Salient Object Detection in the Distributed Cloud-Edge Intelligent Network

Zhifan Gao, Heye Zhang, Shizhou Dong, Shanhui Sun, Xin Wang, Guang Yang, Wanqing Wu, Shuo Li, and Victor Hugo C. de Albuquerque

Abstract

Intelligent network is crucial in the building of telecom networks because it utilizes artificial intelligent technologies to improve the performance. Salient object detection has increasingly attracted interest from intelligent network research since estimating human attention to objects is a crucial step in various surveillance applications. However, the computational-consuming and memory-consuming detection model is still less effective when it is deployed only either on the cloud or on the edge. In this article, we propose a specially-designed cloud-edge distributed framework for salient object detection based on the intelligent network. This framework can overcome the difficulty to transmit massive data in the cloud-only deployment scheme, as well as the difficulty to analyze massive data in the edge-only deployment scheme. However, the traditional cloud-edge distributed schemes are unsuitable to salient object detection task because of two challenges: 1) balance between the within-semantic knowledge and cross-semantic knowledge for the model training in different servers; 2) contradiction between extracting the semantic knowledge with global contextual information and local detailed information. To tackle the first challenge, our framework enables a hierarchical information allocation strategy in the cloud. It can prompt the salient object detection model in the edge to learn more from the similar scenes or semantics with where the edge server is located, while preserving the generalization ability of the model in the different scenes. To address the second challenge, our framework proposes a novel pyramidal deep learning model. It can effectively capture the global contextual features of the salient object, while preserving its local detailed features. The extensive experiments performed on six commonly-used public datasets can demonstrate the effectiveness of our framework and its superiority to 11 state-of-the-art approaches.

Introduction

Intelligent network is the advanced telecom network by adding artificial intelligent technologies, which has been shown to facilitate the IoT-based surveillance application [1]. Detecting the interested objects is crucial in varieties of surveillance applications like human identification and anomalous behavior detection. More advanced than traditional object detection [2], salient object detection (SOD) can capture the object on the focus position of human attention when glimpsing at surveillance images [3]. Traditional SOD methods in surveillance applications are less efficient in capturing the semantic concept and the saliency distribution in various complex visual scenes as the hand-crafted image feature extraction. Benefitting from the development of modern artificial intelligent, the deep neural network (DNN) technologies can largely put forward the performance of SOD [4], because it can effectively extract, represent, and integrate image features without manual intervention. In order to process the massive data efficiently in the DNN-based methods, two commonly-used schemes (i.e., cloud-only and edge-only) allocate the computing resources in the surveillance system. The cloud-only scheme deploys a large amount of computational units in the remote servers to perform the surveillance task, while the local devices aim to acquire and send the media data. In contrast, the edge-only scheme puts the data processing and analysis in the local devices.

However, both the cloud-only scheme and the edge-only scheme have their own shortcomings in the SOD. In the cloud-only scheme, cloud servers receive massive raw media data directly from on-site devices like cameras. This data transmission suffers from the capacity-limited and time-relayed network link (especially for Internet-of-Things devices), while the high-speed network deployment requires a very high cost [5], [6]. In the edge-only scheme, it is also unaffordable to configure large memory and high performance computational units (required by DNN-based models) in every edge server. Moreover, a single edge server usually processes the media data with similar semantic knowledge (e.g., various house) because they are collected from a limited region (i.e., a community). Thus, in varieties of visual scenes, these media data lack global contextual information (i.e., the information describing the relationship of the target object with other objects and the background in the scene). This may lead to the weak generalization ability of DNN-based SOD models.

In this article, we propose a framework to combine edge and cloud computing for salient object detection (SOD) in surveillance applications for

Zhifan Gao and Shuo Li are with Western University, Canada; Heye Zhang and Wanqing Wu are with the School of Biomedical Engi.neering, Sun Yat-sen University, China; Shizhou Dong is with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China, and also with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences, China; Shanhui Sun and Xin Wang are with Shenzhen Keya Medical Technology Corporation, China; Guang Yang is with National Heart and Lung Institute, Imperial College London, UK; Victor Hugo C. de Albuquerque is with the Graduate Program in Applied Informatics, University of Fortaleza, Brazil. The corresponding authors are Wanqing Wu and Shuo Li. Zhifan Gao and Heye Zhang are co-first authors.

Digital Object Identifier:
10.1109/MNET.001.1900260

addressing the above problems from cloud-only and edge-only schemes. This framework divides the SOD model into the inference task and the training task, and distributes them in the edge servers and cloud servers, respectively, shown in Fig. 1. It can reduce the computing requirement of the edge server with respect to the edge-only scheme, because the training task in the cloud server occupies the most computations. Moreover, it can satisfy the real-time requirement of the SOD-based surveillance tasks without being trapped in limited network bandwidth since the inference task runs in the edge server. In addition, the inference process of SOD in the edge server can be optimized by the global contextual information which is sent back from the cloud server. However, simply splitting the SOD model into the edge and cloud servers is difficult to obtain a good balance between learning the within-semantic knowledge (i.e., the similar semantic information which can represent the same visual contents in images) and the cross-semantic knowledge (i.e., the difference between the semantic information in images representing different visual contents). This prevents the further improvement of the specialization and versatility of SOD models. To overcome this challenge, our framework proposes a specially-designed hierarchical cloud computing strategy to make the SOD model suitable for multiple servers in the cloud-edge intelligent network. It can assure that the inference model in the edge server is scene-specific, as well as suitable for other environments. Furthermore, our framework proposes a novel pyramidal neural network (i.e., with a pyramidal-like structure) to tackle the contradiction between extracting the global contextual information and local detailed information in the SOD model. Finally, our framework is validated by extensive experiments on six commonly-used public datasets. The experimental results in comparison with eleven state-of-the-art approaches demonstrate the outperformance of our framework in the cloud-edge deployment of the SOD task.

