Multi-Level Context Aggregation Network With Channel-Wise Attention for Salient Object Detection

ZIHUI JIA, ZHENYU WENG, FANG WAN, AND YUESHENG ZHU, (Senior Member, IEEE)
School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School, Shenzhen 518055, China
Corresponding author: Yuesheng Zhu (zhuys@pku.edu.cn)

This work was supported in part by the Shenzhen Municipal Development and Reform Commission through the Disciplinary Development Program for Data Science and Intelligent Computing, in part by the Research and Development Program in Key Areas of Guangdong Province under Grant 2019B010137001, and in part by the NSFC-Shenzhen Robot Jointed Founding under Grant U1613215.

ABSTRACT
Fully convolutional neural networks (FCNs) have shown their advantages in the salient object detection task. However, the prediction results do not perform well in most existing FCN-based methods, such as coarse object boundaries or even getting wrong predictions, which resulted from ignoring the difference between multi-level features during feature aggregation or underutilizing the spatial details suitable for locating boundaries. In this paper, we propose a novel end-to-end multi-level context aggregation network (MLCANet) to solve the problem mentioned-above, in which both bottom-up and top-down message passing can cooperate in a joint manner. The bottom-up process that aggregates low-level fine details features into high-level semantically-richer features would enhance high-level features, and in turn the top-down process that passes refined features from deeper layers to the shallower ones could benefit from the enhanced high-level features. Also by considering that the features from different layers may not be equally important, a multi-level feature aggregation mechanism with channel-wise attention is proposed to aggregate multi-level features by flexibly adjusting their contributions and absorbing useful information to refine themselves. The features after message passing which simultaneously encode semantic information and spatial details are used to predict saliency maps in our network. Extensive experiments demonstrate that our method can obtain high quality saliency maps with clear boundaries, and perform favorably against the state-of-the-art methods without any pre-processing and post-processing.

INDEX TERMS
Salient object detection, feature aggregation, channel-wise attention.

I. INTRODUCTION
The goal of salient object detection (SOD) is to find one or more objects which attract most attention in the given image or video and then segment these objects out. It has received widespread attention recently and played an important role in many computer vision tasks, such as visual tracking [1], content-aware image editing [2], and robot navigation [3]. The core challenge of this task is how to keep a high accuracy when detecting salient objects in different complex scenes (e.g. have a low contrast with the background or in cluttered scenes). Therefore, it is necessary to design an effective network to achieve accurate prediction.

Traditional methods mostly use hand-crafted features (e.g. colors, texture, contrast, or others) to capture local details and global context separately or simultaneously. However, these hand-crafted features can hardly capture high-level semantic relations and context information, thus they usually fail to detect salient objects with complex scenes.

Recently, fully convolutional neural networks (FCNs) [4] have been adopted for salient object detection for they providing abundant and discriminative image representations. FCN-based methods have achieved remarkable performance in salient object detection, outperforming the traditional methods. Nevertheless, FCN-based methods either focus on aggregating multi-level features while ignoring the spatial details which is suitable for locating boundaries, or directly combine features of different levels while ignoring the difference
between themselves which usually introduces noise and leads to some wrong predictions when positioning and segmenting salient objects. To address the first problem, there has been increasing interest in jointly refining salient object and boundary. Some methods consider both salient object features and boundary features to improve the segmentation results: adding auxiliary boundary detection branch [5]–[7], designing boundary aware loss [8], and considering the complementarity between the salient object features and boundary features [9]. Since there are lots of pixels near the boundaries prone to wrong predictions, these methods may confront the problem of inaccurate boundary features though they can effectively improve the segmentation result. Fig. 1 shows some visual examples.

