Specificity-preserving RGB-D Saliency Detection

Tao Zhou[1], Huazhu Fu[2], Geng Chen[2], Yi Zhou[3], Deng-Ping Fan[2]*, Ling Shao[2]

[1]PCA lab, The Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of MoE,
School of Computer Science and Engineering, Nanjing University of Science and Technology, China

[2]IIAI, Abu Dhabi, UAE   [3]School of Computer Science and Engineering, Southeast University, China

Abstract

RGB-D saliency detection has attracted increasing attention, due to its effectiveness and the fact that depth cues can now be conveniently captured. Existing works often focus on learning a shared representation through various fusion strategies, with few methods explicitly considering how to preserve modality-specific characteristics. In this paper, taking a new perspective, we propose a specificity-preserving network (SP-Net) for RGB-D saliency detection, which benefits saliency detection performance by exploring both the shared information and modality-specific properties (e.g., specificity). Specifically, two modality-specific networks and a shared learning network are adopted to generate individual and shared saliency maps. A cross-enhanced integration module (CIM) is proposed to fuse cross-modal features in the shared learning network, which are then propagated to the next layer for integrating cross-level information. Besides, we propose a multi-modal feature aggregation (MFA) module to integrate the modality-specific features from each individual decoder into the shared decoder, which can provide rich complementary multi-modal information to boost the saliency detection performance. Further, a skip connection is used to combine hierarchical features between the encoder and decoder layers. Experiments on six benchmark datasets demonstrate that our SP-Net outperforms other state-of-the-art methods. Code is available at: https://github.com/taozh2017/SPNet.

1. Introduction

Saliency detection aims to locate the most visually prominent object(s) in a given scene [46]. It has been widely applied in various vision-related tasks, such as image understanding [75], video/semantic segmentation [55, 58], action recognition [51], [55], and person re-identification [67]. Although significant progress has been made, it is still challenging to accurately locate salient objects in com-

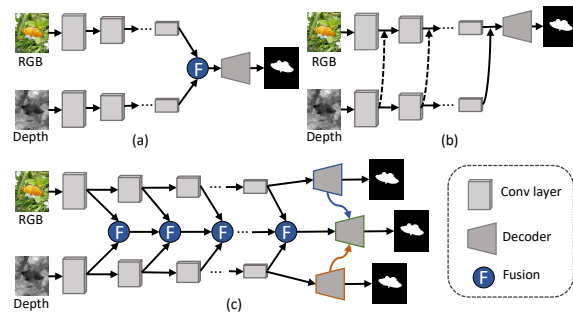*Corresponding author: *Deng-Ping Fan (dengpfan@gmail.com).*

Figure 1. Comparison between existing RGB-D saliency detection frameworks and our proposed model. (a) RGB and depth images are fed into two separate network streams and then the fused high-level features are fed into a decoder (*e.g.*, [4,5,25,37]). (b) Depth features are integrated into the RGB network using a tailor-maid subnetwork (*e.g.*, [6, 66, 72]). (c) Our method explicitly explores the shared information and modality-specific characteristics. Then, the features learned from the modality-specific decoders are integrated into the shared decoder to boost saliency detection performance.

plex scenes, such as instance cluttered background or low-contrast lighting conditions. Recently, with the large availability of depth sensors in smart devices, depth maps have been introduced to provide geometric and spatial information to improve the saliency detection performance. Consequently, fusing RGB and depth images has gained increasing interest in the saliency detection community [39,64,69].

For RGB-D saliency detection, it is critical to effectively fuse RGB and depth images. Some works utilize an **early fusion** strategy via a simple concatenation for this. For example, these models [41,46,52,56] directly integrate RGB and depth images to form a four-channel input. However, this type of fusion does not consider the distribution gap between the two modalities, which may lead to inaccurate feature fusion. Besides, various models based on a **late fusion** strategy use two parallel network streams to generate independent saliency maps for RGB and depth data, then the two maps are fused to obtain a final prediction map [15,24,57]. However, it is difficult to capture the complex interactions between the two modalities with this type of fusion.

