Computational Visual Media
https://doi.org/10.1007/s41095-022-0268-6 Vol. 9, No. 2, June 2023, 297–317

Research Article

Specificity-preserving RGB-D saliency detection

Tao Zhou[1,2], **Deng-Ping Fan**[3] (✉), **Geng Chen**[4], **Yi Zhou**[5], and **Huazhu Fu**[6]

Abstract Salient object detection (SOD) in RGB and depth images has attracted increasing research interest. Existing RGB-D SOD models usually adopt fusion strategies to learn a shared representation from RGB and depth modalities, while few methods explicitly consider how to preserve modality-specific characteristics. In this study, we propose a novel framework, the specificity-preserving network (SPNet), which improves SOD performance by exploring both the shared information and modality-specific properties. Specifically, we use two modality-specific networks and a shared learning network to generate individual and shared saliency prediction maps. To effectively fuse cross-modal features in the shared learning network, we propose a cross-enhanced integration module (CIM) and propagate the fused feature to the next layer to integrate cross-level information. Moreover, to capture rich complementary multi-modal information to boost SOD performance, we use a multi-modal feature aggregation (MFA) module to integrate the modality-specific features from each individual decoder into the shared decoder. By using skip connections between encoder and decoder layers, hierarchical features can be fully combined. Extensive experiments demonstrate that our SPNet outperforms cutting-edge approaches on six popular RGB-D SOD and three camouflaged object detection benchmarks. The project is publicly available at `https://github.com/taozh2017/SPNet`.

Keywords salient object detection (SOD); RGB-D; cross-enhanced integration module (CIM); multi-modal feature aggregation (MFA)

1 Introduction

Salient object detection (SOD, also called saliency detection) aims to emulate the mechanisms of human visual attention and locate the most visually distinctive object(s) in a given scene [1]. SOD has been widely applied in various vision-related tasks, such as image understanding [2], action recognition [3, 4], video/semantic segmentation [4, 5], and person re-identification [6]. Although significant progress has been made, it is still challenging to accurately locate salient objects in many challenging scenarios, such as images with cluttered backgrounds, low-contrast lighting conditions, and salient object(s) having a similar appearance to the background. Recently, with the ready availability of depth sensors in smart devices, depth maps have been introduced to provide geometric and spatial information to improve SOD performance. Consequently, fusing RGB and depth images has gained increasing interest in the SOD community [7–15], and it is a challenging task to adaptively fuse RGB and depth modalities.

Over past years, various RGB-D SOD methods have been proposed; they often focus on how to effectively fuse RGB and depth images. Existing fusion strategies can be divided into categories using early fusion, late fusion, and intermediate fusion. The *early fusion* strategy often adopts a simple concatenation to integrate the two modalities. For example, methods in Refs. [1, 16–18] directly integrate RGB and depth images to form four-

1 School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: taozhou.ai@gmail.com.

2 Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai, China.

3 Computer Vision Lab, ETH Zürich, Zürich, Switzerland. E-mail: dengpfan@gmail.com, dengpingfan@mail.nankai.edu.cn (✉).

4 School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, China. E-mail: geng.chen.cs@gmail.com.

5 School of Computer Science and Engineering, Southeast University, Nanjing, China. E-mail: yizhou.szcn@gmail.com.

6 Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. E-mail: hzfu@ieee.org.

Manuscript received: 2021-10-19; accepted: 2022-01-01

channel input. However, this type of fusion does not consider the distribution gap between the two modalities, which could result in an inaccurate feature fusion. The *late fusion* strategy uses two parallel network streams to generate independent saliency maps for RGB and depth data, which are fused to obtain a final prediction map [19–21]. However, it is still challenging to capture the complex interactions between the two modalities.

Recent research mainly focuses on *intermediate fusion*, which utilizes two independent networks to learn intermediate features of the two modalities separately, and then the fused features are fed into a subsequent network or decoder (see Fig. 1(a)). Other methods carry out cross-modal fusion at multiple scales [22–28]. As a result, complex correlations can be effectively exploited from the two modalities. Further methods utilize depth information to enhance RGB features via an a auxiliary subnetwork [29–31] (see Fig. 1(b)). For example, Zhao et al. [30] introduced a contrast prior into a CNN-based architecture to enhance the depth information, which was then integrated with RGB features using a fluid pyramid integration module. Zhu et al. [31] utilized an independent subnetwork to extract depth-based features, which were then incorporated into the RGB network. The above methods focus on learning shared representations by fusing them and then use a decoder to generate the final saliency map. Furthermore, there is no supervised decoder to guide the depth-based feature learning [30, 31], which may prevent optimal depth features from being obtained. From a multi-modal learning perspective, several works [34–37] have shown that exploring both the shared information and modality-specific characteristics can improve model performance. However, in the RGB-D SOD community, few methods explicitly exploit modality-specific characteristics.

Thus, in this paper, we propose a novel RGB-D SOD framework, the *specificity-preserving network* (SPNet), which can effectively explore the shared information as well as capture modality-specific characteristics to improve the SOD performance. Two encoder subnetworks are used to extract multi-scale features for the two modalities (i.e., RGB and depth), and a cross-enhanced integration module (CIM) is proposed to integrate cross-modal features in different feature layers. Then, we use a simple U-Net [38] structure to construct a modality-specific decoder, in which skip connections between the encoder and decoder layers are used to combine hierarchical features. In this way, we can learn powerful modality-specific features in each independent decoder, which

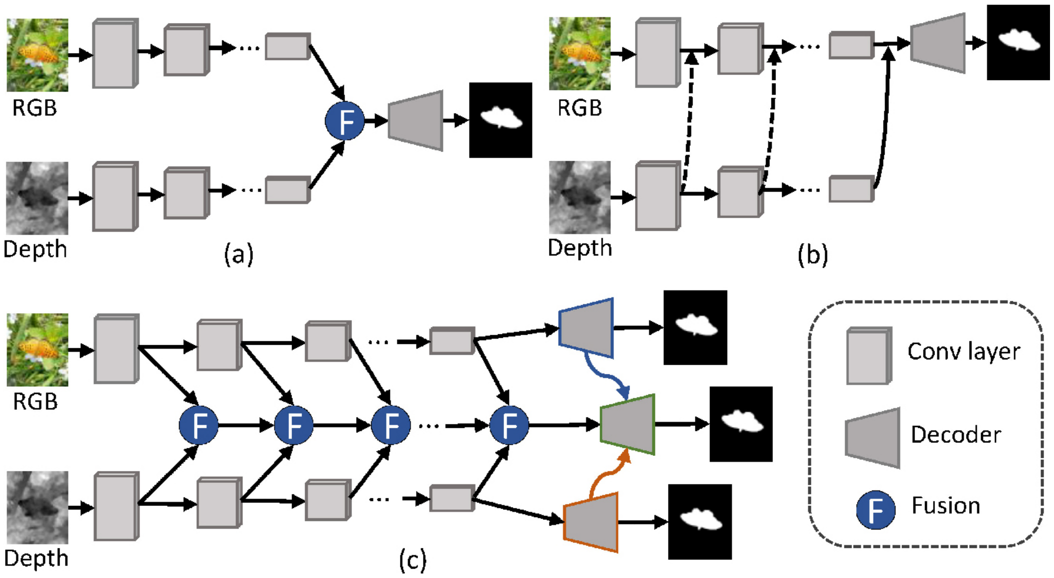


Fig. 1 Comparison of two existing RGB-D salient object detection frameworks and our proposed model. (a) RGB and depth images are fed into two independent network streams, and then fused high-level features are fed into a decoder to predict saliency maps (e.g., Refs. [22–25]). (b) Depth features are integrated into the RGB network using an auxiliary subnetwork (e.g., Refs. [29–33]). (c) Our method adopts two modality-specific networks and a shared learning network to explicitly explore modality-specific characteristics and shared information. Features learned from the modality-specific decoders are integrated into the shared decoder to boost SOD performance.

also captures modality-specific characteristics to provide cross-modal complementarity. Further, we construct a shared decoder to combine hierarchical features from outputs of the previous CIM via a skip connection. To make full use of the modality-specific features, a multi-modal feature aggregation module (MFA) is proposed to integrate them into the shared decoder. Finally, we formulate a unified and end-to-end trainable framework where shared and modality-specific information are simultaneously exploited to boost SOD performance.

The main contributions of our paper in summary are:

- A novel RGB-D salient object detection framework, the specificity-preserving network (SPNet), which explores shared information from RGB and depth images as well as preserving modality-specific characteristics.
- A cross-enhanced integration module (CIM) to integrate cross-modal features and learn shared representations for the two modalities. The output of each CIM is propagated to the next layer to explore rich cross-level information.
- An effective multi-modal feature aggregation (MFA) module to integrate learned modality-specific features. It allows our model to make full use of the features learned in the modality-specific decoder to improve salient object detection.
- Extensive experiments on six public RGB-D SOD and three camouflaged object detection (COD) datasets demonstrate the superiority of our model over other cutting-edge methods. Moreover, we carry out an attribute-based evaluation on various state-of-the-art RGB-D SOD methods under varying conditions (e.g., number of salient objects, indoors or outdoors, lighting, and object scale), which has not been done previously.

This paper significantly extends our previous work in Ref. [39], as follows:

- We discuss differences between (i) our proposed CIM and existing fusion strategies, and (ii) the proposed CIM and MFA.
- We provide further details, including (i) a review of existing RGB SOD methods, (ii) a discussion of the importance of integrating multi-level/scale features, and (iii) better characterisation of our evaluation metrics.
- We provide an additional ablation study and attribute-based evaluation, to validate the effectiveness of the shared decoder, and to examine the effects of different numbers of CIMs. We also show that our model can effectively handle variations in object scale.
- We apply SPNet to a new RGB-D task: COD, and demonstrate its superiority over existing methods.

2 Related work

In this section, we review three types of works most related to the proposed model, i.e., RGB salient object detection, RGB-D salient object detection, and multi-modal learning.

2.1 RGB salient object detection

Early salient object detection methods were based on hand-crafted features and various saliency priors, such as a background prior [40], color contrast [41], a compactness prior [42], and a center prior [43]. However, the generalizability and effectiveness of these traditional methods are limited. With the breakthrough of deep learning in the field of computer vision, various deep learning-based salient object detection methods have been developed with promising results. For example, Hou et al. [44] proposed a novel salient object detection method by introducing short connections to the skip-layer structures within the holistically-nested edge detector architecture. Wang et al. [45] proposed a recurrent fully convolutional network framework for salient object detection with promising results. Liu et al. [46] proposed to hierarchically embed global and local context modules into the top–down pathway, which can generate attention over context regions for each pixel. Deng et al. [47] proposed a recurrent residual refinement network with residual refinement blocks to accurately detect salient objects. Further methods can be found in a survey [48]. Scale variation is a key challenge in the SOD task, so several methods have been proposed to integrate multi-level or scale features [49–52] to improve SOD results. In our method, we consider how to effectively combine cross-modal (RGB and depth) features, and how multi-level information can be exploited via a cross-enhanced integration module.

2.2 RGB-D salient object detection

Early RGB-D based SOD methods often extracted hand-crafted features from the input RGB-D data. For example, Lang et al. [53] in the first RGB-D SOD work utilized Gaussian mixture models to model the distribution of depth-induced saliency. Subsequently, several methods explored different principles, such as center-surround difference [19, 54], contrast [1, 16, 55], a center/boundary prior [56, 57], and background enclosure [58]. However, these methods typically provide poor results due to the limited expressivity of handcrafted features. Benefiting from the rapid development of deep convolutional neural networks (CNNs), several deep learning-based works [7, 12, 30, 59, 60] have recently obtained promising results. For example, Qu et al. [59] used a CNN model to fuse saliency cues from different low levels into hierarchical features to boost SOD abilities. Chen and Li [22] proposed a complementarity-aware fusion module to effectively integrate cross-modal and cross-level features for RGB and depth modalities. Piao et al. [60] proposed a depth-induced multi-scale recurrent attention network to enhance cross-modality feature fusion. Fan et al. [7] designed a depth purification unit to remove some low-quality depth maps. Most other models [23–26, 61, 62] employ cross-modal fusion at multiple scales using different integration strategies.

