable learning without class imbalance problem.

3) Extensive experiments show that the proposed method compares favorably to the state-of-the-art methods, both in terms of visual quality and in terms of different metrics.

## II. RELATED WORKS

### A. ATTENTION MECHANISM

Recently, attention mechanism, which makes computation resource concentrated on the informative region of the image, is applied in various deep neural networks. Over the last few years, the attention mechanism has been studied in natural language processing [28]–[30]. Mnih *et al*. proposed a method to adaptively select a region of interest in an image through a recurrent attention model [19]. To the best of authors' knowledge, this is the first attempt to apply the attention mechanism to the computer vision tasks. However, training the network including the recurrent attention model is a challenging problem since it is not easy for the attention model to focus on a definite point in the image, which is called *hard attention* problem. To solve that problem, Bahdanau *et al*. proposed a soft attention model, which calculates attention weights of all input features [31]. This allowed the RNN encoder-decoder network to overcome the limitations of containing all the sentence information in a fixed-length vector. This method significantly improves performance in machine translation.

In recent years, attention mechanisms have been introduced into various computer vision applications. Xu *et al*. applied the recurrent attention model to the field of image captioning by highlighting the area corresponding to each word of the sentence describing the given image [32]. Sermanet *et al*. enhanced performance of image classification by extracting discriminative regions in the image through the recurrent attention model [33]. Chen *et al*. replaced average-pooling and max-pooling for multi-scale features by the attention module to increase performance of the semantic segmentation [34]. Li *et al*. applied the region of interest (ROI) to the object detection field through the attention model [35]. These studies proved that the attention mechanism successfully assigned higher weights to informative regions to increase performance of object detection. Liu *et al*. proposed pixel-wise contextual attention network (PiCANet) to apply the attention mechanism to saliency detection [36]. In PiCANet, an attention-guided network selectively integrates multi-level contextual information to alleviate distraction of cluttered features. This method is robust to background changes and cases successfully detects objects in most cases. However, it cannot preserve high-level features with semantic information, resulting in a blurred boundary of the object. To solve this problem, Krähenbühl and Koltun used conditional random field (CRF) in post-processing [37]. Feng *et al*. used boundary-enhanced loss (BEL) with the attention feedback module to detect salient objects [38], where the context of the object is learned through attention, and the boundary of the object is learned through BEL.

The proposed network is different from feature integration-based approaches described above in that our bilateral network can separately obtain both spatial and context information from different paths to preserve the advantages of both paths.

### B. ENCODER-DECODER ARCHITECTURE

In computer vision, image segmentation is the process of assigning a pixel-by-pixel label to an entire image, and its performance depends on the ability to preserve multi-scale features. Most existing multi-scale feature handling networks are based on an encoder-decoder architecture. The encoder of the network compresses the information of the object through the layer, where the high-level layer contains detailed information of the object, and the low-level layer contains the context information of the object. Most encoders used a pretrained network to extract general features in an efficient manner using a small amount of training data sets. The feature vectors compressed by the encoder are then reconstructed by the decoding layer. Through this structure, more generalized results can be obtained.

Badrinarayanan *et al*. proposed SegNet which uses an encoder-decoder structure for semantic segmentation [39]. This is first attempt to apply the encoder-decoder structure to the pixel-wise prediction task. SegNet showed higher localization than simple upsampling based methods. Ronneberger *et al.* proposed an encoder-decoder structure using skip-connection, called *U-net* [40]. Skip-connection concatenates pairs of encoder and decoder layers of the same size. A successive layer can then learn to assemble a more precise output based on this information. Saliency detection is distinguished from object detection and segmentation tasks in that the shape of an object is not constant. It is important to create a more generalized network since there is no fixed shape for a salient object. The proposed network uses the U-net structure to obtain generalized spatial information.

## III. PROPOSED METHOD

The proposed network consists of *spatial* and *context paths* as shown in Fig. 3. The spatial path performs semantic segmentation whereas the context path generates the contextual attention vector of the object. Since the low-level layer in the spatial path has high-resolution spatial information, it is not suitable to find the context of the object through the attention mechanism. Therefore, the context attention block (CAB) is applied to the compressed feature map through convolution and pooling. In order to preserve the characteristics of each path, feature map of each path is concatenated through the feature fusion block (FFB) at the last layer of the decoder.

### A. SPATIAL PATH

To extract features, we use the pretrained ResNet-50 as the encoder in the spatial path [41]. This network is modified to be fully convolutional to produce dense feature maps while preserving spatial location. More specifically, we replaced the

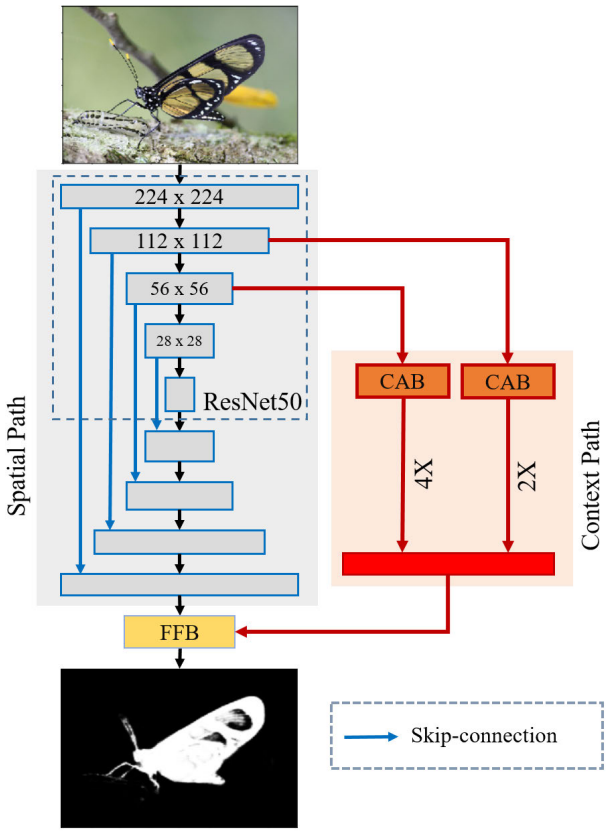



**FIGURE 3.** Architecture of the proposed SdBAN.

last fully-connected layers of the original ResNet-50 by four deconvolution blocks to reconstruct features. The reason for using four deconvolution blocks is that the spatial decimation factor of the ResNet-50 is 16 when four max-pooling layers of stride 2 are employed. In addition, skip-connection is applied between pairs of encoder and decoder layers of the same scale to preserve multi-scale features.

### B. CONTEXT PATH

One of the problems of the saliency detection task is inconsistent prediction result where the background is complex or the difference between background and object is low. These problems are mainly due to the lack of context. Global average pooling can be used to find global contexts [42], [43]. However, global context just has the high level semantic information, which is not helpful for recovering the spatial information. Therefore, a multi-scale receptive view is needed to restore spatial information successfully. To accurate guide multi-scale features, we design a context attention block (CAB) as shown in Fig. 4. A CAB calculates the channel attention vector for each scale feature. Both high- and low-level features provide a consistent guidance and discrimination information of features. In this way, the channel attention vector can select discriminative features.

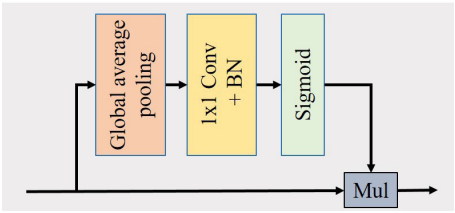



**FIGURE 4.** Components of the context attention block (CAB).

#### 1) CONTEXT ATTENTION BLOCK

In the FCN architecture, the convolution operator has a score map as an output. The score map is interpreted as the probability of a class for each pixel. Let $s$ be the scale of the feature map, the score $y_s$ is the sum of all feature maps as

$$y_s = F(x; w) = \sum_{i=1}^{S} w_i x_i, \tag{1}$$

where $S$ represents the largest scale, $x_i$ the $i$-th scale feature map, and $w_i$ the $i$-th scale convolution kernel.

