Engaging Part-whole Hierarchies and Contrast Cues for Salient Object Detection

Zhang, Qiang; Duanmu, Mingxing; Luo, Yongjiang; Liu, Yi ; Han, Jungong

Published in:
IEEE Transactions on Circuits and Systems for Video Technology

DOI:
10.1109/TCSVT.2021.3104932

Publication date:
2022

Citation for published version (APA):
Zhang, Q., Duanmu, M., Luo, Y., Liu, Y., & Han, J. (2022). Engaging Part-whole Hierarchies and Contrast Cues for Salient Object Detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(6), 3644-3658. https://doi.org/10.1109/TCSVT.2021.3104932

Engaging Part-whole Hierarchies and Contrast Cues for Salient Object Detection

Journal:	*IEEE Transactions on Circuits and Systems for Video Technology*
Manuscript ID	TCSVT-06173-2021.R1
Manuscript Type:	Transactions Papers - Regular Issue
Date Submitted by the Author:	16-Jun-2021
Complete List of Authors:	Zhang, Qiang; Department of Automatic Control, Xidian University Duanmu, Mingxing; Department of Automatic Control, Xidian University Luo, Yongjiang; Xidian University Liu, Yi; Changzhou University, School of Computer Science and Artificial Intelligence, and Aliyun School of Big Data Han, Jungong; Aberystwyth University, computer science
EDICS:	2.6.7 ☐ Scene Understanding < 2.6 ☐ Detection, Recognition and Classification < 2 ☐ IMAGE/VIDEO ANALYSIS AND COMPUTER VISION

Engaging Part-whole Hierarchies and Contrast Cues for Salient Object Detection

Qiang Zhang, Mingxing Duanmu, Yongjiang Luo, Yi Liu* and Jungong Han*

***Abstract*—Real-world scenes always exhibit objects with clutter backgrounds, posing great challenges for deep salient object detection models. In this paper, we propose salient object detection by engaging two saliency cues, *i.e.*, the part-whole hierarchies and contrast cues, resulting in a PWHCNet. Specifically, two branches, which consists of a Dynamic Grouping Capsules (DGC) branch and a DenseHRNet branch, are put in place to learn the part-whole hierarchies and contrast cues, respectively. Moreover, to help highlight the whole salient object in complex scenes, a Background Suppression (BS) module is proposed to guide the shallow features of DenseHRNet with the aid of the part-whole relational cues captured by DGC. Subsequently, these two saliency cues are integrated via a Self-Channel and Mutual-Spatial (SCMS) attention mechanism. Experimental results on five benchmarks demonstrate that the proposed PWHCNet achieves state-of-the-art performance while obtaining the whole salient objects with fine details.**

***Index Terms*—Salient object detection, part-whole hierarchies, contrast, attention.**

I. Introduction

SALIENT Object Detection (SOD) highlights and segments out the most visually appealing objects or regions in natural images [1]–[3]. Acting as a preprocessing step, SOD has been applied in many computer vision fields in recent years, *e.g.*, weakly-supervised image semantic segmentation [4], visual tracking [5], object recognition [6], image retrieval [7] and video compression [8].

Hand-crafted features (*e.g.*, color, texture, *etc*.) dominate the development of earlier salient object detectors [9]–[11]. However, given the limited representation abilities of these features, these traditional methods encounter a performance bottleneck. In light of its powerful representation abilities, Convolutional Neural Networks (CNNs) have been successfully applied for salient object detection and achieved substantial performance improvements [12]–[14].

Despite impressive preliminary results have been achieved by CNNs, these methods still face some challenges. Existing CNNs based salient object detection approaches [14]–[16] predict the saliency map of an entire image mainly depending on the learned contrast information of each image region. Due to the ignorance of correlations between different object parts, these methods struggle to extract the whole objects from clutter scenes, which is demonstrated in the columns 3 and 4 of Fig. 1.

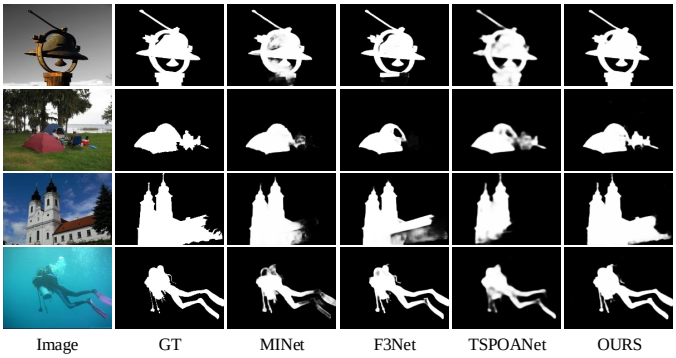


Fig. 1. Illustrations for sample results of our method compared with others. TSPOANet [17]: saliency detector based on part-whole relationships; MINet [14] and F3Net [16]: saliency detectors based on contrast information.

To alleviate the above problem, Liu *et al*. [17] investigated the role of part-whole relationships in salient object detection with the aid of the Capsule Network (CapsNet) [18]. Here, the salient object in a scene can be segmented out from the complicated background by discovering familiar object parts via exploring the part-whole relationships in the scene. As shown in the front two rows of Fig. 1, TSPOANet [17] can detect the whole salient objects from the backgrounds.

However, only part-whole relational cues may not be sufficient to segment complete objects from extremely complex scenes. For example, as illustrated in the last two rows of Fig. 1, some object regions are missed by TSPOANet [17], which may be attributed to the explored inaccurate part-whole hierarchies. This issue may arise from the noisy capsule assignments in TSPOANet [17], where the adopted two-stream strategy directly divides the capsules into two groups for capsules routing. Surprisingly, those missing object regions can be identified by such contrast based methods, *e.g.*, MINet [14] and F3Net [16], which demonstrates the contrast cues provide more exploration of local details compared to the part-whole relational cues. Based on the above observation, the two saliency cues, including the part-whole relational and contrast cues, can complement and reinforce each other for more robust salient object detection.

Considering that, in this paper, we propose a PWHCNet for salient object detection by interacting two saliency cues,

Qiang Zhang and Mingxing Duanmu are with Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China and also with Center for Complex Systems, School of Mechano-Electronic Engineering, Xidian University, Xi'an, Shaanxi 710071, China. Email: qzhang@xidian.edu.cn and duanmu@stu.xidian.edu.cn.

Yongjiang Luo is with the School of Electronic Engineering, Xidian University, Xi'an Shaanxi 710071, China. Email: yjluo@mail.xidian.edu.cn.

Yi Liu is with School of Computer Science and Artificial Intelligence, and Aliyun School of Big Data, Changzhou University, Changzhou, Jiangsu 213164, China. Email: liuyi0089@gmail.com.

Jungong Han is with Computer Science Department, Aberystwyth University, SY233FL, UK. Email: jungonghan77@gmail.com.

*Corresponding authors: Yi Liu and Jungong Han.

including part-whole hierarchies and contrast cues. Concretely, two branches are put in place to explore the part-whole hierarchies and contrast cues, respectively. In order to achieve the complementary information between these two saliency cues, we embed these two cues in a Self-Channel and Mutual-Spatial (SCMS) attention module. Specifically, in SCMS, the self-channel attention mechanism for one specific saliency cue is achieved via the channel weights computed on this cue itself, which helps to promote those informative channels while suppressing un-important ones. The mutual-spatial attention mechanism provides the spatial importance for one specific saliency cue with the aid of another saliency cue. The combination of self-channel and mutual-spatial attentions improves semantics for salient object detection.

Besides, to alleviate the problem of inaccurate part-whole relationships caused by the noisy capsule assignments, a Dynamic Grouping Capsules Routing (DGCR) strategy is proposed in the part-whole hierarchies exploration branch. Specifically, highly-correlated capsules are encouraged to be clustered into the same group for further capsules routing under the guidance of the proposed DGCR strategy. Such a dynamical grouping mechanism divides the capsules representing the same entity into the same group, which helps to alleviate noisy capsule assignments to some extent and thereby explores more accurate part-whole relational cues.

Similarly, to learn primitive contrast cues, a DenseHRNet framework is proposed on top of HRNet [19] to capture multi-scale context information with different receptive fields from the input image. The filtered results of different sub-layer convolutions are integrated through dense residual connections. In the meanwhile, a Background Suppression (BS) module is put at the head of the DenseHRNet sub-network, which aims to use the part-whole relational cues to guide the primitive contrast extraction. The resultant contrast cues will highlight the object regions well while suppressing the background region. As shown in Fig. 1, our model can produce more precise saliency maps in complex scenes, compared with other methods.

In summary, our contributions are summarized as follows:

1). A PWHCNet is proposed for salient object detection, which embeds the part-whole hierarchies and contrast cues into a SCMS attention mechanism to complement the information between them. To the best of our knowledge, it is the first attempt to simultaneously adopt the two saliency cues for salient object detection.

2). A DGC strategy is proposed to dynamically divide capsules with high correlations into a group for capsules routing, which helps to alleviate noisy capsule assignments and thereby explore more accurate part-whole relationships.

3). A DenseHRNet framework is designed to obtain more primitive contrast information with multiple scales while improving the flow of information and gradients throughout the network. Besides, under the guidance of the part-whole relational cues, the DenseHRNet sub-network pays more attention to the object regions.

The composition of this paper is described as follows. Sec. II reviews the works related to our method. Sec. III details the proposed network. Sec. IV conducts lots of experiments and analyses to evaluate the proposed method. Sec. V concludes this paper.

II. Related Work

A. Saliency detection

Traditional saliency detection methods [20]–[22] usually rely on hand-crafted priors. An overall review about these methods can be referred to [23]. Due to difficulties in capturing high-level semantics, these methods encounter a performance bottleneck. CNNs have broken this performance bottleneck because of their powerful representation abilities. For example, Li *et al.* [24] mined multi-scale deep features for high-precision visual saliency. In [25], a label decoupling framework was proposed for salient object detection by decoupling the saliency label into subject mapping and detail mapping. Zhang *et al.* [26] improved the accuracy of saliency detection by constructing an uncertain ensemble of internal feature units in specific convolutional layers. Cong *et al.* [27] proposed a depth-guided transformation model from RGB to RGBD saliency by capturing the explicit and implicit information from the depth map. In order to improve the performance of SOD, BASNet [28], EGNet [13] embedded boundary cues into the models to highlight the boundary regions of salient objects. In order to drive the network to discover complement object regions and details, Wang *et al*. [29] aggregated multi-scale salient context information by fusing those of multiple sub-regions. Chen *et al.* [30] proposed a reverse attention module in the top-down pathway to guide residual saliency learning.

In addition, deep contextual information has proved to be effective for SOD [31]. Zhang *et al*. [32] proposed a multi-level feature aggregation network to better integrate global contexts and local contexts by concatenating feature maps from both high levels and low levels directly. Wang *et al.* [33] used a weighted sum algorithm to integrate the estimated local saliency with a set of searched global salient regions to construct the final saliency map. In order to construct informative contextual features, Liu *et al*. [34] hierarchically embedded global and local context modules into a top-down pathway. Zhu *et al*. [35] aggregated the attentional dilated features by exploring the complementary information between the global and local context. Zhang *et al*. [36] gradually integrated multi-level contextual information through an attention guided network. Pang *et al*. [14] integrated the features from adjacent levels to obtain more efficient multi-scale features. Readers can gain a comprehensive understanding about these methods from [37].

The above mentioned methods try to extract more perceptual contexts for salient object detection. However, they ignore the fact that a target is composed of several geometric parts [38], which will lead to incomplete segmentation of the salient object. To address this problem, Liu *et al*. [17] proposed a part-whole relational saliency by involving the part-whole relational property in SOD with the aid of the Capsule Network (CapsNet) [18]. Specifically, in [17] the activation value of the capsule was used as the saliency value for each position. On top of that, a TSPOANet was proposed in [17] to get the whole saliency map through capsules routing, which was

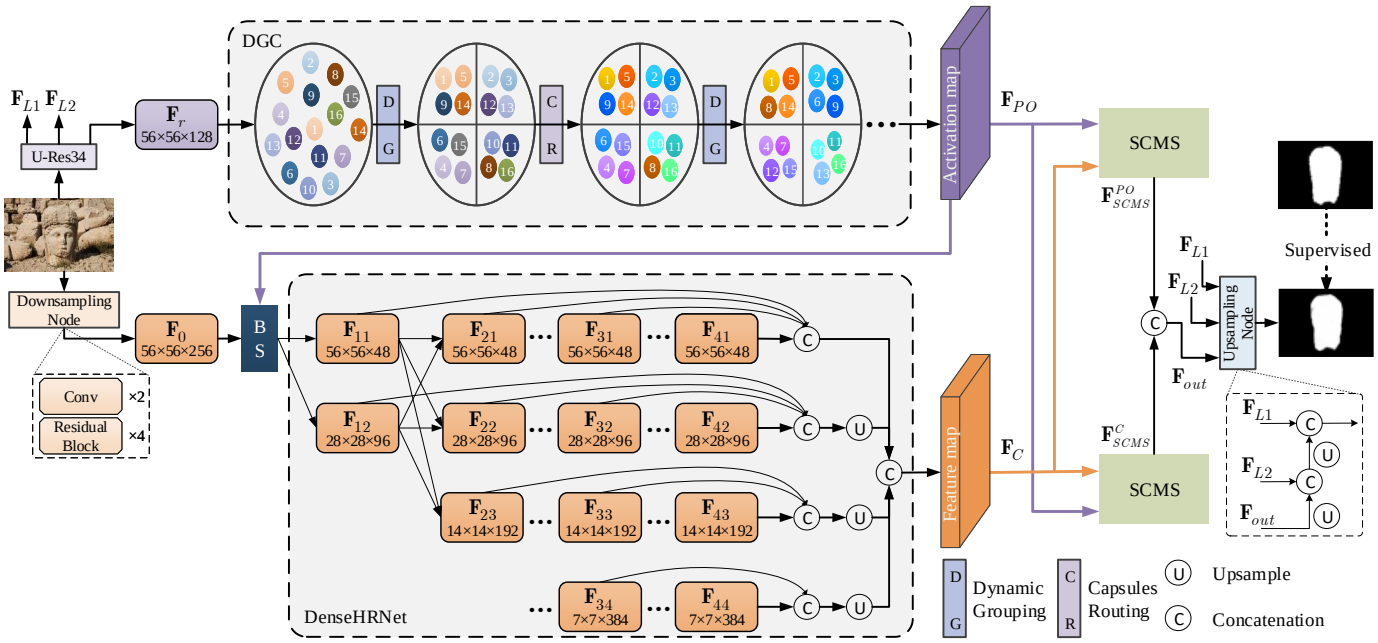


Fig. 2. The overall architecture of our proposed PWHCNet for salient object detection, which consists of a DGC sub-network and a DenseHRNet sub-network to capture the part-whole hierarchies and contrast cues from input images, respectively. The part-whole relational cues are additionally used to guide the feature extraction of DenseHRNet at the shallow layer via a BS module. On top of that, the above two saliency cues are interacted by a SCMS attention module to achieve more primitive saliency semantics $\mathbf{F}_{out}$, which are further used to predict the final saliency map. More details are provided in the text body.

implemented by using two streams for more accurate part-whole relationships while reducing the network parameters and noisy capsule assignments.

Different from the existing SOD methods, in this paper, two saliency cues, including contrast and part-whole hierarchies, are jointly used to infer the saliency map. This mechanism allows to obtain the whole saliency map with complete local details.