Motivated by the above observations, we propose a novel grid-like network named Multi-level Context Aggregation Network (MLCANet) for salient object detection in this paper. Instead of adding auxiliary boundary detection branch or designing boundary aware loss, we design a efficient multi-level feature aggregation mechanism to reduce the impact of inconsistency between features of different levels, and make full use of both low-level fine details feature and high-level semantically-rich features to predict high quality saliency maps with clear boundaries. Firstly, as low-level features preserve rich spatial details as well as background noises, and the high-level features carry semantically-rich information, to mitigate the discrepancy between multi-level features, we design a feature aggregation module with channel-wise attention which aggregates different level features by allowing efficient information exchange and absorb useful information to refine themselves. Compared with common feature aggregation strategies like addition and concatenation, our feature aggregation module is able to flexibly adjust the contributions from different level features. Secondly, our model integrates both top-down and bottom-up message passing in a joint and cooperative manner. The spatial details of low-level features are aggregated into high-level features through bottom-up message passing process, and the semantic information of enhanced high-level features is aggregated into low-level features through top-down message passing process. As a result, the semantic information and spatial details are incorporated at each level. We also conduct a series of ablation experiments to investigate the effectiveness of different combinations of message passing mechanisms in SOD task. Thirdly, we introduce the joint multi-level coarse-to-fine saliency inference module that makes full use of complementary and powerful multi-level features. We consider top-layer saliency to be a globally harmonious interpretation of a visual scene, which is used to guide fine-grained saliency estimation like a spatial attention map in a top-down manner, and gradually integrate of inferior-layer features with upper-layer saliency estimates in a coarse-to-fine manner. In summary, our method not only takes into account the difference between multi-level features, but also makes full use of all level features for saliency estimation.

To demonstrate the performance of MLCANet, we report our experimental results on five popular SOD datasets and compare the results of our method with other 13 state-of-the-art salient object detection networks. A series of ablation studies are conducted to evaluate the effect of each module. Quantitative indicators and visual results show that MLCANet can obtain significantly better local details and high quality saliency maps. In short, our main contributions can be summarized as follows:

- We propose a novel grid-like Multi-level Context Aggregation Network that integrates both bottom-up and top-down message passing in a joint and cooperative manner, whose saliency estimation inference is in a coarse-to-fine manner by gradually integrating upper-layer saliency estimates with inferior-layer features.
- To the best of our knowledge, this is the first paper to investigate the effectiveness of different combinations and configurations of message passing mechanisms in SOD task. The conclusions drawn in this paper are the joint bottom-up and top-down message passing can achieve the best results and only one round of one-way message passing or one directional message passing is not enough to pass sufficient content.
- We design a feature aggregation module with channel-wise attention to aggregate features of different levels, which is able to flexibly adjust the contributions from features of different level to avoid introducing too much redundant information and allow efficient information exchange.
- Experimental results demonstrate that the proposed model MLCANet achieves the state-of-the-art performance on five datasets, which proves the effectiveness and superiority of the proposed method.

This rest of paper is organized as follows: Section 2 briefly presents the related work. Section 3 describes details of the proposed method. We decouple the network structure and each module is described in detail in different subsections. Section 4 provides the details of implementation and the results of numerical evaluations in comparison with other state-of-the-art SOD methods. We also conduct several ablation experiments to demonstrate that all modules improve our method’s performance. Finally, Section 5 draws conclusions.
II. RELATED WORK

Early SOD methods mainly rely on intrinsic cues to estimate saliency maps, such as contrast [11], center prior [12], boundary background [13] and texture [14], which mainly focus on low-level handcrafted features and hardly capture rich contextual semantic information. Recently, deep learning based salient object detection methods that we focus on have achieved the state-of-the-art performances.

A. FCN-BASED METHODS

Early FCN-based methods [15] detect salient objects by combining traditional saliency prior information and have achieved better prediction than other traditional methods. Meanwhile, there are some works attempting to fuse features at different levels to produce high-precision saliency map. In [16], a strongly supervised short-connection structure on the basis of HED [17] is proposed, which enables the model to utilize high-level semantic features and low-level detail features effectively. In [18], [19], the authors both advance the U-shape structures and utilize multiple levels of context information for accurate detection of salient objects. In [20], attention mechanisms are embedded into U-shape models to guide the feature integration. In [21], a controlled bi-directional passing of features between shallow and deep layers is proposed to obtain accurate predictions. A general deep framework called Deepside that can be deeply supervised to incorporate hierarchical CNN features is proposed in [22]. In [23], the authors develop a recurrent saliency detection model that transfers global information from the deep layer to shallower layers by a multi-path recurrent connection. A recurrent FCN is designed in [15] for saliency detection by correcting prediction errors iteratively. In [24], the authors propose a novel recurrent residual refinement network (R3Net) to progressively refine the saliency map by constructing a sequence of residual refinement blocks. A simple yet effective framework named progressive feature polishing network is proposed in [25]. By progressively polishing the multi-level features, this network can detect salient objects accurately with fine details without any post-processing.