The main contributions of this article are summarized as follows:

1. We are the first to introduce the deep learning for SOD into the cloud-edge distributed computing environment. To the best of our knowledge, it is better able to capture the objects of interest than the traditional object detection technologies in the cloud/edge-based surveillance task

2. We propose a hierarchical information allocation strategy to layer the computing servers in the cloud according to the scene semantics for allocating the media data to all servers for the SOD model training. It can facilitate the edge server being expertise in detecting the salient objects in the scene where it is located, as well as being applicable well in the difference scenes.

3. We propose a novel pyramidal deep neural network architecture to be aware of the global and local information in the images captured by the Internet-of-Things devices in order to effectively perceive the saliency regions.

The remainder of this article is organized as follows. The next section introduces two key challenges and our solutions in the proposed framework. Then, we present the detailed implementation procedures. In the next section, we perform extensive experiments to validate the effectiveness of the proposed framework, as well as its superiority to the state-of-the-art salient object detection approaches. Finally, we present the conclusion of this article.

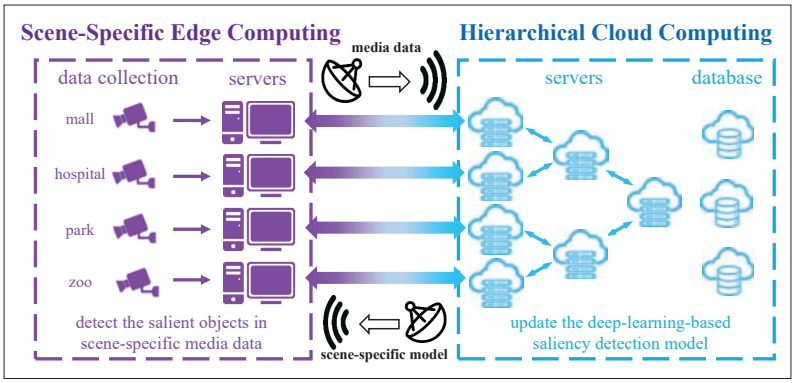

FIGURE 1. The proposed distributed framework based on the cloud computing and edge computing for salient object detection. The cloud is built as a hierarchical structure for allocating the media data to different servers for training the detection model. The edge is structured by various on-site servers containing the detection models to scene-specifically inference the saliency region from media data.

Related Works

Early research in saliency detection focus on exploring the hand-crafted image features (i.e., designed by users according to their experience on the feature extraction research and observation on the specific task) which can effectively represent the saliency information in an image [7]. Two popular features, contrast and background prior, are widely used in varieties of SOD models. The contrast feature indicates that salient objects should have the significantly different appearance with the background region, such as the color, intensity, object orientation and spatial coherence with the nearby regions. From a different view, the background prior models the image background rather than the salient object. It estimates the non-salient region based on the hypothesis that the salient region excludes the image boundary. However, the SOD models based on these low-level hand-crafted features are only suitable in simple scenes, because their robustness may be degraded by the intervention from complex non-salient objects and image backgrounds.

To overcome this drawback, deep learning is introduced in the saliency detection for extracting and utilizing the high-level features from the images. Most DNN-based approaches concentrate on how to effectively infer the coarse location and fine boundary of the salient objects. Locating the salient objects requires the high-level semantic information of the images (i.e., the visual content of an image based on its appearance like color intensity, gradient, and contrast). A commonly-used scheme is the fully convolutional neural network [3], which enhances the semantic region and suppresses the background region through the image encoding-decoding process. However, the image feature extracted from a single level of the neural network contains insufficient semantic information for saliency detection. Thus, the multi-level semantic features (i.e., the output of multiple layers of neural networks with seman-

tic information) produced by the neural network are fused for predicting the saliency map. A side effect of the semantic information extraction in the neural network is the loss of image details owing to the convolution and pooling operations (i.e., a non-linear image down-sampling method for outputing the local maximum of the input). This motivates the development of the strategies for obtaining the accurate boundaries of salient objects [8, 9].

The implementation of deep learning models in the SOD task have the same drawbacks in other tasks like the traditional object detection, whether it is deployed only in the cloud or in the edge. For large-scale surveillance systems, the edge-only deployment scheme requires the high-performance computing ability of every end device, and thus brings in the unaffordable deployment cost. Moreover, training the deep learning models only in edge servers has the potential to learn the insufficient global knowledge for SOD. For example, in the traffic control network, the images collected by surveillance devices on vehicles or streets only contain the information in local regions rather than the entire traffic system. In contrast, the cloud-only scheme is difficult to satisfy the real-time requirement of data processing and analysis owing to the bandwidth-limited data transmission, although it is able to relieve the computational burden of the end devices. For example, in pedestrian tracking, the tracking accuracy may be largely decreased when transmitting the surveillance data to the cloud with low image sampling rate and resolution. However, the high image sampling rate and resolution may make the amount of data to be transmitted beyond the bandwidth of wireless channels, and thus prevent the data processing in the cloud from achieving the real-time requirement. Thus, it is natural to develop the distributed computation of the deep learning model in the cloud/edge network. This distributed scheme has been successfully applied in many tasks [10]–[14], such as traditional object detection and crowd sensing. Specific in the SOD task, our framework proposes a novel cloud-edge distributed scheme to improve the scene-specific saliency awareness of the edge servers as well as preserve their general ability in other surveillance scenes.

Challenges and Solutions

In this section, we discuss two key challenges of the SOD task based on the cloud and edge computing: 1) balance between the within-semantic knowledge and cross-semantic knowledge used for the SOD training in different servers; 2) contradiction between extracting the global contextual information and local detailed information for the SOD training. As regards these challenges, we introduce our contributions including the hierarchical training information allocation and the dilated densely-connected pyramidal neural network, as well as how they can address the two challenges.

Challenge 1: Balance between the within-semantic knowledge and cross-semantic knowledge for SOD training in different servers.