B. Attention mechanism

Attention mechanism has been widely applied in many fields, including machine translation [39], visual question answering [40], semantic segmentation [41] and image captioning [42]. In view of its advantages, the attention mechanism has also been used for SOD. For example, Cheng *et al.* [9] proposed a regional contrast algorithm to evaluate the global contrast differences and spatial coherence for saliency prediction. Kuen *et al.* [43] designed an attention network to identify the salient objects based on the spatial transformer and recurrent network. Liu *et al.* [34] proposed a pixel-wise contextual attention network by generating a contextual relevant spatial weight map to selectively attend the informative pixels for salient object detection. Li *et al.* [44] proposed an attention steered interweave fusion network for salient object detection, which progressively integrated cross-modal and cross-level complementarity from the RGB image and corresponding depth map. In [45], a top-down reverse attention mechanism was designed to guide a residual learning by using spatial weight convolution features, which was further embedded into each side output for residual refinement to detect the salient object. Chen *et al.* [46] designed a gated multi-modality attention module to capture long-range dependencies from a cross-modal perspective for RGB-D saliency detection. In order to utilize more useful features, some methods also try to combine channel and spatial attentions. Zhang *et al.* [36] proposed a progressive attention guided network, which generated attentive features by channel-wise and spatial attention mechanisms sequentially to selectively integrate multi-level contextual information for saliency detection. Zhao *et al.* [47] proposed a pyramid attention based salient object detection network via capturing the semantic high-level features and enhancing the low-level spatial structural features by a channel-wise attention module and a spatial attention module, respectively.

Different from the previous attention based SOD methods, we will design a new attention mechanism to well exploit the interaction information between the contrast cues and the part-whole hierarchies for SOD by simultaneously considering the intra-cues channel interaction and the inter-cues spatial interaction.

III. Proposed Method

Fig. 2 illustrates the overall architecture of the proposed salient object detection network, which fuses part-whole hierarchies and contrast cues to deal with the issue of inaccurate segmentation of salient objects in cluttered scenes. Specifically, a Dynamic Grouping CapsNet (DGC) sub-network and a DenseHRNet sub-network are proposed to capture the part-whole hierarchies and contrast cues from the input images, respectively. Additionally, the explored part-whole relational semantics are utilized to design a Background Suppression (BS) module to guide the shallow feature extraction in the

DenseHRNet sub-network. On top of that, the above two saliency cues are fully interacted by a Self-Channel and Mutual-Spatial (SCMS) attention mechanism to predict the final saliency map.

A. Exploring part-whole relationships stream

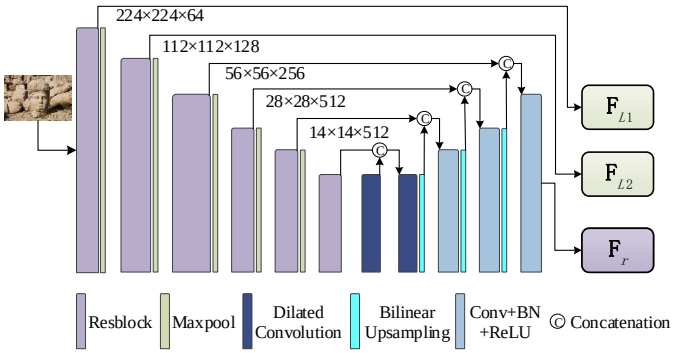


Fig. 3. Details of U-Res34. $\mathbf{F}_r$ will be used for capsule construction in our proposed model, while $\mathbf{F}_{L1}$ and $\mathbf{F}_{L2}$ will be used to recover salient object boundaries in the final saliency prediction stage.

1) Feature extraction for capsules construction:

Before the capsules routing, a U-Res34 unit (as shown in Fig. 3) is used to extract the deep semantic features $\mathbf{F}_r$ from the input images. As observed from Fig. 3, the randomly cropped input image $(224 \times 224 \times 3)$ is first fed into six basic res-blocks. To further capture the global information, a bridge block composed of a dilation convolution layer (dilation rate = 2) is added between the encoder and the decoder. For the decoder, the input of each block is the concatenation of previous upsampled feature maps and their corresponding encoded feature maps, which is able to integrate high-level contexts and low-level details. On top of that, the features $\mathbf{F}_r$ are transformed into multiple types of matrix capsules[1] (16 capsules in this paper), which is implemented by a Primary Capsule (PrimaryCaps) layer, as in [17]. In addition to $\mathbf{F}_r$, as shown in Fig. 3, another two sets of shallow features, *i.e.*, $\mathbf{F}_{L1}$ and $\mathbf{F}_{L2}$, will also be generated from the U-Res34 unit, which will be further used to restore the boundaries of salient objects in the final saliency inference stage.

2) Dynamic grouping for capsules routing:

Considering that CapsNet has the ability of capturing part-whole relationships [17], [18], we also adopt CapsNet [18] to explore the part-whole relational cues for saliency prediction as in [17]. However, the direct grouping strategy in [17] encounters noisy capsule assignments, which may cause inaccurate part-whole relationships and subsequent unsatisfactory results. Alternatively, taking into account the capsules correlations, we involve a dynamic grouping strategy for CapsNet to explore more accurate part-whole relationships in complex scenes. The details will be illustrated in the following contents.

As shown in the top branch of Fig. 2, small circles of different colors indicate distinct types of capsules. The dynamic grouping strategy is implemented before capsules routing to facilitate high-correlated capsules grouping for capturing more accurate part-whole relationships. In essential, capsules from the same object will have high familiarities. Therefore, highly familiar capsules are encouraged to be clustered into the same group for further routing within the group by virtue of the proposed dynamic grouping strategy, which will reduce some noisy capsule assignments. Specifically, the proposed dynamic grouping strategy consists of three steps, *i.e.*, calculating capsule correlation matrix, determining initial capsules in each group, and putting similar capsules into the same group.

Step 1: Calculating capsule correlation matrix: The property of a capsule is represented by its pose matrix. Thus, we measure the correlation among capsules by calculating the Manhattan distance (*i.e.*, L1 norm) among the pose matrices of different capsules. Concretely, the correlation $\mathbf{L}_{m,n}$ between capsules of type m and type n is expressed as follows:

$$\mathbf{L}_{m,n} = \|\sigma\left(Caps_m\right) - \sigma\left(Caps_n\right)\|_1, \tag{1}$$

where $Caps_{m/n}(m, n = 1, 2, ..., K)$ represents the attribute information for the capsule of type m or type n. K denotes the total number of capsule types and is set to 16 in this paper as in [17]. Here, we use the Sigmoid activation function (*i.e.*, $\sigma(*)$) to compress the value of $\mathbf{L}_{m,n}$ to $(0, 1)$, thus making the calculation process easier. After splicing $\mathbf{L}_{m,n}$, the capsule correlation matrix $\mathbf{L} \in \mathbb{R}^{K\times K}$ is thus obtained.

Step 2: Determining initial capsules in each group: As discussed in Step 1, the correlation coefficient $\mathbf{L}_{m,n}$ in the correlation matrix $\mathbf{L} \in \mathbb{R}^{K\times K}$ represents the similarity between the capsules of type m and type n. The larger the correlation coefficient, the higher the dissimilarity between the two capsules is. Then the horizontal and vertical coordinates, o_1 and o_2, of the maximum value in $\mathbf{L} \in \mathbb{R}^{K\times K}$ indicate the serial numbers of two types of capsules with the farthest similarity, *i.e.*,

$$[o_1, o_2] = \arg\max_{m,n}\left(\mathbf{L}_{m,n}\right), \mathbf{L} \in \mathbb{R}^{K\times K}, \tag{2}$$

where $\arg\max$ provides the indexes for the maximum value in the matrix $\mathbf{L}$. Correspondingly, $Caps_{o_1}$ and $Caps_{o_2}$ are defined as the initial capsules of two capsule groups to be constructed.

Step 3: Putting similar capsules into the same group: The values in the one-dimensional vector, $\mathbf{L}_m \in \mathbb{R}^{1\times K}(m = 1, 2, ..., K)$, for the m-th row of the correlation matrix $\mathbf{L} \in \mathbb{R}^{K\times K}$ represent the correlation coefficients between the capsule of type m and those of other types. The group with the initial capsule $Caps_{o_i}(i = 1, 2)$ that a capsule $Caps_p$ belongs to can be determined by

$$Caps_p \in G_{Caps_{o_j}}, where\ o_j = \arg\min_{i=1,2}\left(\mathbf{L}_{p,o_i}\right), \tag{3}$$

where $\mathbf{L}_{p,o_i}(p = 1, 2, ..., 16, p \neq o_i, i = 1, 2)$ represents the correlation coefficient between the remaining 14 capsules and the 2 initial capsules. $\arg\min$ returns the index for the smaller one between $\mathbf{L}_{p,o_1}$ and $\mathbf{L}_{p,o_2}$. With this step, we may dynamically divide the capsules into two groups G_1 and G_2.

By performing the same steps mentioned above on G_1, we

[1]Each capsule contains a 4×4 pose matrix M and an activation value a.

may further obtain two new capsule groups. Similarly, we obtain another two new capsule groups by performing the same steps on G_2. Thus, we finally obtain four capsule groups, *i.e.*, Go_1, Go_2, Go_3, Go_4, with strong correlation within each group.

Capsules routing. There is a 4×4 trainable transformation matrix $\mathbf{W}_{ij}$ between each capsule $i (i \in \Omega_N)$ in layer N and each capsule $j (j \in \Omega_{N+1})$ in layer $N+1$. Ω_N denotes the set of capsules in layer N. The pose matrix $\mathbf{M}_i$ of capsule i is transformed by $\mathbf{W}_{ij}$ to cast a vote $\mathbf{V}_{ij} = \mathbf{M}_i \mathbf{W}_{ij}$ for the pose matrix $\mathbf{M}_j$ of capsule j. $\mathbf{V}_{ij}$ and a_i are utilized for routing to obtain the poses and activations of all capsules in the $N+1$ layer, which is achieved through an iterative Expectation-Maximization (EM) algorithm [18]. More details can be seen in [18].

In this way, the part-whole relationships within the image are obtained by assigning associated parts to their familiar wholes. Similar to [17], the activation values from the last convolutional capsule layer are used as the final feature maps $\mathbf{F}_{PO}$ for the next stage.

B. Extracting contrast information stream

1) Initial feature extraction for contrast cues:

In order to facilitate the extraction of contrast cues, as shown in Fig. 2, a set of initial features $\mathbf{F}_0$ are first extracted in the DenseHRNet branch via a Downsampling Node, which is constructed by two convolutional layers and four residual blocks.

2) BS module for highlighting the foreground regions:

Although local details are captured by contrast information, salient objects in cluttered or low-contrast scenes, *e.g.*, low-contrast between foreground and background, are still difficult to be segmented out from the background accurately just according these local details. Notably, the position of the salient object can be located through the part-whole relational cues. Considering that, a Background Suppression (BS) module is further appended on the Downsampling Node to guide the primitive contrast extraction, which aims to produce more fine details while effectively suppressing complex backgrounds and highlighting the salient object regions.

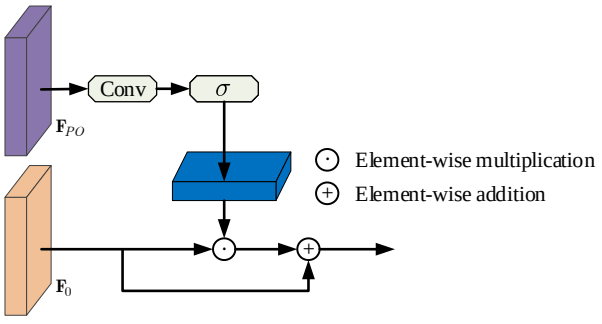


Fig. 4. The architecture of BS module.

Fig. 4 illustrates the details of the proposed BS module, in which the objectness prior maps learned by the DGC sub-network are utilized to generate channel-wise spatial attention. The entire process is formulated as follows:

$$\mathbf{F}_{bs} = \mathbf{F}_0 \odot \left[1 + \sigma\left(\mathrm{Conv}\left(\mathbf{F}_{PO}; \beta^1\right)\right)\right], \tag{4}$$

where $\mathbf{F}_{bs}$, $\mathbf{F}_0$ and $\mathbf{F}_{PO}$ represent the outputs of the BS module, the Downsampling Node and the DGC sub-network in Fig. 2, respectively. $\odot$ means the operation of the element-wise multiplication. $\mathrm{Conv}(*; \beta^1)$ denotes a convolutional block with its parameters β^1, which is responsible for transforming the channel number of $\mathbf{F}_{PO}$ into the same as that of $\mathbf{F}_0$. The value of the spatial weight map is activated by the Sigmoid operation, *i.e.*, $\sigma(*)$.

3) DenseHRNet for contrast information:

To make the potential spatial features more precise, we propose the DenseHRNet sub-network based on HRNet [19] to maintain high-resolution representations while ensuring the maximum information flow between the network output layer and the middle layers. As shown in the bottom branch of Fig. 2, dense residual connections are embedded to integrate the filtering results of different sub-layer convolution kernel operations in the proposed DenseHRNet sub-network. This embedding of such residual connections improves the flow of information and gradients throughout the network, which makes them easy to train.

Actually, the DenseHRNet sub-network is similar to HRNet [19]. While, the difference between them is whether the features of middle sub-layers are used. The small modification leads to substantially different behaviors between the two networks. As shown in Fig. 2, $\mathbf{F}_C$ and $\mathbf{F}_{u,v}\,(u, v = 1, 2, 3, 4)$ represent the final output of the network and the features of each layer. The output of the original HRNet can be written as:

$$\mathbf{F}_C = \mathrm{Cat}\left(\mathbf{F}_{4,v}\right), where\ v = 1, 2, 3, 4, \tag{5}$$

where Cat denotes concatenating feature maps along the channel dimension. Differently, the output of the DenseHRNet sub-network can be formulated as:

$$\mathbf{F}_C = \mathrm{Cat}\left(\mathbf{F}_{u,v}\right), where\ u, v = 1, 2, 3, 4. \tag{6}$$

Due to such dense residual connections, the final features not only integrate the features of different layers, but also aggregate all the features of the previous layers at different scales. The feature maps learned by any of the DenseHRNet layers can be accessed by the last layer. Besides, when the gradient is propagated back, partial information can directly reach each middle layer without going through the deep network. This forces the middle layer to learn more distinguishable features. Therefore, more accurate contrast information can be obtained by the proposed DenseHRNet sub-network.

C. Attention fusion mechanism for two cues integration

Considering different characteristics of the two cues, *i.e.*, contrast cues prefer to capture object details and part-whole relational cues prefer to detect the object wholeness, they can complement to each other to improve the saliency prediction.

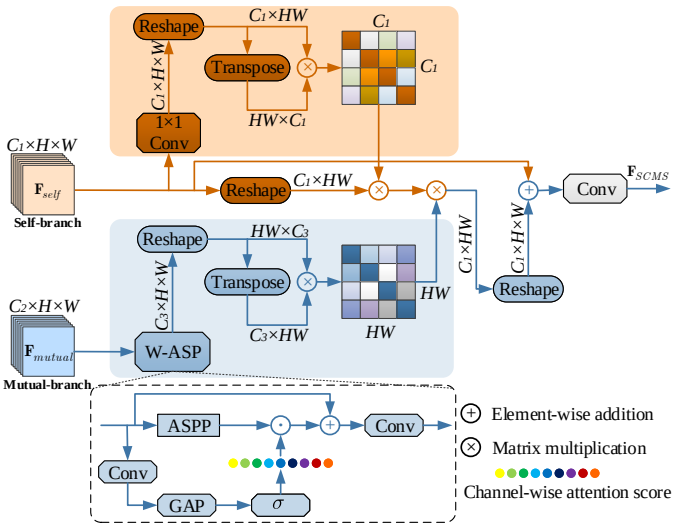


Fig. 5. The architecture of SCMS module. Shadow regions marked by brown and blue colors represent the SCC unit and the MWSA unit, respectively. 'W-ASP' refers to the Weighted Atrous Spatial Pyramid (W-ASP) sub-module.