B. DEEP BOUNDARY- AWARE METHODS

More recently, some researchers also attempt to use boundary information to achieve better representation of segmentation feature. A Coarse-to-Fine Network (BASNet) is proposed in [8]. It designs a novel hybrid loss to enhance the prediction near the boundary by a larger loss. In [26], a boundary-enhanced loss is presented for learning fine boundaries and worked with the cross-entropy loss for saliency detection. In [6], the boundary maps are predicted by an external edge labels and then incorporate it with U-Net architecture to detect salient objects. An edge guidance network is proposed in [9] which considers the complementarity between salient objects and boundary. In [27], a stacked cross refinement network is proposed to mutually complement the information of salient object and boundary while recurrently refine these information.

Instead of adding auxiliary boundary detection branch or designing boundary aware loss, our proposed method continues to tap the potential of multi-level feature aggregation. Our method clearly differs from the above-mentioned FCN-based methods in four aspects. Firstly, as context information is very important for SOD, some methods (e.g. [19], [21]) design a specific module to acquire large receptive field. These methods perform context information extraction for all level features, but do not consider that this operation may destroy the spatial details in the low-level features, which results in predicting saliency maps with unclear boundaries. Our method only employs Multi-scale Context-Aware Feature Extraction Module (MCFEM) in [21] to capture the multi-receptive-field context information of high-level features and use a convolution block to process low-level features so as not to destroy their rich spatial details. Secondly, most FCN-based methods ignore the difference between multi-level features during feature aggregation which usually introduces noise and leads to some wrong predictions. In contrary, our method employs our proposed feature aggregation module to aggregate multi-level feature by flexibly adjusting their contributions and absorbing useful information to refine themselves. Thirdly, Most methods simply aggregate multi-level features from one direction message passing, i.e. bottom-up or up-down, there are also some methods (e.g. [19], [21]) to employ one round of bi-directional message passing to pass the spatial details of low-level features and the semantic information of high-level features to all level features, but we have found through ablation experiments that only one round of bi-directional message passing or one directional message passing is not enough to pass sufficient information and two or more rounds of bi-directional message passing can significantly improve network performance. Finally, all previous methods predict saliency maps in three forms: 1) fuse all level features to predict the final output; 2) learn the residuals from the initial high-level prediction to the final output; 3) fuse the multi-level predictions to get the final output. Our method proposes a new form of prediction, considering the initial high-level prediction as a globally harmonious interpretation of a visual scene to guide fine-grained saliency estimation like spatial attention map in a top-down manner, and gradually integrate inferior-layer complementary and robust features with upper-layer saliency prediction in a coarse-to-fine manner.

III. NETWORK ARCHITECTURE

In this paper, we design a novel grid-like network structure with five rows and four columns as shown in Fig. 2. In terms of the whole network, each row corresponds to a different feature scale and consists of five residual dense blocks (RDBs) [10] that keep the number of feature maps unchanged. Each column can be regarded as a bridge that connects different scale via upsampling/downsampling modules.
FIGURE 2. The architecture of the proposed multi-level context aggregation network. Our network consists of four main parts: feature extraction, bottom-up message passing, top-down message passing and multi-level saliency inference. We employ residual dense block (RDB) [10] as basic feature processing block.

In this section, we decouple the network structure and each module is described in detail in different sub-sections.

