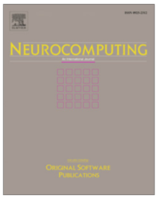
Neurocomputing 428 (2021) 92–103

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



Exploring multi-scale deformable context and channel-wise attention for salient object detection

Yi Liu [a,b], Mingxing Duanmu [b], Zhen Huo [b], Hang Qi [c], Zuntian Chen [d], Lei Li [c,*], Qiang Zhang [b,*]

[a] Changzhou University, Changzhou, Jiangsu 213164, China
[b] Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China
[c] Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing 100074, China
[d] Xi'an Institute of Electromechanical Information Technology, Xi'an, Shaanxi 710065, China

ARTICLE INFO

Article history:
Received 22 May 2020
Revised 29 October 2020
Accepted 14 November 2020
Available online 28 November 2020
Communicated by Steven Hoi

Keywords:
Salient object detection
Multi-cale
Deformation
Attention

ABSTRACT

Contextual information has played an important role in salient object detection. However, due to the fixed geometric structures of convolution kernels employed by existing Convolutional Neural Networks (CNNs) based methods, it is difficult to extract meaningfully visual contexts for those salient objects with varying sizes and non-rigid shapes. To address this problem, in this paper, we propose a Multi-Scale Deformation Module (MSDM) to capture multi-scale visual cues and varying shapes of salient objects. Moreover, most existing CNNs based methods treat all channels of feature maps equally, which tends to differ from the fact that different channels actually contribute differently to saliency prediction. For that, we involve a novel Channel-Wise Attention Mechanism (CWAM) after MSDM to highlight those informative channels while suppressing those confusing ones. Experimental results on five benchmark datasets demonstrate the superiority of the proposed method over the state-of-the-art approaches.

1. Introduction

Salient object detection aims to identify the most visually distinctive objects or regions in a given image. It has been used as a preprocessing step to reduce the complexity by a large margin in a wide range of applications, including person identification [1], visual tracking [2,3], robot navigation [4], visual question answering [5], image segmentation [6,7], image fusion [8,9], image retrieval [10], video segmentation [11], image quality assessment [12], etc.

Traditional methods [13–17] infer the salient object via inferring hand-crafted features, e.g., color, texture, etc. These low-level cues are trivial so that the methods relying on them unavoidably encounter the bottleneck for performance improvements. In recent years, CNNs have successfully broken the limits of traditional methods and achieved impressive results due to their powerful representative ability [18–22]. For instance, to better understand the image, Luo et al. [20] learned non-local deep features, Zhang et al. [21] learned deep uncertain convolutional features, and Zhang et al. [19] combined different levels of deep features to integrate low-level spatial details and high-level semantic knowledge. These learned deep features substantially improve the saliency detection performance. However, there still exist some problems for CNNs based methods to be resolved.

First, most existing CNNs based salient object detection models [19–21] learn deep features based on the grid convolution kernels such that they generally lack the capability of modeling geometric transformations due to the fixed geometric structures of convolution kernels. Failure to do so makes those models unable to capture the desired visual context for non-rigid salient objects with diverse shapes accurately. For example, as shown in the top two rows of Fig. 1, previous CNNs based methods failed to detect the complete objects due to their irregular shapes.

Secondly, all the channels of feature maps are not equally important for the saliency prediction. To give a clear illustration for this problem, we display 4 channels of feature maps[1] in Fig. 2, on which it can be seen that different channels have different importance for saliency prediction. To be specific, the feature map of Fig. 2 (c) cannot distinguish the foreground and background regions, thus failing to identify the salient object from the background. The feature maps of Fig. 2(d) and (e) possess high responses on background regions and foreground regions, respectively, helping to predict the background and foreground regions in the saliency map. The feature map of Fig. 2(f) almost locates the whole salient object with a complete object shape, which greatly helps to predict the salient object.

* Corresponding authors.
E-mail addresses: univer1@sina.com (L. Li), qzhang@xidian.edu.cn (Q. Zhang).

[1] These 4 feature maps are learned by the model with only MSDM.

https://doi.org/10.1016/j.neucom.2020.11.022
0925-2312/© 2020 Elsevier B.V. All rights reserved.

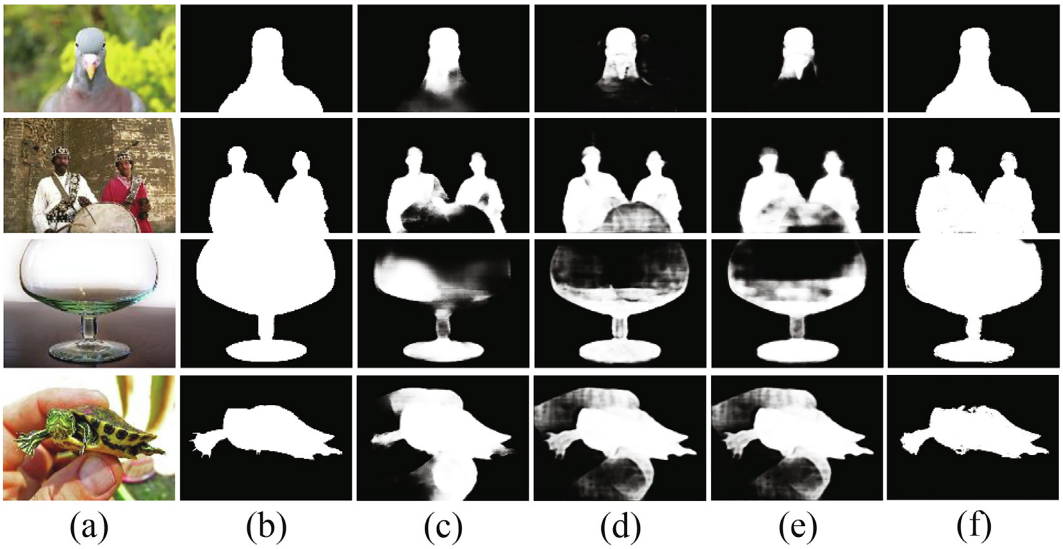


Fig. 1. Visual examples of existing CNNs based methods: (a) Images; (b) Ground truth; (c) NLDF [20]; (d) UCF [21]; (e) Amulet [19]; (f) OURS.

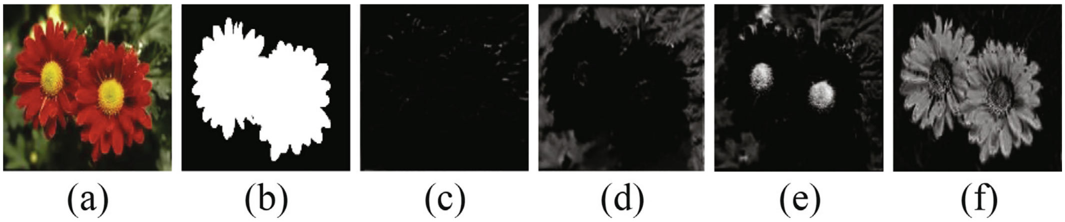


Fig. 2. Visualization for feature maps of our model without considering the attention module. (a) Images; (b) Ground truth; (c)–(f): four channels of feature maps. Different channels have different importance for saliency prediction. For example, (c) cannot identify the salient object from backgrounds, while (f) greatly helps to predict the salient object.

In view of the different importance of different channels, treating all channels equally may degrade the saliency prediction performance.

In this paper, we propose a deep end-to-end salient object detection network with the aim to solve the above two problems. Concretely, to solve the former problem, we design a Multi-Scale Deformation Module (MSDM) to capture more primitive visual context information of those non-rigid salient objects. For each stage, we obtain multiple feature maps by embedding the deformable convolution [23] in a multi-scale structure, leading to deformable receptive fields across different scales. The obtained feature maps are concatenated together to capture diverse shapes of the non-rigid salient object at multiple scales. For the latter problem, we introduce a Channel-Wise Attention Mechanism (CWAM) to boost those more informative channels whereas suppressing those less important ones. For each channel, we calculate a saliency prediction score based on its feature map. These prediction scores are in turn utilized as the channel-wise attention scores to measure the importance of different channels. As such, multiplying those feature maps of MSDM with their channel-wise attention scores enables to highlight those remarkable feature maps for subsequent saliency prediction. In summary, the cooperative combination of MSDM and CWAM makes the network be truly capable of segmenting out those non-rigid salient objects accurately and completely, as can be obviously observed from Fig. 1(f).