Since the convolution operation takes all input feature maps with an equal weight, the predicted output may become incorrect when background is noisy or the object is relatively small. To solve this problem, we use a weighting parameter $\alpha$, which becomes large for a highly discriminative region of the object.

To determine the optimal value of $y_s$, global average pooling (GAP) is performed before the max-pooling layer in the encoder and $d \in \{1, \ldots, D\}$ scores of feature maps are obtained. The final weight vector $\alpha_d$ is then obtained by $1 \times 1$ convolution operation which maps the score between 0 and 1 through the softmax function.

$$\alpha_d = \frac{\exp(y_d)}{\sum_{i=1}^{D} \exp(y_i)}. \tag{2}$$

Finally, we construct an attended contextual feature $y_A$ as a weighted sum of $\alpha_d$ and the original feature map as

$$y_A = \sum_{d=1}^{D} \alpha_d y_d. \tag{3}$$

As shown in Fig. 5, the proposed context attention block (CAB) weights the discriminative region in the feature maps. The output of the CAB can be regarded as the heat map of attention.

Components of the CAB look similar to the squeeze-and-excitation (SE) block proposed by Hu *et al.* [44]. The CAB is different from the SE block in that the intermediate fully connected layer is replaced with a $1 \times 1$ convolution layer to preserve spatial relationship and reduce computational overhead.

### C. DICE SIMILARITY COEFFICIENT LOSS

The size of a salient object is often much smaller than that of background. This makes the learning process get

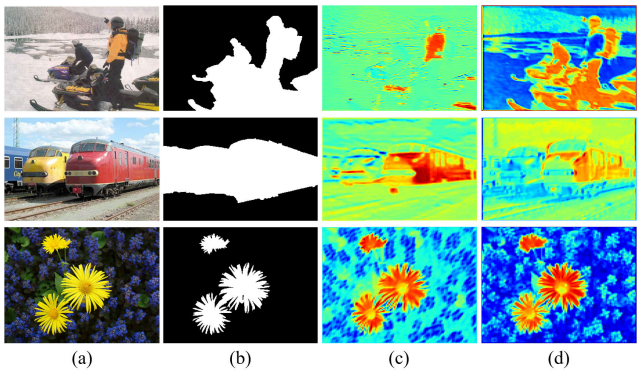



**FIGURE 5.** Attention maps extracted by the context path: (a) input images, (b) saliency masks (ground truth), and (c-d) two most close attention maps to the ground truth.

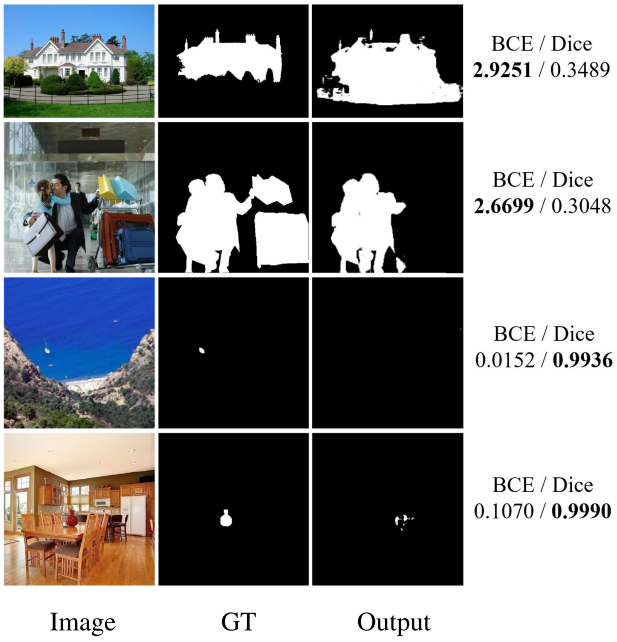



**FIGURE 6.** Comparison of binary cross entropy and dice coefficient values for different size of salient objects. The cross entropy is sensitive to the size of the salient object, while the dice coefficient is less sensitive because it measures the overlap between objects.

trapped in a local minimum of the loss function yielding a network whose predictions are strongly biased towards the background. To solve this problem, weighted cross-entropy loss, class-balanced cross-entropy loss [45] are used in [23], [46].

Weighted cross entropy (WCE) is a variant of CE where all positives get weighted by coefficient $\beta$ and defined as

$$L_{WCE} = -\frac{1}{N}\sum_{i}^{N}(\beta g_i \log(p_i) + (1-g_i)\log(1-p_i)), \quad (4)$$

where $p_i \in P$ be the predicted saliency map, and $g_i \in G$ the corresponding ground truth. If $\beta$ is larger than the unity, the foreground gets more weights, and vice versa.

For the pixel-wise prediction, Xie and Tu used a simpler strategy called class-balanced cross-entropy (CBCE) that adaptively weights positives and negatives as [45]

$$L_{CBCE} = -\frac{1}{N}\sum_{i}^{N}(\beta g_i \log(p_i) + (1-\beta)(1-g_i)\log(1-p_i)), \quad (5)$$

where $\beta = |N_-|/|N|$ and $1-\beta = |N_+|/|N|$. $|N_-|$ and $|N_+|$ represent the saliency and non-saliency maps, respectively.

This simple approach can solve the class imbalance problem. However, Deng *et al.* argued that the CBCE loss causes the 'thickness' in the edge detection task [47]. This is due to the nature of the cross entropy loss. More specifically, the cross entropy loss is calculated as the average of per-pixel loss, and the per-pixel loss is independently calculated without considering whether its adjacent pixels are salient or not. As a result, the cross entropy loss considers loss in a local sense rather than the global sense.

Milletari *et al.* proposed another objective function that maximize the dice coefficient between images [48] to solve class-imbalance problem. The dice coefficient is an index that measures the overlap between the ground truth and the prediction output in segmentation-like tasks. the dice coefficient denoted as $D$ can be computed as

$$D = \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}. \quad (6)$$

In saliency detection tasks, the ground truth and predicted saliency maps can be viewed as two sets. In (6), the denominator considers the total number of saliency maps at the global scale, while the numerator considers the overlap between the two sets at a local scale. Therefore, the dice coefficient loss considers the loss information in both local and global manners. The dice coefficient in (6) is minimized when its gradient with respect to $p_i$ is equal to zero as

$$\frac{\partial D}{\partial p_j} = 2\left[\frac{g_j\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) - 2p_j(\sum_i^N p_i g_i) + \epsilon}{\left(\sum_i^N p_i^2 + \sum_i^N g_i^2\right) + \epsilon}\right], \quad (7)$$

where $\epsilon$ is a smoothing term to avoid division by zero. To make the converged become zero, we modified the loss as

$$L_D = 1 - \frac{2\sum_i^N p_i g_i}{\sum_i^N p_i^2 + \sum_i^N g_i^2}. \quad (8)$$

The class imbalance problem can be solved by minimizing (8). However, since the dice coefficient loss can only learn about object, the learning process is unstable due to the high variance. To learn both object and background, we used binary cross-entropy loss

$$L_{CE} = -\frac{1}{N}\sum_{i}^{N}(g_i \log(p_i) + (1-g_i)\log(1-p_i)). \quad (9)$$