A. FEATURE EXTRACTION
The VGG-16 [28] network suggested by other deep learning based methods [16], [18] is applied to extract multi-scale features. From the backbone network, we can obtain five side features Conv1-2, Conv2-2, Conv3-3, Conv4-3, Conv5-3. We take Conv1-2 and Conv2-2 in VGG-16 as the basic low-level features because they preserve better spatial details, and take Conv3-3, Conv4-3 and Conv5-3 as the basic high-level features because they carry more high-level semantics, with less spatial details. Because the salient objects are variant in shapes, scales and positions, stacking multiple convolutional and pooling layers may not be effectively to handle these complicated variations. In order to make the final extracted high-level features invariant to the scales and shapes of objects and capture multi-receptive-field context information, we introduce Multi-scale Context-aware Feature Extraction Module (MCFEM) in [21] which contains four dilated convolutions with different dilation rates. Specifically, the four different dilation rates are set to 1, 3, 5 and 7 respectively. The details of MCFEM is shown in Table. 1. For low-level features, different from [21], we just use a convolution block to process them so as not to destroy their rich spatial details. In this paper, low-level features with high-resolution have a lower number of channels, while high-level features with low-resolution have a higher number of channels. The number of channels is set to 64 and 128 respectively in our network.

B. BASIC FEATURE PROCESSING BLOCK
RDB [10] is one important component of our network. Fig. 3 provides a detailed structure of the RDB block. It is employed to improve the information flow and avoid gradient vanishing since our network does not use any normalization measures like batch normalization (BN) [29]. Each RDB block consists of five convolutional layers: the first four convolutional layers aim to increase the number of feature maps while the last layer reduces the number of feature maps to the original number and fuses these feature maps. The output is then combined with the input of this RDB block via channel-wise addition. The growth rate in RDB is set to 16 in this paper.

C. JOINT BOTTOM-UP AND TOP-DOWN MESSAGE PASSING
The consensus is that top layers of the network carry semantically-rich information, while bottom layers involve

![Diagram](image-url)
low-level details. Joint bottom-up and top-down message passing aims to generate robust and complementary multi-level features which will be used to predict accurate saliency maps with clear boundaries. First, we pass the low-level details through a bottom-up process and gradually refine the high-level features. Then the refined high-level semantic knowledge or contextual information helps the bottom layers better locate the salient regions through a top-down message passing in return.

The central hypothesis is that the bottom layers transmit more spatial details to top-network layers to refine high-level contextual information which is used to provide a more accurate high-level saliency estimate without background noises through a bottom-up process; the refined high-level features further enable more efficient top-down message passing. In this way, these two processes are performed in a joint manner and multi-level features could cooperate with each other to generate more accurate results. Fig. 4 shows a visual example of different levels features after message passing. Besides, considering that the multi-level features are of various resolutions, upsample and downsample module are added during the process of top-down and bottom-up message passing. Here upsample/downsample module is realized by a bilinear interpolation and a $3 \times 3$ convolutional layer that adjusts the feature channels.

![Visualization of different levels features after message passing.](image)

(a) Input (b) Low-level (c) Enhanced (d) Output

(e) Ground Truth (f) High-level (g) Enhanced (h) Output

**FIGURE 4.** Visualization of different levels features after message passing. (b) and (f) represent the low-level and high-level features respectively. (c) and (g) mean the enhanced low-level and high-level features after top-down message passing and bottom-up message passing respectively. (d) and (h) are the saliency estimates of the rows corresponding to low-level and high-level features. Best viewed in color mode.

**D. FEATURE AGGREGATION MODULE WITH CHANNEL-WISE ATTENTION**

As we know, low-level features preserve rich spatial details as well as background noises due to the restriction of the receptive field. These features usually contain clear boundaries which play a significant role in generating accurate saliency maps. In contrast, high-level features have coarse boundaries as a result of multiple downsamplings. Despite of losing lots of spatial details, high-level features still have consistent semantics and clear background. There is huge statistical discrepancy between these two kinds of features.

Given that feature maps from different levels may not have the same importance, we propose a channel-wise attention mechanism to weighted multi-level multi-receptive-field features. Let $F^l_i$ and $F^h_i$ denote the $i$th feature channel from the row stream and the column stream respectively, and let $\alpha^l_i$ and $\alpha^h_i$ represent their associated channel-wise attention weights. Both $F^l_i$ and $F^h_i$ have same channel $C$. The output of channel-wise attention mechanism can be expressed as

$$\hat{F}^i = \alpha^l_i F^l_i + \alpha^h_i F^h_i$$  \hspace{1cm} (1)$$

where $\hat{F}^i$ stands for the aggregated feature in the $i$th channel. This attention mechanism enables our network to flexibly adjust the contributions from different levels in feature aggregation. By multiple feature aggregations, low-level features and high-level features will gradually absorb useful information from each other to complement themselves.