To sum up, the contributions of our work are as follows:

(1) A Multi-Scale Deformation Module (MSDM) is proposed to capture various receptive fields, leading to more primitive visual context covering different scales and different shapes of the non-rigid salient objects.
(2) A Channel-Wise Attention Mechanism (CWAM) is designed to compute an importance factor for each channel of feature map, and further to promote those informative channels whereas suppressing those unimportant ones.
(3) Numerous experiments on five benchmark datasets demonstrate the superiority of our proposed salient object detection model over the state-of-the-art methods.

The rest of the paper is organized as follows: Section 2 provides an overview of related works. Section 3 describes the details of our salient object detection model. Section 4 discusses the performance of our model compared to other state-of-the-art methods. Section 5 concludes this paper.

2. Related work

The overview of traditional salient object detection methods based on hand-crafted features can be referred to [24]. In this section, we focus on the most related works, including CNNs based salient object detection methods and the attention mechanism.

2.1. CNNs based salient object detection

Early CNNs-based methods usually over-segment the input image into small regions. These regions are then fed into a deep network with fully connected layers to extract high-level features. Li et al. [18] extracted multi-scale deep features for every image region from three nested and increasingly larger rectangular windows. Wang et al. [25] integrated both local features and global cues to generate a weighted sum of salient regions. Lee et al. [26] exploited two subnetworks to encode low-level and high-level features separately. Wang et al. [27] proposed two complementary branches to capture effective semantic features and visual contrast information for saliency inference.

However, high-level features with low resolutions extracted from CNNs usually loose the spatial information because of the existence of the fully connected layers. Recently, several Fully Convolutional Network (FCN)-based methods have been proposed to overcome this problem by integrating multi-level information. Luo et al. [20] combined multi-level information through a multi-resolution 4×5 grid structure. Zhang et al. [19] integrated multi-level convolutional features to multiple resolutions and combined the deeper-level prediction and the shallower-level feature maps via a weighted addition.

Recently, some CNNs based methods attempt to capture rich contextual features to effectively detect the salient objects that have large variations in scale, shape and position. Wang et al. [28] proposed a pyramid pooling module to extract multi-scale features for saliency detection. Chen et al. [29] proposed an inception-segmentation module, in which input features were simultaneously filtered with different-size kernels. Wang et al. [30] used three context filters with different-size kernels to obtain multi-scale contextual information in the contextual weighting module. Different from their works, we introduce the deformable convolution into multi-scale convolution streams to build our proposed MSDM to effectively capture those non-rigid salient objects. With the aid of learned kernel offsets, the deformable convolution is able to achieve adaptive receptive fields, which can further grab irregular shapes of non-rigid objects. More analyses can be found in [31], which provides a survey on deep learning based salient object detection.

2.2. Attention mechanism

There exist two types of attention mechanisms including spatial attention and channel attention. Spatial attention is modeled to mimic the human ability of focusing on informative regions in visual scenes [32–34]. Channel attention is adopted to selectively promote those informative channels [35–37]. The attention mechanism has shown its efficiency in various vision tasks. Yu et al. [38] designed a channel attention block to guide the selection of low-level features with the help of high-level features for the task of sematic segmentation. In the task of image captioning, Chen et al. [35] jointly exploited the spatial-wise attention model and the channel-wise attention model to encode the attentive spatial locations and the attentive channels, respectively. Woo et al. [36] extracted those informative features by designing the attention models along the spatial and channel dimensions. In order to enhance the representation power of the network, Hu et al. [37] adopted an attention model by exploring the channel relationships.

Recently, the attention mechanism has been adopted for salient object detection. Kuen et al. [34] used a recurrent attention model to select local regions to refine their saliency maps. Li et al. [39] learned pixel-level attentional weights along with saliency maps based on different-scale versions of the same image. Zhang et al. [40] exploited a gate function to make the message adaptively pass among multi-level features. Liu et al. [41] designed a pixel-wise attention network to selectively attend to informative context locations. Zhang et al. [22] proposed a spatial attention mechanism to highlight the salient regions while suppressing the background ones, and a channel-wise attention mechanism to assign larger weights to those channels with higher responses to salient objects. Wang et al. [42] exploited a pyramid attention model for discriminative saliency representations with multi-scale feature learning and extended receptive fields. In [43], a channel-wise attention module and a spatial attention module were designed for high-level features and low-level features to capture informative context information for saliency prediction, respectively. Wang et al. [44] predicted the salient object from the fixation map. Differently, we design a novel channel-wise attention mechanism (i.e., CWAM) based on the global context prediction of each channel to promote those informative channels but suppress those unimportant ones.

3. Proposed method

Fig. 3 illustrates the overall architecture of the proposed salient object detection network. The framework consists of the backbone network, MSDM, CWAM, and saliency inference, which are corresponding to each row from top to bottom in Fig. 3, respectively. The details of each part will be elaborated in the following.

3.1. Backbone network

In our proposed framework, ResNet-50 network [45] is adopted as the backbone network. We modify ResNet-50 by removing the last pooling layer and the fully connected layers for the task of salient object detection. As shown in Fig. 3, the input image is fed into the modified ResNet-50 to obtain five levels of feature maps, which are represented as $\{\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_5\}$.

3.2. MSDM for multi-scale deformable visual cues

To obtain more primitive visual context, we design a MSDM at each level following the backbone network to learn multi-scale deformable visual cues. Basically, our MSDM consists of two steps: multi-scale context extraction and deformable context extraction. Regarding the former, three convolution streams with different-size kernels are designed to produce three-scale receptive fields, which are committed to capturing different-scale visual cues and thus detecting different-size salient objects. With respect to the latter, a deformable convolution designed for non-rigid object bounding boxes locations in [23] is embedded in the above multi-scale structure by adding a 2D offset to the regular grid sampling locations in the standard convolution, resulting in adaptive-shape geometric structures of the convolution kernels. Those irregular shapes of those non-rigid salient objects can be further explored. Combining these two steps, MSDM is able to robustly learn various receptive fields to capture more primitive visual context, covering different scales and diverse shapes of those non-rigid salient objects. The details of these two steps will be illustrated in the following.

3.2.1. Step 1: Multi-scale context extraction

Salient objects in different scenes usually have different scales. To extract different-scale context, we learn multi-scale features at each level except the deepest level[2]. Specifically, three streams of different scales are constructed by three convolutional layers with different-scale kernels, i.e., $1 \times 1, 3 \times 3$, and 5×5. Considering that convolutions with larger spatial filters tend to cause expensively computational complexity, we replace the convolutional layer of the kernel of 5×5 with two convolutional layers of the kernel of 3×3. Moreover, in order to reduce the number of channels, a convolutional layer with the kernel of 1×1 is added before the two convolutional streams with large filter kernels. The details are clearly

[2] Since the deepest-level feature maps are with a very small resolution but a large receptive field, which can capture high-level semantic knowledge. Therefore, we use a convolutional layer to obtain the global context ($11 \times 11 \times 1024$) instead of applying MSDM at this level.

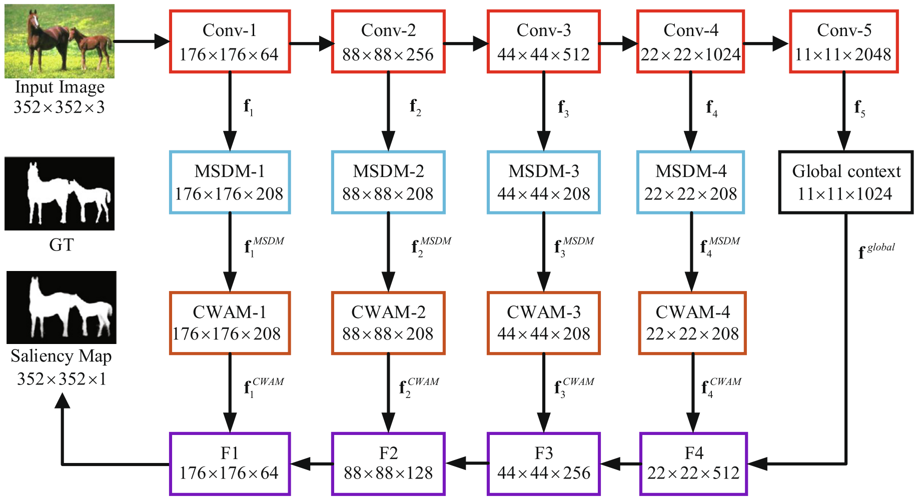


Fig. 3. Architecture of the proposed salient object detection network.

illustrated by the blue boxes in Fig. 4. In such way, these three streams possess different-scale receptive fields, and thus are able to capture different-scale visual context of salient objects.