2.3 Multi-modal learning

Recently, multi-modal (or multi-view) learning has attracted increasing attention: much data can be collected from multiple sources or represented using different types of features. One traditional strategy directly concatenates feature vectors from such multi-modal data into a feature vector. However, this may fail to exploit the complex correlations within multi-modal data. Thus, several multi-modal learning methods have been developed to explicitly fuse the complementary information from different modalities to improve model results. These popular methods can be divided into three types: (i) co-training [63, 64] tries to minimize the disagreement between different modalities, (ii) multiple kernel learning [65] utilizes a predefined set of kernels for multiple modalities and integrates these modalities using the learned kernel weights, and (iii) subspace learning [66, 67] assumes that a latent subspace exists shared by different modalities, with one underlying latent representation. To effectively fuse multi-modal data, several deep learning-based models have also been explored. For example, Ngiam et al. [68] proposed to learn a shared representation from audio and video inputs. Eitel et al. [69] adopted two separate CNN streams for RGB and depth, combining them using a late fusion network for RGB-D object recognition. Hu et al. [34] presented a shared and individual multi-view learning algorithm to explore further properties of multi-modal data. Lu et al. [35] presented a shared-specific feature transfer framework to perform a cross-modal person ReID task.

3 Methodology

In this section, we first present the overall SPNet. Then we describe the two key components in our model, the modality-specific learning network and shared learning network, and finally provide the overall loss function.

3.1 Overview

Figure 2 shows the framework of our proposed specificity-preserving network for RGB-D SOD. First,

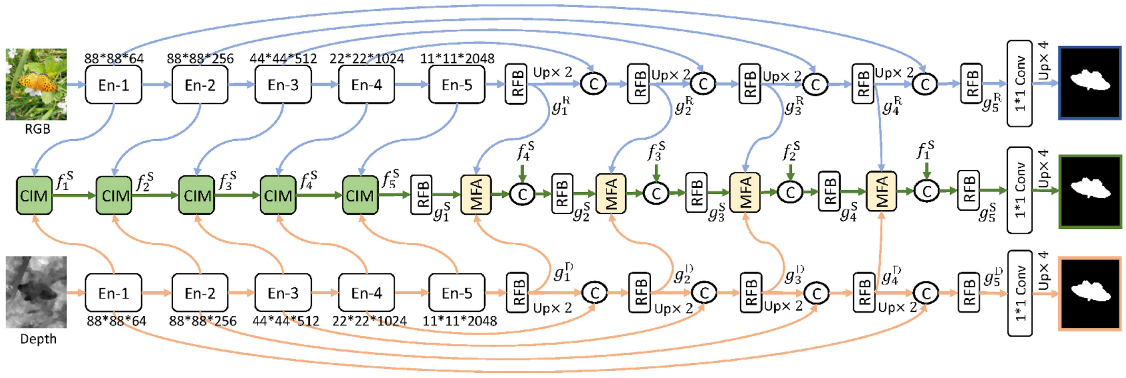


Fig. 2 Architecture of SPNet, consisting of two modality-specific learning networks and a shared learning network. The former preserve individual properties for RGB or depth, while the shared network fuses cross-modal features and explores complementary information. Skip connections combine hierarchical features between encoder and decoder layers. Learned features from the modality-specific decoder are integrated into the shared decoder to provide rich multi-modal complementary information, boosting saliency detection. C denotes feature concatenation.

RGB and depth images are fed into two stream modality-specific learning networks to obtain their multi-level feature representations, and a CIM learns their shared feature representation. Secondly, the individual and shared decoder subnetworks are each utilized to generate saliency prediction maps. The original features from the encoder networks are integrated into the decoder via skip connections. Finally, to make full use of the features learned by using the modality-specific decoder, an MFA module integrates these features into the shared decoder. We detail each key part below.

3.2 Modality-specific learning network

As Fig. 2 shows, the modality-specific subnetwork is built upon Res2Net-50 [70], pretrained on the ImageNet [71] dataset. Thus, there are five multi-level features, i.e., $F^{\mathrm{R}} = \{f_m^{\mathrm{R}}, m = 1, \cdots, 5]\}$ and $F^{\mathrm{D}} = \{f_m^{\mathrm{D}}, m = 1, \cdots, 5\}$, in the modality-specific encoder subnetworks for RGB and depth, respectively. The input resolution of the modality-specific encoder subnetwork is $W \times H$. Thus, we have a feature resolution of $(H/8) \times (W/8)$ for the first layer, and a general resolution of $(H/2^m) \times (W/2^m)$ (for $m > 1$). The number of channel features in the m-th layer is denoted C_m, where $C_m = [64, 256, 512, 1024, 2048]$.

After obtaining the high-level features f_5^{R} and f_5^{D}, they are then fed into the modality-specific decoder subnetworks to generate individual saliency maps. We further utilize a U-Net [38] structure to construct the modality-specific decoder, where the skip connections between encoder and decoder layers are used to combine hierarchical features. Moreover, the concatenated features (only f_5^{R} and f_5^{D} in the first stage of the decoder subnetwork) are fed to the receptive field block (RFB) [72] to capture global context information. This modality-specific learning network enables us to learn effective and powerful individual features for each modality by retaining its specific properties. These features are then integrated into the shared decoder subnetwork to improve saliency detection.

3.3 Shared learning network

3.3.1 Structure

As Fig. 2 shows, in the shared learning network, we fuse the cross-modal features from the RGB and depth modalities to learn their shared representation, which is fed into the shared decoder to generate the final saliency map. We again adopt skip connections between the encoder and decoder layers to combine hierarchical features. Moreover, to make full use of the features learned by the modality-specific decoder, we integrate them into the shared decoder to improve saliency detection.

3.3.2 Cross-enhanced integration module

Our CIM is used to effectively fuse cross-modal features. Let the width, height, and number of channels for the m-th layer be denoted W_m, H_m, and C_m, respectively. Taking $f_m^{\mathrm{R}} \in \mathbb{R}^{W_m \times H_m \times C_m}$ and $f_m^{\mathrm{D}} \in \mathbb{R}^{W_m \times H_m \times C_m}$ as an example, we use a 1×1 convolutional layer to reduce the number of channels to $C_m/2$ for speed. The CIM has two parts, for cross-modal feature enhancement and adaptive feature fusion. First, we use a cross-enhanced strategy to exploit correlations between the two modalities by learning their enhanced features. Specifically, as shown in Fig. 3, the two features are fed into a 3×3 convolutional layer with a sigmoid activation function to obtain the normalized feature maps, $w_m^{\mathrm{R}} = \sigma(\mathrm{Conv}_3(f_m^{\mathrm{R}})) \in [0, 1]$ and $w_m^{\mathrm{D}} = \sigma(\mathrm{Conv}_3(f_m^{\mathrm{R}})) \in [0, 1]$, where σ is the logistic sigmoid activation function. To exploit correlations between the two modalities, the normalized feature maps can be regarded as feature-level attention maps to adaptively enhance the feature representation. In this way, the feature map from one modality can be used to enhance the other modality. To preserve the original information of each modality, a residual connection is used to combine the enhanced features with the original features. Thus, the cross-enhanced feature representations for the two modalities are as Eq. (1):

$$\begin{cases} f_m^{\mathrm{R}'} = f_m^{\mathrm{R}} + f_m^{\mathrm{R}} \otimes w_m^{\mathrm{D}} \\ f_m^{\mathrm{D}'} = f_m^{\mathrm{D}} + f_m^{\mathrm{D}} \otimes w_m^{\mathrm{R}} \end{cases} \tag{1}$$

where $\otimes$ denotes element-wise multiplication.

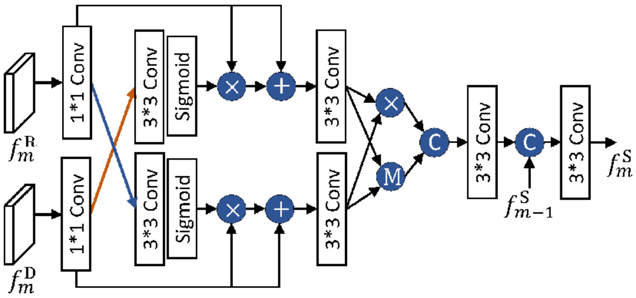


Fig. 3 Cross-enhanced integration module (CIM). C, +, ×, and M denote feature concatenation, element-wise addition, multiplication, and maximization, respectively.

Having obtained the cross-enhanced feature representations $f_m^{\mathrm{R}'}$ and $f_m^{\mathrm{D}'}$, the critical task is to effectively fuse them. Various strategies can be used to fuse features from different modalities, including element-wise multiplication and maximization. However, it is unclear which is best for specific tasks. In order to benefit from the advantages of different strategies, we apply element-wise multiplication and maximization, and concatenate the results. Specifically, the two features $f_m^{\mathrm{R}'}$ and $f_m^{\mathrm{D}'}$ are first fed into a 3×3 convolutional layer to obtain smooth representations, and then we carry out element-wise multiplication and maximization, giving

$$\begin{cases} p_{\mathrm{mul}} = \mathrm{BConv}_3(f_m^{\mathrm{R}'}) \otimes \mathrm{BConv}_3(f_m^{\mathrm{D}'}) \\ p_{\max} = \max(\mathrm{BConv}_3(f_m^{\mathrm{R}'}), \mathrm{BConv}_3(f_m^{\mathrm{D}'})) \end{cases} \tag{2}$$

where $\mathrm{BConv}(\cdot)$ is a sequential operation that applies 3×3 convolution followed by batch normalization, then a ReLU function. Then, we concatenate the results as $p_{\mathrm{cat}} = [p_{\mathrm{mul}}, p_{\max}] \in \mathbb{R}^{W_m \times H_m \times C_m}$, and obtain $p_{\mathrm{cat}}^1 = \mathrm{BConv}_3(p_{\mathrm{cat}})$ through a BConv_3 operation to adaptively weigh the two parts. Further, the output p_{cat}^1 is concatenated with the previous output f_{m-1}^{S} of the $(m-1)$-th CIM, and fed into the second BConv_3 operation. Finally, we obtain the output f_m^{S} of the m-th CIM. Note that, when $m = 1$, we do not need to use a 1×1 convolutional layer to reduce the number of channels. Furthermore, there is no previous output f_{m-1}^{S} when $m = 1$, so we only feed the concatenated features into a BConv_3 operation.

We note that our CIM can effectively exploit correlations between the two modalities via cross-enhanced feature learning, and fuse them by adaptively weighting the different feature representations. The fused feature representation f_m^{S} is propagated to the next layer to capture and integrate cross-level information. Some works [1, 17, 18] directly integrate RGB images and depth maps to form four-channel input (cascaded operation), and other methods carry out cross-modal fusion strategies, e.g., using attention-based fusion modules [24, 26], fusion-refinement modules (e.g., using summation) [23], etc. Unlike these methods, our proposed CIM mainly exploits the correlation between RGB and depth images, and then adaptively integrates enhanced cross-modal features to obtain a fused feature representation.

3.3.3 Multi-modal feature aggregation

To make full use of the features learned in the modality-specific decoder, we propose a simple but effective MFA module to integrate them into the shared decoder. Specifically, in the m-th layer of the shared decoder, we have the shared representation g_m^{S}, and the learned features g_m^{R} and g_m^{D} in the modality-specific decoder. As Fig. 4 shows, two features g_m^{R} and g_m^{D} are multiplied by the shared features of the current layer: $g_m^{\mathrm{RS}} = g_m^{\mathrm{S}} \otimes g_m^{\mathrm{R}}$ and $g_m^{\mathrm{DS}} = g_m^{\mathrm{S}} \otimes g_m^{\mathrm{D}}$. The two features are further concatenated ($[g_m^{\mathrm{DR}}, g_m^{\mathrm{DS}}]$) and then fed into a $\mathrm{BConv}(\cdot)$ operation to obtain g_m^{Sc}. Finally, we obtain the output of the MFA module to combine the convolutional feature g_m^{Sc} with the original feature g_m^{S} via an addition operation.

In the MFA, the learned modality-specific features are used to enhance the shared features and provide rich and complementary cross-modal information. Specifically, we use the two modality-specific features g_m^{R} and g_m^{D} to enhance g_m^{S}. More importantly, the modality-specific decoder is given a supervision signal to guide feature learning for modality-specific property preservation, which benefits the final prediction results when integrating them in the shared decoder. We also note the differences between the CIM and the MFA: the CIM is used to learn the fused multi-modal (RGB and depth) feature representation, while the MFA utilizes the learned modality-specific features to form an aggregate feature representation in the shared decoder.