Specifically, as show in Fig. 5, our feature aggregation module (FAM) contains two branches, one for $F^l_i$ and one for $F^h_i$. We take a branch for $F^l_i$ as example. First, a average pooling is applied to each $F^l_i$ to get an embedding of the global distribution of channel-wise feature vector $v^l_i \in \mathbb{R}^C$. After that, we apply two fully-connected (FC) layers to fully capture channel-wise dependencies. As [30], we encode the channel-wise feature vector by forming a bottleneck with two FC layers around the non-linearity to limit model complexity and aid generalisation. Then, through employing a sigmoid activation function, the value of encoded channel-wise feature vector is constrained between 0 and 1 to indicate the importance of each channel.

$$\alpha = \sigma(f_2(\delta(f_1(v^l_i, W_1), W_2)) \hspace{1cm} (2)$$

where $W_1$ and $W_2$ refer to parameters in channel-wise attention block, $\sigma$ refers to sigmoid activation function, $f_1$ and $f_2$ refer to FC layers, $\delta$ refers to the ReLU function.

Our experimental results show that the performance of our proposed network can be greatly improved with the introduction of this feature aggregation mechanism.
process is summarized as follows: where BCEs to facilitate sufficient training. The MLS is defined as: we add multi-level supervision (MLS) consisting of multiple the most widely used loss function. As we mentioned above, In saliency object detection, binary cross entropy (BCE) is P and progressively transmitted to inferior-layers. And we take processing. See Fig. 6 for specific visual examples.

Using (3), upper-layer saliency estimates are hierarchically operation among channel axis and r

Conv where module takes the feature map saliency estimates in a coarse-to-fine manner. Our inference visual scene, which is used to guide fine-grained saliency estimation gets closer to the ground truth, which leads to the inconspicuous separation of foreground and background. To address this issue, we introduce F-measure based loss function (FLoss) [31] which can hold considerable gradients even in the saturated area, resulting in ‘high-contrast’ and polarized predictions. FLoss [31] can be defined as:

\[
FLoss = 1 - \frac{(1 + \beta^2)TP}{H}
\] (5)

where \( H = \beta^2(TP + FN) + (TP + FP) \). In [31], the true positive, false positive and false negative are reformulated based on the continuous posterior \( Y \):

\[
\begin{align*}
TP(\hat{Y}, Y) &= \sum_i \hat{y}_i \cdot y_i, \\
FP(\hat{Y}, Y) &= \sum_i \hat{y}_i \cdot (1 - y_i), \\
FN(\hat{Y}, Y) &= \sum_i (1 - \hat{y}_i) \cdot y_i,
\end{align*}
\] (6)

where \( Y \) is the ground-truth. In our experiment, we add FLoss [31] to two low-level saliency maps.

\[
\mathcal{L}_F = \frac{2}{r} \sum_{r=1}^N \theta_r \cdot FLoss
\] (7)

where \( \theta_r \) are the weights for different level.

The total loss \( \mathcal{L} \) of our network is defined as:

\[
\mathcal{L} = \mathcal{L}_{MLS} + \mathcal{L}_F
\] (8)

IV. EXPERIMENTS

In this section, we elaborate upon: (1) our implementation details, (2) datasets and evaluation metric, (3) ablation study, in which we investigate the effectiveness of each key module in our network, (4) the experimental comparison, in which we compare the performance of our method with other state-of-the-art SOD methods.

A. IMPLEMENTATION DETAILS

VGG-16 [28] pre-trained on Imagenet [32] is used as the backbone network. As the previous works do, the proposed model is trained on DUTS [33] which contains 10553 images. We implement the proposed model in PyTorch and employ Adam [34] for gradient descent optimization. The learning rate is set as \( 10^{-4} \) until the losses converge. Then it is adjusted to \( 10^{-5} \) to fine-tune the networks until convergence. All the weights of newly added convolution layers are initialized by normal distribution with 0.01 standard deviation and zero mean value. Some data augmentation techniques such as random brightness, saturation and contrast changing, and random horizontal flipping are adopted to improve the robustness of the model. During both training and testing, the sizes of the input images remain unchanged. So the batch-size is set to 1 and BN [29] is not adopted in our network. \( \lambda \)’s are set to 1.0, 0.8, 0.8, 0.6, 0.6 and \( \theta \)'s are set to 1.0, 1.0.
B. DATASETS AND EVALUATION METRIC

