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].

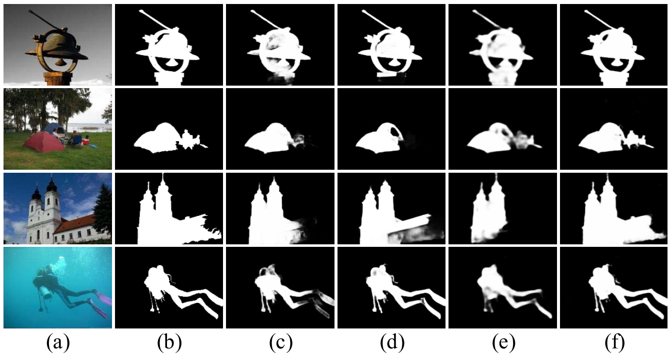


Fig. 1. Illustrations for sample results of our method compared with others. (a) Image; (b) GT; (c) MINet [14]; (d) F3Net [16]; (e) TSPOANet [17]; (f) Ours. TSPOANet [17]: saliency detector based on part-whole relationships; MINet [14] and F3Net [16]: saliency detectors based on contrast information.

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.

To alleviate the above problem, Liu and Yu [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

Manuscript received April 8, 2021; revised June 16, 2021 and July 16, 2021; accepted August 9, 2021. Date of publication August 16, 2021; date of current version June 6, 2022. This work was supported by the National Natural Science Foundation of China under Grant 61773301 and Grant 62001341. This article was recommended by Associate Editor C. Yang. *(Corresponding authors: Yi Liu; Jungong Han.)*

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

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

Yi Liu is with the School of Computer Science and Artificial Intelligence, and Aliyun School of Big Data, Changzhou University, Changzhou, Jiangsu 213164, China (e-mail: liuyi0089@gmail.com).

Jungong Han is with the Computer Science Department, Aberystwyth University, Aberystwyth SY23 3FL, U.K. (e-mail: jungonghan77@gmail.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2021.3104932.

Digital Object Identifier 10.1109/TCSVT.2021.3104932

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, 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