Currently, various **middle fusion** methods utilize 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 (as shown in Fig. 1 (a)). Besides, other methods carry out cross-modal fusion at multiple scales [4, 5, 7, 25, 27, 37]. As a result, the complex correlations can be effectively exploited from the two modalities. Moreover, several methods utilize depth information to enhance RGB features via a tailor-maid subnetwork [6, 66, 72] (as shown in Fig. 1 (b)). For example, Zhao *et al.* [66] introduced a contrast prior into a CNN-based architecture to enhance the depth information, and then the enhanced depth was integrated with RGB features using a fluid pyramid integration module. Zhu *et al.* [72] utilized an independent subnetwork to extract depth-based features, which were then incorporated into the RGB network. It should be noted that the above methods mainly focus on learning shared representations by fusing them and then use a decoder to generate the final saliency map. What is more, there is no decoder with supervision to guide the depth-based feature learning [66, 72], which may prevent optimal depth features from being obtained. From a multi-modal learning perspective, several works [26, 42, 70, 71] have shown that exploring both the shared information and modality-specific characteristics can improve the model performance. However, few RGB-D saliency detection models explicitly exploit modality-specific characteristics.

To this end, we propose a novel specificity-preserving network for RGB-D saliency detection (termed *SP-Net*), which not only explores the shared information but also exploits modality-specific characteristics to boost saliency detection performance. Specifically, two encoder subnetworks are used to extract multi-scale features for the two modalities, and a cross-enhanced integration module (CIM) is proposed to fuse cross-modal features. Then, we use a U-Net [53] 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. Moreover, we also construct a shared decoder to combine hierarchical features from previous CIMs using skip connections. To make full use of the modality-specific features, we propose a multi-modal feature aggregation (MFA) to integrate them into the shared decoder. Finally, we formulate a unified and end-to-end trainable framework to achieve RGB-D saliency detection.

Our main **contributions** are summarized as follows:

- We propose a novel specificity-preserving network for RGB-D saliency detection (*SP-Net*), which can explore the shared information as well as preserve modality-specific characteristics.

- We propose a cross-enhanced integration module

(CIM) to fuse the cross-modal features and learn shared representations for the two modalities. The output of each CIM is then propagated to the next layer to capture cross-level information.

- We propose a simple but effective multi-modal feature aggregation (MFA) module to integrate these learned modality-specific features. It makes full use of the features learned in the modality-specific decoder to boost the saliency detection performance.

- Extensive experiments on six public datasets demonstrate the superiority of our model over thirty benchmarking methods. Moreover, we carry out an attribute-based evaluation to study the performance of many state-of-the-art RGB-D saliency detection methods under different challenging factors (*e.g.*, number of salient objects, indoor or outdoor environments, and light conditions), which has not been done previously by existing studies.

2. Related Work

2.1. RGB-D Saliency Detection

Early RGB-D based SOD models typically extracted handcrafted features from the input RGB-D data. For example, Lang *et al.* [31] proposed the first RGB-D SOD work, which utilized Gaussian mixture models to model the distribution of depth-induced saliency. After that, several methods were explored based on different principles, such as center-surround difference [24, 30], contrast [13, 46, 52], center/boundary prior [36, 74], and background enclosure [19]. However, these methods suffer from unsatisfactory performance due to the limited expression ability of handcrafted features. Benefiting from the rapid development of deep convolutional neural networks (CNNs), several deep learning-based works [18, 48, 50, 63, 66] have recently been developed and obtained promising results. For example, Qu *et al.* [50] developed a CNN model to integrate different low-level saliency cues into hierarchical features to boost the saliency detection performance. Chen *et al.* [4] proposed a complementarity-aware fusion module to effectively integrate cross-modal and cross-level features for RGB and depth modalities. Piao *et al.* [48] proposed a depth-induced multi-scale recurrent attention network to enhance the cross-modality feature fusion. Fan *et al.* [18] designed a depth depurator unit to filter out some low-quality depth maps. Most other models [5, 7, 25, 33, 35, 37] employ cross-modal fusion at multiple scales using different integration strategies.