3.2.2. Step 2: Deformable context extraction

In real scenes, there exist many non-rigid salient objects with various shapes. These non-rigid visual cues are challenging for the standard convolution with a fixed-shape kernel, which is inherently limited to model irregular geometric structures. To solve this issue, we capture the deformable visual context via the deformable convolution [23], which learns a 2D offset[3] for each kernel element based on the preceding feature maps. As such, adaptive receptive fields can be learned to capture diverse object shapes, leading to deformable context extraction of the non-rigid salient object.

As shown by the magenta boxes in Fig. 4, deformable context extraction is performed in the two streams with large filter kernels. It is displayed in Fig. 4 that based on the preceding feature maps, 4 groups of 2D offsets are learned via a series of convolutional layers, resulting in the same spatial resolution with the preceding feature map. The preceding feature maps and the learned 2D offset are then fed into the deformable convolution network. In such way, we can achieve diverse-shape receptive fields, which are able to robustly concentrate on different-shape visual cues of the non-rigid salient objects.

Fig. 5 visualizes the deformable receptive fields on a real image. As shown in Fig. 5, due to the involvement of the deformable convolution, the network can learn a primitive receptive field with respect to the background region and objects, further helping the model to locate background regions and objects. More specifically, receptive fields can cover different objects discriminatively.

Difference to MCFEM in BMP [40]. The major difference between our MSDM and MCFEM in [40] lies in that: The dilated convolution kernels with different dilated rates employed by MCFEM [40] indeed can capture different-scale context information, while they are with regular shapes, which are still trivial for diverse-shape visual context of those salient objects with irregular shapes. Differently, our MSDM adopts the deformable kernel that is able to adaptively learn a 2D offset to achieve a kernel with an adaptive shape, which can effectively capture non-rigid context information. Actually, the dilated kernel is a special case of the deformable kernel. At the stage of the above discussions, our MSDM is more general and robust than MCFEM in BMP [40] to extract multi-scale deformable visual context.

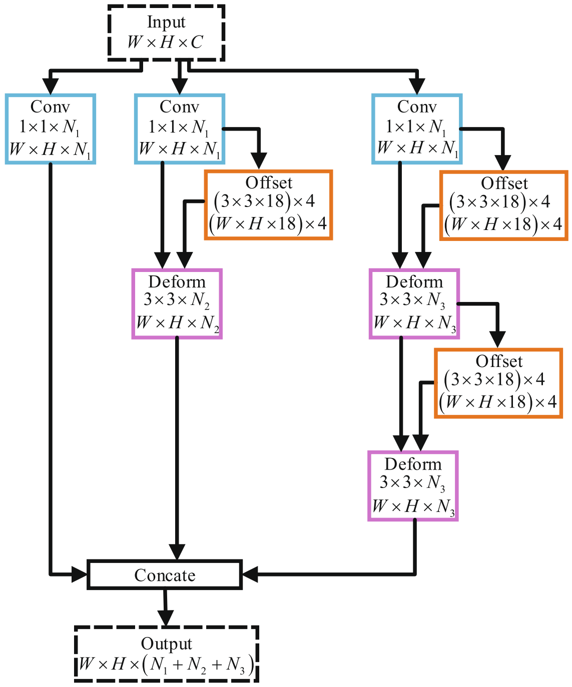


Fig. 4. Architecture of MSDM. The dotted-line boxes and solid-line ones represent data blocks and functional ones, respectively. In the data blocks, the top and bottom rows represent the name and output of the data block, respectively. In the functional block, the top, middle, and bottom rows represent the name, kernel, and output of the functional block, respectively. It is noted that 4 groups of 2D offsets are learned for each kernel. The kernel of 3 × 3 contains 9 elements, each of which has 2 offsets including horizontal and vertical offsets. Therefore, 18 offsets are computed for the kernel.

Comparison to GoogLeNet [46]. The similarity between our MSDM and GoogLeNet [46] lies in the inception structure, which utilizes multiple branches with different-size kernels and thereby different receptive fields. In contrast, the differences between our MSDM and GoogLeNet [46] mainly lie in two folds. First, our MSDM embeds the deformable convolution in two branches to achieve rich context with adaptive receptive fields. Secondly, we remove the pooling operation in GoogLeNet [46] to preserve clear feature representations (such as object boundaries) for the pixel-level segmentation.

[3] A horizontal displacement and a vertical displacement are learned for each kernel element, resulting in a 2D offset.

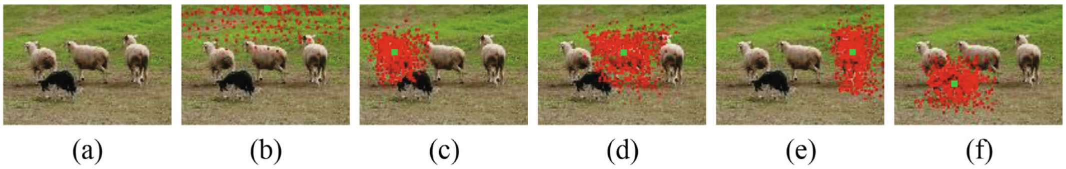


Fig. 5. Visualizations of the deformable receptive field. (a) Image; (b)–(f): Sampling locations (9^3 red points in each image) in three levels of 3 × 3 deformable filters for one activation unit (green point) on (b): background and (c)–(f): different objects, respectively.

3.3. CWAM for channel boosting and suppression

Most methods deal with different channels of feature maps without distinction. Actually, different channels capture different semantics and thus contribute unequally to the saliency prediction. In order to selectively emphasize those informative channels but suppress less useful ones, we design a novel channel-wise attention mechanism, i.e., CWAM, based on the global context of each channel. This is achieved by calculating a saliency score for each channel based on its feature maps. The prediction scores are utilized as the channel-wise attention scores to weigh the importance of different channels, which is achieved by a series of convolutions and a depth-wise convolution [47].

As shown in Fig. 6, the preceding feature maps of MSDM, denoted as $\mathbf{f}_i^{MSDM}$, are first downsampled to the resolution 11 × 11 via a series of convolutional layers. Then, we achieve a global context score (a real number) for each channel via the depth-wise convolutional layer [47] that performs a spatial convolution independently over each channel of feature map. In such way, we will compute a score number for each channel. These scores across all the channels are normalized by the sigmoid function, resulting in the attention scores $\{\mathbf{a}_i\}$ (as shown by the circles of different colors in Fig. 6) to measure the importance for different channels of the preceding feature maps. Finally, the preceding feature maps of MSDM are multiplied by these obtained channel-wise attention scores to achieve the attended feature maps $\mathbf{f}_i^{CWAM}$, which enables to highlight those informative channels of MSDM whereas weakening those less useful ones. The details of CWAM are illustrated in Table 1.

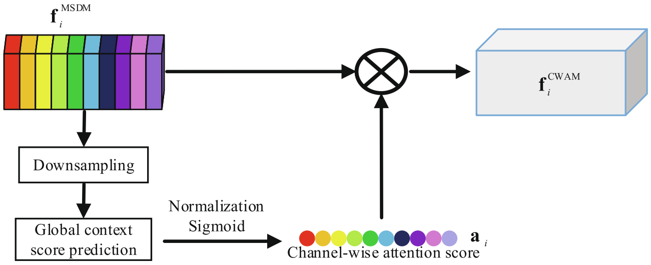


Fig. 6. Architecture of CWAM.