While simple addition or cascading operation cannot fully extract enough useful information for the saliency map. Besides, the features of the same cue usually are affluent in spatial or channel aspect, and also include redundant information. To overcome such issues, a Self-Channel and Mutual-Spatial (SCMS) attention module is designed to automatically select those important features for the prediction of salient regions. The SCMS attention module consists of two units: a Self-branch Channel Correlation (SCC) unit and a Mutual-branch Weighted Spatial Attention (MWSA) unit. The structure of SCMS is shown in Fig. 5.

SCC. Different channels of features in CNNs generate various responses for different semantics and perform differently for highlighting the salient object [48]. This is significant to filter inaccurate features and focus more on valuable features. For that, we assign larger weights to those channels that show higher responses on salient regions by calculating the correlation matrix among channels. In this way, long-range dependencies along the channel dimension will be well exploited, thus capturing more comprehensive channel characteristics for the feature selection. This is different from the traditional channel-wise attention module, where the weight for each channel is calculated in a channel-independent way.

The orange regions shaded in Fig. 5 show the detailed structure of the proposed SCC. First, we apply 1×1 convolution and reshape operations to transform the self-branch input features $\mathbf{F}_{self} \in \mathbb{R}^{C_1 \times H \times W}$ to $\mathbf{W}_q \in \mathbb{R}^{C_1 \times HW}$. After that, a channel correlation matrix is generated by performing matrix multiplication and normalization operations on $\mathbf{W}_q$ and its transpose. Negative values in the correlation matrix are suppressed by ReLU activation function. Finally, the output features $\mathbf{F}_{SCC}$ of SCC are obtained by the matrix multiplication of the channel correlation matrix with the original self-branch input features. The entire process is written as:

$$\mathbf{W}_q = \mathrm{Nor}\left(\mathrm{Reshape}\left(\mathrm{Conv}\left(\mathbf{F}_{self}; \beta^2\right)\right)\right), \tag{7}$$

$$\mathbf{F}_{SCC} = \left(\mathrm{Nor}\left(\mathrm{ReLU}\left(\mathbf{W}_q \times \mathbf{W}_q{}^{\mathrm{T}}\right)\right)\right) \times \mathrm{Reshape}\left(\mathbf{F}_{self}\right), \tag{8}$$

where $\mathrm{Nor}(*)$ means normalizing the values in the channel correction matrix to $[0, 1]$. $\mathrm{Reshape}(*)$ means to transform $\mathbf{F}_{self}$ from the size $C_1 \times H \times W$ to $C_1 \times HW$.

MWSA. The two cues from the two-stream network contain different semantic information. The part-whole hierarchies are responsible for the whole saliency regions, while the contrast cues provide precise details. In order to effectively combine the semantic features from the above two cues, we design an MWSA unit to capture the long-range spatial dependencies across the two cues, as shown in the blue regions shaded in Fig. 5. Specifically, a spatial attention map is generated from MWSA by using some atrous convolutional pyramid operations to further provide spatial guidance for the output of SCC $\mathbf{F}_{SCC}$. More specifically, the input features $\mathbf{F}_{mutual} \in \mathbb{R}^{C_2 \times H \times W}$ of the mutual branch are first fed into a Weighted Atrous Spatial Pyramid (W-ASP) sub-module to extract their enhanced multi-scale contextual information $\mathbf{F}_{W-ASP} \in \mathbb{R}^{C_2 \times H \times W}$. Then, similar to that in SCC, a 1×1 convolution layer and a reshape operation are performed on $\mathbf{F}_{W-ASP}$, thus obtaining $\mathbf{W}_a \in \mathbb{R}^{HW \times C_3}$. After that, a spatial correlation matrix is generated by performing matrix multiplication and normalization operations on $\mathbf{W}_a$ and its transpose. The output features $\mathbf{F}_{MWSA}$ of MWSA are thus obtained by the matrix multiplication of the spatial correlation matrix with the output of SCC $\mathbf{F}_{SCC}$.

Especially, as shown in Fig. 5, an Atrous Spatial Pyramid Pooling (ASPP) operation with the same structure as in [49] but with different dilation rates (*i.e.*, 1, 3, 5 and 7) is first employed to capture some initial multi-scale contextual information $\mathbf{F}_{ASP} \in \mathbb{R}^{C_2 \times H \times W}$ from the input features $\mathbf{F}_{mutual}$ in the W-ASP sub-module. Then a 3×3 convolutional layer together with a global averaging pooling (GAP) layer is performed on the input features $\mathbf{F}_{mutual}$ to generate a set of channel-wise weights $\mathbf{F}_{weight} \in \mathbb{R}^{C_2}$. With the channel-wise weights $\mathbf{F}_{weight}$, enhanced multi-scale contextual information $\mathbf{F}_{E-ASP} \in \mathbb{R}^{C_2 \times H \times W}$ is obtained by performing a channel-wise multiplication operation on the extracted $\mathbf{F}_{ASP}$. By doing so, the useful multi-scale features in $\mathbf{F}_{ASP}$ will be enhanced while some disturbing information will be suppressed. The final output features $\mathbf{F}_{W-ASP}$ of W-ASP is obtained by further performing a convolution layer on the addition of $\mathbf{F}_{E-ASP}$ with the original input features $\mathbf{F}_{mutual}$. Mathematically, the whole process of the proposed MWSA unit can be expressed as follows:

$$\mathbf{F}_{ASP} = \mathrm{ASP}\left(\mathbf{F}_{mutual}\right), \tag{9}$$

$$\mathbf{F}_{weight} = \sigma\left(\mathrm{GAP}\left(\mathrm{Conv}\left(\mathbf{F}_{mutual}\right)\right); \beta^3\right), \tag{10}$$

$$\mathbf{F}_{E-ASP} = \mathbf{F}_{weight} \odot \mathbf{F}_{ASP}, \tag{11}$$

$$\mathbf{F}_{W-ASP} = \mathrm{Conv}\left(\mathbf{F}_{E-ASP} + \mathbf{F}_{mutual}; \beta^4\right), \tag{12}$$

$$\mathbf{W}_a = \mathrm{Nor}\left(\mathrm{Reshape}\left(\mathrm{Conv}\left(\mathbf{F}_a; \beta^5\right)\right)\right) \in \mathbb{R}^{HW \times C_3}, \tag{13}$$

$$\mathbf{F}_{MWSA} = \mathbf{F}_{SCC} \times \left(\mathrm{Nor}\left(\mathrm{ReLU}\left(\mathbf{W}_a \times \mathbf{W}_a{}^{\mathrm{T}}\right)\right)\right), \quad (14)$$

where GAP refers to the global average pooling operation. ASP is the operation of stacked dilation convolutions with different dilation rates of 1, 3, 5, and 7. Finally, we add $\mathbf{F}_{MWSA}$ and $\mathbf{F}_{self}$ to obtain the final output features $\mathbf{F}_{SCMS}$ of the proposed SCMS module so that the original self-branch input features are retained, which can be written as:

$$\mathbf{F}_{SCMS} = \mathrm{Conv}\left(\mathrm{Reshape}^{'}\left(\mathbf{F}_{MWSA}\right) + \mathbf{F}_{self}\right), \quad (15)$$

where $\mathrm{Reshape}^{'}$ denotes the inverse process of Reshape.

As shown in Fig. 2, two SCMS modules are applied to integrate the features of two cues. When $\mathbf{F}_{PO}$ is the self-branch features and $\mathbf{F}_{C}$ is the Mutual-branch features (*i.e.*, $\mathbf{F}_{self}$, $\mathbf{F}_{mutual}$ and $\mathbf{F}_{SCMS}$ are $\mathbf{F}_{PO}$, $\mathbf{F}_{C}$ and $\mathbf{F}^{PO}_{SCMS}$, respectively), the local details of the part-whole hierarchies are enhanced based on the contrast cues. Similarly, when $\mathbf{F}_{C}$ is the self-branch features and $\mathbf{F}_{PO}$ is the Mutual-branch features (*i.e.*, $\mathbf{F}_{self}$, $\mathbf{F}_{mutual}$ and $\mathbf{F}_{SCMS}$ are $\mathbf{F}_{C}$, $\mathbf{F}_{PO}$ and $\mathbf{F}^{C}_{SCMS}$, respectively), the object wholeness of the contrast cues are enhanced based on the part-whole hierarchies. Finally, the final output features $\mathbf{F}_{out}$ from the two SCMS modules are obtained by concatenating $\mathbf{F}^{PO}_{SCMS}$ and $\mathbf{F}^{C}_{SCMS}$, *i.e.*,

$$\mathbf{F}_{out} = \mathrm{Cat}\left(\mathbf{F}^{PO}_{SCMS}, \mathbf{F}^{C}_{SCMS}\right). \quad (16)$$

Different from previous attention mechanism algorithms. Here we mainly discuss the uniqueness of the proposed SCMS module compared to the attention mechanisms in [50] and [51].

1) Comparison with non-local operation in [50]. Non-local operations in [50] can calculate the dependencies among all spatial positions, but the correlation among different channels is not considered. Differently, we focus on spatial attention while considering channel correlation, which can highlight regions and channels that are critical to the saliency map. Besides, the spatial correlation obtained by the proposed MWSA is more accurate than that obtained in [50] because of the introduction of the W-ASP structure, which can better suppress confusing information while maintaining multi-scale contextual information than the traditional ASPP module.

2) Comparison with DANet in [51]. The similarity between our SCMS module and DANet in [51] lies in the simultaneous application of channel and spatial attention. While, the differences between them mainly lie in the following two folds. First, our SCMS module embeds the W-ASP structure in MWSA to capture multi-scale contextual information. Secondly, we use the spatial weights generated by the two cues to interactively guide feature extraction for better mining the complementary advantages of the two cues.

D. Saliency inference

In addition to the output features $\mathbf{F}_{out}$ from the two SCMS modules mentioned above, the shallow features $\mathbf{F}_{L1}$ and $\mathbf{F}_{L2}$ from the U-Res34 unit are also exploited via a Upsampling Node to assist the prediction of final saliency maps for accurate boundaries. As shown in Fig. 2, the Upsampling Node is constructed by stacking upsampling and concatenation operations, and the process can be mathematically expressed by

$$\mathbf{F}_{mid} = \mathrm{Conv}\left(\mathrm{Cat}\left(\mathrm{Up}\left(\mathbf{F}_{out}\right), \mathbf{F}_{L2}\right); \beta^{6}\right), \quad (17)$$

$$\mathbf{P} = \mathrm{Conv}\left(\mathrm{Cat}\left(\mathrm{Up}\left(\mathbf{F}_{mid}\right), \mathbf{F}_{L1}\right); \beta^{7}\right), \quad (18)$$

where $\mathbf{P}$ refers to the final saliency map. Up means upsampling operation by bilinear interpolation.

E. Loss function

For training the network, the cross-entropy loss function in [52] and the IoU boundary loss function in [53] are used to train the saliency prediction. The cross-entropy loss function is defined as:

$$\mathcal{L}_{CE} = -\frac{1}{H \times W} \sum_{m=1}^{H} \sum_{n=1}^{W} [\mathbf{G}(m,n) \log \mathbf{P}(m,n) + (1 - \mathbf{G}(m,n)) \log(1 - \mathbf{P}(m,n))], \quad (19)$$

where $\mathbf{G}(m,n) \in \{0,1\}$ is the ground truth label for the pixel (m,n). $\mathbf{P}(m,n)$ is the predicted probability of being salient object for the pixel (m,n). W and H represent the width and height of the input image, respectively.

IoU is originally proposed for measuring the similarity of two sets [54] and has been used for saliency detection in [53]. It can be defined as:

$$\mathcal{L}_{iou} = 1 - \frac{\sum_{m=1}^{H} \sum_{n=1}^{W} \mathbf{P}(m,n)\, \mathbf{G}(m,n)}{\sum_{m=1}^{H} \sum_{n=1}^{W} [\mathbf{P}(m,n) + \mathbf{G}(m,n) - \mathbf{P}(m,n)\, \mathbf{G}(m,n)]}, \quad (20)$$

The final joint loss function that is used to train our proposed model is constructed by combining the cross-entropy loss function and the IoU Boundary loss function, *i.e.*,

$$\mathcal{L}_{joint} = \mathcal{L}_{CE} + \mathcal{L}_{iou}. \quad (21)$$

IV. Experiments

A. Datasets

We comprehensively evaluate our model on five benchmarks: DUTS [55], HKU-IS [24], ECSSD [56], DUT-OMRON [57] and PASCAL-S [58]. The DUTS is a challenging dataset, which consists of 10,553 training images and 5,019 testing images in complicated scenes. ECSSD contains 1000 images of high content varieties. HKU-IS consists of 4447 images with multiple disconnected objects. The images in this dataset have diverse spatial distributions, and the similar appearances between the foreground regions and the background regions make it more difficult to distinguish the salient objects. DUT-OMRON is composed of 5168 images with different sizes

TABLE I
COMPARISONS OF THE PROPOSED METHOD AND OTHER 13 METHODS ON FIVE BENCHMARK DATASETS IN TERMS OF MAXIMUM AND MEAN F-MEASURE (LARGER IS BETTER), E-MEASURE (LARGER IS BETTER), S-MEASURE (LARGER IS BETTER) AND MAE (SMALLER IS BETTER). THE BEST THREE RESULTS ARE HIGHLIGHTED IN RED, GREEN AND BLUE, RESPECTIVELY."-R" MEANS THE RESULTS ARE ACHIEVED WITH THE RESNET-50/101 BACKBONE ON THIS METHOD.