2.2. Multi-modal Learning

Recently, multi-modal (or multi-view) learning has attracted increasing attention, since most data can be obtained
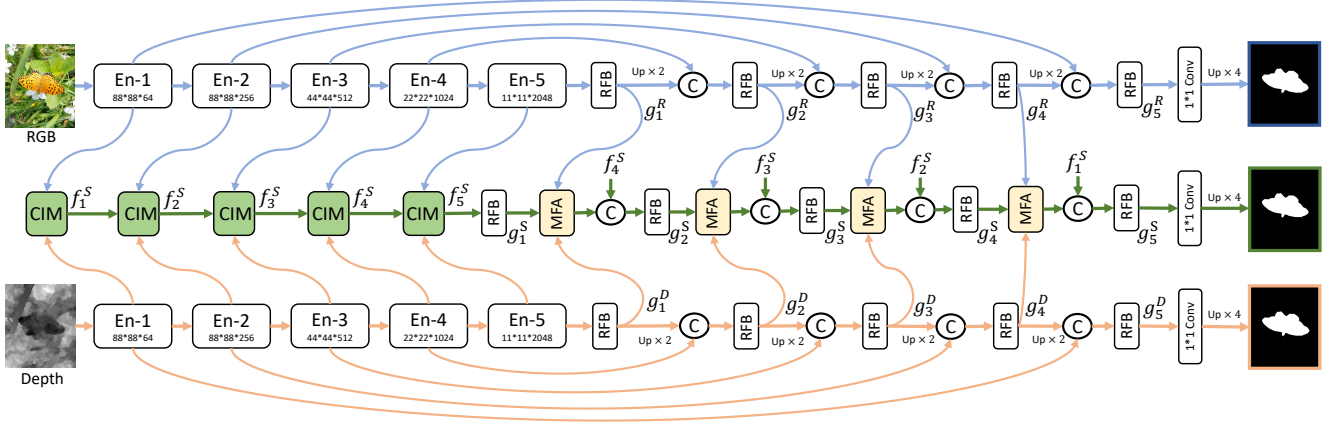
Figure 2. The overall architecture of the proposed *SP-Net*. Our model consists of two modality-specific learning networks and a shared learning network. The modality-specific learning networks are used to preserve the individual properties for each modality (*i.e.*, RGB or depth), while the shared network is used to fuse cross-modal features and explore their complementary information. Skip connections are adopted to combine hierarchical features between the encoder and decoder layers. The learned features from the modality-specific decoder are integrated into the shared decoder to provide rich multi-modal complementary information for boosting saliency detection performance. Here, "C" denotes feature concatenation.

from multiple sources or represented with different types of features. One common strategy is to directly concatenate the feature vectors from such multi-modal data into one long vector. However, this simple concatenation may fail to exploit the complex correlations among multi-modal data. As such, several multi-modal learning methods have been developed to explicitly fuse the complementary information from different modalities to improve model performance. These popular methods can be divided into three following types. 1) Co-training [3, 14] tries to minimize the disagreement between different modalities, 2) Multiple kernel learning [23] utilizes a predefined set of kernels from multiple modalities and integrates these modalities using the learned weights of the kernels, and 3) Subspace learning [60,62] assumes that there exists a latent subspace shared by different modalities and that multiple modalities can originate from one underlying latent representation. Besides, to effectively fuse multi-modal data, several deep learning-based models have also been explored. For example, Ngiam *et al.* [44] proposed to learn a shared representation from audio and video inputs. Eitel *et al.* [16] utilized two separate CNN streams for RGB and depth, respectively, and then combined them using a late fusion network to achieve RGB-D object recognition. Besides, Hu *et al.* [26] presented a shareable and individual multi-view learning algorithm to explore more properties of multi-modal data. Lu *et al.* [42] developed a shared-specific feature transfer framework for the cross-modal person ReID task.

3. Proposed Method

Fig. 2 shows the framework of the proposed specificity-preserving network for RGB-D saliency detection. The RGB and depth images are first fed into two-stream

modality-specific learning networks to obtain their multi-level feature representations, and a CIM is proposed to learn their shared feature representation. Then the specific and shared decoder subnetworks are adopted to generate saliency prediction maps, respectively. Besides, the original features from the encoder networks are integrated into the decoder via a skip connection. To make full use of the features learned by the modality-specific decoder, we propose an MFA module to integrate these features into the shared decoder for boosting the saliency detection performance. We give the details of each key part below.