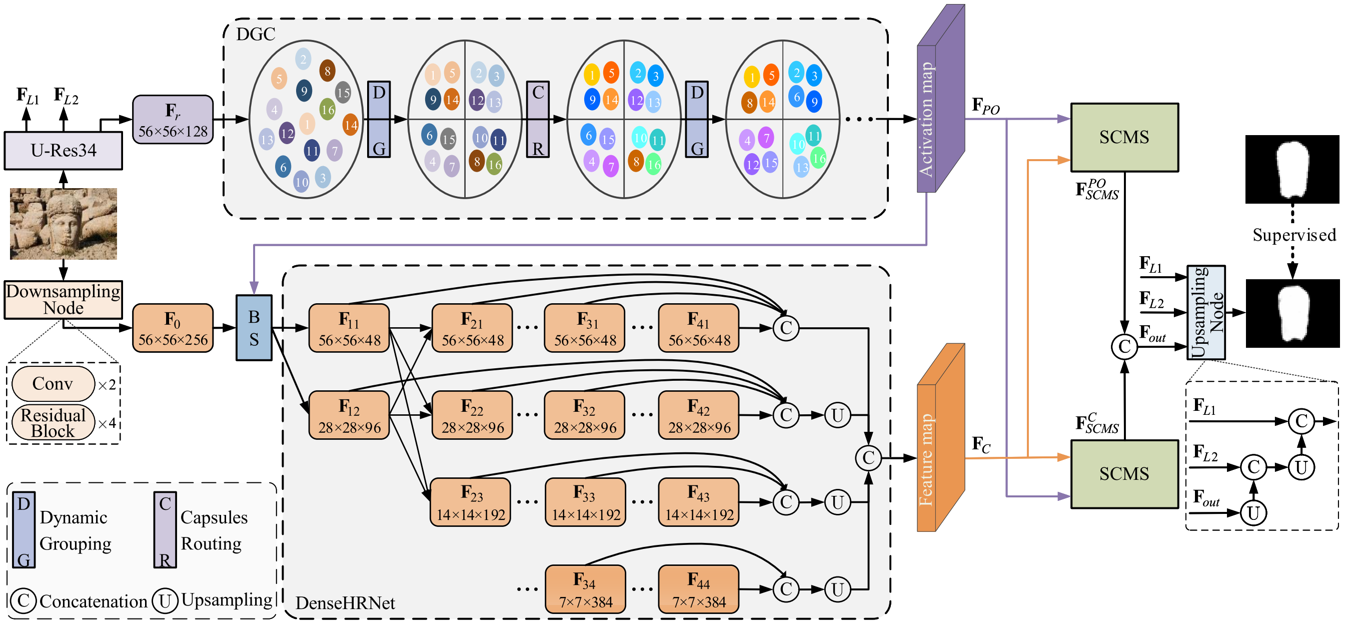


Fig. 2. The overall architecture of our proposed PWHCNet for salient object detection, which consists of a DGC sub-network and a DenseHRNet sub-network to capture the part-whole hierarchies and contrast cues from input images, respectively. The part-whole relational cues are additionally used to guide 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.

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 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

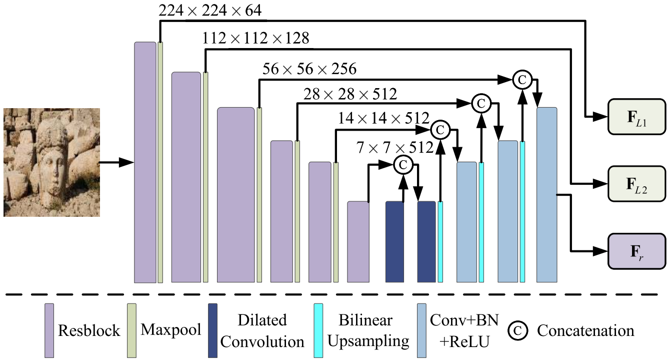


Fig. 3. Details of U-Res34. $\mathbf{F}_r$ will be used for capsule construction in our proposed model, while $\mathbf{F}_{L1}$ and $\mathbf{F}_{L2}$ will be used to recover salient object boundaries in the final saliency prediction stage.

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

1) Feature Extraction for Capsules Construction: Before the capsules routing, similar to the paradigm in [17], 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 × 224 × 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:

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

where $Caps_{m/n}(m, n = 1, 2, \ldots, K)$ represents the attribute information for the capsule of type m or type n. K denotes the total number of capsule types and is experimentally 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.*,

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

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, $\mathbf{L}_m \in \mathbb{R}^{1\times K}$ $(m = 1, 2, \ldots, K)$, 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

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

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

where $\mathbf{L}_{p,o_i} (p = 1, 2, \ldots, 16, p \neq o_i, i = 1, 2)$ 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 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 $i (i \in \mathbf{\Omega}_N)$ in layer N and each capsule $j (j \in \mathbf{\Omega}_{N+1})$ in layer $N + 1$. $\mathbf{\Omega}_N$ 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 and similar to that in [19], 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 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.*, $\sigma(*)$.

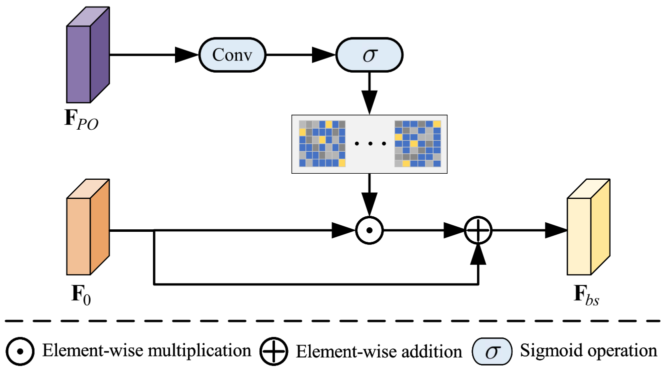


Fig. 4. The architecture of BS module. $\mathbf{F}_{PO}$ and $\mathbf{F}_0$ represent the outputs of the DGC sub-network and the Downsampling Node in Fig. 2, respectively.