This challenge comes from the difference of the semantic knowledge within the information received by the edge server and the cloud server. As shown in Fig. 1, the edge server receives the media data related to the environment where it is installed (e.g., mall, hospital and park), and cannot receive the media data from other environments. Thus, the media data for each server have their own semantic knowledge (i.e., within-semantic knowledge), such as some elements like humans in white and ambulances often appear in the "hospital" scene. In contrast, the cloud server, connected with massive edge servers, can receive the media data with varieties of semantic knowledge (i.e., cross-semantic knowledge). In the traditional cloud-edge deployment of the surveillance model [13], [14], all edge servers share the same model parameters that are provided by the model training based on all media data in the cloud servers. This scheme can improve the generalization ability of the model because the trained model has good ability to capture the common saliency characteristic of different objects. However, it may weaken the detection ability of the model to the semantic-specific pattern of the objects. For instance, the edge server installed in the hospital receives the doctor image with much higher probability than the salesperson image. The edge server will obtain the degraded performance in detecting the doctor when the cloud server trains the detection model on the doctor and salesperson images with equal probability. Therefore, it is challenging to improve the generalization ability of the detection model in complex environments and preserving its ability in the semantic-specific environment.

Our solution. To address this challenge, we propose a hierarchical information allocation strategy to update the SOD model. This strategy reorganizes the computing units in cloud servers into a multi-branching tree structure, shown in Fig. 2. The structure layers the semantics of the environments where all edge servers are installed from fine to coarse. In particular, the cloud servers in the different levels of the tree structure process the media data with different semantics. For example, in Fig. 2, the "road" node in the second level processes all media data from the "road" scenes, while the "highway" node in the first level processes the media data just from "highway" scenes rather than "lane" scenes. Then the media data can be allocated based on the semantic similarity to the different cloud servers with unshared parameters for model training. This can ensure that the SOD model in each edge server mainly learns from the media data with similar semantic environment, and also learns from a small amount of media data from a different semantic environment. Thus, the proposed strategy can lead the SOD model to gain from both the within-semantic and cross-semantic knowledge. The motivation to use the tree structure comes from its two characteristics. First, the levels of the tree structure can be used to divide the levels of semantics, i.e., the level closer to the leaf has finer semantics, while the level closer to the root has coarser semantics. Second, the father node in the tree structure receives the data from all child nodes. Thus the SOD model in the father node can learn more cross-semantic knowledge, while the model in the child node can learn more within-semantic knowledge.

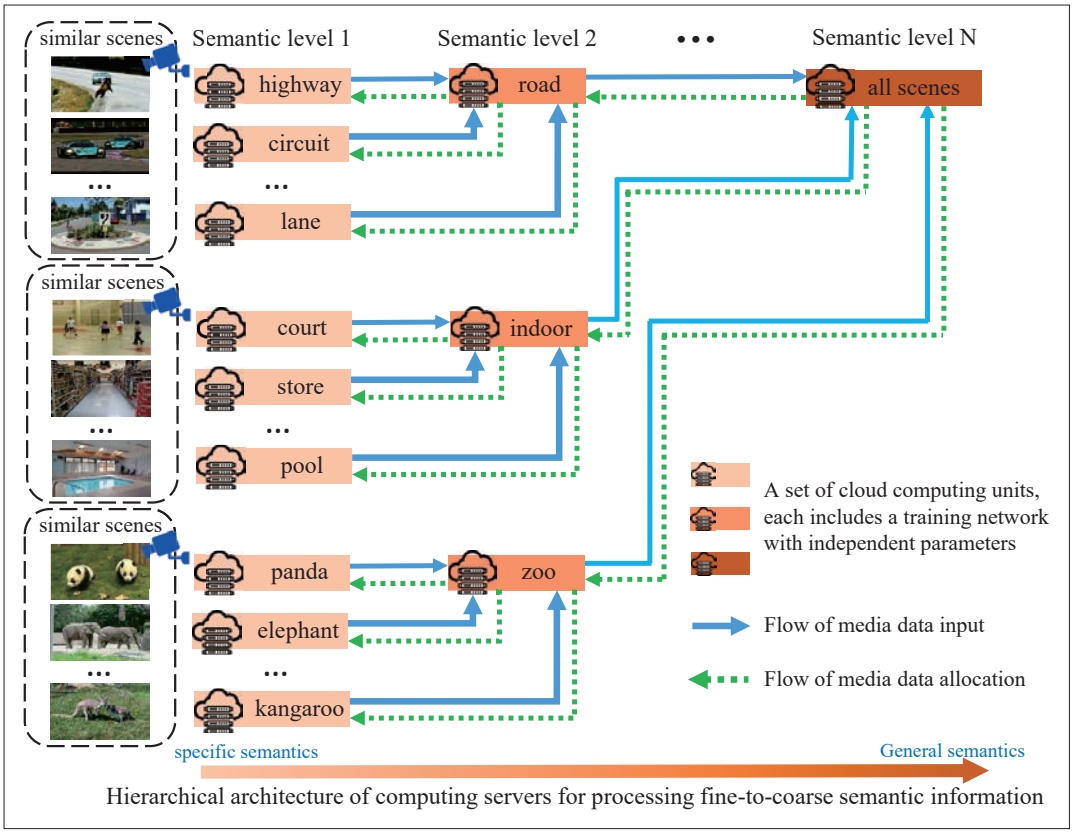

FIGURE 2. The schematic diagram of the hierarchical cloud servers. Each rectan.gle block is a set of computing units or servers in the cloud, and configured a salient object detection model with same neural network architecture but different parameters. The tree structure is multi-branching. From Level 1 to Level N, the block receives and processes the media data with more general semantics. For instance, the block "road" includes the blocks "highway", "circuit" and "lane" from a semantic perspective. In the level N, the root block processes the media data in all possible scenes collected by the surveillance devices. Each block allocates the received medica data to its children block for training the detection model in this children block. The allocation criterion is based on the semantic similarity, that is the father block allocate the media data to the child block with a high probability when the media data contains similar semantic contents with that in the child block. The blue solid arrows show the path of the media data delivery from low level to high level. The green dashed arrows show the path of the media data allocation from high level to low level.

Challenge 2: Contradiction between extracting the semantic knowledge with global contextual information and local detailed information.