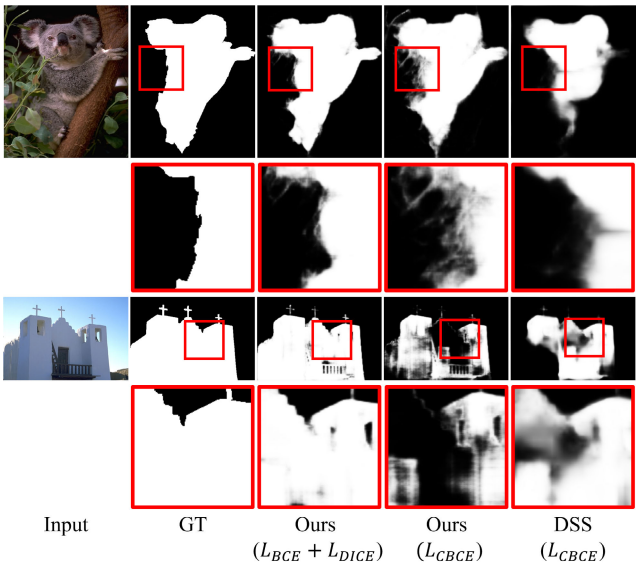



**FIGURE 7.** Subjective comparison of saliency detection performance using different combinations of losses. The proposed method using both binary cross-entropy and dice losses generated better saliency detection result than using only CBCE loss and DSS [23] using CBCE loss.

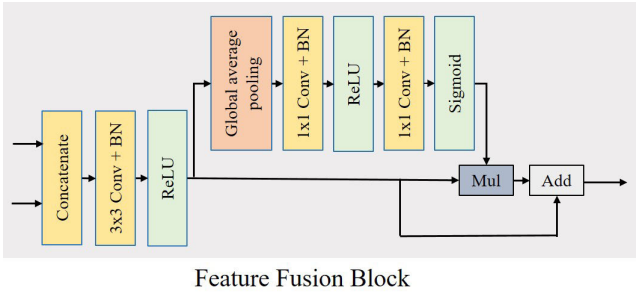



**FIGURE 8.** Components of the feature fusion block (FFB).

The total loss function, denoted as $L_T$, is the sum of the dice coefficient and the binary cross entropy losses as

$$L_T = \tau L_D + (1 - \tau) L_{CE}, \tag{10}$$

where $\tau$ is the weighting parameter to balance the effect of $L_D$ and $L_{CE}$.

### D. FEATURE FUSION BLOCK

Features from the proposed dual path network have different types of representation. Therefore, a simple concatenation deteriorates the performance. The information captured by the spatial path encodes most of rich detail information. On the other hand, the information captured by the context path mainly encodes context information. In other words, the spatial path extracts a low-level feature map, whereas the context path extracts a high-level feature map. Therefore, we present a feature fusion block (FFB) that concatenates features of different levels without loss of information as shown in Fig. 8.

**TABLE 1.** Network architecture table of proposed method.

| Name | size | Spatial Path | Context Path |
|---|---|---|---|
| Input | 224 * 224 | | |
| Conv1S<br>Conv1C | 112 * 112 | 7 * 7, 64, stride 2 | global average pool<br>$\begin{bmatrix} 1*1, 64 \\ multifly \end{bmatrix}$<br>2x interpolation |
| Conv2S<br>Conv2C | 56 * 56 | 3 * 3 max pool, stride 2<br>$\begin{bmatrix} 1*1, 64 \\ 3*3, 64 \\ 1*1, 256 \end{bmatrix}$ x 3 | global average pool<br>$\begin{bmatrix} 1*1, 256 \\ multifly \end{bmatrix}$<br>4x interpolation |
| Conv3 | 28 * 28 | 3 * 3 max pool, stride 2<br>$\begin{bmatrix} 1*1, 128 \\ 3*3, 128 \\ 1*1, 512 \end{bmatrix}$ x 4 | |
| Conv4 | 14 * 14 | 3 * 3 max pool, stride 2<br>$\begin{bmatrix} 1*1, 256 \\ 3*3, 256 \\ 1*1, 1024 \end{bmatrix}$ x 6 | |
| Conv5 | 7 * 7 | 3 * 3 max pool, stride 2<br>$\begin{bmatrix} 1*1, 512 \\ 3*3, 512 \\ 1*1, 2048 \end{bmatrix}$ x 3 | |
| Deconv1 | 14 * 14 | 3 * 3, 2048, stride 2<br>concat[1 * 1, 512, Conv4] | |
| Deconv2 | 28 * 28 | 3 * 3, 1024, stride 2<br>concat[1 * 1, 256, Conv3] | |
| Deconv3 | 56 * 56 | 3 * 3, 512, stride 2<br>concat[1 * 1, 128, Conv2S] | |
| Deconv4 | 112 * 112 | 3 * 3, 256, stride 2<br>concat[1 * 1, 64, Conv1S] | |
| Deconv5S<br>Deconv5C | 224 * 224 | 3 * 3, 128, stride 2<br>concat[1 * 1, 32, Input] | concat $\begin{bmatrix} Conv1C \\ Conv2C \end{bmatrix}$ |
| FFB | 224 * 224 | concat[Deconv5S, Deconv5C]<br>3 * 3, 128<br>1 * 1, 128 | global average pool<br>1 * 1, 128 |

Given various levels of features, we first concatenate the output features of both spatial and context paths. Next, the integrated features obtained by convolution operation and batch normalization. We also obtain the attention weight vector through the softmax function after global average pooling in the integrated feature in the similar way of SENet [44]. The weight vector guides the correct feature selection in the integrated feature.

The network architecture of the proposed method is summarized in Table 1.

## IV. EXPERIMENT RESULTS

### A. DATASETS

We used five popular saliency benchmark datasets to evaluate the performance of our method. SOD dataset has 300 images with complex background and multiple objects per image [49]. HKU-IS dataset consists of 4,447 low-contrast images with multiple objects [50]. DUT-O dataset consists of 5,168 challenging images with complex background and one or more objects per image [51]. DUTS

dataset consists of DUTS-TR consisting of 10,553 images for training and DUTS-TE consisting of 5,019 challenging images for testing [52]. ECSSD dataset consists of 1,000 images of various types and sizes [53].

### B. EVALUATION METRICS

The performance was evaluated using mean absolute error (MAE), precision-recall (PR) curve, F-measure [50], weighted F-measure [54], and S-score [55] which are commonly used in salient object detection.

#### 1) MAE

The mean absolute error (MAE) between the predicted output $p_i$ and the ground truth $g_i$ is defined as

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |p(x, y) - g(x, y)|, \tag{11}$$

where $W$ and $H$ respectively represent the width and height of images.

#### 2) PRECISION-RECALL (PR) CURVE

The PR curves are calculated from the precision and recall values of predicted output $p_i$ and ground truth $g_i$ given a pre-specified threshold between 0 and 255. Specifically, the PR curve reflects the object retrieval performance in the sense of both precision and recall by binarizing the final saliency map using different thresholds.

#### 3) F-MEASURE ($F_\beta$)

The F-measure, denoted as $F_\beta$, is an overall performance measurement, and is computed by the weighted product of precision $P$ and recall $R$ as

$$F_\beta = \frac{(1 + \beta^2)P \times R}{\beta^2 P + R}, \tag{12}$$

where $\beta^2$ is set to 0.3 according to previous researches to assign a higher weight on precision than recall. More specifically, maximum F-measure, denoted as $maxF_\beta$, is associated with the maximum F-measure value computed from the PR curve, while average F-measure, denoted as $meanF_\beta$ uses the adaptive threshold for binarization.