3.4 Loss function

We may now formulate a unified, end-to-end trainable framework. The overall loss function has two parts, $\mathcal{L}_{\mathrm{sp}}$ and $\mathcal{L}_{\mathrm{sh}}$, for the modality-specific and decoders, respectively. For convenience, S_{R} and S_{D} denote the prediction maps for RGB and depth images, respectively, S_{sh} denotes the prediction map using their shared representation, and G denotes the ground

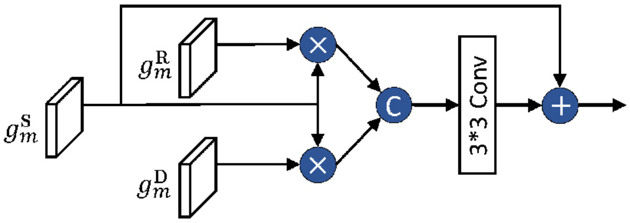


Fig. 4 Multi-modal feature aggregation (MFA) module. C, +, and × denote feature concatenation, element-wise addition, and element-wise multiplication respectively.

truth map. Therefore, the overall loss function can be formulated as Eq. (3):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sh}}(S_{\text{sh}}, G) + \mathcal{L}_{\text{sp}}(S_{\text{R}}, G) + \mathcal{L}_{\text{sp}}(S_{\text{D}}, G) \tag{3}$$

Here, we utilize the pixel position-aware loss [73] for $\mathcal{L}_{\text{sp}}$ and $\mathcal{L}_{\text{sh}}$, which can pay different attention to hard and easy pixels to improve results.

4 Experimental results and analysis

In this section, we first give the experimental setup, including datasets, evaluation metrics, and implementation details. Then we carry out a quan-titatively and qualitatively evaluation, as well as conducting ablation studies to validate the effectiveness of each key component. Finally, we conduct an attribute-based evaluation to show the effectiveness of our model in dealing with different challenges.

4.1 Experimental setup

4.1.1 Datasets

To validate the effectiveness of the proposed model, we have evaluated it on six public RGB-D SOD datasets: NJU2K [54], NLPR [1], DES [74], SSD [75], STERE [76], and SIP [7]. Details of each dataset can be found at `https://github.com/taozh2017/RGBD-SODsurvey`.

For a fair comparison, we utilized the same protocol to form the training and test sets as introduced in Refs. [7, 60]. The training set includes 2195 samples in total, with 1485 samples from NJU2K [54] and 700 samples from NLPR [1]. The remaining samples from NJU2K (500) and NLPR (300), and the entire DES (135), SSD (80), STERE (1000), and SIP (929) datasets were used for testing.

4.1.2 Evaluation metrics

We adopt four widely used metrics to evaluate the effectiveness of the proposed model. Their definitions are as follows.

- **Structure Measure**. The S-measure S_α [77] assesses the structural similarity between regional perception (S_r) and object perception (S_o), and is defined as

$$S_\alpha = \alpha S_\text{o} + (1-\alpha) S_\text{r} \tag{4}$$

where $\alpha \in [0,1]$ is a trade-off parameter, set to 0.5 by default [77].

- **F-measure**. Given a saliency map S, we convert it to a binary map M, and then compute the Precision and Recall [41] using

$$\text{Precision} = \frac{|M \cap G|}{|M|}, \quad \text{Recall} = \frac{|M \cap G|}{|G|} \tag{5}$$

where G denotes the ground truth. A popular strategy is to partition S using a set of thresholds varying from 0 to 255. For each threshold, we calculate a pair of recall and precision scores, and then combine all scores to obtain a PR curve.

The F-measure F_β [41] combines both precision and recall, via a weighted harmonic mean:

$$F_\beta = \left(1+\beta^2\right) \frac{\text{Precision} \times \text{Recall}}{\beta^2 \text{Precision} + \text{Recall}} \tag{6}$$

where β^2 is set to 0.3 to emphasize precision [41]. We use different fixed $[0, 255]$ thresholds to compute the F-measure. This yields a set of F-measure values; we report the maximum F_β values from our experiments.

- **Enhanced-alignment Measure**. E_ϕ [78] is used to capture image-level statistics and local pixel matching information. It is defined as

$$E_\phi = \frac{1}{WH} \sum_{i=1}^{W} \sum_{i=1}^{H} \phi_{FM}(i,j) \tag{7}$$

where ϕ_{FM} denotes the enhanced-alignment matrix [78].

- **Mean Absolute Error** ($\mathcal{M}$). It is adopted to evaluate the average pixel-level relative error between the ground truth (i.e., G) and normalized prediction (i.e., S), which is defined by

$$\mathcal{M} = \frac{1}{W*H} \sum_{i=1}^{W} \sum_{i=1}^{H} |S(i,j) - G(i,j)| \tag{8}$$

where W and H denote the width and height of the map, respectively. $\mathcal{M}$ estimates the similarity between the saliency map and the ground-truth map, and normalizes it to $[0,1]$.

4.1.3 Implementation details

Our proposed model was implemented with the PyTorch library, and trained on an nVidia Tesla V100 GPU with 32 GB memory. Res2Net-50 [70], pre-trained on ImageNet [71], was used as the backbone network. Since RGB and depth images have different numbers of channels, the input channel for the depth encoder was modified to 1. We utilized the Adam algorithm to optimize the proposed model. The initial learning rate was set to 10^{-4} and divided by 10 every 60 epochs. The input RGB and depth images were resized to 352×352. To enhance the generalizability of the proposed learning algorithm, we adopted multiple data augmented strategies: random flipping, rotation, and border clipping. The batch size was set to 20 and the model was trained over 200 epochs.

For testing, the RGB and depth images were first resized to 352×352 and then fed into the model to obtain the predicted saliency map. The predicted saliency map was then resized back to the original size of the input images. The output of the shared decoder is regarded as the final prediction of our model.

4.2 Comparison

4.2.1 Models compared

We compared our proposed SPNet with 31 RGB-D saliency detection methods, including 8 handcrafted traditional models: LHM [1], ACSD [54], LBE [58], DCMC [80], SE [19], MDSF [17], CDCP [56], and DTM [81], and 23 deep models: DF [59], CTMF [25], PCF [22], AFNet [20], CPFP [30], MMCI [29], TANet [24], DMRA [60], cmSalGAN [82], ASIFNet [83], ICNet [84], A2dele [85], JLDCF [11], S2MA [86], UCNet [12], SSF [87], HDFNet [88], Cas-GNN [89], CMMS [61], D3Net [7], CoNet [90], DANet [91], and PGAR [92]. See also the survey in Ref. [10].

4.2.2 Quantitative evaluation

As Table 1 shows, our method is superior to the eight traditional methods LHM [1], ACSD [54], LBE [58], DCMC [80], SE [19], MDSF [17], CDCP [56], and DTM [81] by a large margin, on all six datasets. Our method furthermore outperforms all compared state-of-the-art methods and obtains the best performance in terms of the four evaluation metrics on NJU2K, DES, and SIP datasets. It is worth noting that our model obtains better results on STERE and NLPR than most compared RGB-D saliency detection methods. Our model is also comparable with CoNet on the STERE dataset, and JLDCF and PGAR on the NLPR dataset. Overall, our proposed SPNet obtains promising results in locating salient object(s) in a given scene. We further show PR curves in Fig. 5 and F-measure curves in Fig. 6, giving results for

Table 1 Benchmarking results using 8 representative traditional models and 23 deep models on six public RGB-D saliency detection datasets using four widely used evaluation metrics: S_α [77], max E_ϕ [78], max F_β [41], and $\mathcal{M}$ [79]). ↑,↓ indicate that larger or smaller is better. The subscript for each model denotes the publication year. Best results are highlighted in bold

Model	NJU2K [54]				STERE [76]				DES [74]				NLPR [1]				SSD [75]				SIP [7]			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$\mathcal{M}\downarrow$
LHM_{14} [1]	0.514	0.632	0.724	0.205	0.562	0.683	0.771	0.172	0.562	0.511	0.653	0.114	0.630	0.622	0.766	0.108	0.566	0.568	0.717	0.195	0.511	0.574	0.716	0.184
ACSD_{14} [54]	0.699	0.711	0.803	0.202	0.692	0.669	0.806	0.200	0.728	0.756	0.850	0.169	0.673	0.607	0.780	0.179	0.675	0.682	0.785	0.203	0.732	0.763	0.838	0.172
LBE_{16} [58]	0.695	0.748	0.803	0.153	0.660	0.633	0.787	0.250	0.703	0.788	0.890	0.208	0.762	0.745	0.855	0.081	0.621	0.619	0.736	0.278	0.727	0.751	0.853	0.200
DCMC_{16} [80]	0.686	0.715	0.799	0.172	0.731	0.740	0.819	0.148	0.707	0.666	0.773	0.111	0.724	0.648	0.793	0.117	0.704	0.711	0.786	0.169	0.683	0.618	0.743	0.186
SE_{16} [19]	0.664	0.748	0.813	0.169	0.708	0.755	0.846	0.143	0.741	0.741	0.856	0.090	0.756	0.713	0.847	0.091	0.675	0.710	0.800	0.165	0.628	0.661	0.771	0.164
MDSF_{17} [17]	0.748	0.775	0.838	0.157	0.728	0.719	0.809	0.176	0.741	0.746	0.851	0.122	0.805	0.793	0.885	0.095	0.673	0.703	0.779	0.192	0.717	0.698	0.798	0.167
CDCP_{17} [56]	0.669	0.621	0.741	0.180	0.713	0.664	0.786	0.149	0.709	0.631	0.811	0.115	0.669	0.621	0.741	0.180	0.603	0.535	0.700	0.214	0.595	0.505	0.721	0.224
DTM_{20} [81]	0.706	0.716	0.799	0.190	0.747	0.743	0.837	0.168	0.752	0.697	0.858	0.123	0.733	0.677	0.833	0.145	0.677	0.651	0.773	0.199	0.690	0.659	0.778	0.203
DF_{17} [59]	0.763	0.804	0.864	0.141	0.757	0.757	0.847	0.141	0.752	0.766	0.870	0.093	0.802	0.778	0.880	0.085	0.747	0.735	0.828	0.142	0.653	0.657	0.759	0.185
CTMF_{18} [25]	0.849	0.845	0.913	0.085	0.848	0.831	0.912	0.086	0.863	0.844	0.932	0.055	0.860	0.825	0.929	0.056	0.776	0.729	0.865	0.099	0.716	0.694	0.829	0.139
PCF_{18} [22]	0.877	0.872	0.924	0.059	0.875	0.860	0.925	0.064	0.842	0.804	0.893	0.049	0.874	0.841	0.925	0.044	0.841	0.807	0.894	0.062	0.842	0.838	0.901	0.071
AFNet_{19} [20]	0.772	0.775	0.853	0.100	0.825	0.823	0.887	0.075	0.770	0.729	0.881	0.068	0.799	0.771	0.879	0.058	0.714	0.687	0.807	0.118	0.720	0.712	0.819	0.118
CPFP_{19} [30]	0.878	0.877	0.923	0.053	0.879	0.874	0.925	0.051	0.872	0.846	0.923	0.038	0.888	0.867	0.932	0.036	0.807	0.766	0.852	0.082	0.850	0.851	0.903	0.064
MMCI_{19} [29]	0.859	0.853	0.915	0.079	0.873	0.863	0.927	0.068	0.848	0.822	0.928	0.065	0.856	0.815	0.913	0.059	0.813	0.781	0.882	0.082	0.833	0.818	0.897	0.086
TANet_{19} [24]	0.878	0.874	0.925	0.060	0.871	0.861	0.923	0.060	0.858	0.827	0.910	0.046	0.886	0.863	0.941	0.041	0.839	0.810	0.897	0.063	0.835	0.830	0.895	0.075
DMRA_{19} [60]	0.886	0.886	0.927	0.051	0.886	0.886	0.938	0.047	0.900	0.888	0.943	0.030	0.899	0.879	0.947	0.031	0.857	0.844	0.906	0.058	0.806	0.821	0.875	0.085
cmSalGAN_{20} [82]	0.903	0.896	0.940	0.046	0.900	0.894	0.936	0.050	0.913	0.899	0.943	0.028	0.922	0.907	0.957	0.027	0.791	0.735	0.867	0.086	0.865	0.864	0.906	0.064
ASIFNet_{20} [83]	0.889	0.888	0.927	0.047	0.878	0.878	0.927	0.049	0.934	0.935	0.974	0.019	0.906	0.888	0.944	0.030	0.857	0.834	0.884	0.056	0.857	0.859	0.896	0.061
ICNet_{20} [84]	0.894	0.891	0.926	0.052	0.903	0.898	0.942	0.045	0.920	0.913	0.960	0.027	0.923	0.908	0.952	0.028	0.848	0.841	0.902	0.064	0.854	0.857	0.903	0.069
A2dele_{20} [85]	0.871	0.874	0.916	0.051	0.878	0.879	0.928	0.044	0.886	0.872	0.920	0.029	0.898	0.882	0.944	0.029	0.802	0.776	0.861	0.070	0.828	0.833	0.889	0.070
JLDCF_{20} [11]	0.903	0.903	0.944	0.043	0.905	0.901	0.946	0.042	0.929	0.919	0.968	0.022	0.925	0.916	0.962	0.022	0.830	0.795	0.885	0.068	0.879	0.885	0.923	0.051
S2MA_{20} [86]	0.894	0.889	0.930	0.053	0.890	0.882	0.932	0.051	0.941	0.935	0.973	0.021	0.915	0.902	0.953	0.030	0.868	0.848	0.909	0.052	0.872	0.877	0.919	0.057
UCNet_{20} [12]	0.897	0.895	0.936	0.043	0.903	0.899	0.944	0.039	0.933	0.930	0.976	0.018	0.920	0.903	0.956	0.025	0.865	0.854	0.907	0.049	0.875	0.879	0.919	0.051
SSF_{20} [87]	0.899	0.896	0.935	0.043	0.893	0.890	0.936	0.044	0.904	0.884	0.941	0.026	0.914	0.896	0.953	0.026	0.845	0.824	0.897	0.058	0.876	0.882	0.922	0.052
HDFNet_{20} [88]	0.908	0.911	0.944	0.038	0.900	0.900	0.943	0.041	0.926	0.921	0.970	0.021	0.923	0.917	**0.963**	0.023	**0.879**	0.870	**0.925**	0.045	0.886	0.894	0.930	0.047
Cas-GNN_{20} [89]	0.911	0.903	0.933	0.035	0.899	0.901	0.930	0.039	0.905	0.906	0.947	0.028	0.919	0.904	0.947	0.028	0.872	0.862	0.915	0.047	0.875	0.879	0.919	0.051
CMMS_{20} [61]	0.900	0.897	0.936	0.044	0.895	0.893	0.939	0.043	0.937	0.930	0.976	0.018	0.915	0.896	0.949	0.027	0.874	0.864	0.922	0.046	0.872	0.877	0.911	0.058
CoNet_{20} [90]	0.895	0.893	0.937	0.046	**0.908**	0.905	**0.949**	0.040	0.909	0.896	0.945	0.028	0.908	0.887	0.945	0.031	0.853	0.840	0.915	0.059	0.858	0.867	0.913	0.063
DANet_{20} [91]	0.899	0.910	0.935	0.045	0.901	0.892	0.937	0.043	0.924	0.928	0.968	0.023	0.915	0.916	0.953	0.028	0.864	0.866	0.914	0.050	0.875	0.892	0.918	0.054
PGAR_{20} [92]	0.909	0.907	0.940	0.042	0.907	0.898	0.939	0.041	0.913	0.902	0.945	0.026	**0.930**	0.916	0.961	0.024	0.865	0.838	0.898	0.057	0.876	0.876	0.915	0.055
D3Net_{21} [7]	0.900	0.900	0.950	0.041	0.899	0.891	0.938	0.046	0.898	0.885	0.946	0.031	0.912	0.897	0.953	0.030	0.857	0.834	0.910	0.058	0.860	0.861	0.909	0.063
SPNet (ours)	**0.925**	**0.935**	**0.954**	**0.028**	0.907	**0.915**	0.944	**0.037**	**0.945**	**0.950**	**0.980**	**0.014**	0.927	**0.925**	0.959	**0.021**	0.871	**0.883**	0.915	**0.044**	**0.894**	**0.916**	**0.930**	**0.043**