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.

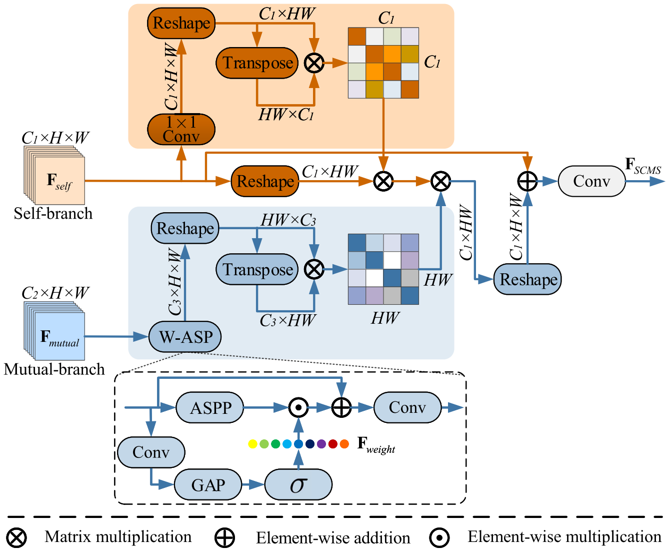


Fig. 5. The architecture of SCMS module. Shadow regions marked by brown and blue colors represent the SCC unit and the MWSA unit, respectively. 'W-ASP' refers to the Weighted Atrous Spatial Pyramid (W-ASP) sub-module.

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. 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.

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

2) 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:

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

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

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

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

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

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

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:

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

where 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.*,

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

3) 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

The resolutions of the output features $\mathbf{F}_{out}$ from the two SCMS modules mentioned above are 56 × 56. Simply using operations, *e.g.*, linear interpolations, to upsample $\mathbf{F}_{out}$ to the size of 224 × 224 (*i.e.*, the size of ground truths) will cause object boundary blurs. While, this can be alleviated with the aid of shallow features that usually possess higher resolutions and contain more details about the input images. For that, 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 in our proposed model, since $\mathbf{F}_{L1}$ and $\mathbf{F}_{L2}$ contain more details about the input images than the features extracted from the Downsampling Node. As shown in Fig. 2, the Upsampling Node is constructed by stacking upsampling and concatenation operations, and the process can be mathematically expressed by

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

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

where **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:

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

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:

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

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.*,

$$\mathcal{L}_{joint} = \mathcal{L}_{CE} + \mathcal{L}_{iou}. \tag{21}$$

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 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].

1) 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.

2) F-Measure: F_β is formulated as the weighted harmonic mean of precision and recall, *i.e.*,

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

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}).

3) 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.

4) 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.

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

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

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

1) 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.

2) 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

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				
	F_{max}	F_{avg}	E_m	S_m	MAE	F_{max}	F_{avg}	E_m	S_m	MAE	F_{max}	F_{avg}	E_m	S_m	MAE	F_{max}	F_{avg}	E_m	S_m	MAE	F_{max}	F_{avg}	E_m	S_m	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
IITSD [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