#### 4) WEIGHTED F-MEASURE ($F_\beta^w$)

Margolin *et al*. proposed weighted F-measure, denoted as $F_\beta^w$, to compensate for the drawback of the original F-measure by considering both pixel dependency and pixel importance with an appropriate weight as [54].

$$F_\beta^w = \frac{\left(1 + \beta^2\right) P^w \times R^w}{\beta^2 P^w + R^w}, \tag{13}$$

where $P^w$ and $R^w$ respectively represegnt the weighted precision and recall. The weighted F-measure is different from the original F-measure in that it directly compares a non-binary map using a binary ground truth without thresholding to avoid the interpolation flaw. $\beta^2 = 0.3$ is used to give more weight the precision more than recall. More details about this metric can be found in [54].

**TABLE 2.** max $F_\beta$, $F_\beta^w$, and mean absolute error (MAE) of the baseline model trained with and without dice coefficient loss, context attention block and feature fusion block on the HKU-IS dataset [50].

| Dataset | HKU-IS [50] | | | DUT-O [51] | | |
|---|---|---|---|---|---|---|
| Method | max $F_\beta$ | $F_\beta^w$ | MAE | max $F_\beta$ | $F_\beta^w$ | MAE |
| Baseline [40] | 0.886 | 0.779 | 0.052 | 0.741 | 0.592 | 0.085 |
| w/o Dice | 0.895 | 0.844 | 0.042 | 0.752 | 0.677 | 0.068 |
| w/o CAB | 0.890 | 0.820 | 0.041 | 0.755 | 0.635 | 0.068 |
| w/o FFB | 0.907 | 0.850 | 0.038 | 0.758 | 0.688 | 0.066 |
| Ours(Dice + CAB + FFB) | **0.913** | **0.869** | **0.035** | **0.764** | **0.698** | **0.064** |

#### 5) S-MEASURE ($S_\alpha$)

Fan *et al*. proposed S-measure, denoted as $S_\alpha$, to quantify the spatial structure similarities (SSIM) of the saliency map, which is widely used in the quality assessment (IQA) field, and is defined as

$$S_\alpha = \alpha * S_0 + (1 - \alpha) * S_r, \tag{14}$$

where the weighting coefficient $\alpha$ controls the balance between two terms, and $\alpha = 0.5$ was used according to previous researches. $S_0$ and $S_r$ respectively represent the object-aware and region-aware structural similarities. More details about this metric can be found in [55].

### C. IMPLEMENTATION DETAILS

The proposed model was implemented on the TensorFlow framework with a single GTX 1080 Ti GPU for acceleration. For a fair comparison with previous works, the proposed model was trained with the DUTS-TR dataset [30]. For data augmentation, we resized each image to 256 × 256, and then used random cropping and random mirror-flipping for training. We trained our model using Adam optimizer [56], with initial learning rate 0.002 decayed down to 0.00003 per epoch, 200 epochs and mini-batch size 16. It took about 6 hours to converge.

For the test, we resized the image to 224 × 224 to get the prediction result, and then restore it back to the original size. Using Resnet-50 as a backbone [41], it took 0.03 second to predict one image.

### D. ABLATION STUDY

#### 1) EFFECTIVENESS OF THE SdBAN

To demonstrate the effectiveness of the proposed network, we investigated each component in the proposed network as shown in Table 2, where *Baseline* represents the U-Net [40] without SdBAN. In Table 2, we designed our ablation study using three different settings. w/o Dice means only use cross-entropy loss, w/o CAB means the baseline with harmonic loss function, and w/o FFB means simple concatenation of feature maps of two paths. The ablation study demonstrated that each component contributed fairly

**TABLE 3.** Comparison of the saliency detection performance of 14 methods including ours in the sense of max $F_\beta$, mean $F_\beta$, and MAE. The best and second-best results are highlighted in red and blue, respectively.

| Dataset | SOD [49] | | | HKU-IS [50] | | | DUT-O [51] | | | DUTS-TE [52] | | | ECSSD [53] | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Metric | *max* $F_\beta$ | *mean* $F_\beta$ | MAE | *max* $F_\beta$ | *mean* $F_\beta$ | MAE | *max* $F_\beta$ | *mean* $F_\beta$ | MAE | *max* $F_\beta$ | *mean* $F_\beta$ | MAE | *max* $F_\beta$ | *mean* $F_\beta$ | MAE |
| VGG-16 backbone [60] | | | | | | | | | | | | | | | |
| DS [61] | 0.784 | 0.698 | 0.190 | 0.865 | 0.788 | 0.080 | 0.745 | 0.603 | 0.120 | 0.777 | 0.633 | 0.090 | 0.882 | 0.826 | 0.122 |
| MDF [50] | 0.787 | 0.721 | 0.159 | 0.861 | 0.784 | 0.129 | 0.694 | 0.644 | 0.092 | 0.730 | 0.673 | 0.094 | 0.832 | 0.807 | 0.105 |
| RFCN [62] | 0.805 | 0.751 | 0.161 | 0.895 | 0.835 | 0.079 | 0.747 | 0.627 | 0.094 | 0.786 | 0.712 | 0.090 | 0.898 | 0.834 | 0.097 |
| DCL [21] + CRF | 0.823 | 0.741 | 0.141 | 0.885 | 0.853 | 0.072 | 0.739 | 0.684 | 0.097 | 0.782 | 0.714 | 0.088 | 0.890 | 0.829 | 0.088 |
| UCF [59] | 0.803 | 0.699 | 0.164 | 0.886 | 0.808 | 0.074 | 0.734 | 0.613 | 0.132 | 0.771 | 0.629 | 0.117 | 0.911 | 0.840 | 0.078 |
| DHS [22] | 0.823 | 0.774 | 0.128 | 0.892 | 0.855 | 0.052 | - | - | - | 0.815 | 0.724 | 0.065 | 0.905 | 0.872 | 0.062 |
| Amulet [58] | 0.806 | 0.755 | 0.141 | 0.895 | 0.839 | 0.052 | 0.742 | 0.647 | 0.098 | 0.778 | 0.676 | 0.085 | 0.915 | 0.870 | 0.059 |
| DSS [23] + CRF | 0.844 | 0.795 | 0.121 | 0.910 | **0.895** | 0.041 | 0.771 | 0.729 | 0.066 | 0.825 | 0.791 | 0.057 | 0.916 | 0.901 | 0.052 |
| SdBAN_VGG | 0.845 | 0.790 | 0.111 | 0.904 | 0.887 | 0.038 | 0.754 | 0.740 | 0.066 | 0.820 | 0.813 | 0.055 | 0.923 | 0.908 | 0.045 |
| ResNet50 backbone [41] | | | | | | | | | | | | | | | |
| SRM [57] | 0.843 | **0.800** | 0.127 | 0.906 | 0.874 | 0.046 | 0.769 | 0.707 | 0.069 | 0.827 | 0.757 | 0.059 | 0.917 | 0.892 | 0.054 |
| RAS [63] | 0.846 | 0.792 | 0.124 | 0.908 | 0.871 | 0.045 | 0.786 | 0.713 | 0.065 | 0.824 | 0.755 | 0.060 | 0.921 | 0.889 | 0.056 |
| PAGR [64] | 0.823 | 0.774 | 0.128 | 0.892 | 0.855 | 0.052 | - | - | - | 0.815 | 0.724 | 0.065 | 0.905 | 0.872 | 0.062 |
| CKTS [65] | - | - | - | 0.883 | - | 0.051 | 0.733 | - | 0.079 | 0.790 | - | 0.066 | 0.896 | - | 0.059 |
| PiCANet [36] | **0.853** | 0.791 | 0.109 | **0.919** | 0.870 | 0.043 | **0.794** | 0.710 | 0.068 | **0.851** | 0.755 | 0.054 | **0.931** | 0.884 | 0.047 |
| SdBAN_ResNet | 0.850 | 0.796 | **0.102** | 0.910 | 0.892 | **0.035** | 0.764 | **0.750** | **0.064** | 0.829 | **0.815** | **0.049** | 0.930 | **0.911** | **0.041** |