Model	DUT-OMRON					DUTS-TE					HKU-IS					ECSSD					Pascal-S				
	Fmax	*Favg*	*Em*	*Sm*	*MAE*	*Fmax*	*Favg*	*Em*	*Sm*	*MAE*	*Fmax*	*Favg*	*Em*	*Sm*	*MAE*	*Fmax*	*Favg*	*Em*	*Sm*	*MAE*	*Fmax*	*Favg*	*Em*	*Sm*	*MAE*
Ours	**0.802**	**0.786**	**0.876**	**0.850**	**0.055**	**0.884**	**0.864**	**0.930**	**0.898**	**0.035**	**0.937**	**0.918**	**0.957**	**0.929**	**0.026**	**0.945**	**0.928**	**0.953**	**0.932**	**0.031**	**0.859**	**0.838**	**0.900**	**0.866**	**0.062**
F3Net [16]	0.778	0.766	**0.864**	0.838	**0.053**	**0.872**	**0.852**	**0.920**	**0.888**	**0.035**	0.925	**0.910**	**0.952**	0.917	**0.028**	0.935	**0.925**	**0.948**	0.924	**0.033**	0.848	**0.835**	**0.898**	0.861	**0.061**
ITSD [65]	**0.792**	**0.768**	**0.865**	**0.840**	0.061	0.868	0.840	0.914	0.885	0.041	**0.926**	0.903	**0.947**	0.917	0.031	**0.939**	0.921	0.947	**0.925**	**0.034**	**0.855**	0.831	0.895	0.859	0.066
MINet-R [14]	0.769	0.757	0.860	0.833	**0.056**	0.865	**0.844**	**0.917**	0.884	**0.037**	**0.926**	**0.909**	**0.952**	**0.919**	**0.029**	**0.938**	**0.923**	**0.950**	**0.925**	**0.033**	0.846	0.830	**0.896**	0.856	0.064
GCPANet [66]	0.775	0.756	0.853	0.839	**0.056**	**0.869**	0.841	0.912	**0.891**	**0.039**	**0.927**	0.901	0.945	**0.920**	0.031	0.936	0.916	0.944	**0.927**	0.035	0.849	0.829	0.895	**0.864**	**0.062**
EGNet-R [13]	0.778	0.760	0.857	**0.841**	**0.053**	0.866	0.839	0.907	0.887	**0.039**	0.924	0.902	0.944	0.918	0.031	0.936	0.918	0.943	**0.925**	0.037	0.841	0.823	0.881	0.852	0.074
SCRN [67]	0.772	0.749	0.848	0.837	**0.056**	0.864	0.833	0.900	0.885	0.040	0.921	0.894	0.935	0.916	0.034	0.937	0.916	0.939	**0.927**	0.037	**0.856**	**0.833**	0.892	**0.869**	**0.063**
CPD-R [15]	0.754	0.742	0.847	0.825	**0.056**	0.840	0.821	0.898	0.869	0.043	0.911	0.892	0.938	0.905	0.034	0.931	0.913	0.942	0.918	0.037	0.833	0.819	0.882	0.848	0.071
AFNet [68]	0.759	0.742	0.846	0.826	0.057	0.839	0.812	0.893	0.867	0.046	0.910	0.888	0.934	0.905	0.036	0.924	0.905	0.935	0.913	0.042	0.844	0.824	0.883	0.849	0.070
BASNet [28]	**0.779**	**0.767**	**0.865**	0.836	**0.056**	0.838	0.823	0.895	0.866	0.048	0.919	0.902	0.943	0.909	0.032	0.931	0.917	0.943	0.916	0.037	0.835	0.818	0.879	0.838	0.076
MLMSNet [69]	0.734	0.710	0.831	0.809	0.064	0.828	0.792	0.883	0.862	0.049	0.910	0.878	0.930	0.907	0.039	0.917	0.890	0.927	0.911	0.045	0.835	0.807	0.876	0.844	0.074
TSPOANet [17]	0.750	0.728	0.840	0.818	0.061	0.828	0.800	0.885	0.860	0.049	0.909	0.884	0.932	0.902	0.038	0.919	0.899	0.928	0.907	0.046	0.830	0.809	0.872	0.842	0.077
PAGE [70]	0.758	0.743	0.849	0.824	0.062	0.815	0.793	0.883	0.854	0.052	0.907	0.884	0.935	0.903	0.037	0.924	0.904	0.936	0.912	0.042	0.830	0.811	0.878	0.842	0.076
JointCRF [71]	0.755	0.737	0.838	0.821	0.057	0.793	0.764	0.854	0.836	0.059	0.905	0.879	0.925	0.903	0.039	0.914	0.888	0.921	0.907	0.049	0.827	0.792	0.852	0.841	0.082

and complex structures. PASCAL-S includes 850 challenging images.

B. Evaluation criteria

We use five metrics to evaluate the proposed method, *i.e.*, Precision-Recall (PR) curve, F-measure [59], E-measure [60], S-measure [61] and Mean Absolute Error (MAE) [62].

PR curves. Precision and recall values are computed by comparing the binary saliency map with the ground truth to plot the PR curve with different thresholds in the range of $[0, 255]$. Specifically, $Precision = TP/(TP + FP)$ and $Recall = TP/(TP+FN)$, where TP, FP and FN represent true-positive, false-positive and false-negative, respectively. The larger the area under the PR curve, the better the performance is.

F-measure. F_β is formulated as the weighted harmonic mean of precision and recall, *i.e.*,

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot \mathrm{Pr}\, ecision \times \mathrm{Recall}}{\beta^2 \cdot \mathrm{Pr}\, ecision + \mathrm{Recall}}, \tag{22}$$

where β^2 is set to 0.3 to emphasize the precision over recall as suggested in [59]. Here, we report the maximum F-measure (F_{max}) computed from all precision-recall pairs and use an adaptive threshold that is twice the mean value of the prediction to calculate the mean F-measure (F_{avg}).

E-measure. E_m combines local pixel values with the image-level mean value to jointly evaluate the similarity between the prediction and the ground truth.

S-measure. S_m computes the object-aware and region-aware structure similarities between the prediction and the ground truth, which can be written as:

$$S_m = \alpha \cdot S_o + (1-\alpha) \cdot S_r, \tag{23}$$

where α is set to 0.5. S_o and S_r represent the prediction and the ground truth, respectively.

MAE. MAE is defined as the average pixel-wise absolute difference between the normalized prediction and the ground truth:

$$MAE = \frac{1}{H \times W} \sum_{m=1}^{H} \sum_{n=1}^{W} |\mathbf{P}\left(m,n\right) - \mathbf{G}\left(m,n\right)|, \tag{24}$$

where $\mathbf{P}$ and $\mathbf{G}$ represent the saliency maps and the ground truth, respectively.

C. Implementation details

We implement our model on Pytorch 1.0.0. An NVIDIA GTX 1080 Ti GPU (with 11GB memory) is used for both training and testing. The DUTS training dataset containing 10553 images is used to train the network. Before training, the dataset is augmented by horizontal flipping to avoid the over-fitting problem. During the training stage, each image is first resized to 256×256 and randomly cropped to 224×224. The U-Res34 is initialized from the ResNet-34 model [63]. The DenseHRNet sub-network parameters are initialized by the weights pretrained on the ImageNet. Other convolutional layers are initialized by Xavier [64]. The stochastic gradient descent (SGD) model is adopted to train our model, where the initial learning rate, momentum and weight_decay are set to 1e-3, 0.9 and 0.0005, respectively. We adopt the exponential decay strategy with base 0.95 to gradually decrease the learning rate. Our network is trained with a mini-batch of 4. The whole training process takes about 65 hours. *The code and results will be released.*

D. Comparison with state-of-the-arts

We compare the proposed algorithm with 13 state-of-the-art salient object detection methods, including F3Net [16], ITSD [65], MINet [14], GCPANet [66], EGNet [13], SCRN [67], CPD [15], AFNet [68], BASNet [28], MLMSNet [69], TSPOANet [17], PAGE [70] and JointCRF [71]. For fair comparisons, all the saliency maps of the above methods are

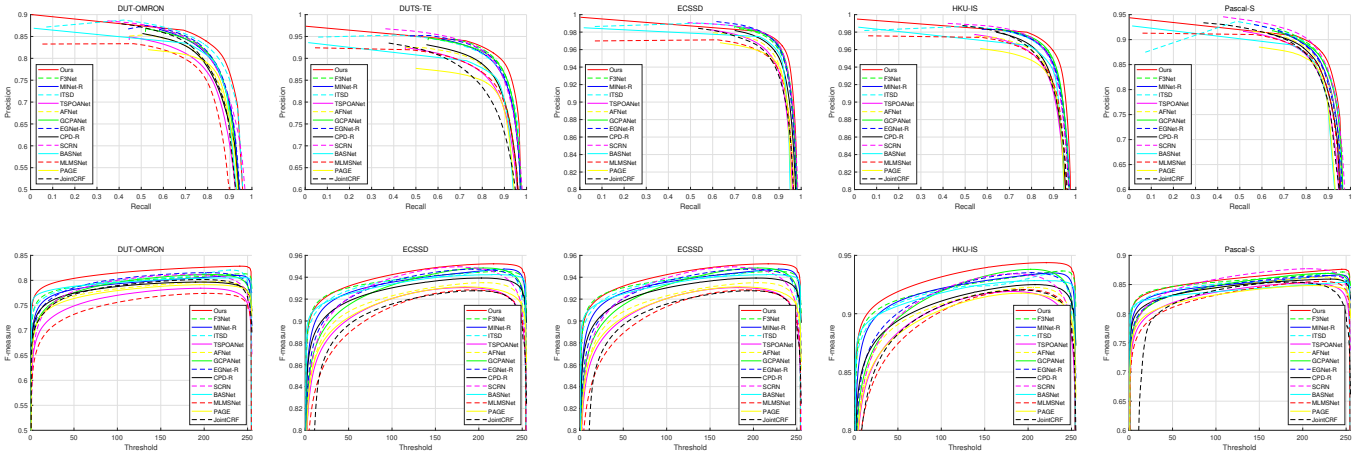


Fig. 6. PR curves (1^{st} row) and F-measure curves (2^{nd} row) on the five saliency datasets.

generated by running their source codes or pre-computed by their authors.

Quantitative Comparison. To fully compare the proposed method with state-of-the-art approaches, we report the detailed experimental results in terms of the five metrics, which are listed in Table I. As can be seen clearly, the proposed algorithm consistently performs better than the competitors across all of the five metrics on most datasets. In particular, in terms of F_{avg} and S_m, the performance is improved by more than 1% on the three most challenging data datasets (*i.e.*, DUT-OMRON, DUTS and HKU-IS). This indicates our model achieves good structural similarities with the ground truth.

In addition, we display PR curves and F-measure curves in Fig. 6. In terms of both PR curves and F-measure curves, our approach (red solid line in Fig. 6) keeps the best results on DUT-OMRON, DUTS-TE, HKU-IS and ECSSD, and is also competitive with others on PASCAL-S.

Furthermore, we compare the floating point operations (*i.e.*, FLOPs), the number of parameters (*i.e.*, Params) and the inference time (*i.e.*, Time) with other popular methods in Table II. Input sizes of different methods are set according to their released codes. The comparisons in Table II show that our model is slightly more complicated than other methods, which may be owe to the complex capsule routing algorithm in DGC sub-network.

Qualitative Evaluation. To further illustrate the superior performance of our method, Fig. 7 shows the visual comparisons of our model and other methods by displaying some images covering different scenarios, including low contrast, similar backgrounds, small objects and multiple objects. It can be easily seen that our proposed method can highlight the whole salient objects with satisfactory uniformity. In contrast, the methods using contrast cues (*i.e.*, (e)-(l) in Fig. 7) just detect parts of the salient objects and fail to capture the whole objects in low contrast scenes or similar backgrounds (as shown in the first six rows of Fig. 7). Furthermore, the objects and the backgrounds cannot be well distinguished by these methods, resulting in poor saliency maps with background noise interference in complex scenes (as illustrated in the 6^{th}, 7^{th} and 8^{th} rows of Fig. 7). Besides, for those scenes with multiple objects, the compared methods miss some salient object parts, while our approach can locate all the salient objects and predict complete object shapes. This results from the fact that these methods ignore the correlation among different object parts. Fortunately, our method can effectively suppress background noise while detecting the whole salient objects in various scenes. This owes to the fact that the part-whole hierarchies are added in our proposed model to infer the saliency maps.

TABLE II
THE NUMBER OF PARAMETERS, FLOPS AND INFERENCE TIME COMPARISONS OF OUR METHOD WITH SOME STATE-OF-THE-ART NETWORKS.

Method	Input size	FLOPs (G)	Params (M)	Time (s)
F3Net [16]	352 × 352	16.43	25.54	0.022
ITSD [65]	288 × 288	15.94	26.07	0.022
GPACNet [66]	320 × 320	54.31	67.06	0.020
BASNet [28]	256 × 256	127.40	87.06	0.032
MINet-R [14]	320 × 320	87.03	162.38	0.036
EGNet-R [13]	~380 × 320	287.67	111.66	0.091
Ours	256 × 256	137.64	153.26	0.167

In addition, although TSPOANet can also obtain the whole salient objects for some scenes, the problem of blurred edges is not well solved (as illustrated in the 1^{st}, 2^{nd}, 10^{th} and 11^{th} rows of Fig. 7(d)). Differently, more accurate prediction maps can be obtained by adding contrast cues in our method. As well, in the scenes with similar backgrounds or low contrast (*e.g.*, the 3^{rd}, 4^{th} and 5^{th} rows in Fig. 7), TSPOANet cannot predict the complete salient objects. But our method shows perfect performance. This may owe to the proposed dynamic grouping strategy for capsules routing in our proposed model, which can better reduce the noise distribution of capsules than the fixed grouping strategy in TSPOANet. As a result, the proposed method can consistently produce more accurate and complete saliency maps with sharp boundaries and coherent

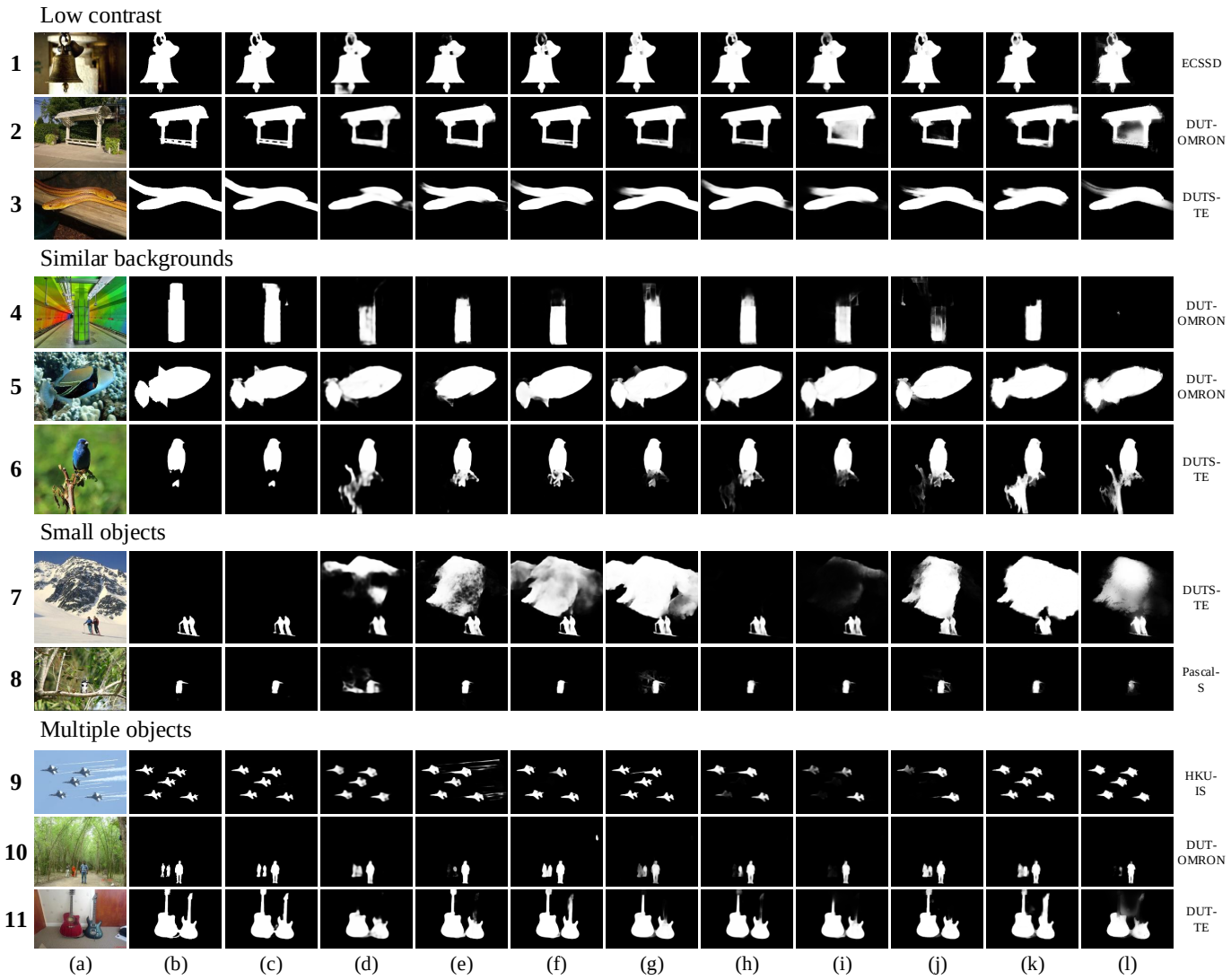


Fig. 7. Visual comparisons of different methods. (a) Image; (b) GT; (c) Ours; (d) TSPOANet [17]; (e) MINet [14]; (f) F3Net [16]; (g) EGNet [13]; (h) GCPANet [66]; (i) SCRN [67]; (j) AFNet [68]; (k) PAGE [70]; (l) JointCRF [71].The right side indicates the name of the dataset for each image.

details in these challenging scenes than TSPOANet, as shown in Fig. 7.