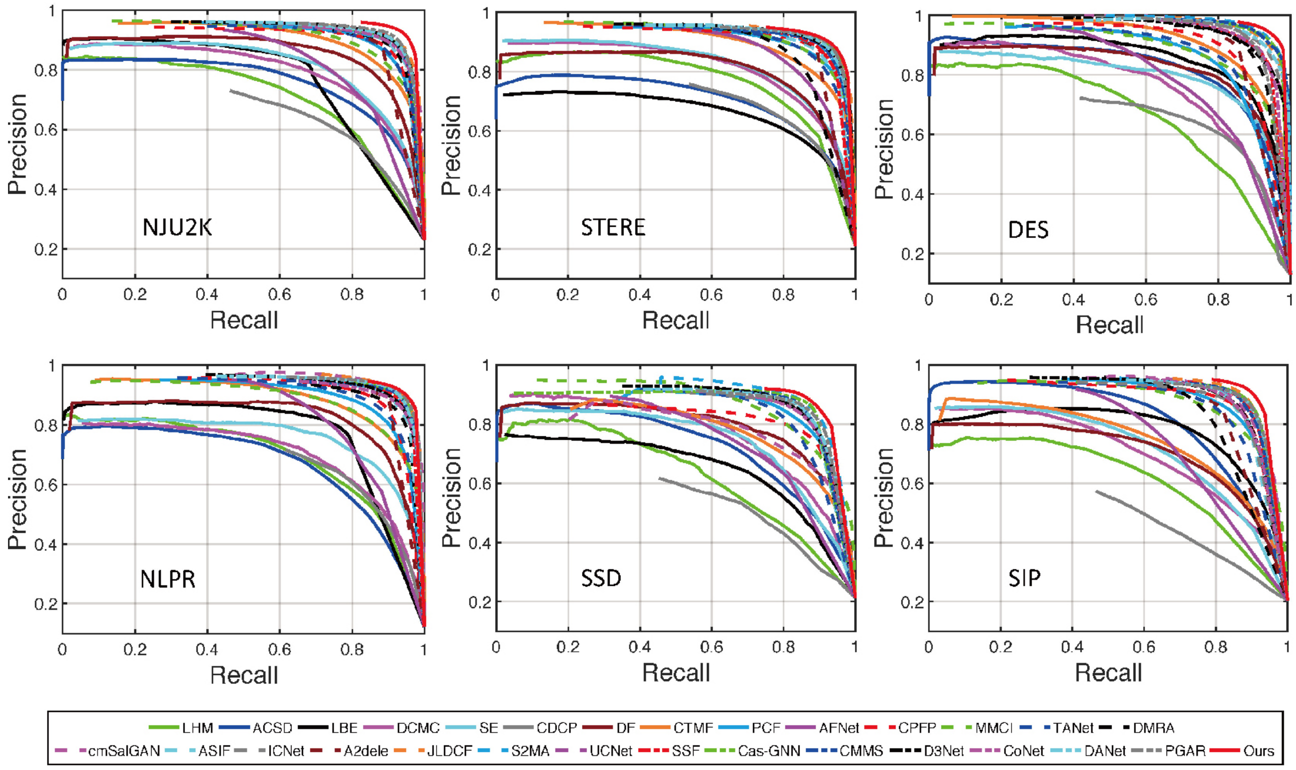


Fig. 5 PR curves for six datasets: NJU2K [54], STERE [76], DES [74], NLPR [1], SSD [75], and SIP [7].

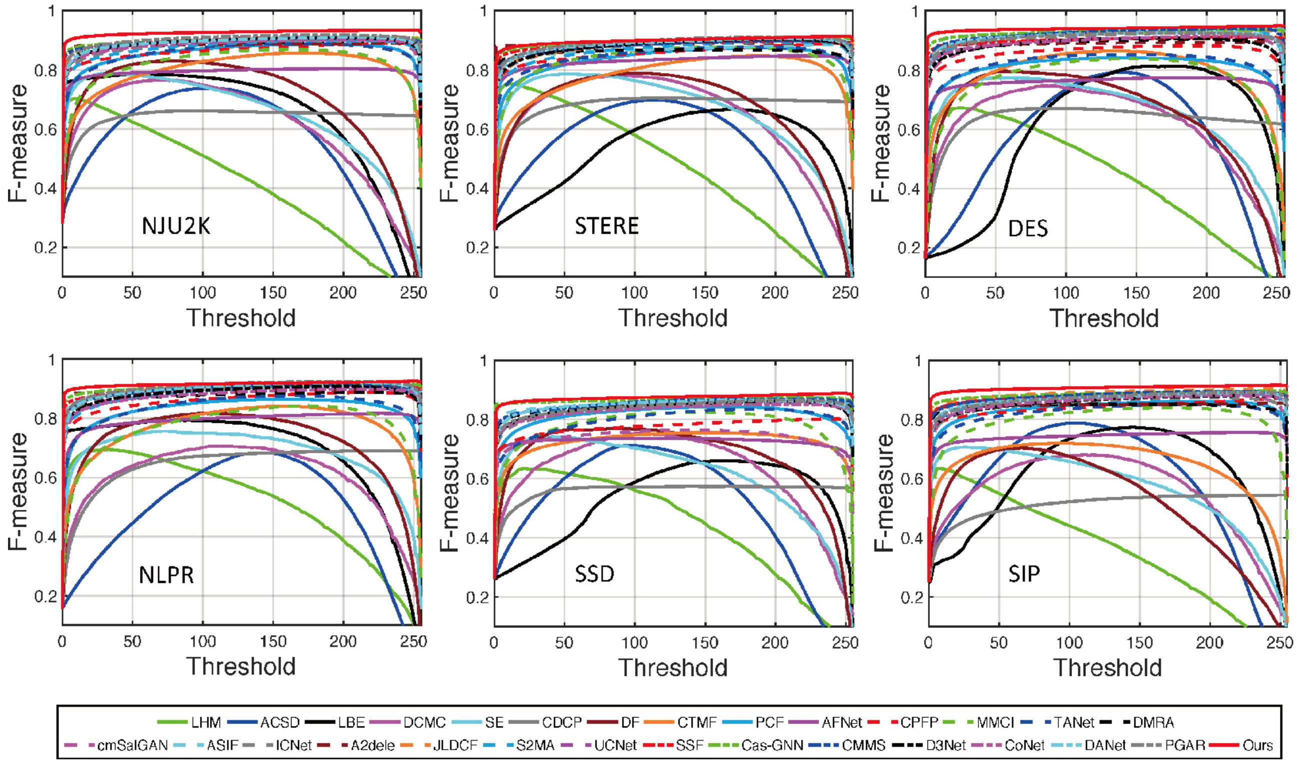


Fig. 6 F-measure curves for different thresholds, for NJU2K [54], STERE [76], DES [74], NLPR [1], SSD [75], and SIP [7].

29 RGB-D saliency detection methods, including 28 state-of-the-art models with complete saliency maps. The superiority of our model is clearly visible on these datasets.

In addition, we compared our SPNet to 13 recent state-of-the-art models on the ReDWeb-S dataset. Results for the other methods are from `https://github.com/nnizhang/SMAC`, while results for our method were obtained by testing the model (trained using NJU2K [54] and NLPR [1]) on the ReDWeb-S dataset. The comparison is shown in Table 2. Our method works better than most compared methods, and is comparable to UCNet and JLDCF on the ReDWeb-S dataset.

We further compared using different backbone networks in the proposed model, with the results shown in Table 3. The proposed model works better when using Res2Net-50 as the backbone, yet the model using ResNet-50 as backbone still performs better than other methods (see Table 1).