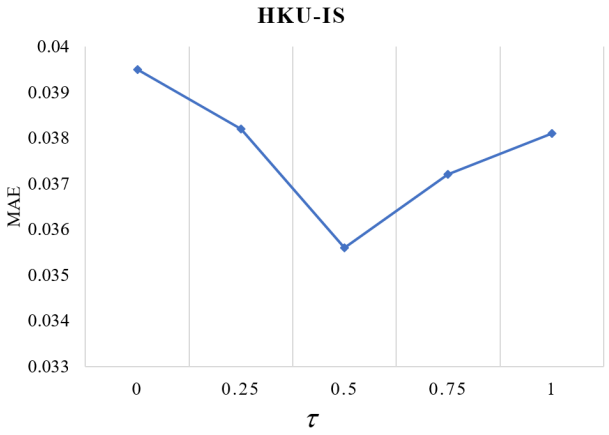



**FIGURE 9.** Quantitative evaluation of HKU-IS dataset [50] using the MAE with different $\tau$ values.

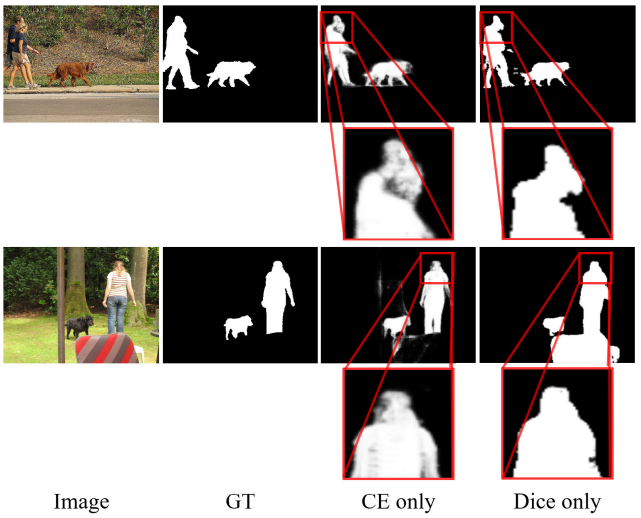



**FIGURE 10.** Prediction results based on binary cross entropy loss and dice coefficient loss respectively. The binary cross entropy loss has a soft boundary but takes into account the context of the object. The dice coefficient loss has a sharp boundary, but does not consider the context of the object.

to the overall performance. In particular, the CAB made the significant contribution in the sense of $F_\beta^w$.

### 2) EFFECTIVENESS OF THE PROPOSED LOSS FUNCTION

Fig. 9 shows the result of using different values of $\tau \in \{0, 0.25, 0.5, 0.75, 1\}$ that balances each loss. $\tau = 0$ means that only binary cross entropy loss is used for learning, whereas $\tau = 1$ means that only dice coefficient loss is used. We found $\tau = 0.5$ was the optimal by experiment. Since the dice coefficient loss is more effective than the binary cross entropy loss, a higher $\tau$ tends to give higher performance.

Fig. 10 shows results of learning with only one of the two loss functions. The difference in MAE between only cross entropy loss and only dice coefficient was not significant. However, qualitative evaluation shows the characteristics of each loss.

### 3) EFFECTIVENESS OF THE CONTEXT PATH

We used frequently used networks as backbone. Specifically, we used VGG16 [57] and ResNet50 [41] with six multi-scale feature maps 224, 112, 56, 28, 14, and 7. We did not consider 224 because it is so close to the input that the receptive field becomes very small. Figs. 12 (c) and 12 (d) respectively show CAB outputs of 112 and 56. Instead of using the single path that separately adds attention maps of each feature scale, we added the bilateral path as a separate path to learn the concatenated multi-scale attention maps. To concatenate the attention maps, 28 and 14 require $\times 8$ and $\times 16$ upsampling, respectively. This scheme can neither obtain fine information,