We evaluate the proposed method on five widely used public benchmark datasets: ECSSD [14], PASCAL-S [35], DUT-OMRON [13], HKUIS [36], DUTS [33]. ECSSD [14] consists of 1000 images with even more complex and semantically meaningful salient objects on the scenes. PASCAL-S [35] contains 850 images in which different salient objects are labeled with different saliencies. DUT-OMRON [13] has 5168 more difficult and challenging images and most images contain one or more salient objects with fairly complex background. HKUIS [36] consists of 4447 challenging images and most of them have low color contrast or multiple disconnected salient objects. DUTS [33] is the largest saliency detection benchmark dataset which contains 10553 training images and 5019 testing images.

We evaluate the performance of our proposed method as well as related comparison salient object detection methods using three metrics: F-measure, mean absolute error (MAE) and S-measure [37]. Precision denotes the ratio of ground truth salient pixels in the predicted saliency map. And recall denotes the percentage of the detected salient pixels and the all ground truth salient pixels. The precision and recall are calculated by comparing the binary image under different thresholds (from 0 to 255) between the predicted saliency map and corresponding ground truth. The F-measure is an overall evaluation standard, computed by the weighted combination of precision and recall with the formulation as follows:

$$F_{\beta} = \frac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}}$$  \hspace{1cm} (9)$$

where $\beta^2 = 0.3$, which is consistent with previous work [13], to emphasize precision more than recall. In this paper, We report the Max F-measure from all precision-recall pairs.

The MAE score indicates the average pixel-wise absolute difference between a predicted saliency map $P$ and the ground truth $G$:

$$\text{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |P(i, j) - G(i, j)|$$  \hspace{1cm} (10)$$

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

S-measure [37] focuses on evaluating the spatial structure similarities of saliency maps. S-measure could be computed as:

$$S = \alpha \ast S_o + (1 - \alpha) \ast S_r$$  \hspace{1cm} (11)$$

where $\alpha$ is set to 0.5 by default, to balance the object-aware structural similarity($S_o$) and region-aware structural similarity ($S_r$).

C. ABLATION STUDY

In this subsection, we first investigate the effectiveness of our proposed FAM. Then, we conduct experiments on the joint top-down and bottom-up message passing. The ablation experiments are conducted on the DUTS [33] dataset.

1) EFFECTIVENESS OF FAM

To demonstrate the effectiveness of our proposed FAMs, we replace FAMs with two common feature aggregation strategies (addition and concatenation).

2) EFFECTIVENESS OF JOINT BOTTOM-UP AND TOP-DOWN MESSAGE PASSING

To demonstrate the effectiveness of joint bottom-up and top-down message passing, we conduct four experiments on different combinations of top-down and bottom-up message passing. All other configurations are the same.

- Top-down message passing Only.
- Bottom-up message passing Only.
- Joint top-down and bottom-up message passing.
- Joint bottom-up and top-down message passing.

The result evaluated on the DUTS [33] dataset is shown in Table. 2 and visual comparisons are shown in Fig. 7. From the results shown in Table. 2 and Fig. 7, we have the following observations. (1) From the first two row of table, we can conclude that our proposed FAM can significantly improve network performance compared to other two common feature aggregation strategies (addition and concatenation).
The proposed FAM allows our network to avoid introducing too much redundant information and absorb useful information to refine different level features. Therefore, our network can produce accurate and high-quality saliency prediction. (2) Only top-down or bottom-up message passing is not enough to produce detailed saliency estimation. Compared the first and last row in Fig. 7, we find that only bottom-up message passing tends to produce estimates with blur boundary while only top-down message passing tends to produce estimates with unclean background. In contrast, the cooperation of joint bottom-up and top-down message passing can effectively boost the multi-level features integration and yield robust features. (3) The order of the bottom-up and top-down message passing also has a significant impact on the performance of the network.