Difference to existing state-of-the-art attention mechanisms. The main differences lie in two folds: i) Most of existing attention mechanisms (e.g., SEN [37] and PAGRN [22]) apply a Global Average Pooling (GAP) [48] for each channel to compute its attention score, which is computed as the mean of all pixel values on each channel of feature map. In this context, all pixels on each channel of feature map are treated equally, which may degrade some informative pixels. As a departure from GAP [48], CWAM learns a weight for each pixel within each channel and then computes a saliency prediction score as the channel-wise score based on the global context of each channel. As a result, CWAM is more robust to boost those informative channels and suppress those less important ones. ii) Existing attention mechanisms usually adopt the fully connected layers after GAP [48] to achieve channel-wise importance factors, unavoidably leading to high complexity. Instead, our CWAM directly achieves channel-wise scores via the depth-wise convolution, which is a more gentle calculation.

3.4. Saliency inference

3.4.1. Saliency map generation

The feature maps of different levels contain different saliency cues. Deep-level feature maps are good at capturing high-level semantic knowledge, whereas shallow-level feature maps specialize in preserving low-level spatial details. Therefore, we integrate multi-level feature maps to predict the final saliency map. Considering the inconsistent resolutions of multi-level feature maps, we fuse them and generate the saliency map in a coarse-to-fine manner. The fusion process is summarized as follows:

$$\mathbf{F}_i = \begin{cases} Conv(Cat(dec(\mathbf{F}_{i+1}), \mathbf{f}_i^{CWAM}), dim_i), \ i = 1,2,3,4 \\ \mathbf{f}^{global}, \ i = 5, \end{cases} \tag{1}$$

where $Conv(\cdot, dim_i)$ performs a convolution operation on the input and outputs dim_i channels of feature maps. As illustrated in Fig. 3, $dim_1 = 64, \ dim_2 = 128, \ dim_3 = 256, \ dim_4 = 512$. $Cat(\cdot)$ is the cross-channel concatenation. $dec(\cdot)$ is the deconvolution layer with the kernel of 3 × 3. The output details of $\mathbf{F}_i$ can be viewed in the bottom row of Fig. 3. After a step-wise integration in Eq. (1), the integrated feature maps will simultaneously incorporate coarse semantics and fine details.

Finally, the saliency map of the input image can be computed from $\mathbf{F}_1$ by a series of operations, i.e., Conv (3, 3) -> BN -> ReLU -> Deconv -> BN -> ReLU -> Conv (3, 3) -> Sigmoid, where (3, 3) represents the kernel size.

3.4.2. Loss function

The joint loss function, which is obtained by the cross entropy loss function [40] and the IoU boundary loss function, is adopted to train the proposed salient object detection network, i.e.,

$$L_{Joint} = L_{CE} + L_{IoU}, \tag{2}$$

Here, the cross-entropy loss function is defined as

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c\in\{0,1\}}(y(\mathbf{v}_i) = c)(\log(\hat{y}(\mathbf{v}_i) = c)), \tag{3}$$

where $\mathbf{v}_i$ represents the location of pixel i. $y(\mathbf{v}_i)$ and $\hat{y}(\mathbf{v}_i)$ represent saliency values of the pixel i in the ground truth and the predicted saliency map, respectively. N represents the number of pixels in the input image.

The IoU boundary loss function is defined as

$$L_{IoU} = 1 - \frac{2|C_q \cap \hat{C}_q|}{|C_q| + |\hat{C}_q|}, \tag{4}$$

where $\hat{C}_q$ and C_q are the gradient magnitudes of saliency map and ground truth corresponding to region q. The gradient magnitude is computed by using a Sobel operator followed by a tanh activation on the saliency map. $|\cdot|$ represents the number of non-zero entries in a mask.

Table 1
Details of CWAM.

Block	Input	Layer	Kernel	Stride	Zero padding	Output
CWAM-1	MSDM-1	conv	4×4	4	Yes	$44 \times 44 \times 208$
	$(176 \times 176 \times 208)$	conv	4×4	4	Yes	$11 \times 11 \times 208$
		depth-wise conv	11×11	1	Yes	$1 \times 1 \times 208$
CWAM-2	MSDM-2	conv	4×4	4	Yes	$22 \times 22 \times 208$
	$(88 \times 88 \times 208)$	conv	2×2	2	Yes	$11 \times 11 \times 208$
		depth-wise conv	11×11	1	Yes	$1 \times 1 \times 208$
CWAM-3	MSDM-3	conv	2×2	2	Yes	$22 \times 22 \times 208$
	$(44 \times 44 \times 208)$	conv	2×2	2	Yes	$11 \times 11 \times 208$
		depth-wise conv	11×11	1	Yes	$1 \times 1 \times 208$
CWAM-4	MSDM-4	conv	2×2	2	Yes	$11 \times 11 \times 208$
	$(22 \times 22 \times 208)$	depth-wise conv	11×11	1	Yes	$1 \times 1 \times 208$

4. Experiments

4.1. Experimental setup

4.1.1. Datasets

We apply five public benchmark datasets, including ECSSD [49], HKU-IS [25], DUT-OMRON [16], PASCAL-S [50], and DUTS [51], to evaluate the performance of our network and other state-of-the-art methods.

ECSSD [49] contains 1000 complex images with objects of varying sizes. Some images even have multiple objects.

HKU-IS [25] consists of 4447 challenging images, including 3000 training images and 1447 test images. Most of those images in this dataset contain multiple salient objects with low contrast. We evaluate different methods on the test dataset.

DUT-OMRON [16] contains 5168 complex images. Each image in this dataset contains one or more salient objects, some of which are too large or too small. Besides, most images of this dataset are provided with cluttered backgrounds. This dataset is the most challenging for salient object detection at the moment.

PASCAL-S [50] stems from the PASCAL VOC dataset [52] and contains 850 images with various scenes.

DUTS [51] includes 10553 images for training and 5019 images for testing. Those images in this dataset have different scenes and various sizes. We evaluate the proposed salient object detection network on the test dataset.

4.1.2. Evaluation criteria

We adopt four evaluation criteria, including Precision-Recall (PR) curve [53], F-measure curve [53], mean F-measure (F_{mean}) [53], adaptive F-measure (F_{adap}) [53], and Mean Absolute Error (MAE) [54]. For a given continuous saliency map S, we convert it to a binary mask B by using a threshold. Then, its precision and recall values are computed as $precision = |B \cap S|/|B \cap S||B||B|$ and $recall = |B \cap S|/|B \cap S||S||S|$, respectively, where $|\cdot|$ accumulates the non-zero entries in a mask. F-measure is formulated as a weighted combination of Precision and Recall, i.e.,

$$F_{\beta} = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall}. \quad (5)$$

As suggested in [53], $\beta^2 = 0.3$.

PR curve is plotted by a series of pairs of precision and recall values under different thresholds.

F-measure curve of a dataset is plotted by a series of F-measure values, which are computed based on a series of pairs of precision and recall values under different thresholds.