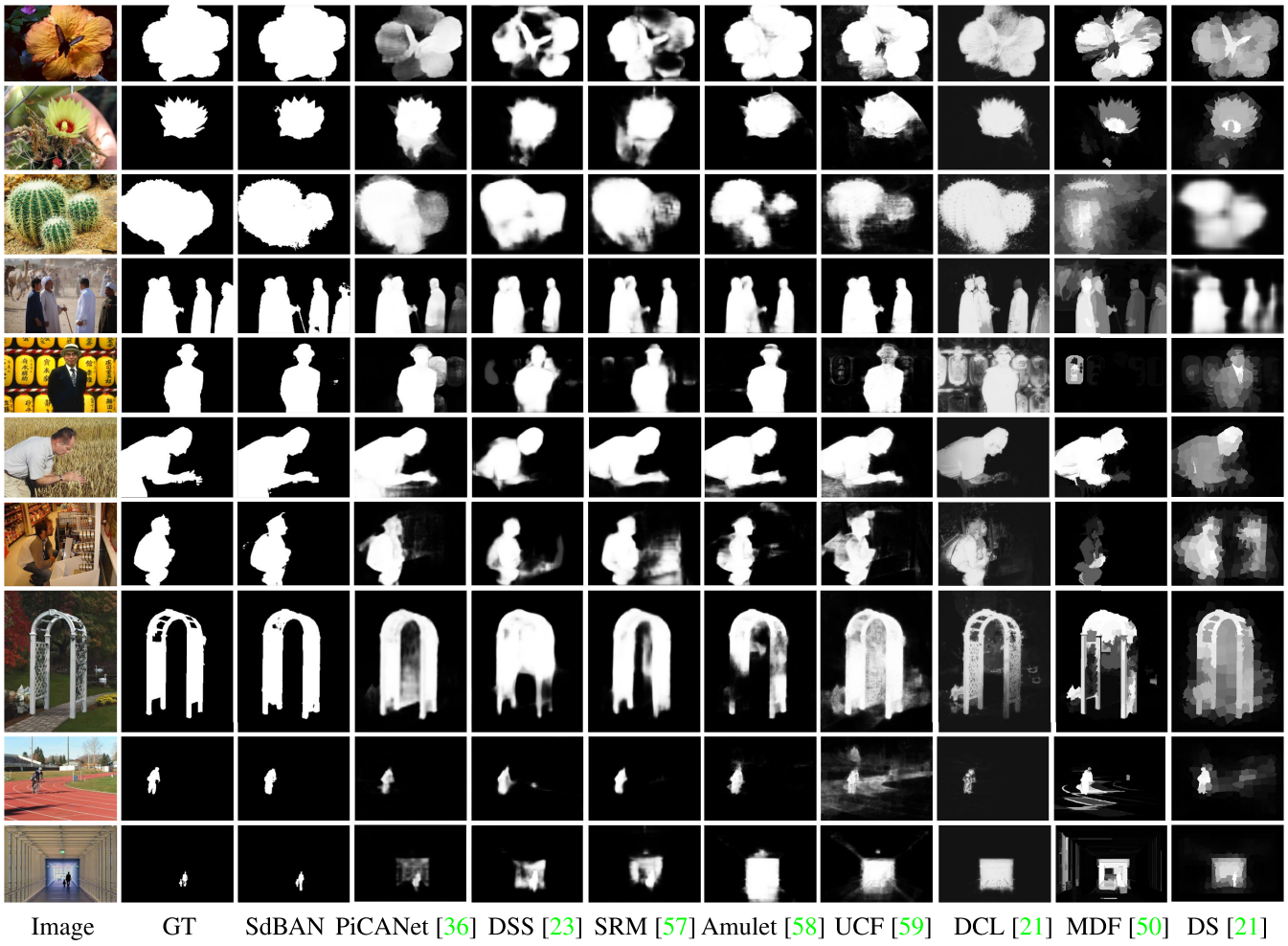



**FIGURE 11.** Qualitative evaluations with previous methods. GT means ground truth image.

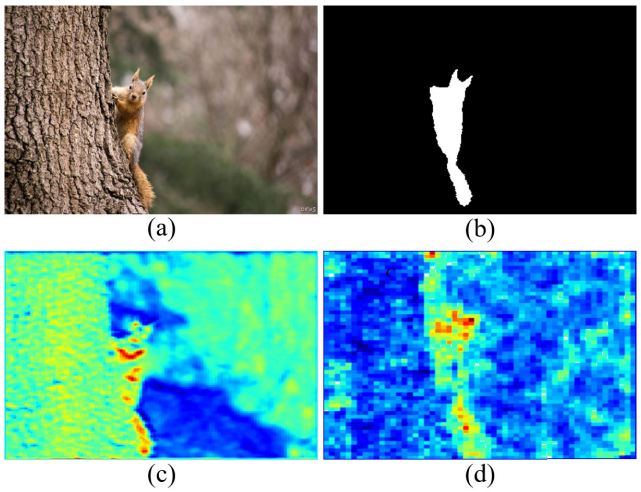



**FIGURE 12.** Attention maps of the context path: (a) input image, (b) ground truth, (c) attention map of CAB-112, and (d) attention map of CAB-56.

nor reduce errors. Experimental results showed that performance improved in 112 and 56, but not from 28.

### *E. COMPARISON WITH STATE-OF-THE-ART METHODS*

We compared our network with 12 deep learning-based state-of-the-art methods, including PiCANet [36], CKTS [65], PAGR [64], RAS [63], DSS [23], SRM [62], Amulet [61], DHS [22], UCF [60], CDL [21], RFCN [59], MDF [50], and DS [21].

#### 1) QUANTITATIVE EVALUATION

Results of quantitative comparison of the proposed network with 12 state-of-the-art methods are shown in Table 3 and Table 4. As shown in the tables, our method outperforms other methods for all the seven benchmark datasets in the sense of MAE. Our method also gives the first or second performance in the sense of $maxF_\beta$, mean $meanF_\beta$, $F_\beta^w$, and $S_\alpha$. F-measures of our method were relatively low compared to MAE. In particular, the $maxF_\beta$ is low since the PR curves of the proposed network are short as shown in Fig. 13. The shorter the PR curve, the better the binarization of the prediction output without blurring. In the F-measure curve of Fig. 13, we can see that the proposed method has a constant value, while the other methods differently behave

**TABLE 4.** Comparison of the saliency detection performance of 14 methods including ours in the sense of $F_\beta^w$ and $S_\alpha$. The best and the second-best results are highlighted in red and blue, respectively.

| Dataset | SOD [49] | | HKU-IS [50] | | DUT-O [51] | | DUTS-TE [52] | | ECSSD [53] | |
|---|---|---|---|---|---|---|---|---|---|---|
| Metric | $F_\beta^w$ | $S_\alpha$ | $F_\beta^w$ | $S_\alpha$ | $F_\beta^w$ | $S_\alpha$ | $F_\beta^w$ | $S_\alpha$ | $F_\beta^w$ | $S_\alpha$ |
| VGG-16 backbone [60] | | | | | | | | | | |
| DS [61] | - | 0.712 | - | 0.852 | - | 0.750 | - | 0.793 | - | 0.821 |
| MDF [50] | 0.501 | 0.679 | - | 0.810 | 0.565 | 0.721 | 0.509 | 0.732 | 0.705 | 0.776 |
| RFCN [62] | 0.592 | 0.730 | 0.718 | 0.858 | 0.562 | 0.774 | 0.587 | 0.792 | 0.727 | 0.852 |
| DCL [21] + CRF | 0.641 | 0.735 | 0.841 | 0.819 | 0.639 | 0.713 | 0.606 | 0.735 | 0.838 | 0.828 |
| UCF [59] | 0.644 | 0.754 | 0.751 | 0.866 | 0.565 | 0.758 | 0.588 | 0.778 | 0.789 | 0.883 |
| DHS [22] | 0.686 | 0.750 | 0.841 | 0.870 | - | - | 0.696 | 0.817 | 0.842 | 0.884 |
| Amulet [58] | 0.686 | 0.758 | 0.795 | 0.802 | 0.592 | 0.780 | 0.630 | 0.803 | 0.832 | 0.894 |
| DSS [23] + CRF | 0.718 | 0.751 | 0.866 | 0.879 | 0.691 | 0.788 | 0.754 | 0.822 | 0.871 | 0.882 |
| SdBAN_VGG | 0.722 | 0.770 | 0.866 | 0.880 | 0.694 | 0.795 | 0.785 | 0.836 | 0.898 | 0.895 |
| ResNet50 backbone [41] | | | | | | | | | | |
| SRM [57] | 0.671 | 0.742 | 0.845 | 0.887 | 0.662 | 0.797 | 0.725 | 0.834 | 0.862 | 0.895 |
| RAS [63] | - | 0.764 | - | 0.887 | - | 0.789 | - | 0.834 | - | 0.893 |
| PAGR [64] | - | - | - | 0.887 | - | 0.775 | - | 0.837 | - | 0.889 |
| CKTS [65] | - | - | - | - | - | - | - | - | - | - |
| PiCANet [36] | 0.721 | 0.772 | 0.847 | **0.904** | 0.691 | **0.815** | 0.748 | **0.850** | 0.865 | **0.910** |
| SdBAN_ResNet | **0.724** | **0.776** | **0.869** | 0.889 | **0.698** | 0.798 | **0.790** | 0.839 | **0.906** | 0.899 |