E. Ablation study

In this section, we carry out a series of experiments to validate the effectiveness of each key component used in our network. The ablation study contains two parts: different components and different capsule grouping strategies. The ablation experiments are conducted on the challenging DUT-OMRON dataset and DUTS-TE dataset.

Different components. To prove the effectiveness of each component in our model, we report the quantitative comparison results in Table III. Here, "B" denotes the common basic model (ResNet-50). "H" and "H^+" represent the original HRNet [19] and the improved HRNet (*i.e.*, DenseHRNet), respectively. "PO" and "PO^+" mean fixed grouping and dynamic grouping strategies adopted in the capsule network, respectively. "H^+ + PO^+" means that the DGC output $\mathbf{F}_{PO}$ and the DenseHRNet output $\mathbf{F}_C$ are integrated by the addition operation (Here, the BS module is not used in this structure). "BS" denotes that the background suppression module is inserted into DenseHRNet. "S-C" denotes the SCMS module in Fig. 5.

As shown in Table III, by comparing the 1^{st} and 2^{nd} rows, we can see that F-measure increases by more than 1% if "H", instead of "B", is used as the baseline. This proves that maintaining high-resolution representations through the whole process can improve the detection performance. By embedding residual connections in HRNet [19], DenseHRNet (*i.e.*, "H^+") has further improved the performance while hardly increasing FLOPs and the number of parameters, which can be illustrated by observing the "H" and "H^+" rows in Table III. Similarly, the comparison of "PO" and "PO^+" indicates that the proposed dynamic grouping capsules strategy can improve performance without increasing FLOPs and the number of parameters. Besides, it can be observed from the comparison between "H + PO" and "H" or "PO" in Table III that the idea of integrating the above two cues is feasible, which

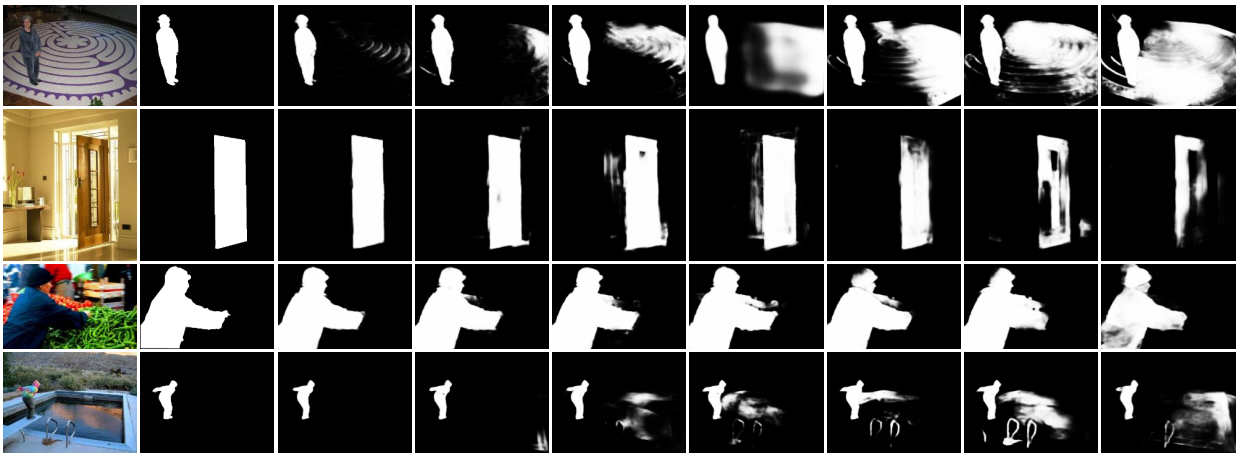


Fig. 8. Visual comparisons with different components. (a) Image; (b) GT; (c) "H^+" + "PO^+" + "BS" + "S-C"; (d) "H^+" + "PO^+" + "BS"; (e) "H^+" + "PO^+"; (f) "PO^+"; (g) "H^+"; (h) "H"; (i) "B".

TABLE III
ABLATION STUDIES OF DIFFERENT COMPONENTS. THE BEST PERFORMANCE IS MARKED BY **BOLD**. "B" REPRESENTS THE COMMON BACKBONE (RESNET-50). "H" AND "H^+" REPRESENT THE HRNET [19] AND THE DENSEHRNET, RESPECTIVELY. "PO" AND "PO^+" MEAN FIXED GROUPING AND DYNAMIC GROUPING STRATEGIES ADOPTED IN THE CAPSULE NETWORK, RESPECTIVELY. "BS" AND "S-C" DENOTE THE BS AND SCMS MODULES, RESPECTIVELY.

Configurations	DUT-OMRON			DUTS-TE			*FLOPs* (G)	*Params* (M)
	F_{max}	F_{avg}	*MAE*	F_{max}	F_{avg}	*MAE*		
B	0.754	0.740	0.057	0.833	0.810	0.040	7.86	33.61
H	0.768	0.753	0.061	0.848	0.828	0.042	26.98	66.32
PO	0.758	0.736	0.065	0.840	0.814	0.046	111.92	86.41
H + PO	0.772	0.761	0.062	0.852	0.836	0.041	137.04	152.68
H^+	0.786	0.762	0.059	0.866	0.838	0.039	27.10	66.77
PO^+	0.758	0.742	0.064	0.848	0.826	0.044	111.92	86.41
H^+ + PO^+	0.792	0.772	0.057	0.870	0.848	0.037	137.14	153.13
H^+ + PO^+ + BS	0.799	0.778	0.056	0.878	0.854	0.036	137.33	153.19
H^+ + PO^+ + BS + S-C	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**	137.64	153.26

can significantly improve the saliency detection performance. Meanwhile, the proposed "$H^+ + PO^+$" achieves consistently higher performance than "$H + PO$" does by integrating "H^+" and "PO^+". On top of "$H^+ + PO^+$", we progressively extend it with different units, including background suppression (*i.e.*, "BS") and SCMS (*i.e.*, "S-C") modules. The results in the last two rows of Table III illustrate the effectiveness of each unit. As can be seen, our PWHCNet architecture achieves the best performance among these configurations. In addition, it can be seen from the columns $FLOPs$ and $Params$ in Table III that a large number of parameters are mainly caused by the DGC sub-network, which covers complex capsule routing. Reducing the complexity of the capsule network to implement an efficient architecture is what we need to optimize further.

Visual comparisons can be found in Fig. 8. As shown in Fig. 8(g-i), the proposed DenseHRNet sub-network can better capture the salient object regions than the traditional basic model and the original HRNet [19] do. Moreover, the whole saliency maps can be well obtained by further combining the part-whole hierarchies with DenseHRNet, as can be shown in Fig. 8(e) and (f). By comparing (d) and (e) in Fig. 8, it can be easily observed that the background noise is suppressed by virtue of the BS module. Besides, it can be also noticed from Fig. 8(c) that the two salience cues can be well integrated by the proposed SCMS module.

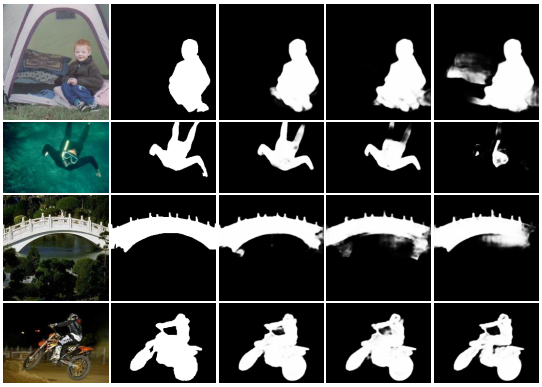
Capsule grouping strategies. To prove the effectiveness of the proposed dynamic grouping algorithm for capsules routing, we report the quantitative comparison results in Table IV. Here, "O" and "T" represent the original CapsNet [18] (*i.e.*, no grouping for capsules routing) and the improved two-stream CapsNet (*i.e.*, directly dividing capsules into two groups without distinction for capsules routing) in [17], respectively. "D_γ" ($\gamma = 2, 4, 8$) denotes that capsules are dynamically divided into γ groups according to the proposed dynamic grouping method.

TABLE IV
ABLATION STUDIES OF DIFFERENT CAPSULE GROUPING STRATEGIES. THE BEST PERFORMANCE IS MARKED BY **BOLD**. "O" DENOTES NO GROUPING STRATEGY. "T" AND "D_γ" ($\gamma = 2, 4, 8$) REPRESENT FIXED GROUPING STRATEGY AND DYNAMIC GROUPING STRATEGIES WITH DIFFERENT GROUP NUMBERS, RESPECTIVELY.

Configurations	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	*MAE*	F_{max}	F_{avg}	*MAE*
H^+ + O	0.769	0.753	0.060	0.847	0.830	0.040
H^+ + T	0.782	0.763	0.058	0.861	0.838	0.039
H^+ + D_2	0.791	0.767	0.058	0.868	0.841	0.039
H^+ + D_4	**0.792**	**0.772**	**0.057**	**0.870**	**0.848**	**0.037**
H^+ + D_8	0.790	0.769	**0.057**	0.867	0.845	0.038

In Table IV, the 1^{st} and 2^{nd} rows show the performance using the fixed grouping strategy (*i.e.*, H^+ + T) and using the dynamic grouping strategy (*i.e.*, "$H^+ + D_2$"). Numerically, the dynamic grouping strategy is effective and further alleviates the noise distribution phenomenon. In addition, we find that the number of groups also has an impact on the performance in the experiment. As shown in the last three rows of Table IV,

dividing capsules into 4 groups (*i.e.*, "$H^+ + D_4$") achieves the best performance. The reason for the performance degradation by dynamically dividing capsules into 8 groups (*i.e.*, "$H^+ + D_8$") may be that a little fewer capsules in each group are not enough to characterize the part-whole hierarchies.



(a) (b) (c) (d) (e)

Fig. 9. Visual comparisons with different capsule grouping strategies. (a) Image; (b) GT; (c) $H^+ + D_4$; (d) $H^+ + T$; (e) $H^+ + O$.

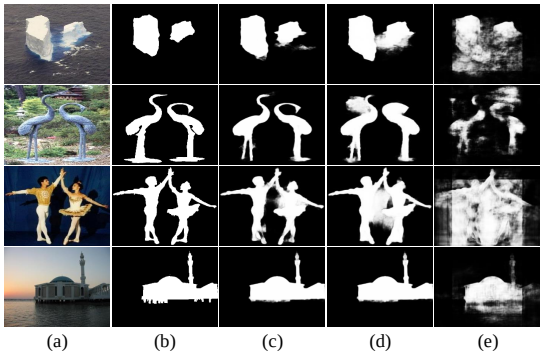
The visualization in Fig. 9 also illustrates the above quantitative results. Allowing each low-level capsule (part) to vote for all the high-level ones (object) will sometimes generate noisy assignment, thus giving rise to performance declines. By comparing (d) and (e) in Fig. 9, the grouping strategy in [17] does predict a better saliency map compared to the original capsule in [18]. Moreover, as seen from Fig. 9(c) and Fig. 9(d), it is obvious that the dynamic grouping strategy can produce better saliency maps by further alleviating the noise distribution phenomenon.

Feature extraction architectures for DGC sub-network. As discussed in TSPOANet [17], the feature extraction stage before capsules routing is critical to explore the part-whole relationships. To demonstrate the validity of U-Res34, we replace U-Res34 in our proposed DGC sub-network with FLNet in [17] or the two Conv+ReLU layers in the original CapsNet [18]. It can be easily observed from Table V that U-Res34 boosts the saliency detection performance of our proposed model significantly. As shown in Fig. 10(c-e), it is obvious that U-Res34 makes the framework possess the ability of identifying the salient object wholly, which is attributed to the rich features learned by U-Res34.

TABLE V
ABLATION STUDIES OF DIFFERENT FEATURE EXTRACTION ARCHITECTURES FOR DGC SUB-NETWORK. THE BEST PERFORMANCE IS MARKED BY **BOLD**. HERE, THE CAPSULES ARE DYNAMICALLY DIVIDED INTO FOUR GROUPS.

Feature Extraction Architectures	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
Two Conv+ReLU layers	0.506	0.452	0.195	0.552	0.482	0.182
FLNet	0.712	0.695	0.071	0.797	0.769	0.055
U-Res34	**0.758**	**0.742**	**0.064**	**0.848**	**0.826**	**0.044**

Different integration strategies. To demonstrate the advantages of the proposed integration strategy (*i.e.*, SCMS module) over Non-local [50] and DA [51] modules, we report the quantitative comparison results in Table VI. As shown in Table VI, it can be seen that the proposed SCMS module can obtain the competitive performance compared with non-local [50] and DA module [51]. Meanwhile, from the last three rows of Table VI, it can be seen that the performance obtained by only using SCC or MWSA is inferior to that obtained by using SCMS. This demonstrates that simultaneously considering the intra-cues channel interaction and the inter-cues spatial interaction indeed helps to improve performance.



(a) (b) (c) (d) (e)

Fig. 10. Visual comparisons with different feature extraction architectures for DGC sub-network. (a) Image; (b) GT; (c) U-Res34; (d) FLNet in [17]; (e) Two Conv+ReLU layers.

TABLE VI
ABLATION STUDIES OF DIFFERENT INTEGRATION STRATEGIES. THE BEST PERFORMANCE IS MARKED BY **BOLD**.

Integration Strategies	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
Baseline (H^+ + PO^+)	0.792	0.772	0.057	0.870	0.848	0.037
+ Non-local [50]	0.799	0.772	0.060	0.880	0.850	0.037
+ DA module [51]	0.800	0.781	0.056	0.881	0.858	**0.035**
+ SCC	0.789	0.772	**0.055**	0.876	0.853	0.036
+ MWSA	0.796	0.782	0.056	0.880	0.859	**0.035**
+ SCMS	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**

F. Failure cases

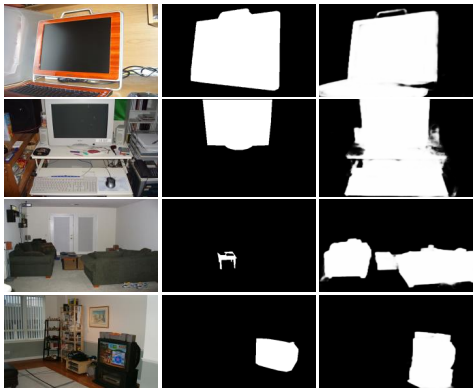


Image GT Ours

Fig. 11. Some failure cases for our proposed method.

Fig. 11 shows some failure cases for our proposed method. The scenes in those images contain some unique scenes. It can be seen that, under the effect of part-whole hierarchies, some objects with certain relations are detected together, *e.g.*, computer and keyboard, table and sofa, television and television cabinet, etc., instead of one individual object as masked by the ground truth. We will study this issue as the future work, which may be solved using scene parsing [72].