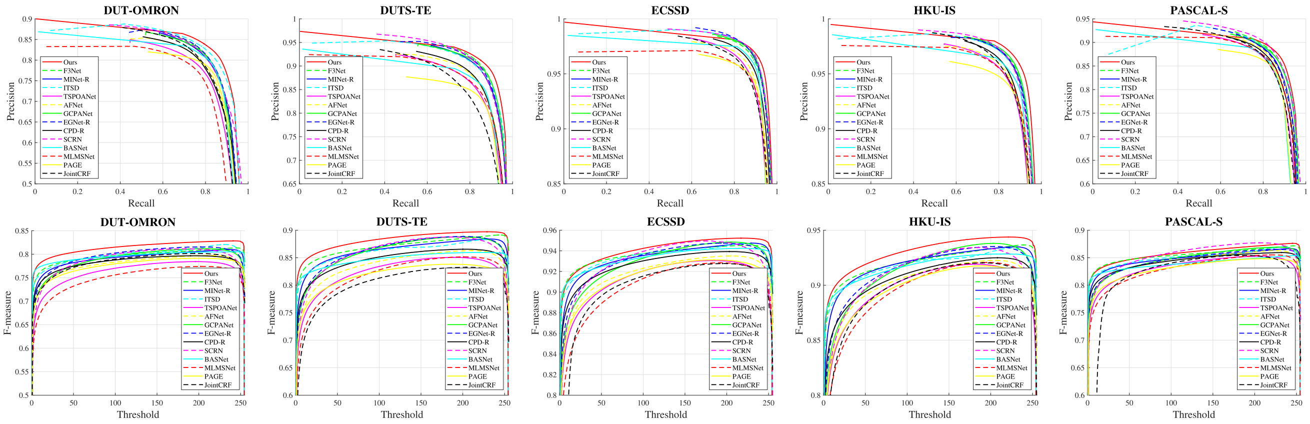


Fig. 6. PR curves (1st row) and F-measure curves (2nd row) on the five saliency detection datasets.

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

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 6th, 7th and 8th 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.

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 1st, 2nd, 10th and 11th 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 3rd, 4th and 5th 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 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