4.2.3 *Qualitative evaluation*

Figure 7 shows several representative samples of results comparing our model with those from eight state-of-the-art methods. The first row shows a scene with a small object. Our method, A2dele, PGAR, and D3Net accurately detect the salient object, while JLDCF, S2MA, SSF, and UCNet predict some non-object regions. Rows 2 and 3 show two examples of scenes with complex backgrounds. Our method and S2MA produce reliable results, while other RGB-D saliency detection models fail to locate the object or confuse the background with a salient object. In row

Table 2 Results from our model and 13 state-of-the art methods: CTMFF [25], PCF [22], AFNet [20], MMCI [29], CPFP [30], DMRA [60], TANet [24], A2dele [85], UCNet [12], JLDCF [11], S2MA [86], SSF [87], and D3Net [7]) on the ReDWeb-S dataset

Model	CTMF	PCF	AFNet	MMCI	CPFP	DMRA	TANet	A2dele	UCNet	JLDCF	S2MA	SSF	D3Net	Ours
$S_\alpha \uparrow$	0.641	0.655	0.546	0.660	0.685	0.592	0.656	0.641	0.713	0.734	0.711	0.595	0.689	0.710
$F_\beta \uparrow$	0.607	0.627	0.549	0.641	0.645	0.579	0.623	0.603	0.710	0.727	0.696	0.558	0.673	0.715
$E_\phi \uparrow$	0.739	0.743	0.693	0.754	0.744	0.721	0.741	0.672	0.794	0.805	0.781	0.710	0.768	0.800
$\mathcal{M} \downarrow$	0.204	0.166	0.213	0.176	0.142	0.188	0.165	0.160	0.130	0.128	0.139	0.189	0.149	0.129

Table 3 Results from our model using different backbone networks

Backbone	NJU2K [54]				STERE [76]				DES [74]				NLPR [1]				SSD [75]				SIP [7]			
	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$F_\beta \uparrow$	$E_\phi \uparrow$	$\mathcal{M} \downarrow$
ResNet-50	0.922	0.934	0.952	0.030	0.904	0.914	0.942	0.037	0.936	0.944	0.974	0.016	0.930	0.931	0.965	0.020	0.869	0.876	0.906	0.044	0.896	0.916	0.934	0.041
Res2Net-50	0.925	0.935	0.954	0.028	0.907	0.915	0.944	0.037	0.945	0.950	0.980	0.014	0.927	0.925	0.959	0.021	0.871	0.883	0.915	0.044	0.894	0.916	0.930	0.043

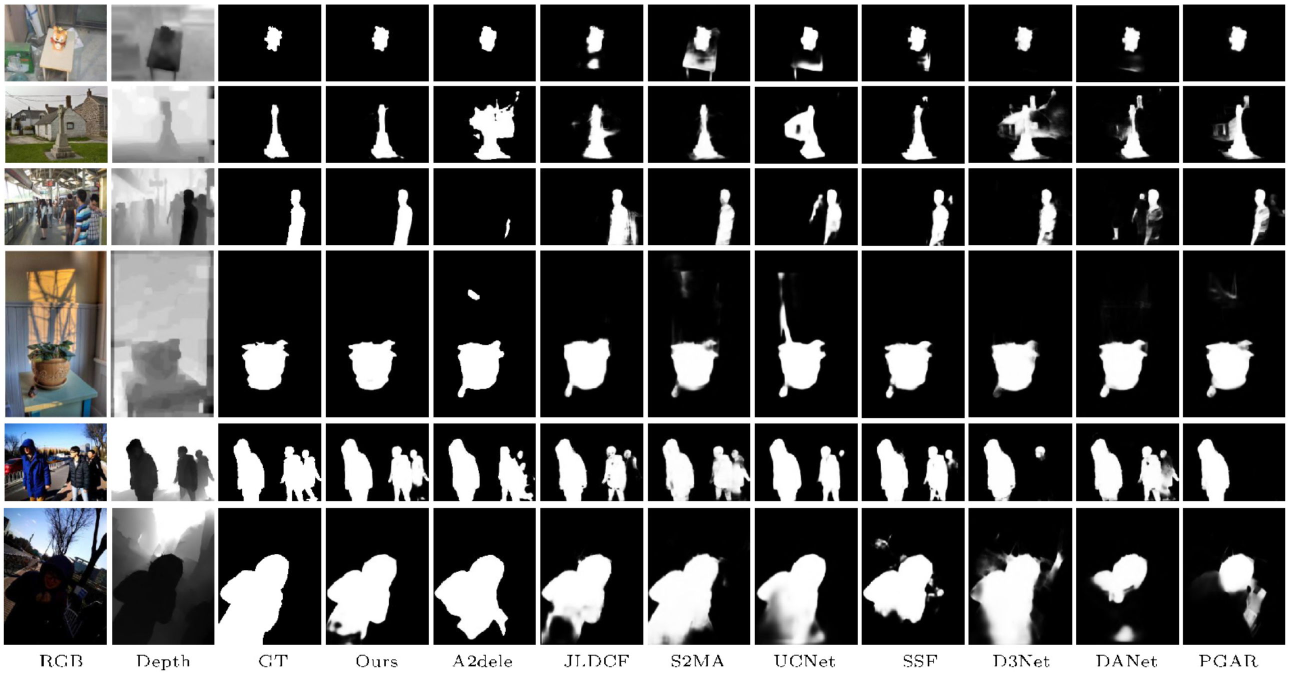


Fig. 7 Visual comparison of results from our method and eight state-of-the-art methods: A2dele [85], JLDCF [11], S2MA [86], UCNet [12], SSF [87], D3Net [7], DANet [91], and PGAR [92].

4, the compared methods other than D3Net locate a non-salient and small object. In row 5, we show an example with multiple salient objects, where it is challenging to accurately locate them all. Our method locates all salient objects and segments them more accurately, generating sharper edges than other approaches. We show an example under low-light conditions in the last row. While some approaches fail to detect the entire extent of the salient object, our model suppresses background distractors and gives good saliency detection results.

4.2.4 Inference time and model size

We tested the inference time for different methods on an NVIDIA TESLA P40 GPU with 24 GB memory. The inference time and model size of different methods, including our SPNet, JLDCF [11], S2MA [86], UCNet [12], SSF [87], and HDFNet [88], are shown in Table 4. Because our model adopts two modality-specific networks and a shared learning network to generate individual and shared saliency prediction maps, it has a relatively large model size and takes more inference time for saliency prediction than other methods. We thus hope to design light-weight networks to improve the efficiency of SPNet in future work.

4.3 Ablation studies

To verify the relative importance of different key components of our model, we conducted ablation studies by removing or replacing them.

4.3.1 Effectiveness of CIM

Since the proposed CIM is used to fuse cross-modal features and learn their shared representation, we compared it to an alternative of a direct concatenation strategy. Specifically, the two features f_m^{R} and f_m^{D} (see Fig. 3) are directly concatenated and then fed into a 3×3 convolutional layer to obtain the fused representation in each layer. We denote this approach as A1 in Table 5, which shows that our model performs better when using the proposed CIM than using a simple feature concatenation strategy. This also indicates the contribution of the CIM in improving the saliency detection results. Going further, there are two parts to CIM: cross-modal feature enhancement and adaptive feature fusion. Thus, to evaluate the contribution of each part, we modified CIM to have only cross-modal feature enhancement or adaptive feature fusion, with results denoted A2 and A3, respectively. When comparing them to the full version of CIM, we can see the effectiveness of the proposed CIM. Moreover, in CIM, the features of the last layer are propagated to the next layer to capture cross-level correlations. To validate the effectiveness of the propagation strategy, we removed this propagation in the CIM, with results denoted A4, showing that the propagation strategy does improve saliency detection results.

Table 4 Comparisons of inference time and model size for different methods

Method	Ours	JLDCF	S2MA
Model size (MB)	175.3	124.5	82.7
Inference time (ms)	91.7	21.8	22.1
Method	UCNet	SSF	HDFNet
Model size (MB)	31.3	32.9	153.2
Inference time (ms)	31.8	45.7	57.1

4.3.2 Effectiveness of MFA

In the proposed framework, the MFA is proposed to make full use of the features learned in the modality-specific decoder, which are then integrated into the shared decoder to provide more multi-modal complementary information. To validate its effectiveness, we deleted this module in an approach denoted B1. We also considered two other feature fusion strategies: see Fig. 8. One provides cross-modal feature enhancement fusion; the other is a simple concatenation strategy. Results for the two strategies are denoted B2 and B3. Table 5 demonstrates, by comparing results of B1 and our full model, the effectiveness of integrating the features learned into the shared decoder. Comparing results of B2 and B3 with our full model, we can see that the MFA module outperforms both other fusion strategies.

4.3.3 Effectiveness of modality-specific decoders

We deleted the two modality-specific decoders, with results shown in C1 in Table 5. Performance degrades when not using the two parts. This indicates the effectiveness of the modality-specific decoders, which provide supervision signals to ensure that modality-specific properties can be learned.

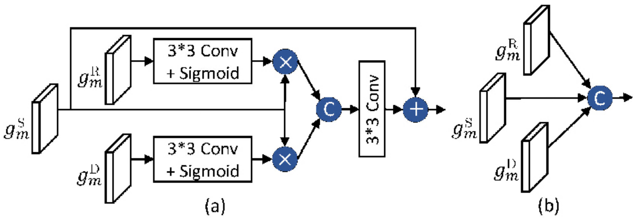


Fig. 8 Comparison of MFA module with other fusion strategies.

Table 5 Quantitative evaluation for ablation studies

	NJU2K [54]		STERE [76]		DES [74]		NLPR [1]		SSD [75]		SIP [7]	
	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$
Ours	**0.925**	**0.028**	**0.907**	**0.037**	**0.945**	**0.014**	**0.927**	**0.021**	**0.871**	**0.044**	**0.894**	**0.043**
A1	0.916	0.034	0.898	0.042	0.939	0.016	0.926	0.022	0.869	0.047	0.892	0.044
A2	0.921	0.031	0.895	0.042	0.938	0.016	0.925	0.022	0.865	0.051	0.896	0.042
A3	0.919	0.032	0.895	0.043	0.938	0.016	0.929	0.020	0.864	0.049	0.887	0.048
A4	0.924	0.029	0.903	0.038	0.930	0.019	0.927	0.023	0.867	0.049	0.888	0.046
B1	0.918	0.034	0.901	0.041	0.939	0.017	0.922	0.024	0.858	0.050	0.885	0.048
B2	0.924	0.029	0.900	0.041	0.941	0.015	0.926	0.022	0.864	0.049	0.893	0.044
B3	0.921	0.031	0.903	0.039	0.938	0.016	0.925	0.022	0.863	0.050	0.891	0.045
C1	0.913	0.037	0.900	0.047	0.935	0.019	0.922	0.025	0.861	0.055	0.880	0.051
C2	0.916	0.034	0.906	0.040	0.923	0.021	0.924	0.022	0.866	0.049	0.882	0.051

To further evaluate the effectiveness of the combination of the two modality-specific decoders, we added an experiment to compare the SOD results when using the output from the shared decoder and the combination of the two modality-specific decoders. Results are shown in C2 of Table 5. We can see that the shared decoder outperforms the combination of the two modality-specific decoders, indicating that the shared decoder can combine multi-modal shared information and modality-specific characteristics to improve SOD results.

4.3.4 Effects of varying numbers of CIMs

To investigate the effects of changing the numbers of CIMs, we compare our full model using five CIMs with two degraded versions, CIM_1, which only applies a CIM to the features from the last layer in the encoder network, and CIM_3, using CIMs on the features from each of the last three layers in the encoder network. Table 6 shows the results; our model with five CIMs works better for most datasets.

4.4 Attribute-based evaluation

There are several challenging factors that may affect results from RGB-D saliency detection models, such as the number of salient objects, indoor versus outdoor environments, lighting conditions, and so on. Thus, we evaluated saliency detection results under different conditions, to show the strengths and weaknesses of state-of-the-art models in handling these challenges.

4.4.1 Single vs. multiple objects

In this evaluation, we constructed a hybrid dataset with 1229 images from the NLPR [1] and SIP [7] datasets. Results using S_α are shown in Fig. 9(a). As can be observed, it is easier to detect a single salient object than several. Our model outperforms other state-of-the-art methods in locating single and multiple objects.

4.4.2 Indoor vs. outdoor

We evaluated the results of different RGB-D SOD models on indoor and outdoor scenes. As DES [74] and NLPR [1] include indoor and outdoor scenes, we constructed a hybrid dataset collected from the two datasets. Results are shown in Fig. 9(b). As can be observed, many models find it harder to detect salient objects in indoor scenes than outdoor scenes, while JLDCF, S2MA, UCNet, ICNet, SSF, DANet, and our model work a little better on outdoor scenes.