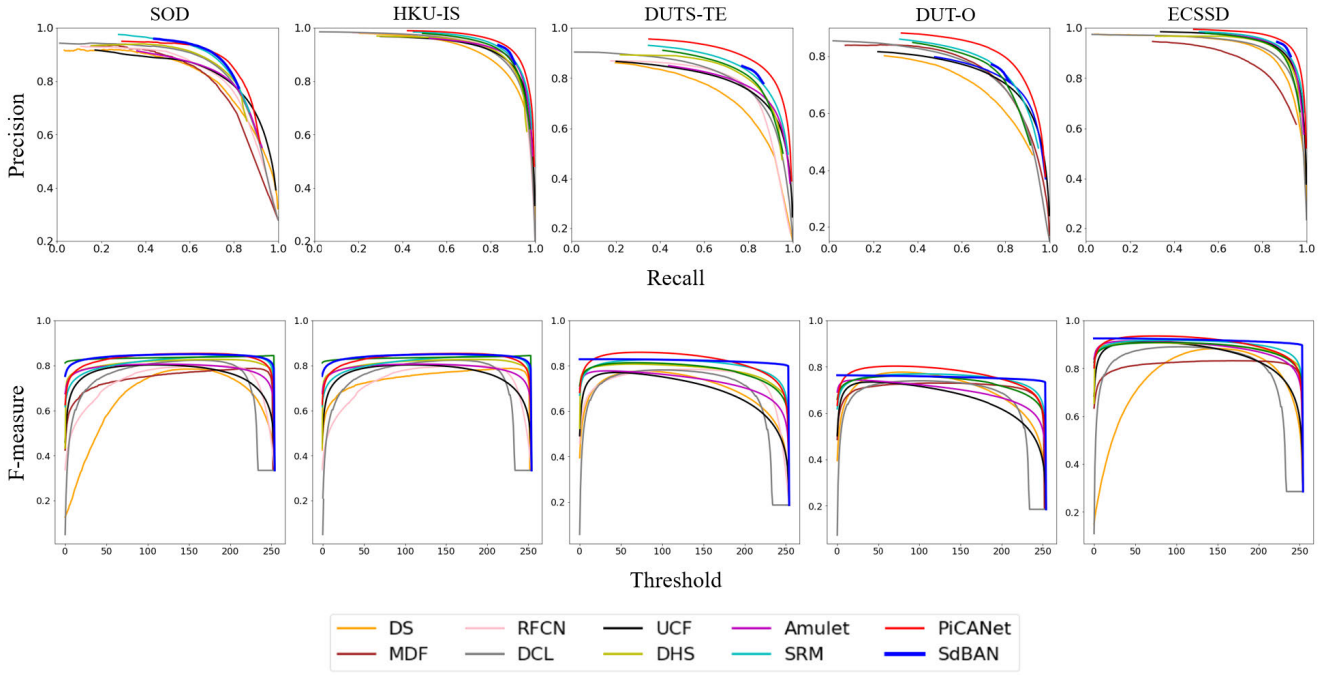



**FIGURE 13.** The PR curves and F-measure curves 10 state-of-the-arts and proposed method across five benchmark datasets.

according to the threshold. Therefore, our average value of the F-measure curve, denoted as $meanF_\beta$, shows the best performance in most cases. As shown in Fig. 4, our method gives the best results over all five datasets in the sense of $F_\beta^w$, while it gives either the best or second-best results in the sense of $S_\alpha$, which was compensated for the flaws of conventional metrics.

### 2) QUALITATIVE EVALUATION

For a subjective evaluation, we compares the saliency detection results of our method over several challenging images with existing state-of-the-art methods. As shown in the column of Fig. 11, our method generated sharper object boundaries and was less sensitive to background clutters than other methods. For example, the first three rows of Fig. 11 show the

**TABLE 5.** Processing speed of seven different methods including ours in frame per second (FPS). All experiments were conducted using a single Nvidia GTX 1080-Ti GPU.

| | SdBAN | PiCANet [36] | SRM [57] | Amulet [58] | UCF [59] | DSS [23] | DHS [22] |
|---|---|---|---|---|---|---|---|
| Size | 224 x 224 | 224 x 224 | 353 x 353 | 256 x 256 | 224 x 224 | 400 x 300 | 224 x 224 |
| FPS | **34** | 7 | 14 | 16 | 23 | 12 | 23 |

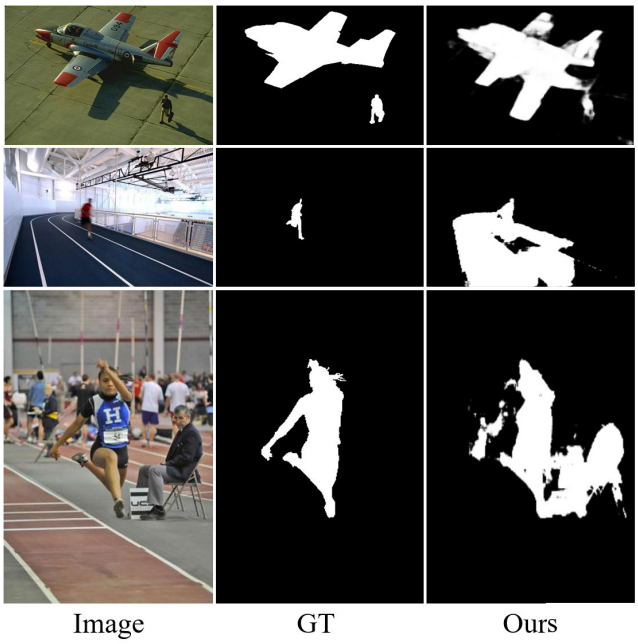



**FIGURE 14.** Failure cases of the proposed method.

case of irregular boundary. Our method accurately detected object boundaries, while other methods produced blurry boundaries. The fourth row include four objects, all of which were correctly detected by our method. Other methods missed the right most person since the existing network is trained with a center-biased training dataset. In other words, conventional learning-based methods cannot successfully detect an off-centered object. On the other hand, our method can detect an object located anywhere in the image by learning the global context of an object. However, since our method learns the global context of an object, it detect outer object well. In the case of complex background represented in the fifth to seventh rows, other methods detect the background as an object. The fifth, sixth, and seventh rows of Fig. 11 show the case of complex background. Our methods correctly detected objects in a robust manner, while others could not. The eighth row of Fig. 11 shows the case of unusually shaped structures, which was correctly detected by our method, but not by others. The ninth and tenth rows of Fig. 11 show the case of small objects, which were correctly detected by our method.

### F. FAILURE CASE

Although the proposed method correctly detected the salient object in most cases, it fails in some cases. In the first low

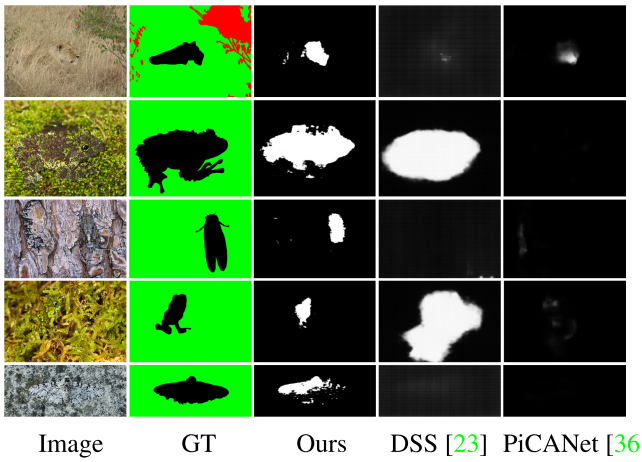



**FIGURE 15.** Qualitative evaluation results of camouflage object dataset [66].

of Fig. 14, the shadow of the plane is erroneously detected. Differentiation of a real object from its shadow is still challenging because it is similar to the object and the contrast also changes drastically. In the second row, the salient object is blurred, while the track lane is distinct. This is the case when the track lane is detected as a salient object instead of the running person. This is because most of the learning data are objects with distinct characteristics. In the third low, the salient object is a person, but the colors of the clothes vary, and the background and the color of the clothes are similar to background clutters. Also, the person to the right of the object can be a salient object in some cases. In the proposed method, the right person is detected as a salient object, and false detection occurs.

In summary, a deep learning based method is also sensitive to changes in contrast. Therefore, we will carry out a study on the regularization term in order to obtain generalized results in the future research.

### G. CAMOUFLAGE CASE

Small objects and complex backgrounds are one of the factors that make salient object detection task difficult. However, if the difference between the object and the background is very small as shown in Fig. 15, it is also a challenging problem even using human eyes. In this case, if the network does not understand the context of the object, it is difficult to obtain the correct prediction result. In Fig. 15, DSS [23] creates a grid artifact when it can not detect a salient object. This problem is the effect of the dilated convolution. Dilated convolution has the advantage that the receptive field can be

increased without increasing the parameter. However, if the difference between the background and the object is small, the background is recognized as an object and vice versa. To alleviate this problem, DSS used CRF [37] as post processing. PiCANet [36] used a similar attention mechanism to the proposed method. However, as shown in the experimental results, the attention mechanism of PiCANet did not find a discriminative part of the image. Unlike the previous two methods, the proposed method sensitively responds to the small context of the object. It is not enough to predict the detail of an object, but the discriminative region of the object is well detected.

### H. PROCESSING TIME

Processing time of the proposed method is compared with other methods as shown in Table 5.