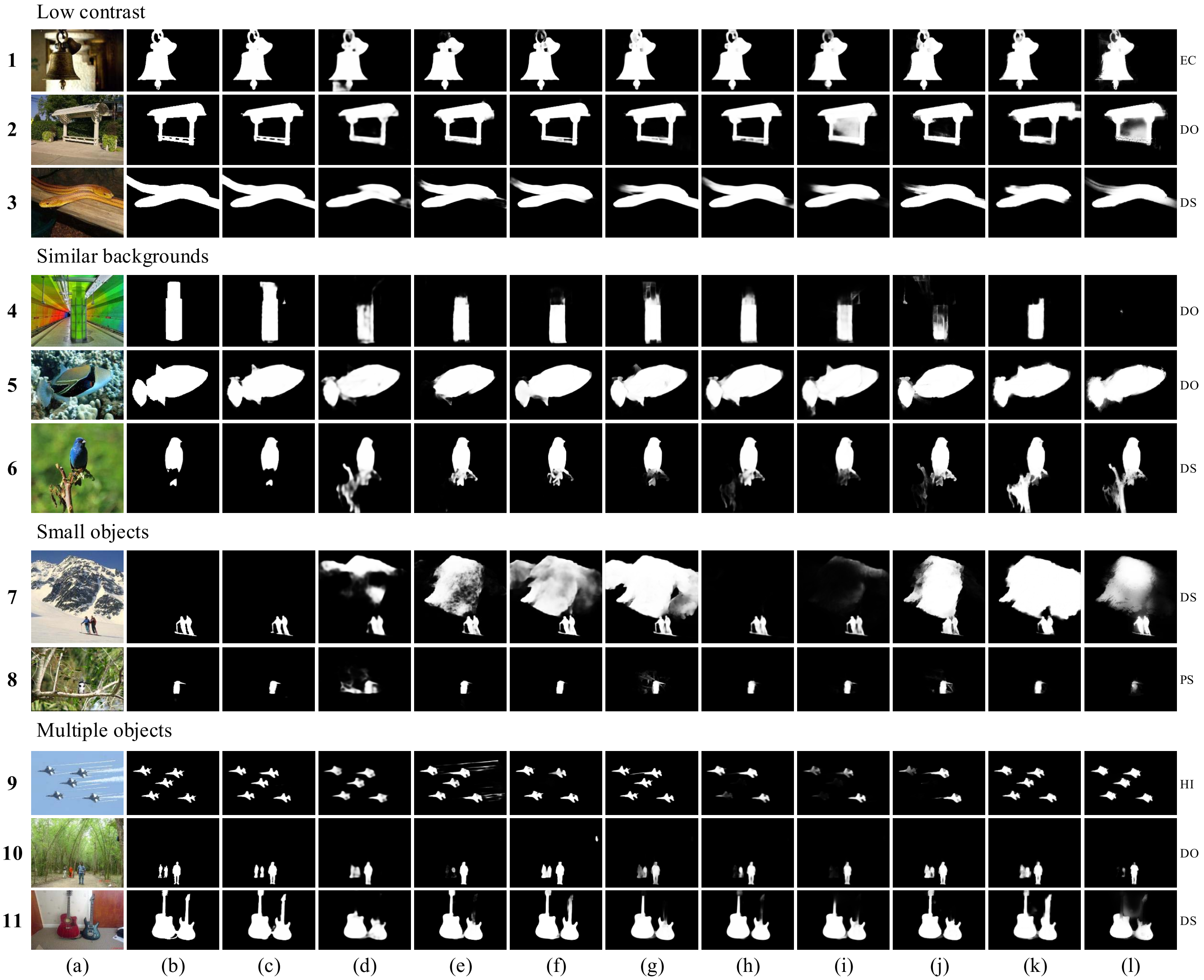


Fig. 7. Visual comparisons of different methods. (a) Image; (b) GT; (c) Ours; (d) TSPOANet [17]; (e) MINet [14]; (f) F3Net [16]; (g) EGNet [13]; (h) GCPANet [66]; (i) SCRN [67]; (j) AFNet [68]; (k) PAGE [70]; (l) JointCRF [71].The right side indicates the name of the dataset for each image, including ECSSD (EC) [56], DUT-OMRON (DO) [57], DUTS-TE (DS) [55], PASCAL-S (PS) [58] and HKU-IS (HI) [24].

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.

1) 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 output $\mathbf{F}_{PO}$ from DGC and the output $\mathbf{F}_C$ from DenseHRNet are integrated by the element-wise addition operation (Here, the BS module is not used in this structure). "H$^+$ + PO$^+$ + BS" denotes that the background suppression module is inserted into DenseHRNet. "H$^+$ + PO$^+$ + BS+ S-C" denotes that two SCMS modules are used to integrate "H$^+$" and "PO$^+$". It should be also noted that the same feature extraction method shown in Fig. 2 (*i.e.*, U-Res34 and Downsampling Node are used before the DGC sub-network and DenseHRNet sub-network, respectively) are used for all of these ablation experiments mentioned here.

As shown in Table III, by comparing the 1st and 2nd 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