4.4.3 Lighting conditions

We carried out this evaluation on the SIP dataset [7], with examples grouped into two categories, sunny and low-light. Results are shown in Fig. 9(c). All models struggle to detect salient objects in low-light conditions, confirming that low-light negatively impacts SOD performance.

Table 6 Results for different numbers of CIMs

	NJU2K		STERE		DES		NLPR		SSD		SIP	
	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
CIM_1	0.918	0.034	0.908	0.039	0.929	0.019	0.928	0.022	0.865	0.047	0.889	0.046
CIM_3	0.920	0.032	0.900	0.041	0.935	0.017	0.928	0.021	0.857	0.049	0.891	0.045
Ours	**0.925**	**0.028**	**0.907**	**0.037**	**0.945**	**0.014**	**0.927**	**0.021**	**0.871**	**0.044**	**0.894**	**0.043**

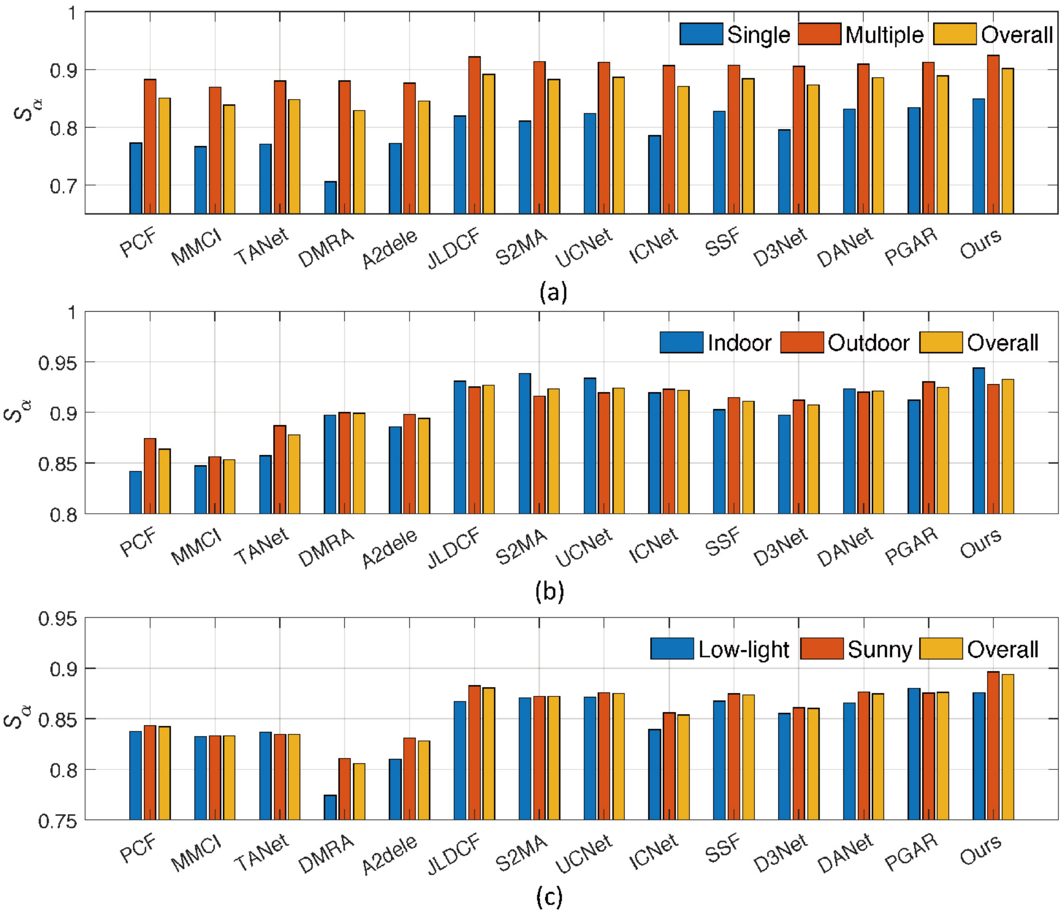


Fig. 9 Attribute-based evaluation with respect to (a) number of salient objects (one or multiple), (b) indoor versus outdoor environments, and (c) lighting conditions (low-light versus sunny).

4.4.4 Object scale

To characterize the scale of a salient object, we compute the ratio r of the size of the salient region to the whole image, and define three object scales: small, when $r < 0.1$, large, when $r > 0.4$, and medium otherwise. To evaluate how different methods handle scale variation, we constructed a hybrid dataset with 2444 images from STERE [76], NLPR [1], SSD [75], DES [74], and SIP [7]. Figure 10 shows results of this attribute-based evaluation with respect to the scales of the salient objects. All methods work better at detecting small salient objects and relatively at

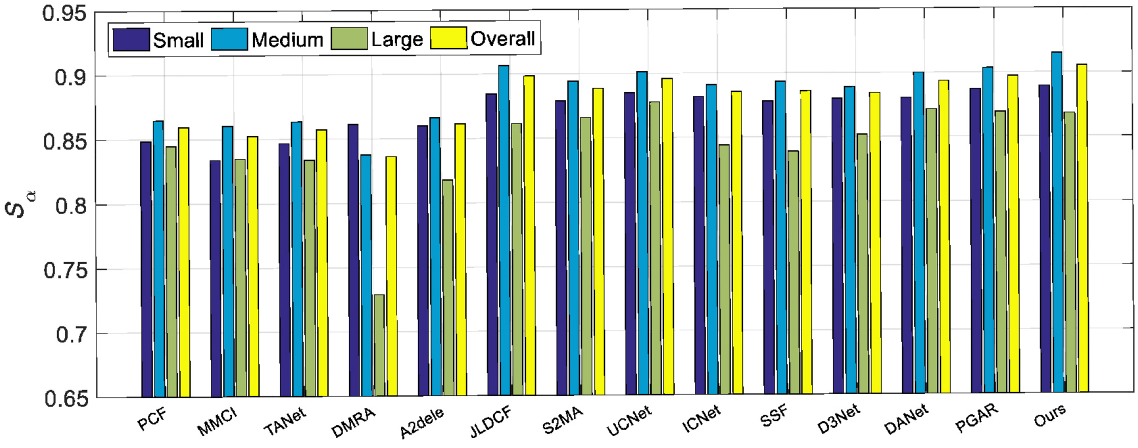


Fig. 10 Attribute-based evaluation with respect to scale of the salient object.

detecting large salient objects. The most recent models, JLDCF, DANet, PGAR, and our model, obtain the promising results.

4.5 Failures and discussion

Our proposed SPNet shows good RGB-D saliency detection in most cases. However, it fails to detect salient objects in some challenging scenes such as those with complex backgrounds and low-quality depth data. Some failures of our model are shown in Fig. 11. In the first row, the depth data quality is very poor, so our model can only roughly locate the boat without fine details. This suggests that it is helpful to enhance or filter depth maps to improve saliency detection results. In the second row, the annotated salient object has a similar appearance to other objects in the scene, so it is challenging to accurately detect the salient object. In the third row, the object has fine details, but our model only locates the main regions without the fine details. There is still considerable room to improve our model to handle such scenes with fine structures.

4.6 Application to RGB-D camouflaged object detection

SPNet was originally designed for the RGB-D SOD task, which can be easily extended to other related RGB-D tasks, e.g., RGB-D based camouflaged object detection (COD). The aim of COD is to identify objects that are "seamlessly" embedded in their background surroundings. This is a very challenging task due to the high intrinsic similarities between the target object and the background [100–102]. Recent research [103] suggests that depth can provide useful spatial information to improve COD results. Thus, we extended SPNet to the RGB-D COD task.

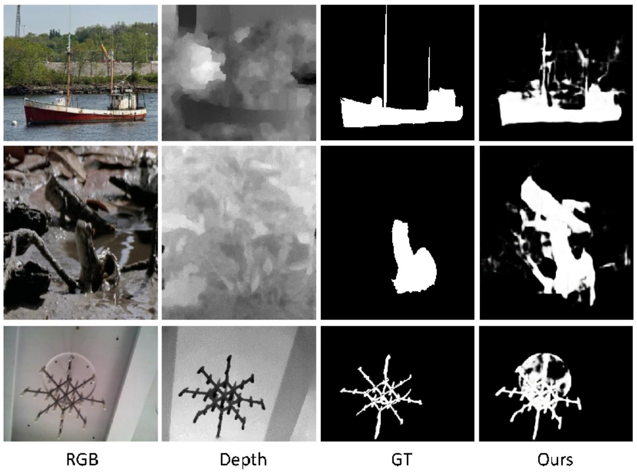


Fig. 11 Cases in which our model fails.

We conducted this experiment on three public benchmark datasets for camouflaged object detection: (i) CHAMELEON [100], consisting of 76 camouflaged images, (ii) CAMO [104], with 1250 images (1000 for training, 250 for testing) in 8 categories, and (iii) COD10K [100], with 5066 camouflaged images (3040 for training, 2026 for testing) in 5 super-classes and 69 sub-classes. Following the same setting in Ref. [105], we divided the training and testing sets and then trained our model on the training set.

We compare our method to other existing COD models, including FPN [93], MaskRCNN [94], PSPNet [95], PiCANet [46], BASNet [96], PFANet [97], CPD [98], EGNet [99], and SINet [105] (results are from Ref. [105]). Since there are few works for RGB-D camouflaged object detection, we also compared two recent RGB-D salient object detection methods, DANet [91] and HDFNet [88], in this experiment. We re-trained the two RGB-D SOD models and our model using RGB and depth images.

Table 7 shows quantitative results for three public datasets. Our model performs better than the other COD methods. Our model and the two RGB-D COD methods use depth cues, and work better than other methods which do not, indicating that depth cues can provide spatial information to improve COD results. Figure 12 shows qualitative results for different COD

Table 7 Results for camouflaged object detection models on benchmark datasets using evaluation metrics S_α [77] and $\mathcal{M}$ [79]. ↑, ↓ indicate that larger or smaller is better

Model	CHAMELEON		CAMO		COD10K	
	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$\mathcal{M}\downarrow$
FPN [93]	0.794	0.075	0.684	0.131	0.697	0.075
MaskRCNN [94]	0.643	0.099	0.574	0.151	0.613	0.080
PSPNet [95]	0.773	0.085	0.663	0.139	0.678	0.080
PiCANet [46]	0.769	0.085	0.609	0.156	0.649	0.090
BASNet [96]	0.687	0.118	0.618	0.159	0.634	0.105
PFANet [97]	0.679	0.144	0.659	0.172	0.636	0.128
CPD [98]	0.853	0.052	0.726	0.115	0.747	0.059
EGNet [99]	0.848	0.050	0.732	0.104	0.737	0.056
SINet [100]	0.869	0.044	0.751	0.100	0.771	0.051
DANet [91]	0.874	0.043	0.752	0.100	0.765	0.051
HDFNet [88]	0.875	0.032	0.778	0.085	0.779	0.045
SPNet (ours)	**0.895**	**0.027**	**0.795**	**0.082**	**0.797**	**0.042**