V. Conclusion

In this paper, we have proposed a PWHCNet for salient object detection by interacting part-whole hierarchies and contrast cues, which consists of two branches, including a part-whole relationships exploration branch and a contrast cues extraction branch. Specifically, the former exploits the dynamic grouping strategy to obtain more accurate part-whole relationships while the latter captures multi-scale contrast information through the DenseHRNet. In addition, the above two cues are interacted and integrated by the proposed BS and SCMS modules to retain useful features for the final saliency map. Extensive experiments validate that our proposed algorithm can well detect the whole salient objects together with their accurate boundaries even in the cluttered scenes. Moreover, our model outperforms some current state-of-the-art methods on five datasets.

It should be also noted that high saliency detection results obtained by our proposed model are at the cost of complex architectures, which limits its applications in some other vision tasks. In the future, we will further reduce the complexity of the capsule network to achieve a smaller architecture for SOD tasks while maintaining the saliency detection accuracy.

Acknowledgment

This work is supported by the National Natural Science Foundation of China under Grant No. 61773301 and 62001341.

References

[1] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Dense and sparse labeling with multidimensional features for saliency detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 5, pp. 1130–1143, 2016.

[2] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 12, pp. 2009–2021, 2013.

[3] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 582–593, 2021.

[4] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 11, pp. 2314–2320, 2016.

[5] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 541–554, 2012.

[6] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 5, pp. 769–779, 2013.

[7] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 141–145, 2005.

[8] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing*, vol. 19, no. 1, pp. 185–198, 2009.

[9] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[10] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 818–832, 2016.

[11] C. Deng, X. Yang, F. Nie, and D. Tao, "Saliency detection via a multiple self-weighted graph-based manifold ranking," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 885–896, 2019.

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

[13] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 8779–8788.

[14] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9413–9422.

[15] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3907–3916.

[16] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.

[17] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 1232–1241.

[18] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *International Conference on Learning Representations*, 2018, pp. 3856–3866.

[19] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[20] L. Yi, Z. Qiang, H. Jungong, and W. Long, "Salient object detection employing robust sparse representation and local consistency," *Image and Vision Computing*, vol. 69, pp. 155–167, 2018.

[21] H. Lu, X. Li, L. Zhang, X. Ruan, and M.-H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1592–1603, 2016.

[22] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform and local spatial support," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 9–23, 2015.

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

[24] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep cnn features," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5012–5024, 2016.

[25] Wei, Jun and Wang, Shuhui and Wu, Zhe and Su, Chi and Huang, Qingming and Tian, Qi, "Label decoupling framework for salient object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 022–13 031.

[26] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 212–221.

[27] Cong, Runmin and Lei, Jianjun and Fu, Huazhu and Hou, Junhui and Huang, Qingming and Kwong, Sam, "Going from RGB to RGBD saliency: a depth-guided transformation model," *IEEE Transactions on Cybernetics*, vol. 50, no. 8, pp. 3627–3639, 2020.

[28] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7479–7489.

[29] Wang, Liansheng and Chen, Rongzhen and Zhu, Lei and Xie, Haoran and Li, Xiaomeng, "Deep sub-region network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 728–741, 2021.
[30] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 234–250.
[31] Hu, Xiaowei and Fu, Chi-Wing and Zhu, Lei and Wang, Tianyu and Heng, Pheng-Ann, "SAC-Net: Spatial attenuation context for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 3, pp. 1079–1090, 2021.
[32] P. Zhang, D. Wang, H. Lu, H. Wang, and 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.
[33] L. Wang, H. Lu, X. Ruan, and 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.
[34] N. Liu, J. Han, and 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.
[35] Zhu, Lei and Chen, Jiaxing and Hu, Xiaowei and Fu, Chi-Wing and Xu, Xuemiao and Qin, Jing and Heng, Pheng-Ann, "Aggregating attentional dilated features for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 10, pp. 3358–3371, 2019.
[36] X. Zhang, T. Wang, J. Qi, H. Lu, and 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.
[37] Cong, Runmin and Lei, Jianjun and Fu, Huazhu and Cheng, Ming-Ming and Lin, Weisi and Huang, Qingming, "Review of visual saliency detection with comprehensive information," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 10, pp. 2941–2959, 2019.
[38] A. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton, "Stacked capsule autoencoders," in *Advances in Neural Information Processing Systems*, 2019, pp. 15 512–15 522.
[39] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 1, pp. 154–163, 2018.
[40] J. Yu, W. Zhang, Y. Lu, Z. Qin, Y. Hu, J. Tan, and Q. Wu, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3196–3209, 2020.
[41] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 603–612.
[42] C. Yan, Y. Hao, L. Li, J. Yin, A. Liu, Z. Mao, Z. Chen, and X. Gao, "Task-adaptive attention for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.
[43] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3668–3677.
[44] Li, Chongyi and Cong, Runmin and Kwong, Sam and Hou, Junhui and Fu, Huazhu and Zhu, Guopu and Zhang, Dingwen and Huang, Qingming, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Transactions on Cybernetics*, vol. 51, no. 1, pp. 88–100, 2021.
[45] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3763–3776, 2020.
[46] Chen, Zuyao and Cong, Runmin and Xu, Qianqian and Huang, Qingming, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection," *IEEE Transactions on Image Processing*, 2020.
[47] T. Zhao and 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.
[48] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked u-shape network with channel-wise attention for salient object detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 1397–1409, 2021.
[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
[50] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7794–7803.
[51] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3146–3154.
[52] L. Zhang, J. Dai, H. Lu, Y. He, and 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.
[53] G. Máttyus, W. Luo, and R. Urtasun, "Deeproadmapper: Extracting road topology from aerial images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 3438–3446.
[54] P. Jaccard, "The distribution of the flora in the alpine zone. 1," *New Phytologist*, vol. 11, no. 2, pp. 37–50, 1912.
[55] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 136–145.
[56] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.
[57] C. Yang, L. Zhang, H. Lu, X. Ruan, and 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.
[58] Y. Li, X. Hou, C. Koch, J. M. Rehg, and 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.
[59] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1597–1604.
[60] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *IJCAI*, 2018.
[61] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "A new way to evaluate foreground maps," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, p. 245484557.
[62] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 733–740.
[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
[64] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010, pp. 249–256.
[65] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9141–9150.
[66] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 599–10 606.
[67] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 7264–7273.
[68] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1623–1632.
[69] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 8150–8159.
[70] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and 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.
[71] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 3789–3798.

[72] Zhao, Hengshuang and Shi, Jianping and Qi, Xiaojuan and Wang, Xiaogang and Jia, Jiaya, "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2881–2890.

Authors' response to the reviewers' comments on "Engaging Part-whole Hierarchies and Contrast Cues for Salient Object Detection" (TCSVT-06173-2021)

We thank the reviewers, the Associate Editor and Editor-in-Chief for their constructive suggestions, which are of great importance for improving the quality of our work as well as for their patience with this revision all along. In the revised version, we have conducted more in-depth analysis and several experiments to address the reviews' comments. We believe that all comments raised in the review report have been carefully accommodated to the best of our knowledge. The main changes are highlighted in blue in the manuscript and the point-by-point responses to all comments/questions raised by the reviews are as follows.

Response to the comments of Editor-in-Chief:

Q1. What are the 3-5 papers published in the IEEE Transactions on Circuits and Systems for Video Technology, which are most closely related to your manuscript?

Response: Thank you very much for your suggestions. We believe that the following three papers are most closely related to our manuscript:

[r1] Lei Zhu et al. "Aggregating attentional dilated features for salient object detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 10, pp. 3358-3371, 2019.

[r2] Liansheng Wang et al. "Deep sub-region network for salient object detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 2, pp. 728-741, 2021.

[r3] Xiaowei Hu et al. "SAC-Net: Spatial attenuation context for salient object detection," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 3, pp. 1079-1090, 2021.

In the revised manuscript, these references and some descriptions about these references have

been added. See Section II.B and References [35, 29, 31] in the revision.

Q2. What is distinctive / new about your current manuscript related to these previously published papers?

Response: The most distinctive point between our proposed method and these previously published papers lies in that we integrate two saliency cues, i.e., part-whole hierarchies and contrast cues, instead of only contrast cues. More specific differences between each of these previous methods can be illustrated separately in the following contents:

(1) [r1] proposes a deep learning model to aggregate the attentional dilated features for salient object detection by exploring the complementary information between the global context and the local context. Unlike [r1] only based on contrast saliency cues, we propose a PWHCNet for salient object detection by interacting two saliency cues, i.e., part-whole hierarchies and contrast cues. This mechanism allows to obtain the whole saliency map with complete local details. What's more, [r1] employs some simple concatenation operations to achieve the fusion of multi-level features. While in our model, a Self-Channel and Mutual-Spatial attention module is designed to achieve the fusion of two types of saliency features (i.e., part whole hierarchies and contrast cues), and further automatically select those important features for the prediction of salient regions.

(2) Similar to [r2], we have also noticed that the importance of multi-scale salient context information for predicting the saliency map. For that, [r2] equips with a sequence of subregion dilated blocks by aggregating multi-scale salient context information of multiple sub-regions, such that the global context information from the whole image and the local contexts from sub-regions are fused together. Different from [r2], we present a DenseHRNet framework to capture multiscale context information with different receptive fields from the input image. The filtered results of different

sub-layer convolutions are integrated through dense residual connections. In addition, our proposed model also embeds the part-whole hierarchies to obtain the whole saliency map.

(3) [r3] presents a saliency detection network based on the spatial attenuation context by integrating the aggregated features with some weights that are learned from an attention mechanism. In contrast to [r3], we focus on employing a Self-Channel and Mutual-Spatial attention module to explore the long-range dependencies along the self-channel dimension and capture the long-range spatial dependencies across the two cues.

Response to the comments of Associate Editor:

Q1. Depending on the review comments, the details are unclear and some comparisons need to be added as well.

Response: Thank you very much for your suggestions. Following your and reviewers' suggestions, we have revised the manuscript carefully, which makes the statements and diagrams more accurate and the experiments richer. The specific modifications are as follows.

(1) To make the details of our algorithm clearer, we have revised some illustrations and statements of the manuscript as follows.

a). To better illustrate the diagram of our proposed the DGC sub-network, as shown in Fig. R1 (See the response to the comments of Reviewer 1), we have modified Fig. 2 and its statements in the revised manuscript. This better illustrates that the dynamic grouping strategy is applied before capsules routing to facilitate the exploration of more accurate part-whole relationships. In addition, as shown in Fig. R3 (See the response to the comments of Reviewer 1), we have modified Fig. 5 and some statements in the revised manuscript. This further clearly describes the inputs and mechanism of the

SCMS module. Please see Section III in the revised version.

b). For Fig. 6 and Fig. 7 in Section IV (i.e., Experimental part) of the original manuscript, we have improved Fig. 6 and Fig.7 in the revised manuscript according to the reviewer 2 suggestions. Please see Section IV.D in the revised version.

(2) In the revised manuscript, we have carried out more ablation studies to demonstrate the advantages of the method that integrates part-whole hierarchies and contrast cues for salient object detection.

a). Some new ablation studies for different components were carried out. Table R3 (See the response to the comments of Reviewer 2) below provides the experimental results. As shown in Table R3, we can see that the effectiveness of the proposed different components in our network. Meanwhile, it can be seen from Table R3 that a large number of parameters are mainly caused by the capsule routing of the DGC sub-network. We will further reduce the complexity of the capsule network to achieve an efficient architecture for SOD tasks in the future. Besides, to better demonstrate the performance of our model, we have carried out some new ablation experiments on DUTS-TE dataset with different scenes and various sizes. Table R6, Table R7, Table R8 and Table R9 (See the response to the comments of Reviewer 2) below provides the new experimental results on DUTS-TE dataset. This indicates that the experimental results of different components on different datasets in the proposed model are generally consistent, which further proves the effectiveness of our method.

b). To prove the validity of the proposed U-Res34, we compared the DGC sub-network that learns features through U-Res34 with its modified versions, which learns features of input images through FLNet in [17] or two Conv+ReLU layers used by the original CapsNet [18]. Table R10 and Fig. R5(See the response to the comments of Reviewer 2) provide the qualitative and quantitative results of

the experiment, respectively. It is obvious that U-Res34 makes the framework possess the ability of identifying the salient object wholly, which is attributed to the rich features learned by U-Res34.

c). In addition, we have also added a new ablation study on the integration strategy of the two cues in the revision. A new experiment was conducted to compare the proposed SCMS with SCC, MWSA, non-local [50] and DA module [51]. As shown in Table R9 (See the response to the comments of Reviewer 2, i.e., Table VI in the revised manuscript), the proposed SCMS module can obtain the best performance compared with non-local [50] or DA module [51]. This demonstrates the superiority of the SCMS module.

(3) Following reviewers' suggestions, as shown in Table R11 (See the response to the comments of Reviewer 2), we compare the floating point operations (i.e., FLOPs), the number of parameters (i.e., Params) and the inference time (i.e., Time) with other popular methods. We have added Table R11 (i.e., Table II in the revised manuscript) and its statements in Section IV.D in the revised manuscript. Please see Section IV.D in the revised version.

Response to the comments of Reviewer #1:

This paper proposed a PWHCNet for salient object detection by interacting part-whole hierarchies and contrast cues. At the cost of complex architectures, the performance is satisfactory and experiment is sufficient. However, the design is not clear enough. Some details should be further clarified.

Response: Thank you very much for your positive comments and constructive suggestions. Our responses to your comments are as follows:

Q1. The top branch structure (DGC) in Fig. 2 is not clearly described. Does it describe the three steps of the dynamic grouping strategy, i.e., calculating capsule correlation matrix, determining initial capsules in each group, and putting similar capsules into the same group? If it is, is it consistent with dividing into 2 groups and then 4 groups as described in the paper; if not, what does the top branch structure (DGC) of Fig. 2 describe?

Response: Thank you very much for your suggestions. We are sorry that vague Fig. 2 in the original manuscript may mislead your understanding on our proposed DGC sub-network. Actually, Fig. 2 describes the whole architecture of the proposed DGC subnetwork, including dynamic grouping and capsule routing, instead of only the dynamic grouping strategy. To avoid misunderstanding, we have modified Fig. 2 to Fig. R1 (i.e., Fig. 2 in the revised manuscript) below. As shown in Fig. R1, the dynamic grouping strategy (i.e., “D G” in Fig. R1) is implemented before capsules routing (i.e., “C R” in Fig. R1) to facilitate high-correlated capsules grouping for capturing more accurate part-whole relationships. Specifically, capsules with similar properties are grouped into the same group for further capsule routing in the group. For example, by virtue of the dynamic grouping strategy, capsule 9 is finally allocated into a group with capsules 2, 3, and 6 (rather than 1, 5, and 14) for routing. We have revised the relevant statements in the revised manuscript. Please see Section III and Fig. 2 in the

revised version.