3.1. Modality-specific Learning Network

As shown in Fig. 2, the modality-specific subnetwork is built using the Res2Net-50 [22], which has been pretrained on ImageNet [54] dataset. Thus, there are five multi-level features, *i.e.*, $F^R = [f^R_m, m = 1, 2, \ldots, 5]$ and $F^D = [f^D_m, m = 1, 2, \ldots, 5]$, in the modality-specific encoder subnetworks for RGB and depth, respectively. In our study, we denote the input resolution of the modality-specific encoder subnetwork as $W \times H$. Thus, we have a feature resolution of $\frac{H}{8} * \frac{W}{8}$ for the first layer, and a general resolution of $\frac{H}{2^m} * \frac{W}{2^m}$ (when $m > 1$). Besides, the channel number of features in the m-th layer is given as C_m ($m = 1, 2, \ldots$), and we have $C = [64, 256, 512, 1024, 2048]$.

Once we obtain the high-level features f^R_5 and f^D_5, they are then fed into the modality-specific decoder subnetworks to generate individual saliency maps. Besides, we utilize a U-Net [53] structure to construct the modality-specific decoder, where the skip connections between the encoder and decoder layers are used to combine hierarchical features. Moreover, the concatenated features (only f^R_5 or f^D_5 in the first stage of the decoder subnetwork) are fed to the
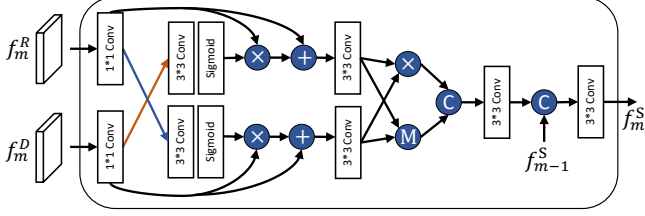
Figure 3. Diagram of the proposed cross-enhanced integration module (CIM). Here, "C" denotes feature concatenation, and "+", "×", "M" denote the element-wise addition, multiplication, and maximization, respectively.

Figure 4. Diagram of the proposed multi-modal feature aggregation (MFA) module.

receptive field block (RFB) [61] to capture global context information. It is worth noting that the 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 boost the saliency detection performance.

3.2. Shared Learning Network

As shown in Fig. 2, 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. Besides, we also 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 the saliency detection performance.

3.2.1 Cross-enhanced Integration Module

We propose a CIM to effectively fuse cross-modal features. Taking $f_m^R \in \mathbb{R}^{W_m * H_m * C_m}$ and $f_m^D \in \mathbb{R}^{W_m * H_m * C_m}$ as an example (for convenience, the width, height, and channel number of the m-th layer are denoted as W_m, H_m and C_m), we use a 1×1 convolutional layer to reduce the channel number to $C_m/2$ for acceleration. The CIM includes two parts, *i.e.*, cross-modal feature enhancement and adaptive feature fusion. First, we use a cross-enhanced strategy to exploit the correlations between the two modalities by learning their enhanced features. Specifically, as shown in Fig. 3, the two features can be fed into a 3×3 convolutional layer with a *Sigmoid* activation function, and then we can obtain the normalized feature maps, *i.e.*, $w_m^R = \sigma(Conv_3(f_m^R)) \in [0, 1]$ and $w_m^D = \sigma(Conv_3(f_m^R)) \in [0, 1]$, where σ is the logistic *Sigmoid* activation function. To exploit the 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 another modality. Besides, to preserve
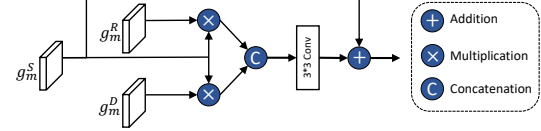
the original information of each modality, a residual connection is adapted to combine the enhanced features with their original features. Thus, we have the cross-enhanced feature representations for the two modalities as follows:

$$\begin{cases} f_m^{R'} = f_m^R + f_m^R \otimes w_m^D, \\ f_m^{D'} = f_m^D + f_m^D \otimes w_m^R, \end{cases} \quad (1)$$

where $\otimes$ denotes element-wise multiplication.