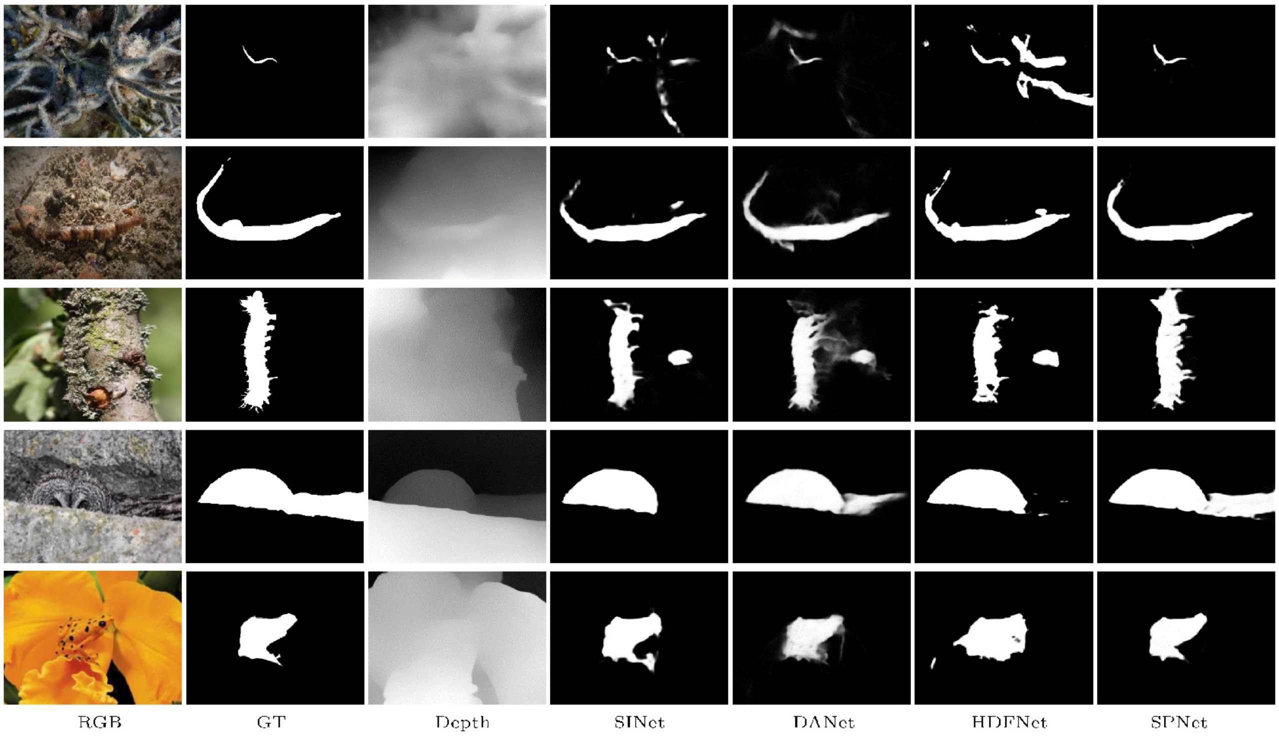


Fig. 12 COD results of our SPNet and three state-of-the-art COD methods: SINet [105], DANet [91], and HDFNet [88].

methods. Compared to other COD models, our SPNet achieves better results by detecting more accurate boundaries of camouflaged objects.

5 Conclusions

In this paper, we have presented a novel RGB-D salient object detection framework, SPNet. Unlike most existing RGB-D SOD methods, which focus on learning shared representations, SPNet not only explores shared cross-modal information but also uses modality-specific characteristics to improve SOD results. To learn the shared representations for the two modalities, we introduce a cross-enhanced integration module (CIM) to fuse the cross-modal features, and the output of each CIM is propagated to the next layer to explore rich cross-level information. We further adopt a multi-modal feature aggregation (MFA) module to integrate the learned modality-specific features to enhance the complementary multi-modal information. Extensive results on benchmark datasets show the effectiveness of our model in comparison to other state-of-the-art RGB-D SOD methods. Moreover, we have thoroughly validated the effectiveness of key components of our framework, and an attribute-based evaluation was conducted to study the ability of many cutting-edge RGB-D SOD approaches to meet different challenges. Finally, we extended SPNet to the recently proposed RGB-D camouflaged object detection task, and its effectiveness was verified.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62172228, in part by an Open Project of the Key Laboratory of System Control and Information Processing, Ministry of Education (Shanghai Jiao Tong University, No. Scip202102).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Peng, H.; Li, B.; Xiong, W.; Hu, W.; Ji, R. RGBD salient object detection: A benchmark and algorithms. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8691*. Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 92–109, 2014.

[2] Zhu, J.-Y.; Wu, J.-J.; Xu, Y.; Chang, E.; Tu, Z. W. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 4, 862–875, 2015.

[3] Rapantzikos, K.; Avrithis, Y.; Kollias, S. Dense saliency-based spatiotemporal feature points for action recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1454–1461, 2009.

[4] Shimoda, W.; Yanai, K. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9908.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 218–234, 2016.

[5] Wang, W. G.; Shen, J. B.; Yang, R. G.; Porikli, F. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 1, 20–33, 2018.

[6] Zhao, R.; Oyang, W.; Wang, X. Person re-identification by saliency learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 2, 356–370, 2017.

[7] Fan, D. P.; Lin, Z.; Zhang, Z.; Zhu, M. L.; Cheng, M. M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems* Vol. 32, No. 5, 2075–2089, 2021.

[8] Zhang, J.; Fan, D.-P.; Dai, Y. C.; Yu, X.; Zhong, Y. R.; Barnes, N.; Shao, L. RGB-D saliency detection via cascaded mutual information minimization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4318–4327, 2021.

[9] Liu, N.; Zhang, N.; Wan, K. Y.; Shao, L.; Han, J. W. Visual saliency transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4702–4712, 2021.

[10] Zhou, T.; Fan, D. P.; Cheng, M. M.; Shen, J. B.; Shao, L. RGB-D salient object detection: A survey. *Computational Visual Media* Vol. 7, No. 1, 37–69, 2021.

[11] Fu, K. R.; Fan, D. P.; Ji, G. P.; Zhao, Q. J.; Shen, J. B.; Zhu, C. Siamese network for RGB-D salient object detection and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi: 10.1109/TPAMI.2021.3073689, 2021.

[12] Zhang, J.; Fan, D.-P.; Dai, Y. C.; Anwar, S., Saleh, F., Aliakbarian, S.; Barnes, N. Uncertainty inspired RGB-D saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi: 10.1109/TPAMI.2021.3073564, 2021.

[13] Chen, H.; Li, Y. F.; Deng, Y. J.; Lin, G. S. CNN-based RGB-D salient object detection: Learn, select, and fuse. *International Journal of Computer Vision* Vol. 129, No. 7, 2076–2096, 2021.

[14] Li, G. Y.; Liu, Z.; Chen, M. Y.; Bai, Z.; Lin, W. S.; Ling, H. B. Hierarchical alternate interaction network for RGB-D salient object detection. *IEEE Transactions on Image Processing* Vol. 30, 3528–3542, 2021.

[15] Zhao, Y. F.; Zhao, J. W.; Li, J.; Chen, X. W. RGB-D salient object detection with ubiquitous target awareness. *IEEE Transactions on Image Processing* Vol. 30, 7717–7731, 2021.

[16] Ren, J. Q.; Gong, X. J.; Lu, Y.; Zhou, W. H.; Yang, M. Y. Exploiting global priors for RGB-D saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 25–32, 2015.

[17] Song, H. K.; Liu, Z.; Du, H.; Sun, G. L.; Le Meur, O.; Ren, T. W. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE Transactions on Image Processing* Vol. 26, No. 9, 4204–4216, 2017.

[18] Liu, Z. Y.; Shi, S.; Duan, Q. T.; Zhang, W.; Zhao, P. Salient object detection for RGB-D image by single stream recurrent convolution neural network. *Neurocomputing* Vol. 363, 46–57, 2019.

[19] Guo, J. F.; Ren, T. W.; Bei, J. Salient object detection for RGB-D image via saliency evolution. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 1–6, 2016.

[20] Wang, N. N.; Gong, X. J. Adaptive fusion for RGB-D salient object detection. *IEEE Access* Vol. 7, 55277–55284, 2019.

[21] Ding, Y.; Liu, Z.; Huang, M. K.; Shi, R.; Wang, X. Y. Depth-aware saliency detection using convolutional neural networks. *Journal of Visual Communication and Image Representation* Vol. 61, 1–9, 2019.

[22] Chen, H.; Li, Y. F. Progressively complementarity-aware fusion network for RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3051–3060, 2018.

[23] Liu, D.; Hu, Y.; Zhang, K.; Chen, Z. Two-stream refinement network for RGB-D saliency detection. In: Proceedings of the IEEE International Conference on Image Processing, 3925–3929, 2019.

[24] Chen, H.; Li, Y. F. Three-stream attention-aware network for RGB-D salient object detection. *IEEE Transactions on Image Processing* Vol. 28, No. 6, 2825–2835, 2019.

[25] Han, J. W.; Chen, H.; Liu, N.; Yan, C. G.; Li, X. L. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics* Vol. 48, No. 11, 3171–3183, 2018.

[26] Chen, H.; Li, Y. F.; Su, D. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 6821–6826, 2018.

[27] Ji, W.; Li, J. J.; Yu, S.; Zhang, M.; Piao, Y. R.; Yao, S. Y.; Bi, Q.; Ma, K.; Zheng, Y.; Lu, H.; et al. Calibrated RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9466–9476, 2021.

[28] Huang, Z.; Chen, H. X.; Zhou, T.; Yang, Y. Z.; Liu, B. Y. Multi-level cross-modal interaction network for RGB-D salient object detection. *Neurocomputing* Vol. 452, 200–211, 2021.

[29] Chen, H.; Li, Y. F.; Su, D. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition* Vol. 86, 376–385, 2019.

[30] Zhao, J.-X.; Cao, Y.; Fan, D.-P.; Cheng, M.-M.; Li, X.-Y.; Zhang, L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3922–3931, 2019.

[31] Zhu, C. B.; Cai, X.; Huang, K.; Li, T. H.; Li, G. PDNet: Prior-model guided depth-enhanced network for salient object detection. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 199–204, 2019.

[32] Fan, D. P.; Zhai, Y.; Borji, A.; Yang, J.; Shao, L. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12357.* Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 275–292, 2020.

[33] Zhai, Y. J.; Fan, D.-P.; Yang, J. F.; Borji, A.; Shao, L.; Han, J. W.; Wang, L. Bifurcated backbone strategy for RGB-D salient object detection. *IEEE Transactions on Image Processing* Vol. 30, 8727–8742, 2021.

[34] Hu, J. L.; Lu, J. W.; Tan, Y. P. Sharable and individual multi-view metric learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 9, 2281–2288, 2018.

[35] Lu, Y.; Wu, Y.; Liu, B.; Zhang, T.; Li, B.; Chu, Q.; Yu, N. Cross-modality person re-identification with shared-specific feature transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13376–13386, 2020.

[36] Zhou, T.; Zhang, C.; Peng, X.; Bhaskar, H.; Yang, J. Dual shared-specific multiview subspace clustering. *IEEE Transactions on Cybernetics* Vol. 50, No. 8, 3517–3530, 2020.

[37] Zhou, T.; Fu, H. Z.; Chen, G.; Shen, J. B.; Shao, L. Hi-net: Hybrid-fusion network for multi-modal MR image synthesis. *IEEE Transactions on Medical Imaging* Vol. 39, No. 9, 2772–2781, 2020.

[38] Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In: *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Lecture Notes in Computer Science, Vol. 9351.* Navab, N.; Hornegger, J.; Wells, W.; Frangi, A. Eds. Springer Cham, 234–241, 2015.

[39] Zhou, T.; Fu, H.; Chen, G.; Zhou, Y.; Fan, D.-P.; Shao, L. Specificity-preserving RGB-D saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4661–4671, 2021.

[40] Zhu, W. J.; Liang, S.; Wei, Y. C.; Sun, J. Saliency optimization from robust background detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2814–2821, 2014.

[41] Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1597–1604, 2009.

[42] Zhou, L.; Yang, Z. H.; Yuan, Q.; Zhou, Z. T.; Hu, D. W. Salient region detection via integrating diffusion-based compactness and local contrast. *IEEE Transactions on Image Processing* Vol. 24, No. 11, 3308–3320, 2015.

[43] Jiang, Z. L.; Davis, L. S. Submodular salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2043–2050, 2013.

[44] Hou, Q. B.; Cheng, M. M.; Hu, X. W.; Borji, A.; Tu, Z. W.; Torr, P. H. S. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 41, No. 4, 815–828, 2019.

[45] Wang, L. Z.; Wang, L. J.; Lu, H. C.; Zhang, P. P.; Ruan, X. Salient object detection with recurrent fully convolutional networks. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence Vol. 41, No. 7, 1734–1746, 2019.

[46] Liu, N.; Han, J.; Yang, M. PiCANet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3089–3098, 2018.

[47] Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; Heng, P.-A. R^{3}Net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 684–690, 2018.

[48] Wang, W.; Lai, Q.; Fu, H.; Shen, J.; Ling, H.; Yang, R. Salient object detection in the deep learning era: An in-depth survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 44, No. 6, 3239–3259, 2022.