This challenge originates from the dilemma of the DNN in enlarging the receptive field and preserving the image resolution. The receptive field is the region of the input space that affects a particular unit of the DNN. The large receptive field (RF) can help DNN to capture global contextual information. The saliency information detection relies on the global contextual information extraction of the objects with different sizes. Thus, it requires that DNN-based SOD model has the ability to perceive the large area in the raw input image (i.e., having the large RF). However, commonly-used RF enlargement schemes (e.g., pooling) are at the cost of image resolution reduction. It inevitably brings in the loss of local detailed information in saliency maps of raw images.

Our solution. To tackle this challenge, we propose a novel pyramidal neural network, shown in Fig. 3. It consists of two elaborately-designed pyramidal structures. The forward pyramid is constructed by multiple dense and dilated convolution blocks, where the dilated convolution can enlarge the RF size in every layer and the dense convolution can fuse the low-level information to the high-level feature maps (i.e., the output of a given filter to represent a certain kind of image features) within a block. The backward pyramid is formed by the combination of the global context information and the multi-scale feature information from the forward pyramid for gradually recovering the resolution of the context-rich feature maps. It inherently fuses the high-level information to the low-level feature maps. Accordingly, the pyramidal structure in our neural network can fuse the feature information in different levels, that is, fuse the local detailed information into the high-level feature maps and fuse the global context detailed information into the low-level feature maps. This characteristic enables our neural network to be simultaneously aware of the global context and local details in the image, and thus can produce the feature map with large RF while preserving the high resolution.

IMPLEMENTATION

In this section, we present the implementation of the proposed framework to address two challenges mentioned in the last section. For achiev-

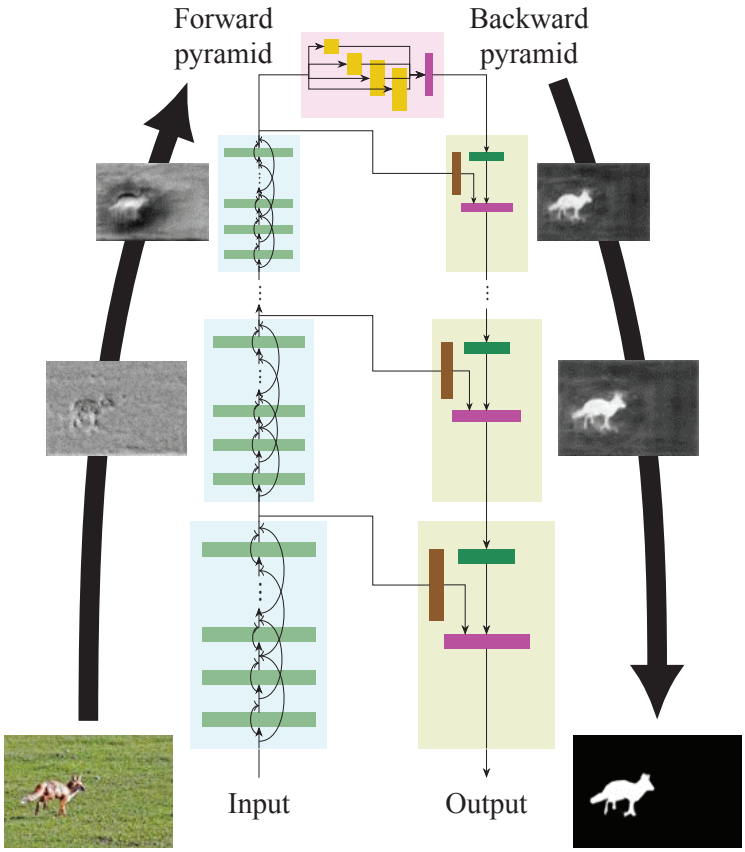

Fig. 3. The architecture of the proposed pyramidal neural network. It consists of a forward pyramid structure (left) and a backward pyramidal structure (right). The forward pyramid includes the cascade of dilated densely-connected convolution blocks (light cyan blocks) for extracting the global contextual information while preserving the local detailed information of the raw image. The backward pyramid aims to fuse the low-level features (light yellow blocks) produced from the different level of the forward pyramid for recovering the detected salient objects in the resolution of the raw input image. The thistle block is the pyramid pooling module, where the gold rectangles are different scales of convolution layers. The green rectangles are the dilated convolution layers. The tan rectangles are the convolution layers. The dark green rectangles are the upsampling operators. The magenta rectangles are concatenate layers.

ing this goal, this implementation contains two key parts:

- Constructing a hierarchical structure within the cloud servers to allocate the training media data for balancing between the within-semantic and cross-semantic knowledge used in training the SOD model.
- Designing an effective neural network architecture for edge and cloud servers in order to detect the location of the saliency region while preserve its local details.

Hierarchical Training Information Allocation for gaining from within-semantic and cross-semantic knowledge.

We introduce the proposed hierarchical information allocation strategy in the cloud with two issues: how to construct the hierarchical structure and how to allocate the information with various semantic meanings according to visual scenes. The hierarchical structure of cloud servers is built as a multi-branching tree structure based on the fine-to-coarse layering of the image semantics. In this tree structure, each node is a subset of all computing units in the cloud server. Each computing set stores and trains a SOD model with the same architecture (introduced in the next subsection). The SOD models in different computing sets have independent trainable parameters (i.e., their weight and bias parameters are unshared). The key idea is that the node at the deep level of the tree have SOD training information with coarse or general semantics, while at the shallow level of the tree it has fine or specific semantics.

Figure 2 shows a schematic diagram to illustrate this tree structure with N levels. In the first level, the leaf nodes aim to train a SOD model for applying specified scenes (e.g., lanes, courts or pandas). Then in the second level, the semantics become more general than in the first level. For instance, the "indoor" node has the more abstract semantics containing the semantics "court" and "pool" in its offspring nodes, because the court and pool images are collected by the surveillance camera installed at the indoor environment. In the level N, the root node has the most abstract semantics (i.e., "all scenes") in the entire hierarchical cloud servers, including all possible scenes collected by the surveillance devices.

The allocation of the training information in the tree structure is based on the semantic similarity. In order to assure the specialization of the SOD model in certain semantics and its generalization in other environment, we propose a strategy to allocate the training information to different nodes on semantics with the following three rules:

1. The training data of all nodes are allocated by their ancestor nodes with different probabilities.
2. The probability to allocate the training data to a node increases with their semantic similarity.
3. The trainable parameters in the leaf node are used to update the SOD model in the corresponding leaf node.