$\boldsymbol{F_{mean}}$ for a dataset is computed by

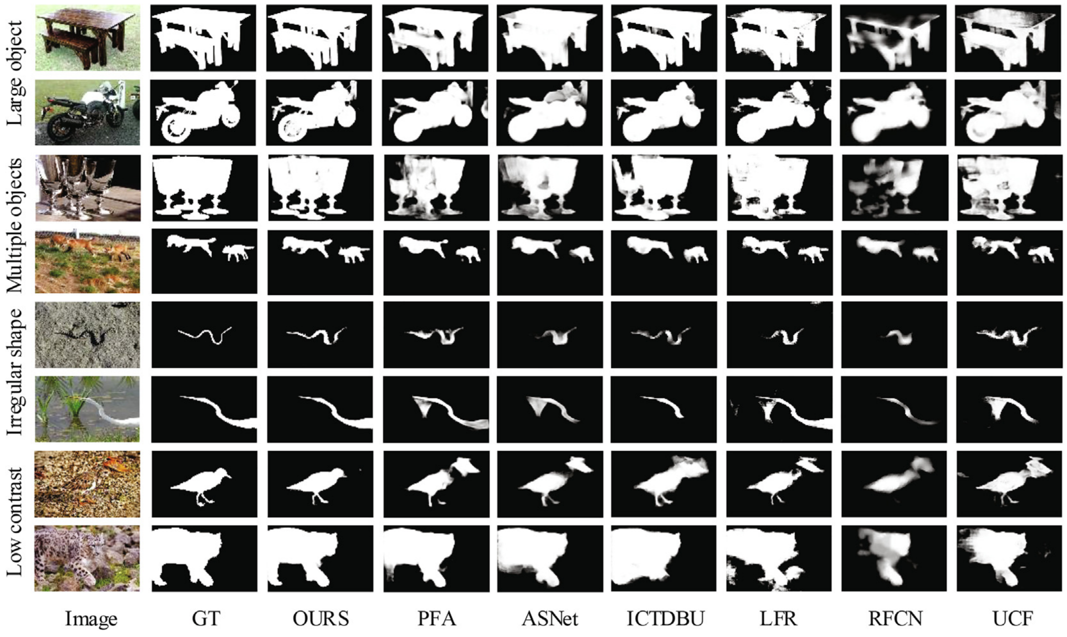


Fig. 7. Detection results of different methods.

$$F_{mean} = \frac{1}{num1}\sum_{i=1}^{num1}\left(\frac{1}{num2}\sum_{j=1}^{num2}(F_\beta)_i^j\right), \tag{6}$$

where $(F_\beta)_i^j$ represents the F_β value for the $i-th$ image under the $j-th$ threshold in a dataset. $num1$ is the number of images within the dataset. $num2 = 256$ is the number of thresholds, which range from 0 to 1 with a uniform distribution.

$\boldsymbol{F_{adp}}$ for a dataset is formulated as

$$F_{adp} = \frac{1}{num2}\sum_{i=1}^{num2}(F_\beta)_{2mean}^i, \tag{7}$$

Here, $(F_\beta)_{2mean}^i$ is the F_β value for the $i-th$ image under a specific threshold, which is set to twice the mean saliency value of the saliency map.

MAE reflects the average pixel-wise absolute difference between the saliency map and the ground truth. It is defined as

$$MAE = \frac{1}{w \times h}\sum_{x=1}^{w}\sum_{y=1}^{h}|S(x,y) - B(x,y)|. \tag{8}$$

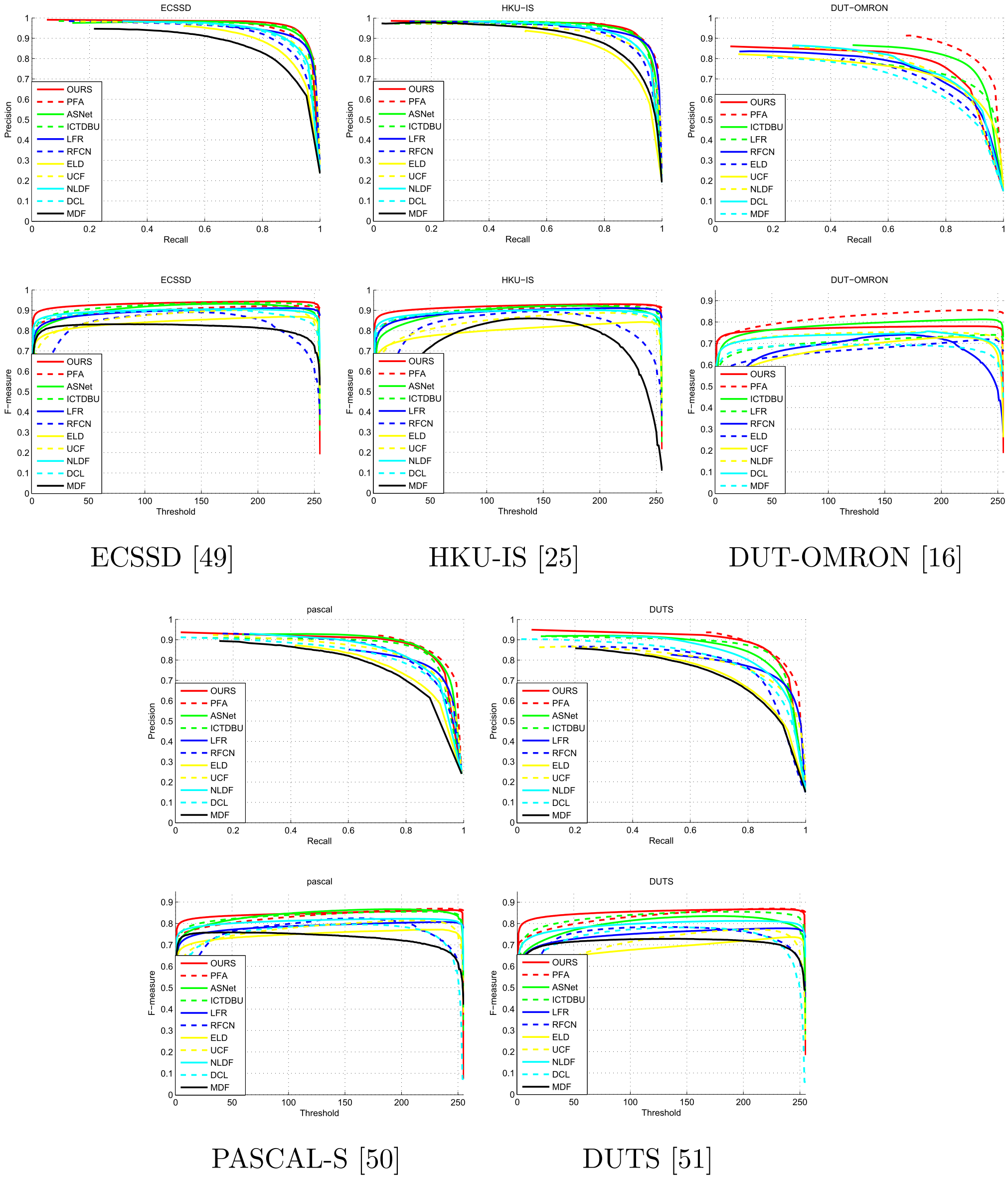
Here, w and h are the height and width of the image, respectively.

4.1.3. Implementation details

The proposed deep salient object detection network is trained on Pytorch with the Stochastic Gradient Descent (SGD) optimizer [55]. We set the initial learning rate to 10^{-3}. The momentum parameter, weight decay, batch size are set to 0.9, 0.0005, and 4, respectively. Our backbone network is initialized by the pre-trained ResNet-50 weights [45] and other convolutional parameters are randomly initialized. The DUTS training dataset [51], which is augmented by mirror reflection and rotation techniques to improve the varieties, is chosen to train the proposed network. The input images are resized to 352 × 352 for training and testing. Training the model to convergence needs about 50 h.

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

Our network is compared with 10 state-of-the-art deep learning based salient object detection methods, including ASNet [44], PFA



ECSSD [49] HKU-IS [25] DUT-OMRON [16]

PASCAL-S [50] DUTS [51]

Fig. 8. PR and F-measure curves of different methods.

Table 2
F_{mean} (larger is better) and MAE (smaller is better) values of different methods. The best three results are displayed in red, green, and blue, respectively. "-" means that the corresponding authors do not provide the detection results of the dataset.