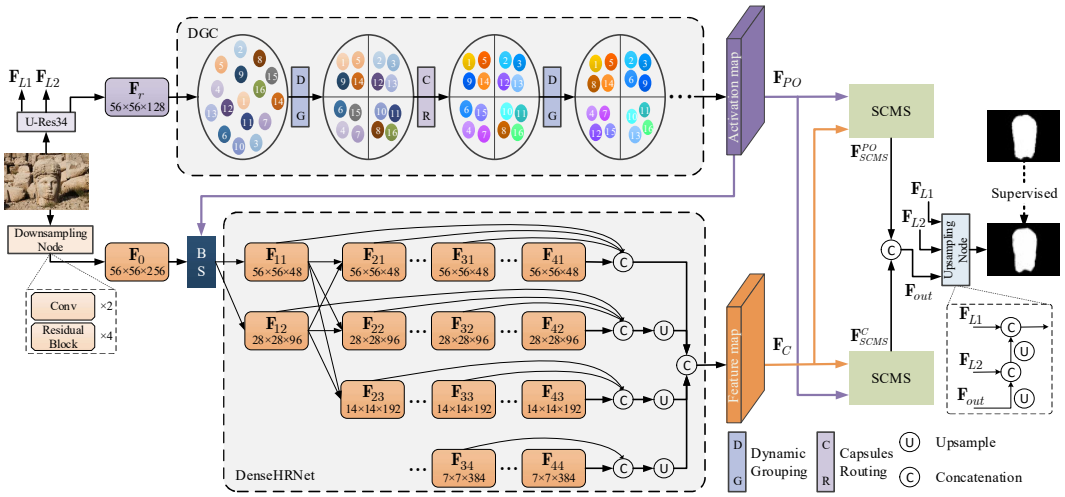


Fig. R1. The overall architecture of our proposed PWHCNet for salient object detection, which consists of a DGC sub-network and a DenseHRNet sub-network to capture the part-whole hierarchies and contrast cues from input images, respectively. The part-whole relational cues are additionally used to guide the feature extraction of DenseHRNet at the shallow layer via a BS module. On top of that, the above two saliency cues are interacted by a SCMS attention module to achieve more primitive saliency semantics $\mathbf{F}_{out}$, which are further used to predict the final saliency map. More details are provided in the text body.

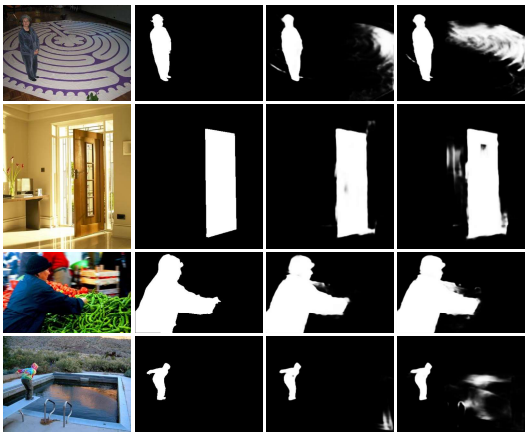
Q2. In the contributions, the author mentions "under the guidance of the part-whole relational cues, the DenseHRNet sub-network pays more attention on the object regions", so whether it can be better explained by the performance of DenseHRNet sub-network without the DGC guidance in the ablation study.

Response: Thanks for your comments. The ablation study results in Table R1 and Fig. R2 (i.e., the 7th and 8th rows of Table III and Fig. 8 (d) and (e) in the revised manuscript) demonstrate the performance with the DGC guidance. "H$^+$ + PO$^+$" and "H$^+$ + PO$^+$ + BS" represent the results without and with the DGC guidance, respectively. From Table R1, we can see that the saliency detection performance has been improved by using the DGC guidance. Besides, from the visual comparisons in Fig. R2 (c) and

(d), it can also be observed that the background noise is well suppressed by virtue of the BS module. In other words, under the guidance of the part-whole relational cues, the DenseHRNet sub-network may pay more attention to the object regions.

Table R1. Ablation studies on the DGC. The best performance is marked by **bold**.

Configurations	F_{max}	F_{avg}	MAE
$H^+ + PO^+$	0.792	0.772	0.057
$H^+ + PO^+ + BS$	**0.799**	**0.778**	**0.056**



(a) (b) (c) (d)

Fig. R2. Visual comparisons for the DGC. (a) Image; (b) GT; (c) "$H^+ + PO^+ + BS$", i.e., the results with the DGC guidance; (d) "$H^+ + PO^+$", i.e., the results without the DGC guidance.

Q3. What is the input of Fig. 5, which is not described clearly. In addition, are the colors in Fig. 2 related to the colors in Fig. 5?

Response: Thank you for your comments and constructive suggestions.

(1) We are sorry for that vague Fig. 5 in the original manuscript. There are two inputs in Fig. 5, i.e., $\mathbf{F}_{self}$ and $\mathbf{F}_{mutual}$. Specifically, to achieve $\mathbf{F}_{SCMS}^{PO}$, $\mathbf{F}_{self}$ and $\mathbf{F}_{mutual}$ are the output features $\mathbf{F}_{PO}$ of the DGC branch and the output features $\mathbf{F}_{C}$ of the DenseHRNet branch, respectively. Similarly, to achieve $\mathbf{F}_{SCMS}^{C}$, $\mathbf{F}_{self}$ and $\mathbf{F}_{mutual}$ are the output features $\mathbf{F}_{C}$ of the DenseHRNet branch and the output features $\mathbf{F}_{PO}$ of the DGC branch, respectively. In order to better illustrate the inputs of Fig. 5, we have added these statements in the revised manuscript. Please see Section III.C and Fig. 5 in the

revised version.

(2) The colors in Fig. 2 are not related to the colors in Fig. 5. To avoid confusion, as shown in Fig. R3 (i.e., Fig. 5 in the revised manuscript), we have modified the colors of Fig. 5. Please see Fig. 5 in the revised version.

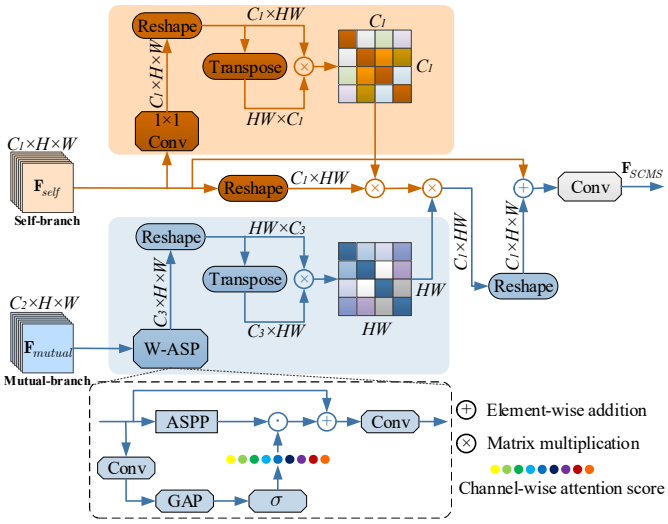


Fig. R3. The architecture of SCMS module. Shadow regions marked by brown and blue colors represent the SCC unit and the MWSA unit, respectively. 'W-ASP' refers to the Weighted Atrous Spatial Pyramid (W-ASP) sub-module.

Q4. It can be seen from Fig. 2 that the model should be very large, how much is it, and how about the running time of the proposed method?

Response: Thank you very much for your suggestions.

(1) The number of parameters (Params) and FLOPs of our proposed model are 153.26M and 137.64G, respectively. For a deeper understanding, Table R2 lists the parameter numbers and FLOPs of different component. It can be seen from Table R2 (i.e., Table III in the revised manuscript) that the large number of parameters of our proposed model is mainly caused by the DGC sub-network (i.e., PO^{+}), which covers complex capsule routing.

Table R2. The FLOPs and the number of parameters for different components. "H^+" and "PO^+" represent the DenseHRNet and DGC sub-network, respectively. "BS" and "S-C" denote the BS and SCMS modules, respectively.

Configurations	*FLOPs(G)*	*Params(M)*
H^+	27.10	66.77
PO^+	111.92	86.41
$H^+ + PO^+$	137.14	153.13
$H^+ + PO^+ +$ BS	137.33	153.19
$H^+ + PO^+ +$ BS + S-C	137.64	153.26

(2) In addition, the running time of the entire model for each image is 0.167 s.

Q5. Some related works of SOD should be discussed in the paper for completeness, such as Review of visual saliency detection with comprehensive information, TCSVT 2019; ASIF-Net: Attention steered interweave fusion network for RGBD salient object detection, TCyb 2021; DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection, TIP 2021; Going from RGB to RGBD saliency: A depth-guided transformation model, TCyb 2020.

Response: Thank you for your suggestions. We have discussed the above related works in the revised manuscript. Please see Section II.B and references [37, 44, 46, 27] in the revised manuscript.

Response to the comments of Reviewer #2:

This paper proposes a PWHCNet for salient object detection by interacting part-whole hierarchies and contrast cues. The part-whole relationships exploration branch innovates on the basis of TSPOANet, and the contrast cues extraction branch innovates on the basis of HRNet. The innovation of this paper is that two salient cues, part-whole and contrast, are simultaneously used for salient object detection. The following points are concerned:

Response: Thank you very much for your positive comments and constructive suggestions. Our responses to your comments are as follows:

Q1. The idea to integrate part-whole and contrast cues for salient object detection is good, but the weakness of the paper is that too many tricks or modules are added into the whole network. Thus, we doubt that the good performance achieved is mainly by the idea or by integrating various tricks.

Response: Thank you very much for your comments. As shown in Table R3 (i.e., Table III in the revised manuscript), the significant performance increase for our proposed network mainly results from the integration of part-whole hierarchies and contrast cues, though we have to admit that some proposed modules also slightly improve the performance of proposed model.

In the revised manuscript, Table R3 illustrates some new ablation studies on different components to demonstrate the effectiveness of the integration of part-whole hierarchies and contrast cues for salient object detection. "H" , "PO" and "H + PO" represent three structures, i.e., contrast cues extracted by the HRNet [19], part-whole hierarchies explored by fixed grouping capsule network, and the combination of the above two cues by the addition operation. By comparing "H + PO" with "H" or "PO" in Table R3, we can see that the idea to integrating the above two cues is feasible, which can significantly improve the saliency detection performance. Meanwhile, the prediction results of "H" and

"PO" are, respectively, improved by the proposed "H^+" (i.e., DenseHRNet sub-network) and "PO^+" (i.e., DGC sub-network). Additionally, it can be seen from Table R3 that "$H^+ + PO^+$" further improves the performance of our proposed model by the combination of the two cues. We have added these statements in the revised manuscript. Please see Section IV.E and Table III in the revised version.

Table R3. Ablation studies of different components. The best performance is marked by **bold**. "B" represents the common backbone (ResNet-50). "H" and "H^+" represent the HRNet [19] and the DenseHRNet, respectively. "PO" and "PO^+" mean fixed grouping and dynamic grouping strategies adopted in the capsule network, respectively. "BS" and "S-C" denote the BS and SCMS modules, respectively.

Configurations	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
B	0.754	0.740	0.057	0.833	0.810	0.040
H	0.768	0.753	0.061	0.848	0.828	0.042
PO	0.758	0.736	0.065	0.840	0.814	0.046
H + PO	0.772	0.761	0.062	0.852	0.836	0.041
H^+	0.786	0.762	0.059	0.866	0.838	0.039
PO^+	0.758	0.742	0.064	0.848	0.826	0.044
$H^+ + PO^+$	0.792	0.772	0.057	0.870	0.848	0.037
$H^+ + PO^+$ + BS	0.799	0.778	0.056	0.878	0.854	0.036
$H^+ + PO^+$ + BS + S-C	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**

Q2. Similar to the above question, the ablation has been done on the effectiveness of different components in this method, however, the comparison to this level is not enough. That's because one single component in this paper actually contains more than one units, such as SCMS module for two different saliency cues integration including SCC and MWSA. More experimental comparisons on the effectiveness about these two units should also be provided. Similar problems also exist in DenseHRNet. The influence by the initial feature extraction to the whole module is also expected.

Response: Thank you for your comments and constructive suggestions.

(1) Following what you have suggested, as shown in Table R4, we compare the proposed SCMS with SCC, MWSA for better understanding the proposed model. It can be seen that the performance

obtained by using SCC or MWSA individually is inferior to that obtained by using SCMS, which demonstrates that simultaneously considering the intra-cues channel interaction and the inter-cues spatial interaction indeed helps to improve performance. We have added these statements in the revised manuscript. Please see Section IV.E and Table VI in the revised manuscript.

Table R4. Ablation studies of different integration strategies. The best performance is marked by **bold**.

Integration Strategies	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
Baseline(H^++PO^+)	*0.792*	*0.772*	*0.057*	0.870	0.848	0.037
+ SCC	0.789	0.772	**0.055**	0.876	0.853	0.036
+ MWSA	0.796	0.782	0.056	0.880	0.859	**0.035**
+ SCMS	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**

(2) In order to demonstrate the influence of the initial feature extraction on DenseHRNet, we have carried out a new experiment by replacing the initial feature extraction block in the original manuscript with a simpler feature extraction block, i.e., two convolutional layers. Table R5 shows the experimental results. We can see that the influence of the initial feature extraction on the performance of the DenseHRNet sub-network is negligible. This structure that is adopted in our proposed model is consistent with that in HRNet [19]. Considering the space limitation, we only showed this result in the response.

Table R5. The experimental results of two feature extraction manners.

Downsampling Node	F_{avg}	MAE
Two convolutional layers	0.760	0.061
Initial feature extraction	**0.762**	**0.059**

Q3. From Table 1, we can see that the proposed method achieved the second best results from Sm and MAE two evaluation metrics on PASCAL-S dataset. It should be explained why this method does not work so well on this dataset.

Response: Thank you for your comments. Compared with other datasets, PASCAL-S dataset contains some unique scenes. It can be seen from Fig. R4 (i.e., Fig. 11 in the revised manuscript) that, under the

effect of part-whole hierarchies, some objects with certain relations are detected together, e.g., computer and keyboard, table and sofa, television and television cabinet, etc., instead of one individual object as masked by the ground truth, which may cause poor S_m and MAE. We will study this issue as the future work, which may be solved using scene parsing [72]. We have added these statements and Fig. R4 in the revised manuscript. Please see Section IV.F and Fig. 11 in the revised manuscript.

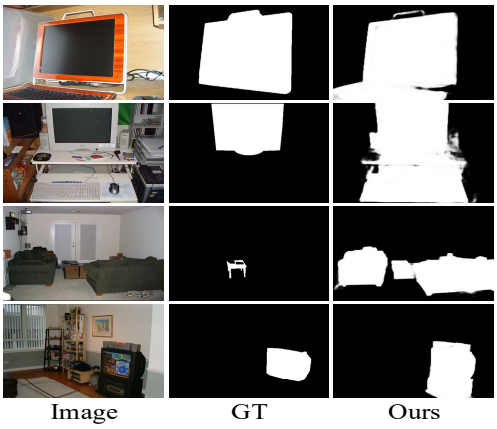


Fig. R4. Some failure cases for our proposed method.

Q4. Fig.7 provides the visually detected SOD results. Better to present the results in groups of the employed datasets. By this way, we can know clearly which dataset each image is from.

Response: Thank you very much for your suggestions. Presenting the results in groups is a very constructive suggestion. However, considering that the starting point of this paper is to solve the saliency detection problem in various complex scenes, we show the visual comparisons of our proposed model and other methods in Fig. 7 of the revised manuscript by displaying some images covering different scenarios (i.e., low contrast, similar backgrounds, small objects and multiple objects), rather than from different datasets. This will better illustrate the superiorities of our proposed method over others in various scenes. As well, to clearly know which dataset each image is from, we marked the name of the dataset for each image on the right part of Fig. 7 in the revised manuscript. Please see Section IV.D and Fig. 7 in the revised version.

Q5. As for Figure 6 of the experimental part, some curves are too closely overlapped to be clearly seen. Try resetting the spacing or the width of the curve to draw the graph.

Response: Thank you for your suggestions. To show the PR curve and F-measure curves of different methods more clearly, we have modified Fig. 6 in the revised manuscript. Please see Section IV.D and Fig. 6 in the revised version.

Q6. The ablation experiment in this paper was only carried out on DUT-OMRON data set. Are there any ablation experiments on other data sets?