We perform further ablation studies by considering different configurations of the message passing module of the proposed MLCANet. Note that each row in our network corresponds to a different feature scale, and the columns in our network serve as bridges to facilitate the information exchange across different feature scales. Table 3 shows how the performance of the proposed MLCANet depends on the number of columns (denoted by col) in the message passing module. It is clear that only one round of bi-directional message passing is not enough to pass sufficient content and the performance gaps between two or more round of bi-directional message passing is not significant but the improvement is significantly more than that of only one round of bi-directional message passing. Thus we set column to 4 in our network.

As we can see, all these modules of our network do improve the model performance. When these modules are combined, the best SOD results are finally obtained.
It demonstrates that all components are necessary and essential for the proposed network.

**D. COMPARISONS TO THE STATE-OF-THE-ARTS**

We compare the results of our method with other 14 state-of-the-art salient object detection networks. They are DCL [38], ELD [39], RFCN [15], DSS [16], NLD [18], Amulet [19], PAG [23], BMPM [21], AFNet [26], MLMSNet [7], PAGENet [40], poolNet [6], EGNet [9] and PFPN [25] (results are not publicly available). For a fair comparison, all of these methods use VGG-16 as backbone network. All the saliency maps are provided by the authors.

1) **QUANTITATIVE EVALUATION**

The comparisons of detection quality between our model and 14 previous state-of-the-art methods in terms of F-measure, MAE and S-measure score are shown in Table 4. It is evident that our MLCANet achieves excellent results for all the datasets. In addition, it is worth noting that MLMSNet [7] and poolNet [6] need extra edge detection dataset to train the boundary detection branch.

2) **VISUAL COMPARISON**

Some saliency maps produced by our model and other approaches are visualized in Fig. 8 to explain the advantages of our approach. It could be seen that our method not only highlights the right salient object regions, but also maintain their clear boundaries. It is good at dealing with various challenging scenarios, including cluttered backgrounds (row 1, 4 and 5), low contrast between object and background (row 3 and 7), multiple objects (row 2), small objects (row 6). Compared with other state-of-the-art methods, the saliency maps produced by our method are more accurate and clearer. Note that our method achieves these results without any pre-processing or post-processing.

**V. CONCLUSION**

In this paper, we propose a novel end-to-end network named Multi-level Context Aggregation Network for salient object detection. To get robust features and fine details, we introduce joint bottom-up and top-down message passing in our network to aggregate multi-level features. Based on the consideration that features of different levels are of different significance, we propose a novel multi-level feature aggregating mechanism with channel-wise attention to aggregate features. It is able to adjust the contributions from features of different level flexibly so that redundant information is avoided to a large extent and efficient information can exchange efficiently. Besides, we design joint multi-level coarse-to-fine saliency inference module which gradually integrate upper-layer saliency estimates used to guide fine-grained saliency estimation with inferior-layer features. The whole network not only takes into account the difference between multi-level features, but also makes full use of all level features for saliency estimation, which make it generate high quality saliency maps with clear boundaries in various challenging scenarios. Experimental results on five datasets demonstrate that MLCANet performs favorably against the state-of-the-art methods.

**REFERENCES**

[1] S. Hong, T. You, S. Kwak, and B. Han, “Online tracking by learning discriminative saliency map with convolutional neural network,” in Proc. Int. Conf. Mach. Learn., 2015, pp. 597–606.

[2] M.-M. Cheng, F.-L. Zhang, N.-J. Mitra, X. Huang, and S.-M. Hu, “Repfinder: Finding approximately repeated scene elements for image editing,” ACM Trans. Graph., vol. 29, no. 4, pp. 1–8, 2010.

[3] C. Craye, D. Filiat, and J.-F. Goudou, “Environment exploration for object-based visual saliency learning,” in Proc. IEEE Int. Conf. Robot. Autom. (ICRA), May 2016, pp. 2303–2309.

[4] 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.

[5] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, “Selectivity or invariance: Boundary-aware salient object detection,” in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 3799–3808.

[6] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, “A simple pooling-based design for real-time salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3917–3926.

[7] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, “A mutual learning method for salient object detection with intertwined multi-supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 8150–8159.