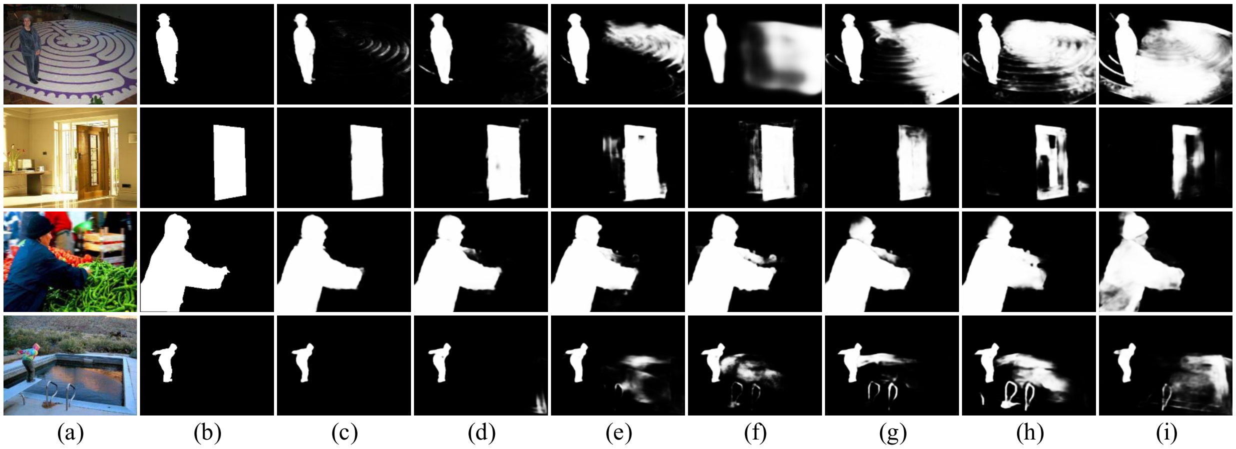


Fig. 8. Visual comparisons with different components. (a) Image; (b) GT; (c) “H^+” + “PO^+” + “BS” + “S-C”; (d) “H^+” + “PO^+” + “BS”; (e) “H^+” + “PO^+”; (f) “PO^+”; (g) “H^+”; (h) “H”; (i) “B”.

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	Params
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE	(G)	(M)
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 “$\text{H}^+ + \text{PO}^+$” achieves consistently higher performance than “$\text{H} + \text{PO}$” does by integrating “H^+” and “PO^+”. On top of “$\text{H}^+ + \text{PO}^+$”, 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.

2) 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.

In Table IV, the 1^{st} and 2^{nd} rows show the performance using the fixed grouping strategy (*i.e.*, $\text{H}^+ + \text{T}$) and using the dynamic grouping strategy (*i.e.*, “$\text{H}^+ + \text{D}_2$”). 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.*, “$\text{H}^+ + \text{D}_4$”) achieves the best performance. The reason for the performance degradation by dynamically dividing capsules into 8 groups (*i.e.*, “$\text{H}^+ + \text{D}_8$”) may be that a little fewer capsules in each group are not enough to characterize the part-whole hierarchies.

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.

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

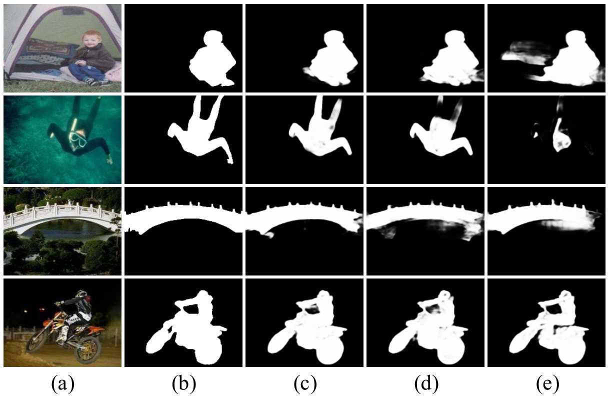


Fig. 9. Visual comparisons with different capsule grouping strategies. (a) Image; (b) GT; (c) H$^+$ + D$_4$; (d) H$^+$ + T; (e) H$^+$ + O.

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	DUT-OMRON			DUTS-TE		
Architectures	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**

3) 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.

4) 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.

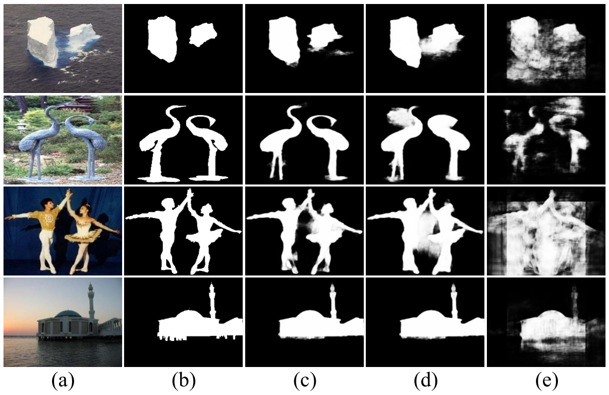


Fig. 10. Visual comparisons with different feature extraction architectures for DGC sub-network. (a) Image; (b) GT; (c) U-Res34; (d) FLNet in [17]; (e) Two Conv+ReLU layers.