	ECSSD [49]		HKU-IS [25]		DUT-OMRON [16]		PASCAL-S [50]		DUTS [51]	
	F_{mean}	MAE	F_{mean}	MAE	F_{mean}	MAE	F_{mean}	MAE	F_{mean}	MAE
Ours	0.9142	0.0373	0.8997	0.0316	0.7262	0.0582	0.8217	0.0652	0.8259	0.0395
PFA [43]	0.8900	0.0448	0.8909	0.0328	0.8009	0.0415	0.8203	0.0655	0.8148	0.0409
ASNet [44]	0.8903	0.0468	0.8732	0.0414	-	-	0.8174	0.0699	0.7609	0.0607
ICTDBU [56]	0.9041	0.0411	0.8852	0.0373	0.7596	0.0605	0.8137	0.0711	0.8042	0.0482
LFR [57]	0.8794	0.0525	0.8786	0.0396	0.6776	0.1030	0.7641	0.1066	0.7211	0.0834
RFCN [58]	0.7894	0.1070	0.7897	0.0882	0.6189	0.1105	0.7247	0.1320	0.6934	0.0900
ELD [59]	0.8218	0.0783	0.7826	0.0719	0.6342	0.0909	0.7149	0.1206	0.6509	0.0924
UCF [21]	0.8453	0.0690	0.8396	0.0612	0.6318	0.1204	0.7410	0.1155	0.6630	0.1122
NLDF [20]	0.8742	0.0626	0.8713	0.0480	0.6825	0.0796	0.7811	0.0977	0.7568	0.0651
DCL [27]	0.8552	0.0679	0.8607	0.0481	0.6762	0.0797	0.7107	0.1257	0.6755	0.0879
MDF [60]	0.7586	0.1050	0.6882	0.1292	0.6177	0.0916	0.6516	0.1420	0.6437	0.0935

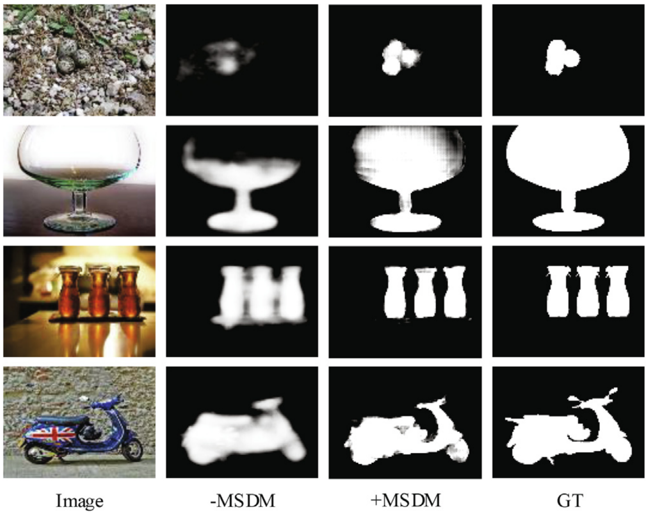


Fig. 9. Visual illustrations for MSDM.

[43], ICTDBU [56], LFR [57], RFCN [58], UCF [21], NLDF [20], DCL [27], ELD [59], and MDF [60]. For fair comparisons, we utilize the saliency detection results provided by their corresponding authors.

4.2.1. Visual comparison

Fig. 7 shows some detection results of different methods on various challenging scenarios, including large object, multiple objects, irregular shape, and low contrast. To be specific, for those large objects, the compared methods hardly detect the complete salient objects. In contrast, our method can detect the whole salient objects with uniform saliency values. For those multiple objects, the compared methods miss some salient object parts, while our approach can locate all the salient objects and predict complete object shapes. For those objects with irregular shapes, most of the compared methods cannot detect the salient objects completely and accurately. By contrast, our model can handle these cases very well, which owes to the primitive visual context captured by the proposed MSDM. For those scenes with low contrast between foreground and background, the state-of-the-art methods easily introduce some confusing noise into the detection results, while our model can effectively suppress the distractions. This benefits from the proposed CWAM that promotes those informative channels of feature maps whereas suppressing those confusing ones.

4.2.2. Quantitative comparison

Fig. 8 displays PR and F-measure curves of different methods. It can be found from Fig. 8 that our model is competitive with the compared methods in terms of PR curves and achieves higher F-measure values across a wider range over the other listed methods with respect to F-measure curves on ECSSD [49], HKU-IS [25],

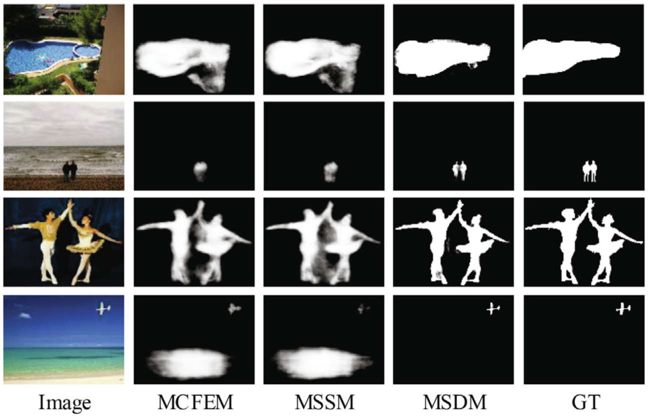


Fig. 10. Visual illustrations for visual context extraction.

Table 3
Performance illustrations for MSDM on DUT-OMRON [16]. "+MSDM" and "MSDM" are our entire framework. The best performance in each group is marked by bold.

	MSDM		Visual context extraction			Deformable kernel	
	+MSDM	-MSDM	MSDM	MCFEM [40]	MSSM	+Deformation	Backbone
F_{adp}	**0.7601**	0.5899	**0.7601**	0.6693	0.6727	**0.5682**	0.5151
MAE	**0.0588**	0.1162	**0.0588**	0.0792	0.0789	**0.1227**	0.1403

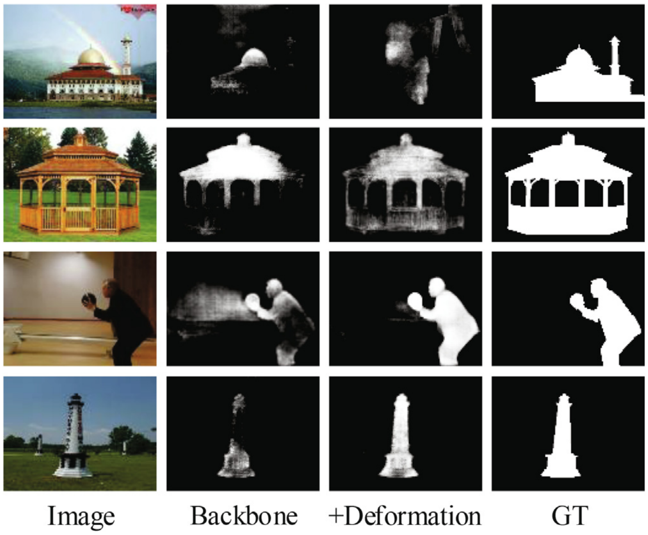


Fig. 11. Visual illustrations for the deformable convolutional kernel.

PASCAL-S [50], and DUTS [51]. On DUT-OMRON [16], our method gets poorer performance than PFA [43] and ICTDBU [56], but beats the other methods.

Table 2 lists F_{mean} and MAE values of different methods. It is obvious that our model achieves the highest F_{mean} values and the lowest MAE values among all the listed methods on ECSSD [49], HKU-IS [25], PASCAL-S [50], and DUTS [51]. On DUT-OMRON [16] that contains images with complex scenes, our method ranks the third and the second in terms of F_{mean} and MAE metrics, respectively. In summary, our method performs robustly for various scenes.

4.3. Ablation analysis

In this section, we conduct a series of experiments to analyze the contributions of the two proposed modules in our model, i.e., MSDM and CWAM. Here we train our model with the MSRA10K dataset [14] just using the cross entropy loss function.

4.3.1. MSDM

To clearly demonstrate the effectiveness of the proposed MSDM, we compare our entire model with a modified version, which is obtained by removing MSDM from our entire model. Fig. 9 and Table 3 show some detection results and the quantitative performance to illustrate the effectiveness of the proposed MSDM, respectively. It can be easily found from Table 3 that the proposed MSDM greatly improve the performance. Besides, as shown in the top two rows of Fig. 9, MSDM can detect small and large salient objects, which is owing to the three streams for multi-scale visual context extraction. As displayed in the bottom two rows of Fig. 9, MSDM can well completely detect those non-rigid salient objects with good backgrounds suppression, which benefits from the adaptively diverse-shape kernels provided by MSDM.