## V. CONCLUSION

In this paper, we proposed a novel saliency detection method using a bilateral attention network (SdBAN) consisting of: i) a spatial path containing an encoder-decoder structure to learn the spatial information of the salient object and ii) a context path containing the attention-module structure to learn the context information. Weight vectors of different scales in the attention network are concatenated through the feature fusion module at the last layer to effectively preserve the information of each path. In addition, effective learning is achieved by incorporating a novel loss function based on the invariant index of the salient object scale and dice coefficient loss along with the cross entropy loss. As a result of the comparison with the state-of-the-art methods for five different datasets, we demonstrate that the proposed network performs best in most cases. The proposed method also outperforms existing methods in the sense of processing speed in frames per second. In addition to the quantitative evaluation, qualitative performance of the proposed method is much better than others especially in camouflage cases.

## REFERENCES

[1] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Oct. 2009, pp. 817–824.

[2] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2. Oct. 2004, p. 2.

[3] A. Abdulmunem, Y.-K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Comput. Vis. Media*, vol. 2, no. 1, pp. 97–106, Mar. 2016.

[4] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[5] Q. Hou, P. Kumar Dokania, D. Massiceti, Y. Wei, M.-M. Cheng, and P. Torr, "Bottom-up top-down cues for weakly-supervised semantic segmentation," 2016, *arXiv:1612.02101*. [Online]. Available: http://arxiv.org/abs/1612.02101

[6] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1568–1576.

[7] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, and S. Yan, "Learning to segment with image-level annotations," *Pattern Recognit.*, vol. 59, pp. 234–244, Nov. 2016.

[8] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 23–30.

[9] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[10] J. Guo, T. Ren, L. Huang, X. Liu, M.-M. Cheng, and G. Wu, "Video salient object detection via cross-frame cellular automata," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 325–330.

[11] H. Jacob, F. L. C. Pádua, A. Lacerda, and A. C. M. Pereira, "A video summarization approach based on the emulation of bottom-up mechanisms of visual attention," *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 193–211, Oct. 2017.

[12] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, p. 194, 2001.

[13] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2214–2219.

[14] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 2083–2090.

[15] Y. Kong, J. Zhang, H. Lu, and X. Liu, "Exemplar-aided salient object detection via joint latent space embedding," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5167–5177, Oct. 2018.

[16] Y. Zeng, M. Feng, H. Lu, G. Yang, and A. Borji, "An unsupervised game-theoretic approach to saliency detection," *IEEE Trans. Image Process.*, vol. 27, no. 9, pp. 4545–4554, Sep. 2018.

[17] L. Zhou, Z. Yang, Z. Zhou, and D. Hu, "Salient region detection using diffusion process on a two-layer sparse graph," *IEEE Trans. Image Process.*, vol. 26, no. 12, pp. 5882–5894, Dec. 2017.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[19] V. Mnih, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[20] S. He, R. W. H. Lau, W. Liu, Z. Huang, and Q. Yang, "SuperCNN: A superpixelwise convolutional neural network for salient object detection," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 330–344, Dec. 2015.

[21] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 478–487.

[22] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[23] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.

[24] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2300–2309.

[25] K. Fu, Q. Zhao, I. Yu-Hua Gu, and J. Yang, "Deepside: A general deep framework for salient object detection," *Neurocomputing*, vol. 356, pp. 69–82, Sep. 2019.

[26] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[27] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7264–7273.

[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[29] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*. [Online]. Available: http://arxiv.org/abs/1810.04805

[30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," 2015, *arXiv:1508.04025*. [Online]. Available: http://arxiv.org/abs/1508.04025

[31] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473

[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," 2015, *arXiv:1502.03044*. [Online]. Available: http://arxiv.org/abs/1502.03044

[33] P. Sermanet, A. Frome, and E. Real, "Attention for fine-grained categorization," 2014, *arXiv:1412.7054*. [Online]. Available: http://arxiv.org/abs/1412.7054

[34] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.

[35] J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, and S. Yan, "Attentive contexts for object detection," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 944–954, May 2017.

[36] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[37] Y. Zhang and T. Chen, "Efficient inference for fully-connected CRFs with stationarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 109–117.

[38] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.

[39] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[40] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* New York, NY, USA: Springer, 2015, pp. 234–241.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[42] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579

[43] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013, *arXiv:1312.4400*. [Online]. Available: http://arxiv.org/abs/1312.4400

[44] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[45] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1395–1403.

[46] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 3048–3060, Jun. 2019.

[47] R. Deng, C. Shen, S. Liu, H. Wang, and X. Liu, "Learning to predict crisp boundaries," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 562–578.

[48] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

[49] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 49–56.

[50] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[51] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[52] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

[53] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.

[54] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[55] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

[56] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[58] 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," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[59] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* New York, NY, USA: Springer, 2016, pp. 825–841.

[60] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.

[61] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[62] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

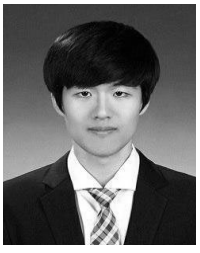
[63] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 234–250.

[64] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 714–722.

[65] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 355–370.

[66] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, and A. Kornacki. (2017). *Chameleon Database—Animal Camouflage Analysis*. [Online]. Available: http://zgwisk.aei.polsl.pl/index.php/en/research/other-research/63-animal-camouflage-analysis

**DONGGOO KANG** was born in Seoul, South Korea, in 1992. He received the B.S. degree in financial economics from Seokyeong University, South Korea, in 2018. He is currently pursuing the M.S. degree with the Department of Image, Chung-Ang University. His research interests include salient object detection and image segmentation.



**SANGWOO PARK** (Student Member, IEEE) was born in Inchen, South Korea, in 1989. He received the B.S. degree in electric and electronic engineering from Soon Chun Hyang University, South Korea, in 2015, and the M.S. degree from the Department of Image, Chung-Ang University, South Korea, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include image translation and object detection.

**JOONKI PAIK** (Senior Member, IEEE) was born in Seoul, South Korea, in 1960. He received the B.S. degree in control and instrumentation engineering from Seoul National University, in 1984, and the M.Sc. and Ph.D. degrees in electrical engineering and computer science from Northwestern University, in 1987 and 1990, respectively. From 1990 to 1993, he was with Samsung Electronics, where he designed image stabilization chipsets for consumer camcorders. Since 1993, he has been a member of the faculty with Chung-Ang University, Seoul, where he is currently a Professor with the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 1999 to 2002, he was a Visiting Professor with the Department of Electrical and Computer Engineering, The University of Tennessee, Knoxville. Since 2005, he has been the Director of the National Research Laboratory in the field of image processing and intelligent systems. From 2005 to 2007, he served as the Dean of the Graduate School of Advanced Imaging Science, Multimedia, and Film. From 2005 to 2007, he was the Director of the Seoul Future Contents Convergence Cluster established by the Seoul Research and Business Development Program. In 2008, he was a full-time Technical Consultant of the System LSI Division, Samsung Electronics, where he developed various computational photographic techniques, including an extended depth of field system. He has served as a member of the Presidential Advisory Board for Scientific/Technical Policy with the Korean Government. He is currently serving as a Technical Consultant of Korean Supreme Prosecutor's Office for computational forensics. He was a two-times recipient of the Chester-Sall Award from the IEEE Consumer Electronics Society, the Academic Award from the Institute of Electronic Engineers of Korea, and the Best Research Professor Award from Chung-Ang University. He has served the Consumer Electronics Society of the IEEE as a member of the Editorial Board, the Vice President of International Affairs, and the Director of Sister and Related Societies Committee.