As shown in Fig. 2, for example, the "panda" node and the "kangaroo" node receive the images separately from the surveillance cameras installed in the panda and kangaroo enclosures, and then send them to the "zoo" node. The "zoo" node sends the panda images back to the "panda" node with a high probability and to the "kangaroo" node with a low probability for training their own SOD models. Meanwhile, some images like pool images, allocated by the "scene" node, are sent to the "panda" and "kangaroo" node with a lower probability for SOD model training.

In summary, the proposed strategy enables that most media data for SOD training in one node contains the information mainly from the scenes with the corresponding semantics, while a small amount of them come from the scenes with other semantics. The former is able to improve the SOD's accuracy in the similar scenes and the latter can preserve the generalization ability in different scenes.

Pyramidal Neural Network for effectively extracting the location and details of the salient object.

The proposed neural network architecture consists of two pyramid structures: the forward

pyramid and the backward pyramid. Figure 3 shows the architecture of the neural network.

Forward Pyramid. The forward pyramid aims to extract the global contextual information and preserve the local detailed information in the raw image. In the forward pyramid, five dense and dilated convolution blocks are cascaded. Each block contains a neural network with dilated densely-connected convolution (referred to as dilated densely-connected block (DDCB)). The dilated convolution in DDCB is able to increase RF by enlarging the spatial gap between the sampling points for convolution. The densely-connected structure in DDCB can preserve the local detailed information by supplementing the spatial information directly from the low-level to high-level network layers. In addition, a convolution layer is added between any two connected blocks. Two pooling layers are added among the first, second and third blocks. From the first to fifth block, the feature maps have increasing channel numbers (i.e., 160, 224, 256, 272, 280). Then, a pyramid-pooling module is connected behind the last block for extracting the global contextual information of the salient object in the largest RF produced by the proposed forward pyramid network. In the pyramid pooling module, the input feature map is convolved in four different scales (i.e., the side length of the convolution kernels are 1, 2, 3, 6), and finally transformed into a four-channel feature map. Finally, the feature maps output from five blocks and the pyramid pooling module [15] (totally including four feature maps) are fed into the subsequent backward pyramid for providing the local detailed information and global contextual information in the different levels.

Backward Pyramid. The backward pyramid aims to fuse the different-level features from the forward pyramid to predict the saliency map of the raw image. In the backward pyramid, all of the six feature maps are concatenated into five feature maps in turn from the high level to the low level. In the first and the second level, an upsampling operator is implemented when two feature maps are concatenated for enlarging the size of the fused feature map. Then, five fused feature map are convolved and upsampled to produce the saliency maps in different levels (denoted by I_1 to I_5). In the saliency map, the pixel value represents the probability (range in [0, 1]) that the pixel is within the salient region. Finally, I_1 to I_5 are summed to obtain a feature map I_0 as the output of the saliency map in the proposed neural network.

Other Details. In the training phase, we produce an objective function to optimize the SOD model in every computing node by measuring the error between the predicted salient region and the ground truth. It consists of three terms including the pixel-wise weighted cross-entropy, the generalized Dice index and the mean absolute errors. Then, the proposed objective function is imposed to supervise the all of the six saliency maps I_1 to I_5. In addition, the values of the dilated rate d in DDCB are set 1, 1, 2, 4 and 8 from the first to the last level, respectively. Then, the stochastic gradient descent is used to optimize the entire neural network. Other configurations include the momentum is 0.9, the weight decay is 0.0001 and the training epoch number is 40.

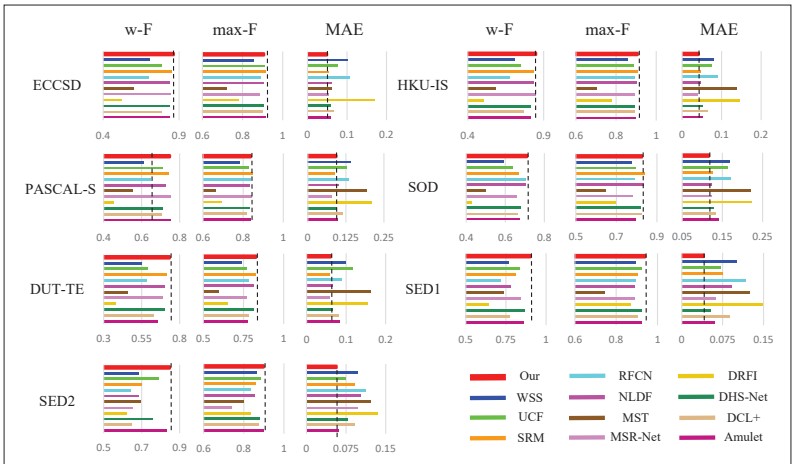

Fig. 4. State-of-the-art performance of our framework evaluated by weighted F-measure (w-F), maximum F-measure (max-F) and mean absolute error (MAE). The black dashed lines show the performance of our framework. We show the superiority of the proposed framework (red) to the eleven state-of-the-art approaches, including WSS (blue), UCF (green), SRM (dark yellow), RFCN (cyan), NLDF (magenta), MST (tan), MSR-Net (violet), DRFI (gold), DHS-Net (sea green), DCL+ (burlywood) and Amulet (violet red).

The learning rate changes from 2×10^{-3} (initial) to 2×10^{-4} (after 25 epochs), and finally set 2×10^{-5} (after 30 epochs).

Performance Evaluation

Datasets and Evaluation Metrics

Datasets. The performance of our framework is trained on three public datasets (DUT-TR, DUT-OMRON, MSRA10K) and then evaluated on six public datasets (ECSSD, HKU-IS, PASCAL-S, SOD, DUT-TE, SED) [4], [8], [9]:

DUT-TR contains 10553 images obtained from ImageNet DET training and validation sets. 50 subjects are asked to annotate the salient object in each image.