[49] Wang, X.; Ma, H.; Chen, X.; You, S. Edge preserving and multi-scale contextual neural network for salient object detection. *IEEE Transactions on Image Processing* Vol. 27, No. 1, 121–134, 2018.

[50] Zhang, P. P.; Wang, D.; Lu, H. C.; Wang, H. Y.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 202–211, 2017.

[51] Zhang, L.; Dai, J.; Lu, H. C.; He, Y.; Wang, G. A bi-directional message passing model for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1741–1750, 2018.

[52] Pang, Y. W.; Zhao, X. Q.; Zhang, L. H.; Lu, H. C. Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9410–9419, 2020.

[53] Lang, C.; Nguyen, T. V.; Katti, H.; Yadati, K.; Kankanhalli, M.; Yan, S. Depth matters: Influence of depth cues on visual saliency. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7573.* Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 101–115, 2012.

[54] Ju, R.; Ge, L.; Geng, W.; Ren, T.; Wu, G. Depth saliency based on anisotropic center-surround difference. In: Proceedings of the IEEE International Conference on Image Processing, 1115–1119, 2014.

[55] Desingh, K.; Krishna, K. M.; Rajan, D.; Jawahar, C. V. Depth really matters: Improving visual salient region detection with depth. In: Proceedings of the British Machine Vision Conference, 98.1–98.11, 2013.

[56] Zhu, C. B.; Li, G.; Wang, W. M.; Wang, R. G. An innovative salient object detection using center-dark channel prior. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 1509–1515, 2017.

[57] Liang, F. F.; Duan, L. J.; Ma, W.; Qiao, Y. H.; Cai, Z.; Qing, L. Y. Stereoscopic saliency model using contrast and depth-guided-background prior. *Neurocomputing* Vol. 275, 2227–2238, 2018.

[58] Feng, D.; Barnes, N.; You, S. D.; McCarthy, C. Local background enclosure for RGB-D salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2343–2350, 2016.

[59] Qu, L. Q.; He, S. F.; Zhang, J. W.; Tian, J. D.; Tang, Y. D.; Yang, Q. X. RGBD salient object detection via deep fusion. *IEEE Transactions on Image Processing* Vol. 26, No. 5, 2274–2285, 2017.

[60] Piao, Y. R.; Ji, W.; Li, J. J.; Zhang, M.; Lu, H. C. Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7253–7262, 2019.

[61] Li, C. Y.; Cong, R. M.; Piao, Y. R.; Xu, Q. Q.; Loy, C. C. RGB-D salient object detection with cross-modality modulation and selection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12353.* Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 225–241, 2020.

[62] Li, G. Y.; Liu, Z.; Ye, L. W.; Wang, Y.; Ling, H. B. Cross-modal weighting network for RGB-D salient object detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12362.* Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 665–681, 2020.

[63] Chaudhuri, K.; Kakade, S. M.; Livescu, K.; Sridharan, K. Multi-view clustering via canonical correlation analysis. In: Proceedings of the 26th Annual International Conference on Machine Learning, 129–136, 2009.

[64] Ding, C. X.; Tao, D. C. Robust face recognition via multimodal deep face representation. *IEEE Transactions on Multimedia* Vol. 17, No. 11, 2049–2058, 2015.

[65] Gönen, M.; Alpaydın, E. Multiple kernel learning algorithms. *The Journal of Machine Learning Research* Vol. 12, 2211–2268, 2011.

[66] White, M.; Yu, Y.; Zhang, X.; Schuurmans, D. Convex multi-view subspace learning. In: Proceedings of the 25th International Conference on Neural Information Processing Systems, Vol. 1, 1673–1681, 2012.

[67] Zhang, C. Q.; Hu, Q. H.; Fu, H. Z.; Zhu, P. F.; Cao, X. C. Latent multi-view subspace clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4333–4341, 2017.

[68] Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A. Multimodal deep learning. In: Proceedings of the 28th International Conference on International Conference on Machine Learning, 689–696, 2011.

[69] Eitel, A.; Springenberg, J. T.; Spinello, L.; Riedmiller, M.; Burgard, W. Multimodal deep learning for robust RGB-D object recognition. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 681–687, 2015.

[70] Gao, S. H.; Cheng, M. M.; Zhao, K.; Zhang, X. Y.; Yang, M. H.; Torr, P. Res2Net: A new multi-scale backbone architecture. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 2, 652–662, 2021.

[71] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* Vol. 115, No. 3, 211–252, 2015.

[72] Wu, Z.; Su, L.; Huang, Q. M. Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3902–3911, 2019.

[73] Wei, J.; Wang, S.; Huang, Q. F^{3}Net: Fusion, feedback and focus for salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 12321–12328, 2020.

[74] Cheng, Y. P.; Fu, H. Z.; Wei, X. X.; Xiao, J. J.; Cao, X. C. Depth enhanced saliency detection method. In: Proceedings of the International Conference on Internet Multimedia Computing and Service, 23–27, 2014.

[75] Li, G.; Zhu, C. B. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 3008–3014, 2017.

[76] Niu, Y. Z.; Geng, Y. J.; Li, X. Q.; Liu, F. Leveraging stereopsis for saliency analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 454–461, 2012.

[77] Cheng, M. M.; Fan, D. P. Structure-measure: A new way to evaluate foreground maps. *International Journal of Computer Vision* Vol. 129, No. 9, 2622–2638, 2021.

[78] Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; Borji, A. Enhanced-alignment measure for binary foreground map valuation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 698–704, 2018.

[79] Perazzi, F.; Krähenbühl, P.; Pritch, Y.; Hornung, A. Saliency filters: Contrast based filtering for salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 733–740, 2012.

[80] Cong, R. M.; Lei, J. J.; Zhang, C. Q.; Huang, Q. M.; Cao, X. C.; Hou, C. P. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters* Vol. 23, No. 6, 819–823, 2016.

[81] Cong, R. M.; Lei, J. J.; Fu, H. Z.; Hou, J. H.; Huang, Q. M.; Kwong, S. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE Transactions on Cybernetics* Vol. 50, No. 8, 3627–3639, 2020.

[82] Jiang, B.; Zhou, Z. T.; Wang, X.; Tang, J.; Luo, B. cmSalGAN: RGB-D salient object detection with cross-view generative adversarial networks. *IEEE Transactions on Multimedia* Vol. 23, 1343–1353, 2021.

[83] Li, C. Y.; Cong, R. M.; Kwong, S.; Hou, J. H.; Fu, H. Z.; Zhu, G. P.; Zhang, D.; Huang, Q. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics* Vol. 51, No. 1, 88–100, 2021.

[84] Li, G.; Liu, Z.; Ling, H. ICNet: Information conversion network for RGB-D based salient object detection. *IEEE Transactions on Image Processing* Vol. 29, 4873–4884, 2020.

[85] Piao, Y. R.; Rong, Z. K.; Zhang, M.; Ren, W. S.; Lu, H. C. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9057–9066, 2020.

[86] Liu, N.; Zhang, N.; Han, J. W. Learning selective self-mutual attention for RGB-D saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13753–13762, 2020.

[87] Zhang, M.; Ren, W. S.; Piao, Y. R.; Rong, Z. K.; Lu, H. C. Select, supplement and focus for

RGB-D saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3469–3478, 2020.

[88] Pang, Y.; Zhang, L.; Zhao, X.; Lu, H. Hierarchical dynamic filtering network for RGB-D salient object detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12370*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 235–252, 2020.

[89] Luo, A.; Li, X.; Yang, F.; Jiao, Z.; Cheng, H.; Lyu, S. Cascade graph neural networks for RGB-D salient object detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12357*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springe Cham, 346–364, 2020.

[90] Ji, W.; Li, J.; Zhang, M.; Piao, Y.; Lu, H. Accurate RGB-D salient object detection via collaborative learning. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12363*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 52–69, 2020.

[91] Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; Zhang, L. A single stream network for robust and real-time RGB-D salient object detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12367*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 646–662, 2020.

[92] Chen, S.; Fu, Y. Progressively guided alternate refinement network for RGB-D salient object detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12353*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 520–538, 2020.

[93] Lin, T. Y.; Dollár, P.; Girshick, R.; He, K. M.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 936–944, 2017.

[94] He, K. M.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, 2980–2988, 2017.

[95] Zhao, H. S.; Shi, J. P.; Qi, X. J.; Wang, X. G.; Jia, J. Y. Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6230–6239, 2017.

[96] Qin, X. B.; Zhang, Z. C.; Huang, C. Y.; Gao, C.; Dehghan, M.; Jagersand, M. BASNet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7471–7481, 2019.

[97] Zhao, T.; Wu, X. Q. Pyramid feature attention network for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3080–3089, 2019.

[98] Wu, Z.; Su, L.; Huang, Q. M. Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3902–3911, 2019.

[99] Zhao, J. X.; Liu, J. J.; Fan, D. P.; Cao, Y.; Yang, J. F.; Cheng, M. M. EGNet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8778–8787, 2019.

[100] Fan, D.-P.; Ji, G.-P.; Cheng, M.-M.; Shao, L. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi: 10.1109/TPAMI.2021.3085766, 2021.

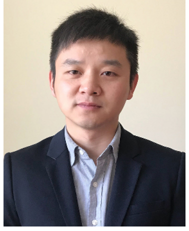
[101] Sun, Y. J.; Chen, G.; Zhou, T.; Zhang, Y.; Liu, N. Context-aware cross-level fusion network for camouflaged object detection. In: Proceedings of the 30th International Joint Conference on Artificial Intelligence, 1025–1031, 2021.

[102] Li, L.; Dong, B.; Rigall, E.; Zhou, T.; Dong, J. Y.; Chen, G. Marine animal segmentation. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 32, No. 4, 2303–2314, 2022.

[103] Zhang, J.; Lv, Y.; Xiang, M.; Li, A.; Dai, Y.; Zhong, Y. Depth confidence-aware camouflaged object detection. *arXiv preprint* arXiv:2106.13217, 2021.

[104] Le, T. N.; Nguyen, T. V.; Nie, Z. L.; Tran, M. T.; Sugimoto, A. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding* Vol. 184, 45–56, 2019.

[105] Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; Shao, L. Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2774–2784, 2020.



Tao Zhou received his Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. From 2016 to 2018, he was a postdoctoral fellow in the BRIC and IDEA lab, University of North Carolina at Chapel Hill. From 2018 to 2020, he was a research scientist at the Inception Institute of Artificial Intelligence (IIAI), United Arab Emirates. He is currently a professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include machine learning, computer vision, and medical image analysis.

Deng-Ping Fan is a postdoctoral researcher at ETH Zürich, Switzerland. He received his Ph.D. degree from Nankai University in 2019. He joined IIAI in 2019. He has published about 30 top journal and conference papers in outlets such as IEEE TPAMI, CVPR, and ICCV. His research interests include computer vision, deep learning, saliency detection, especially co-salient object detection, RGB salient object detection, RGB-D salient object detection, and video salient object detection.

Geng Chen is a professor at Northwestern Polytechnical University, China, where he received his Ph.D. degree in 2016. He was a research scientist at IIAI from 2019 to 2021, and a postdoctoral research associate at the University of North Carolina at Chapel Hill, USA, from 2016 to 2019. He has published over 50 papers in peer-reviewed international conference proceedings and journals. His research interests lie in computer vision and medical image analysis.

Yi Zhou is currently an associate professor at Southeast University. He received his M.Sc. degree from the Department of Electronic and Electrical Engineering, University of Sheffield, UK, in 2014, and his Ph.D. degree from the School of Computing Sciences, University of East Anglia, UK, in 2018. He was a research scientist with IIAI. His research interests include computer vision, pattern recognition, machine learning, and medical imaging.

Huazhu Fu is currently a senior scientist at Inception Institute of Artificial Intelligence, United Arab Emirates. He received his Ph.D. degree from Tianjin University in 2013 and was a research fellow at Nanyang Technological University (NTU) for two years. From 2015 to 2018, he was a research scientist with I2R, A*STAR, Singapore. His research interests include computer vision, machine learning, and AI in healthcare. He serves as an Associate Editor for IEEE TMI and IEEE JBHI, and also served as the Area Chair for MICCAI 2021 and Co-Chair for the OMIA Workshop.

Other papers from this open access journal are available free of charge from http://www.springer.com/journal/41095. To submit a manuscript, please go to https://www.editorialmanager.com/cvmj.