Response: Thank you very much for your comments and suggestions. To provide more comprehensive ablation studies, we have added ablation experiments on DUTS-TE dataset, resulting in ablation experiments on DUT-OMRON and DUTS-TE, i.e., Table R6, Table R7, Table R8 and Table R9. Please see Section IV.E, Table III, Table IV, Table V and Table VI in the revised version.

Table R6. Ablation studies of different components. The best performance is marked by **bold**. "B" represents the common backbone (ResNet-50). "H" and "H^+" represent the HRNet [19] and the DenseHRNet, respectively. "PO" and "PO^+" mean fixed grouping and dynamic grouping strategies adopted in the capsule network, respectively. "BS" and "S-C" denote the BS and SCMS modules, respectively.

Configurations	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
B	0.754	0.740	0.057	0.833	0.810	0.040
H	0.768	0.753	0.061	0.848	0.828	0.042
PO	0.758	0.736	0.065	0.840	0.814	0.046
H + PO	0.772	0.761	0.062	0.852	0.836	0.041
H^+	0.786	0.762	0.059	0.866	0.838	0.039
PO^+	0.758	0.742	0.064	0.848	0.826	0.044
$H^+ + PO^+$	0.792	0.772	0.057	0.870	0.848	0.037
$H^+ + PO^+$ + BS	0.799	0.778	0.056	0.878	0.854	0.036
$H^+ + PO^+$ + BS + S-C	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**

Table R7. Ablation studies of different capsule grouping strategies. The best performance is marked by **bold**. "O" denotes no grouping strategy. "T" and "D_γ" (γ = 2, 4, 8) represent fixed grouping strategy and dynamic grouping strategies with different group numbers, respectively.

Configurations	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
H^+ + O	0.769	0.753	0.060	0.847	0.830	0.040
H^+ + T	0.782	0.763	0.058	0.861	0.838	0.039
H^+ + D_2	0.791	0.767	0.058	0.868	0.841	0.039
H^+ + D_4	**0.792**	**0.772**	**0.057**	**0.870**	**0.848**	**0.037**
H^+ + D_8	0.790	0.769	**0.057**	0.867	0.845	0.038

Table R8. Ablation studies of different feature extraction architectures for DGC sub-network. The best performance is marked by **bold**. Here, the capsules are dynamically divided into four groups.

Feature Extraction Architectures	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
two Conv+ReLU layers	0.506	0.452	0.195	0.552	0.482	0.182
FLNet	0.712	0.695	0.071	0.797	0.769	0.055
U-Res34	**0.758**	**0.742**	**0.064**	**0.848**	**0.826**	**0.044**

Table R9. Ablation studies of different integration strategies. The best performance is marked by **bold**.

Integration Strategies	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
Baseline(H^+ + PO^+)	0.792	0.772	0.057	0.870	0.848	0.037
+ Non-local [50]	0.799	0.772	0.060	0.880	0.850	0.037
+ DA module [51]	0.800	0.781	0.056	0.881	0.858	**0.035**
+ SCC	0.789	0.772	**0.055**	0.876	0.853	0.036
+ MWSA	0.796	0.782	0.056	0.880	0.859	**0.035**
+ SCMS	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**

Q7. Have you done the ablation experiment of U-Res34 in the feature extraction stage?

Response: Thank you very much for your suggestions. To prove the validity of U-Res34, we have carried out a new ablation experiment in the feature extraction stage. We replace U-Res34 in our proposed DGC sub-network with FLNet in [17] or the two Conv+ReLU layers in the original CapsNet [18]. It can be observed from Table R10 (i.e., Table V in the revised manuscript) that U-Res34 boosts the saliency detection performance of our proposed model significantly. As shown in Fig. R5(c-e) (i.e., Fig. 10 (c-e) in the revised manuscript), it is obvious that U-Res34 makes the framework possess the

ability of identifying the salient object wholly, which is attributed to the rich features learned by U-Res34. We have added the above statements, Table R10 and Fig. R5 in the revised manuscript. Please see Section IV.E, Table V and Fig. 10 in the revised version.

Table R10. Ablation studies of different feature extraction architectures for DGC sub-network. The best performance is marked by **bold**. Here, the capsules are dynamically divided into four groups.

Feature Extraction Architectures	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
two Conv+ReLU layers	0.506	0.452	0.195	0.552	0.482	0.182
FLNet	0.712	0.695	0.071	0.797	0.769	0.055
U-Res34	**0.758**	**0.742**	**0.064**	**0.848**	**0.826**	**0.044**

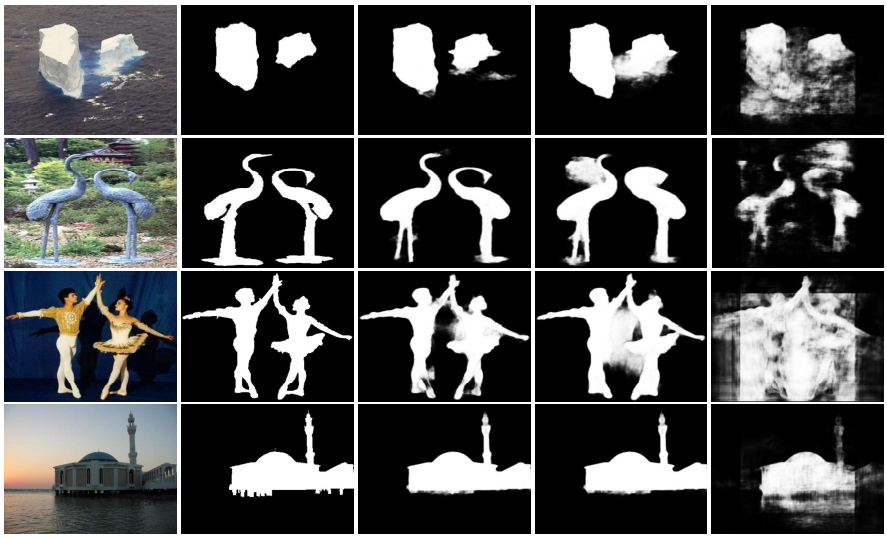


Fig. R5. Visual comparisons with different feature extraction architectures for DGC sub-network. (a) Image; (b) GT; (c) U-Res34; (d) FLNet; (e) Two Conv+ReLU layers.

Q8. What's the number of parameters and run-time compare to other methods, and which module caused the complexity.

Response: Thank you for your comments and constructive suggestions.

(1) Following your suggestion, as shown in Table R11, we compare the floating point operations (i.e., FLOPs), the number of parameters (i.e., Params) and the inference time (i.e., Time) with other popular methods. Table R11 shows that our model is complicated than other methods. This illustrates

that the performance improvement of our proposed model is at the cost of computational complexity. We have added these statements in the revised manuscript. Please see Section IV.D and Table II in the revised version.

Table R11. The number of parameters, FLOPs and inference time comparisons of our method with some state-of-the-art networks.

Method	Input size	FLOPs (G)	Params (M)	Time (s)
F3Net [16]	352 × 352	16.43	25.54	0.022
ITSD [65]	288 × 288	15.94	26.07	0.022
GPACNet [66]	320 × 320	54.31	67.06	0.020
BASNet [28]	256 × 256	127.40	87.06	0.032
MINet-R [14]	320 × 320	87.03	162.38	0.036
EGNet-R [13]	380 × 320	287.67	111.66	0.091
Ours	256 × 256	137.64	153.26	0.167

(2) Table R12 lists parameter numbers and FLOPs of different components. It can be seen from Table R12 (i.e., Table III in the revised manuscript) that the large number of parameters of our proposed model are mainly caused by the DGC sub-network, which covers complex capsule routing. In the future, we will further reduce the complexity of the capsule network to achieve an efficient architecture for SOD tasks.

Table R12. The FLOPs and the number of parameters for different components. "H^+" and "PO^+" represent the DenseHRNet and DGC sub-network, respectively. "BS" and "S-C" denote the BS and SCMS modules, respectively.

Configurations	*FLOPs(G)*	*Params(M)*
H^+	27.10	66.77
PO^+	111.92	86.41
$H^+ + PO^+$	137.14	153.13
$H^+ + PO^+ +$ BS	137.33	153.19
$H^+ + PO^+ +$ BS + S-C	137.64	153.26

Q9. I am wondering when there are multiple salient objects on the image whether the proposed method still can get the good performance.

Response: Thank you very much for your suggestions. Our model can still achieve good performance

for those scenes with multiple salient objects, as illustrated in Fig. R6. Such illustrations have also been added in the revised manuscript. Please see Fig. 7 in the revised version.

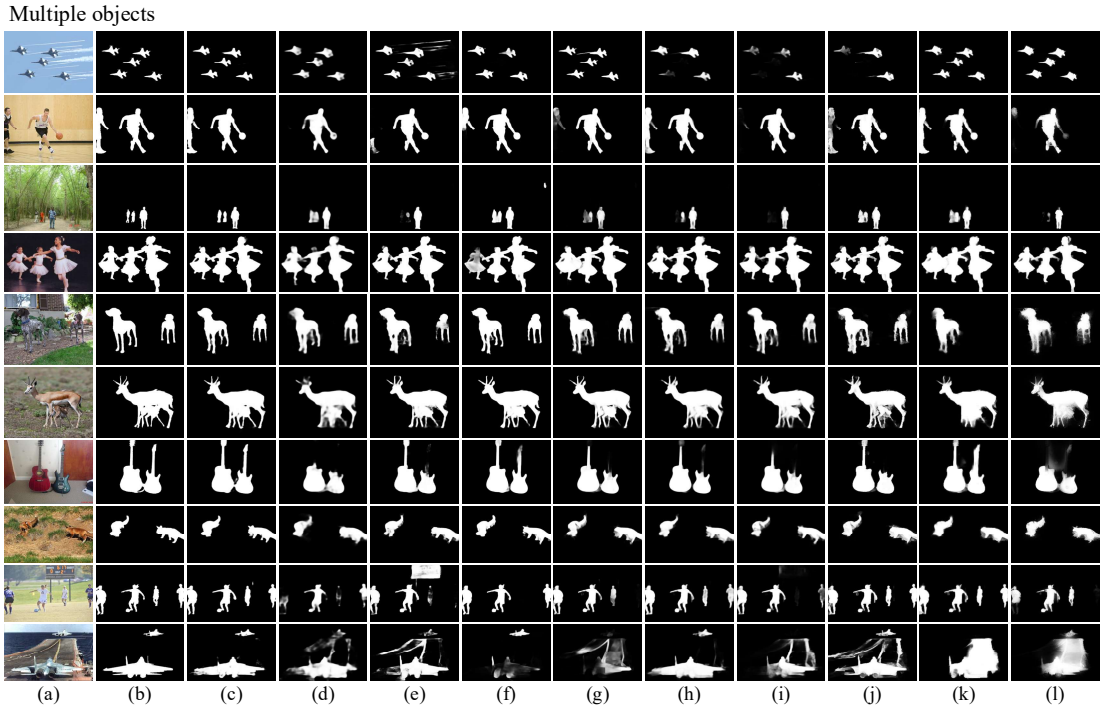


Fig. R6. Visual comparisons of different methods. (a) Image; (b) GT; (c) Ours; (d) TSPOANet [17]; (e) MINet [14]; (f) F3Net [16]; (g) EGNet [13]; (h) GCPANet [66]; (i) SCRN [67]; (j) AFNet [68]; (k) PAGE [70]; (l) JointCRF [71]

Response to the comments of Reviewer #3:

The authors propose salient object detection method by engaging two saliency cues, i.e., the part-whole hierarchies and contrast cues. The part-whole hierarchies is implemented via Capsule Network and they propose a Dynamic Grouping Capsules Routing (DGCR) strategy to alleviate the problem of inaccurate part-whole relationships caused by the noisy capsule assignments. Such technique is reasonable and effective. A DenseHRNet framework is designed to obtain more primitive contrast information with multiple scales. Experimental results validate the effectiveness of the proposed method. However, there are several shortcomings should be improved.

Response: Thank you very much for your positive comments and constructive suggestions. Our responses to your comments are as follows:

Q1. The proposed method is very complicated. Therefore, I would like to see the inference time of the proposed method and other compared methods.

Response: Thank you very much for your suggestions.

Table R13. The number of parameters, FLOPs and inference time comparisons of our method with some state-of-the-art networks.

Method	Input size	FLOPs (G)	Params (M)	Time (s)
F3Net [16]	352 × 352	16.43	25.54	0.022
ITSD [65]	288 × 288	15.94	26.07	0.022
GPACNet [66]	320 × 320	54.31	67.06	0.020
BASNet [28]	256 × 256	127.40	87.06	0.032
MINet-R [14]	320 × 320	87.03	162.38	0.036
EGNet-R [13]	380 × 320	287.67	111.66	0.091
Ours	256 × 256	137.64	153.26	0.167

As what you pointed out, the proposed model is very complicated. As shown in Table R13, we compare the floating point operations (i.e., FLOPs), the number of parameters (i.e., Params) and the inference time (i.e., Time) with other popular methods. This shows that the saliency detection

performance improvement of our proposed model is at the cost of computational complexity. We have added these statements in the revised manuscript. Please see Section IV.D and Table II in the revised version.

Q2. The authors discuss the uniqueness of the proposed SCMS module compared to the attention mechanisms in [45] and [46]. So what if replacing [45] or [46] with SCMS. Can the authors show the superiority of SCMS module over [45] and [46].

Response: Thank you very much for your constructive suggestions. Following what you have suggested, we have added a new ablation study on the integration strategy of the two cues in the revision, i.e., comparing the proposed SCMS with non-local [50] (i.e., [45] in the original manuscript) and DA [51] modules (i.e., [46] in the original manuscript). As shown in Table R14 (i.e., Table VI in the revised manuscript), it can be seen that the proposed SCMS module can obtain the best performance compared with non-local [50] or DA module [51], which demonstrates the superiority of our proposed SCMS module. Please see Section IV.E and Table VI in the revised version.

Table R14. Ablation studies of different integration strategies. The best performance is marked by **bold**.

Integration Strategies	DUT-OMRON			DUTS-TE		
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE
Baseline($H^+ + PO^+$)	0.792	0.772	0.057	0.870	0.848	0.037
+ Non-local [50]	0.799	0.772	0.060	0.880	0.850	0.037
+ DA module [51]	0.800	0.781	0.056	0.881	0.858	**0.035**
+ SCMS	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**

Q3. Can the Downsampling Node be replaced by U-Res34 (sharing weights), such that the inference time may be reduced.

Response: Thank you very much for your suggestions. Following your suggestion, we re-trained the network after replacing the Downsampling Node with U-Res34 (sharing weights). The experimental results are shown in Table R15. It can be seen that although this strategy can slightly reduce the

inference time of the network, it greatly reduces the saliency detection performance of our proposed model. Considering that, we still employ the Downsampling Node in our proposed model to achieve the initial feature extraction for the DenseHRNet branch.

Table R15. Ablation studies of different architectures on DUT-OMRON dataset.

Architectures	F_{max}	F_{avg}	*MAE*	*Time*
Sharing weight	0.789	0.771	0.060	**0.133**
Don't sharing weight	**0.802**	**0.786**	**0.055**	0.167

Response to the comments of missing key references:

Reviewer #1: Missing key ref:

Reviewer #2: Missing key ref:

Reviewer #3: Missing key ref: J. Wei, S. Wang, Z. Wu et al. Label Decoupling Framework for Salient Object Detection. CVPR, 2020.

Response: In the revised manuscript, the missing key references and some descriptions about the references have been added in the revised manuscript. See Section II.B and References [25] in the revision.