DUT-OMRON contains 5172 images such as fruits, birds and flowers. Five subjects draw the salient region in all images.

MSRA10K contains 10,000 images selected from the MSRA dataset provided with pixel-level saliency labeling by three to nine subjects.

ECSSD contains 1000 images acquired from the Internet. Five persons were asked to delineate the ground truth masks individually. The delineation results are selected by averaging five candidate masks with the threshold 0.5.

HKU-IS contains 4447 images with pixel-wise annotations of salient objects with complex shapes including car, horse, pedestrian, etc.

PASCAL-S contains 850 natural images selected from the PASCAL VOC 2010 segmentation challenge. Every image is first fully segmented with 1) not intentionally labeling parts of the images, 2) labeling the disconnected regions in the same object separately, 3) using solid regions to approximate the objects. Then 12 subjects vote for the salient object in each image.

SOD contains 300 images collected from the Berkeley Segmentation Dataset (BSD). Seven subjects select the regions or segments corresponding to the salient objects in a random subset of BSD.

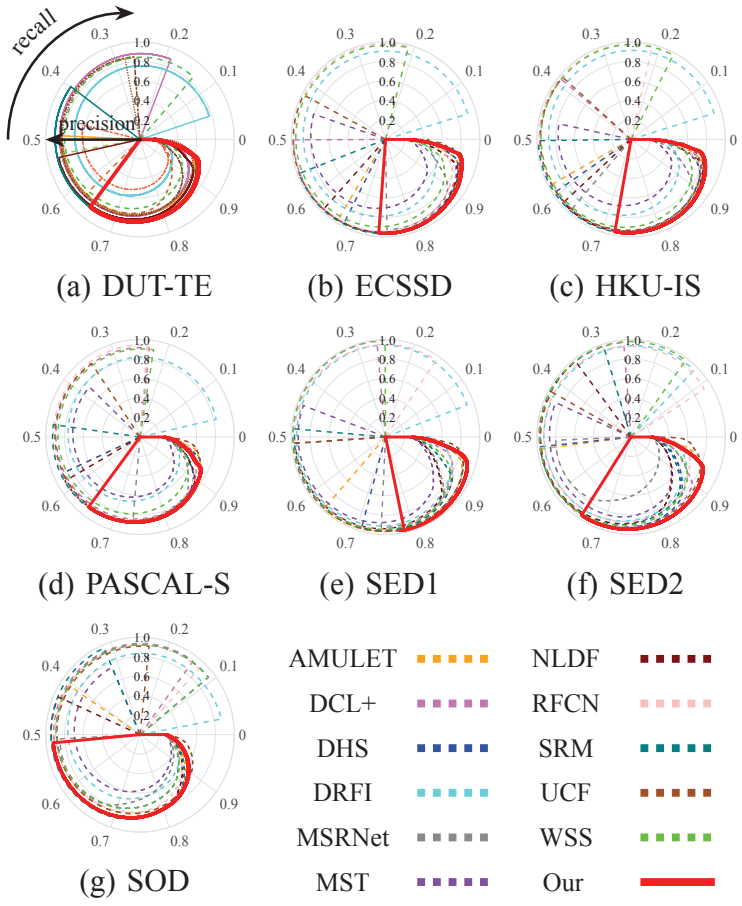

Fig. 5. Better performance of our framework than the state-of-the-art methods evaluated by the performance curve of precision and recall on seven datasets. In each Nightingale rose chart, the radial direction represents the value of the precision ranged from 0 to 1, and the circumferential direction represents the value of the recall ranged from 0 to 1. The closed curves with different colors represent different methods, including the proposed framework (red), Amulet (orange), DCL+ (pink), DHS (blue), DRFI (cyan), MSRNet (gray), MST (purple), NLDF (dark brown), RFCN (rosy), SRM (dark green), UCF (dark brown), and WSS (green). The color legend is at the lower right corner of this figure.

DUT-TE contains 5091 images obtained from the ImageNet DET test set and the SUN data set. The regions of salient objects are manually delineated by 50 subjects.

SED contains two subsets named SED1 and SED2. SED1 has 100 images, where each image has only one salient object. SED2 has 100 images, where each image has two salient objects.

Evaluation Metrics. Five evaluation metrics are applied to validate the performance of salient object detection, including the weighted F-measure, maximum F-measure, mean absolute error, precision and recall [9].

Outperformance over the State-of-the-Art Methods

We have compared the proposed framework with eleven state-of-the-art SOD approaches, including WSS, UCF, SRM, RFCN, NLDF, MST, MSR-Net, DRFI, DHS-Net, DCL+ and Amulet. DRFI and MST utilize the hand-crafted image fea.tures and other approaches are based on DNN. The details of these approaches are shown in [3].

Figure 4 shows the comparative results in weighted F-measure (w-F), maximum F-measure (max-F) and mean absolute error (MAE). The increase of w-F and max-F and the decrease of MAE indicate the performance improvement of salient object detection. The results illustrate that our framework (the red bars) ranks at the top place in the state-of-the-art approaches in all datasets.

Figure 5 presents the comparative results of precision-recall performance by Nightingale chart. Each chart shows the evaluation results tested on a dataset. In the rose chart, the radial direction shows the precision (from 0 to 1), and the circumferential direction shows the recall (from 0 to 1). Our framework is shown in red curves, which fall in the bottom right part of all charts as a whole. This means that the overall values of the precision and recall are higher than other approaches.

Figure 6 illustrates the results of nine representative images by all the comparative approaches. These results can show some good performance of our framework:

1) Robust to the interference of the large background object (e.g., the cylinder in the first row and the water in the sixth row).
2) Able to detect the large and small salient object simultaneously (e.g., the giraffe in the second row).
3) Preserve the details of the salient object (e.g., the cattle legs in the third row and the crosses in the eighth row).
4) Perceivable to the morphologically similar but semantically different salient objects (e.g., sandbag and human in the fourth row).
5) Robust to the inner holes within the saliency region (e.g., the tree in the fifth row).
6) Robust to the influence from uneven illumination (e.g., the face and bottle in the seventh row).
7) Able to detect slender salient region (e.g., the coconut tree in the last row).