Once we have obtained the cross-enhanced feature representations (*i.e.*, $f_m^{R'}$ and $f_m^{D'}$), one 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 then concatenate the results together. Specifically, the two features $f_m^{R'}$ and $f_m^{D'}$ are first fed into a 3×3 convolutional layer to obtain their smooth representations, and then we carry out element-wise multiplication and maximization. Thus, we can obtain:

$$\begin{cases} p_{mul} = Bconv_3(f_m^{R'}) \otimes Bconv_3(f_m^{D'}), \\ p_{max} = Max(Bconv_3(f_m^{R'}), Bconv_3(f_m^{D'})), \end{cases} \quad (2)$$

where $Bconv(\cdot)$ is a sequential operation that combines a 3×3 convolution followed by batch normalization, and a *ReLU* function. Then, we concatenate the results as $p_{cat} = [p_{mul}, p_{max}] \in \mathbb{R}^{W_m * H_m * C_m}$, and obtain $p_{cat}^1 = Bconv_3(p_{cat})$ through a $Bconv_3$ operation to adaptively weight 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 for reducing the channel number. Besides, there is no previous output f_{m-1}^S (when $m = 1$), so we only feed the concatenated features into one $Bconv_3$ operation.

It is worth noting that our CIM can effectively exploit the correlations between the two modalities via cross-enhanced feature learning, and fuse them by adaptively weighting the different feature representations. Besides, the fused feature representation f_m^S is propagated to the next layer to capture and integrate cross-level information.

3.2.2 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 modality-specific features g_m^R and g_m^D in the modality-specific decoder. As shown in Fig. 4, two features g_m^R and g_m^D are multiplied by the shared features of the current layer, $i.e.$, $g_m^{RS} = g_m^S \otimes g_m^R$ and $g_m^{DS} = g_m^S \otimes g_m^D$. The two features are further concatenated ($[g_m^{DR}, g_m^{DS}]$) and then fed into a $Bconv(\cdot)$ 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.

It is worth noting that the learned modality-specific features are used to enhance the shared representation and provide rich and complementary cross-modal information. More importantly, the modality-specific decoder is given a supervision signal to guide feature learning for the modality-specific property preservation, which benefits the final prediction results when integrating them into the shared decoder.

3.3. Loss Function

Finally, we formulate a unified and end-to-end trainable framework. The overall loss function consists of two parts, $i.e.$, $\mathcal{L}_{sp}$ and $\mathcal{L}_{sh}$, for the modality-specific and shared decoders, respectively. For convenience, S_R and S_D denote the prediction maps when using RGB and depth images, respectively, S_{sh} denotes the prediction map using their shared representation, and G denotes the ground truth map. Therefore, the overall loss function can be formulated as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{sh}(S_{sh}, G) + \mathcal{L}_{sp}(S_R, G) + \mathcal{L}_{sp}(S_D, G). \quad (3)$$

In Eq. (3), we utilize the pixel position aware loss [59] for $\mathcal{L}_{sp}$ and $\mathcal{L}_{sh}$, which can pay different attention to hard and easy pixels to boost the saliency detection performance.

4. Experiments

4.1. Experimental Setup

Datasets: To validate the effectiveness of the proposed model, we evaluate it on six public RGB-D SOD datasets, including NJU2K [30], NLPR [46], DES [10], SSD [73], STERE [45] and SIP [18]. Following [18, 48], we select the same 1,485 samples from NJU2K [30] and 700 samples from the NLPR [46], resulting in 2,195 samples in total for training. The remaining samples from NJU2K (500) and NLPR (300), and the whole DES (135), SSD (80), STERE (1,000) and SIP (929), are used for testing.

Evaluation Metrics: We adopt four widely used metrics for quantitative evaluation. 1) S-measure (S_α) [8] is used to assess the structural similarity between the regional perception (S_r) and object perception (S_o). It is defined as $S_\alpha = \alpha * S_o + (1 - \alpha) * S_r$, where $\alpha \in [0, 1]$ is a trade-off parameter and it is set to 0.5 as default [8]. 2) E_ϕ [17] is used to capture image-level statistics and their local pixel matching information, and it is defined as $E_\phi = \frac{1}{W*H} \sum_{i=1}^{W} \sum_{i=1}^{H} \phi_{FM}(i, j)$, where ϕ_{FM} denotes the enhanced-alignment matrix [17]. 3) F-measure [1] (F_β) is used to comprehensively consider both precision and recall, and we can obtain the weighted harmonic mean by $F_\beta = (1 + \beta^2) \frac{P*R}{\beta^2 P + R}$, where β^2 is set to 0.3 to emphasize the precision [1]. We use different fixed $[0, 255]$ thresholds to compute the F-measure. This yields a set of F-measure values for which we report the maximal F_β. 4) Mean absolute error (MAE) [47] is used to evaluate the average pixel-level relative error between the ground truth and the normalized prediction by calculating the mean of the absolute value of the difference.