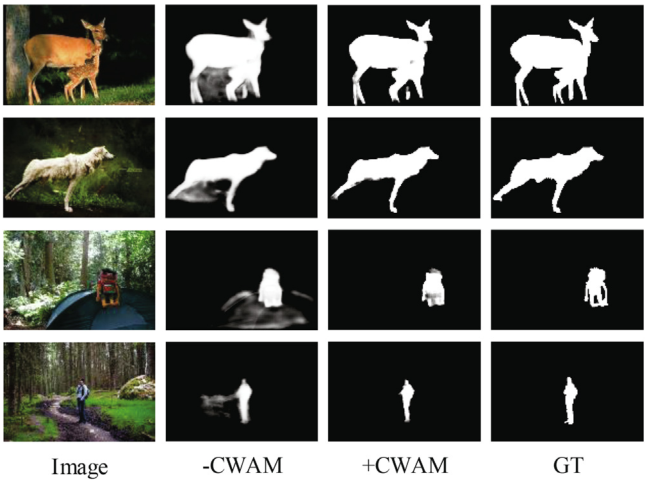


Fig. 12. Visual illustrations for CWAM.

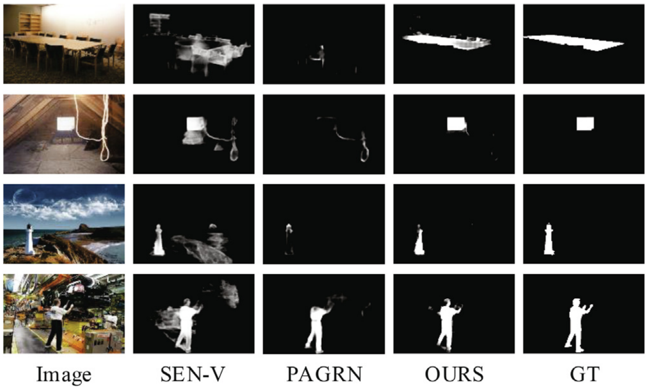


Fig. 13. Visual illustrations for the attention mechanism.

4.3.2. Visual context extraction comparison

To describe the superiority of our visual context extraction mechanism, i.e., MSDM, over MCFEM in BMP [40], we compare our model with a modified version, which is constructed by replacing MSDM with MCFEM [40] in our entire model. This modified version is also called MCFEM for short. As listed in Table 3, our MSDM achieves better F_{adp} and MAE values than MCFEM. As shown in Fig. 10, compared with MCFEM, MSDM is able to extract more accurate shapes and clear boundaries of non-rigid salient objects. The superiority of our MSDM mainly lies in that the deformable convolution kernels used in our MSDM can extract diverse-shape kernels in the real scenes, which is impossible for the dilated convolution kernels used by MCFEM [40].

In order to further illustrate the ability of the deformable kernel for non-rigid visual context extraction, we compare our entire model with another modified version, which is obtained by replacing the deformable convolution with the standard convolution in MSDM. This modified version is called Multi-Scale Standard

Table 4
Performance illustrations for CWAM on DUT-OMRON [16]. “+CWAM” and “CWAM” are our entire framework. The best performance in each group is marked by bold.

	CWAM		Attention mechanism		
	+CWAM	-CWAM	CWAM	PAGRN	SEN-V
F_{adp}	**0.7601**	0.6920	**0.7601**	0.7114	0.6933
MAE	**0.0588**	0.0739	**0.0588**	0.0709	0.0739

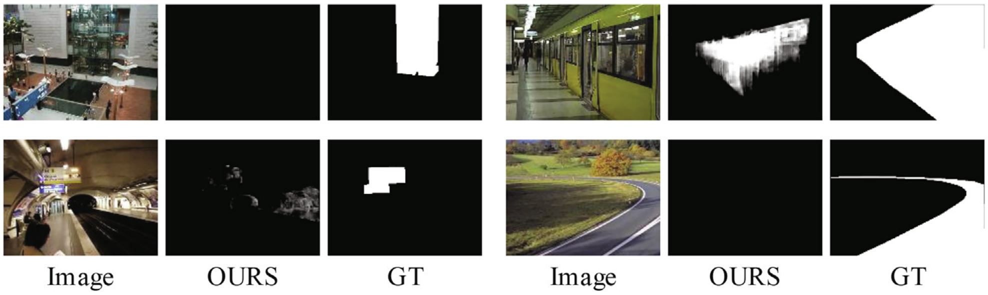


Fig. 14. Some failure cases.

Module (MSSM). As shown in Table 3 and Fig. 9(b), it is obvious that MSDM gets much better performance than MSSM, which indicates the effectiveness of the deformable kernel for robust visual cues extraction.

In addition, to verify the effectiveness of the deformable kernel in a single scale architecture, we compare the backbone network (i.e., ResNet-50) and the modified version by adding the deformable convolution. As shown in Table 3, the deformable convolution can promote the performance to a large margin, compared to the backbone network. Besides, it can be obviously found from Fig. 11 that the deformable convolution can achieve more wholeness and uniformity for the salient object.

4.3.3. CWAM

In this section, we analyze the effectiveness of the proposed CWAM. The quantitative performance and visual illustrations are shown in Table 4 and Fig. 12, respectively. It is obvious from Table 4 that the proposed CWAM improves the detection performance by a large margin. In addition, it can be easily seen from Fig. 12 that CWAM can well suppress the confusing backgrounds. This is because that CWAM can promote those informative channels of feature maps whereas weakening those confusing ones.

4.3.4. Attention mechanisms comparison

In this section, we discuss the superiority of our attention mechanism, i.e., CWAM, over the state-of-the-art mechanisms, including SEN [37] and PAGRN [22]. Specifically, SEN [37] is a state-of-the-art attention mechanism used for classification, and PAGRN [22] is a state-of-the-art attention mechanism based salient object detector. The evaluation performance comparison and visual comparison are shown in Table 4 and Fig. 13, respectively. It can be easily found from Table 4 that our CWAM achieves much better performance than PAGRN and SEN-V[4]. Besides, as shown in Fig. 13, SEN-V introduces some background noise into the detected saliency map, and PAGRN cannot detect those difficult salient objects (e.g., the top three rows of Fig. 13) or misses some salient parts (e.g., the bottom row of Fig. 13). In contrast, our CWAM is able to accurately highlight the salient objects whereas suppressing those backgrounds. This is owing to that the proposed CWAM precisely protrudes those informative channels whereas weakening those unimportant ones.

4.4. Failure cases

Fig. 14 displays some challenging cases for our model. In these cases, the foregrounds and backgrounds have quite low contrast (as shown in the first column of Fig. 14), and the salient objects span over a large range within the image (as shown in the fourth column of Fig. 14). These complicated scenes make our method difficult to identify the salient objects. Scene parsing [61,62] may be a solution for salient object detection under these complex scenes.

5. Conclusion

In this paper, we have presented a novel deep end-to-end salient object detection framework, which consists of a MSDM and a CWAM. Specifically, MSDM is proposed to robustly extract more primitive visual context information, covering different scales and varying shapes of the salient objects. CWAM is designed to compute the channel-wise attention impactors that further promote those informative channels while suppressing those unimportant ones. With the aid of MSDM and CWAM, the proposed model can well detect the salient objects with good accuracy and completeness in various scenes. In the future, we will apply scene parsing for salient object detection to improve the detection performance in the complex scene.

CRediT authorship contribution statement

Yi Liu: Conceptualization, Methodology, Writing - review & editing. **Mingxing Duanmu:** Visualization, Software. **Zhen Huo:** Methodology, Software; Writing - original draft. **Hang Qi:** Validation. **Zuntian Chen:** Investigation. **Lei Li:** Project administration, Funding acquisition. **Qiang Zhang:** Supervision, Conceptualization, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 62001341 and No. 61773301, and by the Science Foundation of Science and Technology on Complex System Control and Intelligence Agent Cooperative Laboratory under Grant No. 181101.

References

[1] S. Bi, G. Li, Y. Yu, Person re-identification using multiple experts with random subspaces, Journal of Image and Graphics 2 (2) (2014) 151–157.
[2] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 597–606.
[3] Y. Su, Q. Zhao, L. Zhao, D. Gu, Abrupt motion tracking using a visual saliency embedded particle filter, Pattern Recognition 47 (5) (2014) 1826–1834.
[4] C. Craye, D. Filliat, J.-F. Goudou, Environment exploration for object-based visual saliency learning, in: Proceedings of the IEEE International Conference on Robotics and Automation, 2016, pp. 2303–2309.
[5] Y. Lin, Z. Pang, D. Wang, Y. Zhuang, Task-driven visual saliency and attention-based visual question answering, arXiv preprint arXiv:1702.06700.