All of these representative results can demonstrate the superiority of our framework to the state-of-the-art approaches.

Conclusion

Distributing the data analysis in the cloud and the edge computing is becoming increasingly popular in varieties of surveillance applications. In this article, we have proposed a salient object detection framework for surveillance applications based on the intelligent network with hierarchical cloud computing and the scene-specific edge computing. Hierarchical cloud computing builds a tree structure connection to allocate the training information according to the semantics similarity. It can make the saliency detection model in the edge server suitable for its own specific scene, and performs well in other environments. Moreover, the proposed framework develops a novel pyramidal deep learning model for effectively extracting the global contextual information and preserving the local detailed information during the salient object detection procedure. The extensive experiments demonstrate the effectiveness of our framework, as well as its superiority to the 11 state-of-the-art approaches.

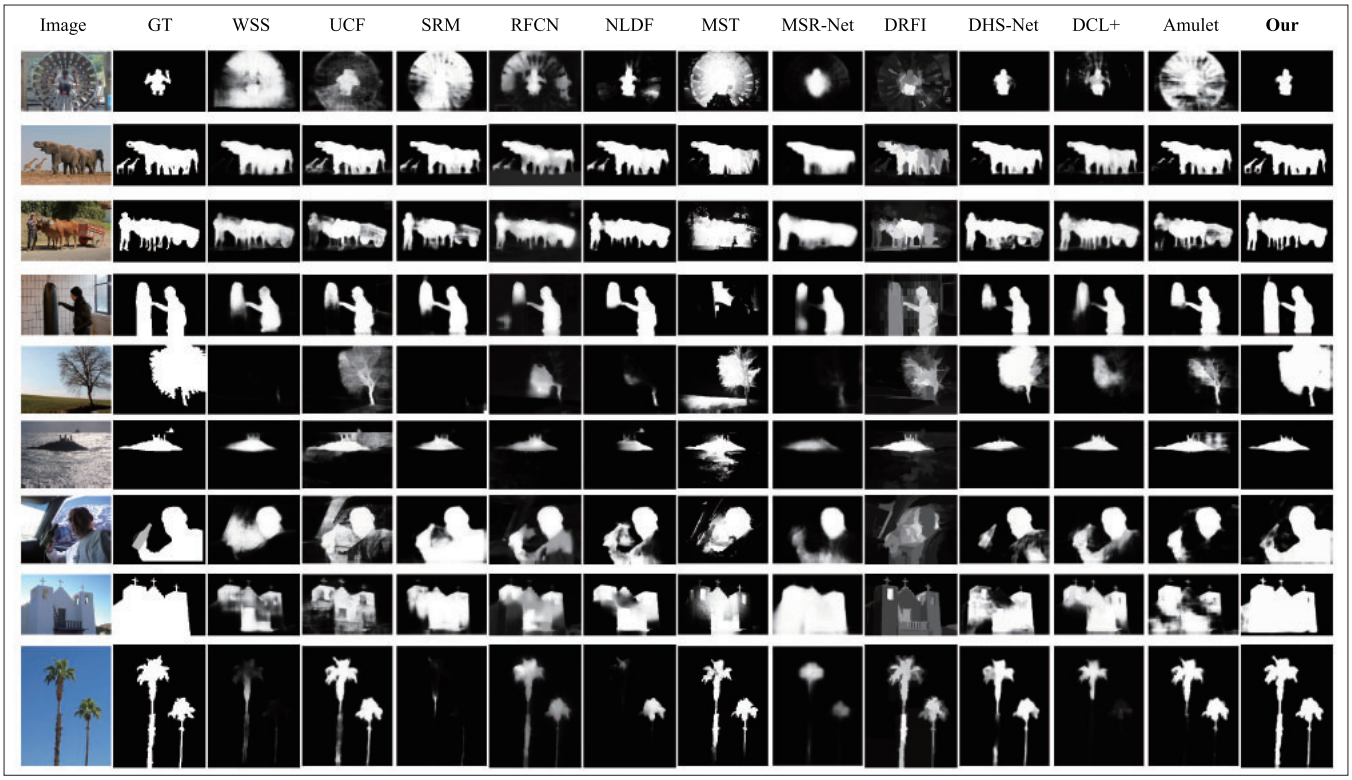

FIGURE 6. Representative results to show the better performance of our framework than the state-of-the-art approaches. The first and second columns show the raw images and the ground truth (GT), respectively. The third to the eleven column show the saliency maps predicted by all comparative approaches. The last column show the results of our framework.

Acknowledgment

This work was supported by the Guangdong Province Science and Technology Planning Project (2018A050506031, 2019B010110001); the Brazilian National Council for Research and Development (CNPq) (Grant No. 304315/2017-6 and 430274/2018-1); the Shenzhen Overseas High Level Talent (Peacock Plan) Project (KQTD2016112809330877); and the National Natural Science Foundation of China under grants (61771464, 61873349, U1801265).