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**

5) Different Initial Feature Extraction Strategies for DGC and DenseHRNet: In addition to our current initial feature extraction strategy (Res34-DN, for short, *i.e.*, U-Res34 for DGC and Downsampling Node for DenseHRNet), we re-trained our proposed model by applying another four different feature extraction strategies, i.e., two Downsampling Nodes with the same structures and shared weights (DNs-identical, for short), two U-Res34s with the same structures and shared weights (Res34s-identical, for short), two Downsampling Nodes with the same structures but different weights (DNs-different, for short) and two U-Res34s with the same structures but different weights (Res34s-different, for short) to extract the initial features for DGC and DenseHRNet branches. The experimental results are shown in Table VII.

As shown in Table VII, we can see that DNs-different, Res34s-different and Res34-DN significantly outperform DNs-identical and Res34s-identical. This indicates that applying two different modules for the initial feature extraction of DGC and DenseHRNet is a more reasonable way than employing two identical modules for the initial feature extraction of

TABLE VII

ABLATION STUDIES OF DIFFERENT INITIAL FEATURE EXTRACTION STRATEGIES. THE BEST PERFORMANCE IS MARKED BY **BOLD**

Strategies	DUT-OMRON			DUTS-TE			Time
	F_{max}	F_{avg}	MAE	F_{max}	F_{avg}	MAE	(s)
DNs-identical	0.796	0.769	0.064	0.872	0.841	0.047	0.129
Res34s-identical	0.789	0.771	0.060	0.871	0.848	0.042	0.133
DNs-different	0.801	0.772	0.063	0.875	0.847	0.040	0.131
Res34s-different	**0.802**	0.784	0.056	0.882	0.861	0.036	0.171
Res34-DN	**0.802**	**0.786**	**0.055**	**0.884**	**0.864**	**0.035**	0.167

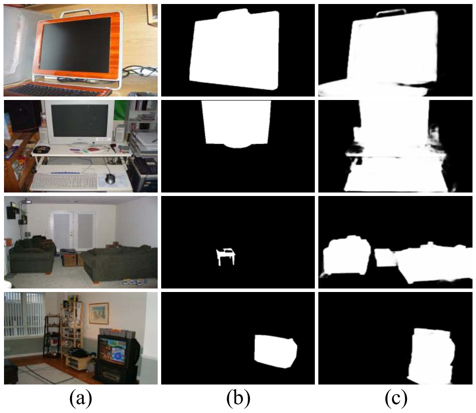


Fig. 11. Some failure cases for our proposed method. (a) Image; (b) GT; (c) Ours.

DGC and DenseHRNet does. This may owe to the following facts. DGC and DenseHRNet focus on capturing different cues for salient object detection. Specifically, DGC focuses on mining the part-whole hierarchies whilst DenseHRNet focuses on mining the contrast cues. Employing different initial features for DGC and DenseHRNet, respectively, may benefit the extraction of different cues for salient object detection. In addition, from Table VII, it can also be easily observed that Res34-DN performs competitively with Res34s-different and outperforms DNs-different in terms of F-measure and MAE. As well, the proposed Res34-DN strategy achieves higher inference efficiency than Res34s-different does.

F. Failure Cases

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.

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 Trans. Circuits Syst. Video Technol.*, vol. 28, no. 5, pp. 1130–1143, May 2018.

[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 Trans. Circuits Syst. Video Technol.*, vol. 23, no. 12, pp. 2009–2021, Dec. 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 Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 582–593, Feb. 2021.

[4] Y. Wei *et al.*, "STC: A simple to complex framework for weakly-supervised semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2314–2320, Nov. 2017.

[5] V. Mahadevan and N. Vasconcelos, "Biologically inspired object tracking using center-surround saliency mechanisms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 3, pp. 541–554, Mar. 2013.

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

[7] J. Han, K. N. Ngan, M. Li, and H.-J. Zhang, "Unsupervised extraction of visual attention objects in color images," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 1, pp. 141–145, Jan. 2006.

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

[9] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[10] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[11] C. Deng, X. Yang, F. Nie, and D. Tao, "Saliency detection via a multiple self-weighted graph-based manifold ranking," *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 885–896, Apr. 2020.

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

[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 *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8779–8788.

[14] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[15] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.