[4] We replace the proposed CWAM with SEN in our entire model for a fair comparison. This version is called SEN-V for short.

[6] M. Donoser, M. Urschler, M. Hirzer, H. Bischof, Saliency driven total variation segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2009, pp. 817–824.
[7] Q. Cai, H. Liu, Y. Qian, S. Zhou, X. Duan, Y.-H. Yang, Saliency-guided level set model for automatic object segmentation, Pattern Recognition 93 (2019) 147–163.
[8] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, S.-M. Hu, Sketch2photo: Internet image montage, ACM transactions on graphics 28 (5) (2009) 124: 1–10.
[9] C. Goldberg, T. Chen, F.-L. Zhang, A. Shamir, S.-M. Hu, Data-driven object manipulation in images, Computer Graphics Forum 31 (2) (2012) 265–274.
[10] H. Liu, L. Zhang, H. Huang, Web-image driven best views of 3d shapes, The Visual Computer 28 (3) (2012) 279–287.
[11] P. Zhang, T. Zhuo, H. Huang, M. Kankanhalli, Saliency flow based video segmentation via motion guided contour refinement, Signal Processing 142 (2018) 431–440.
[12] X. Wang, L. Ma, S. Kwong, Y. Zhou, Quaternion representation based visual saliency for stereoscopic image quality assessment, Signal Processing 145 (2018) 202–213.
[13] D.A. Klein, S. Frintrop, Center-surround divergence of feature statistics for salient object detection, in: Proceedings of the International Conference on Computer Vision, 2011, pp. 2214–2219.
[14] M.-M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (3) (2015) 569–582.
[15] H. Lu, X. Li, L. Zhang, X. Ruan, M.-H. Yang, Dense and sparse reconstruction error based saliency descriptor, IEEE Transactions on Image Processing 25 (4) (2016) 1592–1603.
[16] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
[17] H.-H. Yeh, K.-H. Liu, C.-S. Chen, Salient object detection via local saliency estimation and global homogeneity refinement, Pattern Recognition 47 (4) (2014) 1740–1750.
[18] G. Li, Y. Yu, Visual saliency detection based on multiscale deep cnn features, IEEE Transactions on Image Processing 25 (11) (2016) 5012–5024.
[19] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202–211.
[20] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, P.-M. Jodoin, Non-local deep features for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6609–6617.
[21] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 212–221.
[22] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 714–722.
[23] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, Y. Wei, Deformable convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 764–773.
[24] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, IEEE Transactions on Image Processing 24 (12) (2015) 5706–5722.
[25] L. Wang, H. Lu, X. Ruan, M.-H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3183–3192.
[26] G. Lee, Y.-W. Tai, J. Kim, Deep saliency with encoded low level distance map and high level features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 660–668.
[27] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 478–487.
[28] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4019–4028.
[29] X. Chen, A. Zheng, J. Li, F. Lu, Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 1050–1058.
[30] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: a novel approach to saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3127–3135.
[31] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, R. Yang, Salient object detection in the deep learning era: An in-depth survey, arXiv preprint arXiv:1904.09146.
[32] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 2048–2057.
[33] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, Residual attention network for image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3156–3164.
[34] J. Kuen, Z. Wang, G. Wang, Recurrent attentional networks for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3668–3677.
[35] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.-S. Chua, Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5659–5667.
[36] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, 2018, pp. 3–19.
[37] J. Hu, L. Shen, G. Sun, Squeeze-and-excitation networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7132–7141.
[38] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, Learning a discriminative feature network for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1857–1866.
[39] G. Li, Y. Xie, L. Lin, Y. Yu, Instance-level salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2386–2395.
[40] L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, A bi-directional message passing model for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1741–1750.
[41] N. Liu, J. Han, M.-H. Yang, Picanet: Learning pixel-wise contextual attention for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.
[42] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1448–1457.
[43] T. Zhao, X. Wu, Pyramid feature attention network for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3085–3094.
[44] W. Wang, J. Shen, X. Dong, A. Borji, R. Yang, Inferring salient objects from human fixations, IEEE Transactions on Pattern Analysis and Machine Intelligence 42 (8) (2020) 1913–1927.
[45] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[46] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[47] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[48] M. Lin, Q. Chen, S. Yan, Network in network, in: Proceedings of the International Conference on Learning Representations, 2014.
[49] J. Shi, Q. Yan, L. Xu, J. Jia, Hierarchical image saliency detection on extended cssd, IEEE Transactions on Pattern Analysis and Machine Intelligence 38 (4) (2016) 717–729.
[50] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[51] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, Learning to detect salient objects with image-level supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 136–145.
[52] M. Everingham, S.A. Eslami, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes challenge: a retrospective, International Journal of Computer Vision 111 (1) (2015) 98–136.
[53] R. Achanta, S. Hemami, F. Estrada, S. Süsstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 1597–1604.
[54] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 733–740.
[55] L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of the International Conference on Computational Statistics, 2010, pp. 177–186.
[56] W. Wang, J. Shen, M. Cheng, L. Shao, An iterative and cooperative top-down and bottom-up inference network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 5968–5977.
[57] P. Zhang, W. Liu, H. Lu, C. Shen, Salient object detection with lossless feature reflection and weighted structural loss, IEEE Transactions on Image Processing 28 (6) (2019) 3048–3060.
[58] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Salient object detection with recurrent fully convolutional networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 41 (7) (2018) 1734–1746.
[59] G. Lee, Y.-W. Tai, J. Kim, Eld-net: An efficient deep learning architecture for accurate saliency detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 40 (7) (2018) 1599–1610.
[60] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 5455–5463.
[61] D. Xu, W. Ouyang, X. Wang, N. Sebe, Pad-net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 675–684.
[62] W.-C. Hung, Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, M.-H. Yang, Scene parsing with global context embedding, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2631–2639.

Yi Liu received the B. S. degree from Nanjing Institute of Technology, Nanjing, China, in 2012, the M. S. degree from the Dalian University, Dalian, China, in 2015, and the Ph. D. degree in Control Theory and Control Engineering at Xidian University, China, in 2019. He was a visiting student at Lancaster University from 2018 to 2019. He is currently working at Changzhou University, Changzhou, China. His current research interests include computer vision and salient object detection.

Mingxing Duanmu received the B.S. degree from Henan University of Science and Technology, Luoyang, China, in 2019. He is currently pursuing the M.S. degree in control engineering with Xidian University, Xi'an, China. His research interests include deep learning, computer vision and salient object detection.

Zhen Huo received the B. S. degree from Xi'an University of Technology, Xi'an, China, in 2016, and the M. S. degree from the Xidian University, Xi'an, China, in 2019. His research interests include deep learning and salient object detection.

Hang Qi received the B. S. degree in Electronic Engineering from Tsinghua University, China, in 2017. She is currently working towards the M. S. degree in Navigation Guidance and Control at Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, Beijing. Her current research interests include computer vision and automatic target recognition.

Zuntian Chen received the B.S. degree in Optics and the M.S. degree in Information Optics from Sichuan University, China, in 1988 and 1991, respectively. He is currently a researcher with Xi'an Institute of Electromechanical Information Technology, China. His current research interests include image processing and pattern recognition.

Lei Li received the B.S. degree in electronic information engineering from Wuhan University of Technology in 2002, and received the M.S. degree in pattern recognition and intelligent systems from Beihang University, China, in 2006. He is currently an engineer with Science and Technology on Complex System Control and Intelligent Agent Cooperation Laboratory, China. His current research interests include computer vision and automatic target recognition.

Qiang Zhang received the B.S. degree in automatic control, the M.S. degree in pattern recognition and intelligent systems, and the Ph.D. degree in circuit and system from Xidian University, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with the Center for Intelligent Machines, McGill University, Canada. He is currently a professor with the Automatic Control Department, Xidian University, China. His current research interests include image processing, pattern recognition.