References

[1] Y. Yu *et al.*, "Enabling Secure Intelligent Network with Cloud-Assisted Privacy-Preserving Machine Learning," *IEEE Network*, vol. 33, no. 3, 2019, pp. 82–87.
[2] J. Han *et al.*, "Object Detection in Optical Remote Sensing Images Based on Weakly Supervised Learning and High-Level Feature Learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, 2015, pp. 3325–37.
[3] J. Han *et al.*, "Advanced Deep-Learning Techniques for Salient and Category-Specific Object Detection: A Survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, 2018, pp. 84–100.
[4] N. Liu and J. Han, "A Deep Spatial Contextual Long-Term Recurrent Convolutional Network for Saliency Detection," *IEEE Trans. Image Process.*, vol. 27, no. 7, 2018, pp. 3264–74.
[5] M. Chen *et al.*, "Label-less Learning for Traffic Control in an Edge Network", *IEEE Network*, vol. 32, no. 6, 2018, pp. 8–14.
[6] M. Chen *et al.*, "A Dynamic Service Migration Mechanism in Edge Cognitive Computing," *ACM Trans. Internet Technol.*, vol. 19, no. 2, 2019, pp. 1–15.
[7] J. Han *et al.*, "Unsupervised Extraction of Visual Attention Objects in Color Images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, 2006, pp. 141–45.
[8] J. Han *et al.*, "Background Prior-Based Salient Object Detection via Deep Reconstruc.tion Residual," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 8, 2015, pp. 1309–21.
[9] X. Wang *et al.*, "Edge Preserving and Multi-Scale Contextual Neural Network for Salient Object Detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, 2018, pp. 121–34.
[10] H. Khelif *et al.*, "Bringing Deep Learning at the Edge of Information-Centric Internet of Things," *IEEE Commun. Lett.*, vol. 23, no. 1, 2019, pp. 52–55.
[11] L. Hu and Q. Ni, "IoT-Driven Automated Object Detection Algorithm for Urban Surveillance Systems in Smart Cities," *IEEE Internet Things J.*, vol. 5, no. 2, 2018, pp. 747–54.
[12] H. Li *et al.*, "Learning IoT in Edge: Deep Learning for the Internet of Things with Edge Computing," *IEEE Network*, vol. 32, no. 1, 2018, pp. 96–101.
[13] Z. Zhou *et al.*, "Robust Mobile Crowd Sensing: When Deep Learning Meets Edge Computing," *IEEE Network*, vol. 32, no. 4, 2018, pp. 54–60.
[14] J. Ren *et al.*, "Distributed and Efficient Object Detection in Edge Computing: Challenges and Solutions," *IEEE Network*, vol. 32, no. 6, 2018, pp. 137–43.
[15] H. Zhao *et al.*, "Pyramid Scene Parsing Network," *IEEE Conf. Comput. Vis. Pattern Recognit. 2017*, pp. 6320–39.

Biographies

Zhifan Gao [M'18] (zgao246@uwo.ca) received his B.S. and M.E. degrees in electronics and information engineering from Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2011, respectively, and his Ph.D. degree in pattern recognition and intelligent systems from the University of Chinese Academy of Sciences, China, in 2017. He is currently a postdoctoral fellow at the Western University, London, Canada. His research focuses on medical image processing, computer vision and machine learning.

Heye Zhang [M'17] (zhangheye@mail.sysu.edu.cn) received his B.S. and M.E. degrees from Tsinghua University, Beijing, China, in 2001 and 2003, respectively, and his Ph.D. degree from the Hong Kong University of Science and Technology, Hong Kong, in 2007. He is currently a professor with the School of Biomedical Engineering, Sun Yat-Sen University, China. His research interests include cardiac electrophysiology and cardiac image analysis.

Shizhou Dong (sz.dong@siat.ac.cn) received his B.S. degree from Chongqing University of Posts and Telecommunications, China, in 2017. Currently he is a master student at the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China. His work focuses on computer vision and deep learning.

Shanhui Sun (shanhui.sun@united-imaging.com) is currently the Director of AI and Medical Imaging in UII America Inc., USA.

His contribution in this work was made when he was with the Shenzhen Keya Medical Technology Corporation and the CuraCloud Corporation. His research interests include medical image analysis, computer vision and machine learning

Xin Wang [M'18] (xinw@keyayun.com) is currently a senior machine learning scientist at the Shenzhen Keya Medical Technology Corporation and the CuraCloud Corporation. He received his Ph.D. degree in computer science from the University of Albany, State University of New York in 2015. His research interests are in artificial intelligence, machine learning and computer vision.

Guang Yang [M'15] (g.yang@imperial.ac.uk) obtained his M.Sc. in vision imaging and virtual environments from the Department of Computer Science in 2006 and his Ph.D. in medical image analysis jointly from the CMIC, Department of Computer Science and Medical Physics in 2012 both from University College London. He is currently an honorary lecturer with the Neuroscience Research Centre, Cardiovascular and Cell Sciences Institute, St. George's, University of London. He is also an image processing physicist and honorary senior research fellow working at the Cardiovascular Research Centre, Royal Brompton Hospital, and also affiliated with the National Heart and Lung Institute, Imperial College London. His research interests include pattern recognition, machine learning, and medical image processing and analysis. His current research projects are funded by the British Heart Foundation.

Wanqing Wu (wuwanqing8133@gmail.com) is an associate professor at the School of Biomedical Engineering, Sun Yat-Sen University, China. He graduated from the Department of Computer Engineering at Hunan Normal University with a Gold Medal and received his Master degree from the College of Computer Science of Chongqing University, and his Ph.D. degree from the Pusan National University of Computer Engineering in 2013. He has over 10 years of teaching and research experience. His fields of interest include wearable sensing technology, biosignal processing and multimodal information fusion, etc. He has published over 60 papers in renowned journals and conferences, contributed to the peer review of 15 journals and secured over 15 major competitive research grants.

Shuo Li (slishuo@gmail.com) received his Ph.D. degree from Concordia University, Canada, in 2006. He is currently a professor with Western University, London, Canada. He is also a member of the MICCAI Society board, a research fellow at the Lawson Institute of Health, and a member of the IEEE Engineering in Medicine and Biology Society's Translational Engineering and Healthcare Innovation Technology Committee. He has published more than 100 papers in top international journals and conferences, such as TPAMI, MIA, TNNLS, TMI, CVPR. He is the deputy editor of *Medical Image Analysis* and *Computerized Medical Imaging and Graphics* (CMIG). He is also a guest editor of the TMI, CMIG, CVIU. His research interests include cardiac electrophysiology and cardiac image analysis.

Victor Hugo C. de Albuquerque [M'17, SM'19] (vic.tor.albuquerque@unifor.br) is a professor and senior researcher at the University of Fortaleza, UNIFOR, Brazil. He has a Ph.D. in mechanical engineering from the Federal University of Paraíba, an M.Sc. in teleinformatics engineering from the Federal University of Ceará, and he graduated in mechatronics engineering from the Federal Center of Technological Education of Ceará. He is currently an assistant VI professor in the Graduate Program in Applied Informatics of UNIFOR. He is a specialist, mainly in IoT, machine/deep learning, pattern recognition, and robotics.