[16] J. Wei, S. Wang, and Q. Huang, "F^{3}Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12321–12328.

[17] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 1232–1241.

[18] G. E. Hinton, S. Sabour, and N. Frosst, "Matrix capsules with EM routing," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 3856–3866.

[19] J. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 1, 2020, doi: 10.1109/TPAMI.2020.2983686.

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

[21] H. Lu, X. Li, L. Zhang, X. Ruan, and M. H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Trans. Image Process.*, vol. 25, no. 4, pp. 1592–1603, Apr. 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 Trans. Image Process.*, vol. 25, no. 1, pp. 9–23, Jan. 2016.

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

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

[25] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13022–13031.

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

[27] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and A. Kwong, "Going from RGB to RGBD saliency: A depth-guided transformation model," *IEEE Trans. Cybern.*, vol. 50, no. 8, pp. 3627–3639, Aug. 2020.

[28] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[29] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep sub-region network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 2, pp. 728–741, Feb. 2021.

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

[31] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng, "SAC-Net: Spatial attenuation context for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1079–1090, Mar. 2021.

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

[33] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3183–3192.

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

[35] L. Zhu *et al.*, "Aggregating attentional dilated features for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3358–3371, Oct. 2020.

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

[37] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.

[38] A. Kosiorek, S. Sabour, Y. W. Teh, and G. E. Hinton, "Stacked capsule autoencoders," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15512–15522.

[39] B. Zhang, D. Xiong, and J. Su, "Neural machine translation with deep attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 1, pp. 154–163, Jan. 2020.

[40] J. Yu *et al.*, "Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval," *IEEE Trans. Multimedia*, vol. 22, no. 12, pp. 3196–3209, Dec. 2020.

[41] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[42] C. Yan *et al.*, "Task-adaptive attention for image captioning," *IEEE Trans. Circuits Syst. Video Technol.*, early access, Mar. 19, 2021, doi: 10.1109/TCSVT.2021.3067449.

[43] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3668–3677.

[44] C. Li *et al.*, "ASIF-net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 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 Trans. Image Process.*, vol. 29, pp. 3763–3776, 2020.

[46] Z. Chen, R. Cong, Q. Xu, and Q. Huang, "DPANet: Depth potentiality-aware gated attention network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 7012–7024, 2021.

[47] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 Trans. 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[50] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.

[51] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1741–1750.

[53] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 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, Feb. 1912.

[55] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

[56] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[59] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 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 *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1–7.

[61] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "A new way to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, Art. no. 245484557.

[62] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

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

[64] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9141–9150.

[66] Z. Chen, Q. Xu, and R. Cong, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10599–10606.

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

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

[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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1448–1457.

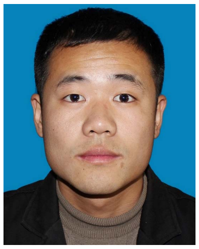
[71] Y. Xu *et al.*, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3789–3798.

[72] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2881–2890.

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

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

Yongjiang Luo received the B.S. degree in automatic control and the M.S. and Ph.D. degrees in circuit and system from Xidian University, China, in 2001, 2004, and 2011, respectively. He was a Visiting Scholar with the University of California at Merced, Merced, USA. He is currently an Associate Professor with the School of Electronic Engineering, Xidian University. His current research interests include wideband signal processing and intelligent information processing.



Yi Liu received the B.S. degree from Nanjing Institute of Technology, Nanjing, China, in 2012, the M.S. degree from Dalian University, Dalian, China, in 2015, and the Ph.D. degree from Xidian University, Xi'an, China, in 2019. He was a Visiting Student at Lancaster University from September 2018 to September 2019. He is currently working at Changzhou University. His current research interests include computer vision and machine learning.

Jungong Han is currently a Full Professor and the Chair of Computer Science with Aberystwyth University, U.K. He has published over 180 papers, including more than 40 IEEE TRANSACTIONS and more than 40 A*STAR conference papers. His research interests span the fields of video analysis, computer vision, and applied machine learning.