Compared RGB-D SOD Models: We compare the proposed model with 30 benchmarking RGB saliency detection methods, including 8 handcrafted traditional models ($i.e.$, LHM [46], ACSD [30], LBE [19], DCMC [12], SE [24], MDSF [56], CDCP [74]), and DTM [11], and 22 deep models ($i.e.$, DF [50], CTMF [25], PCF [4], AFNet [57], CPFP [66], MMCI [6], TANet [5], DMRA [48], cmSal-GAN [29], ASIFNet [32], ICNet [34], A2dele [49], JLDCF [20], S^{2}MA [38], UCNet [63], SSF [65], Cas-GNN [43], CMMS [33], D^{3}Net [18], CoNet [28], DANet [68], and PGAR [9]). Details for these methods are omitted here and readers are referred to the related papers.

Implementation Details: Our model is implemented in PyTorch, and trained on one NVIDIA Tesla V100 GPU with 32 GB memory. The backbone network (Res2Net-50 [22]) is used, which has been pre-trained on ImageNet [54]. Since RGB and depth images have different channels, the input channel of the depth encoder is modified to 1. We adopt the Adam algorithm to optimize the proposed model. The initial learning rate is set to $1e-4$ and is divided by 10 every 60 epochs. The input resolutions of RGB and depth images are resized to 352×352. The training images are augmented using various strategies, including random flipping, rotating, and border clipping. The batch size is set to 20 and the model is trained over 200 epochs. During the testing stage, the RGB and depth images are resized to 352×352 and then fed into the model to obtain prediction maps. Then, the prediction maps can be rescaled to the original size to achieve the final evaluation. Finally, the output of the shared decoder is the final prediction map for our model.

4.2. Performance Comparison

Quantitative Evaluation: As shown in Table 1, our method is superior to eight traditional methods ($i.e.$, LHM [46], ACSD [30], LBE [19], DCMC [12], SE [24], MDSF

Table 1. Benchmarking results of 8 representative traditional models and 22 deep models on six public RGB-D saliency datasets using four widely used evaluation metrics (*i.e.*, S_α [8], max E_ϕ [17], max F_β [1], and M [47]). "↑" & "↓" indicate that larger or smaller is better. The subscript of each model denotes the publication year. The best results are highlighted in **Bold** fonts.

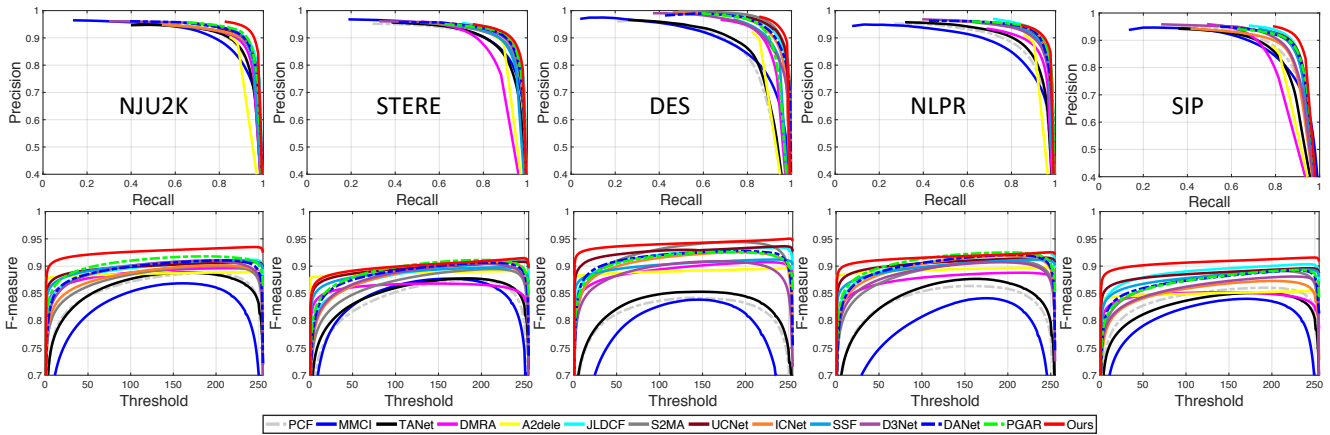| Model	NJU2K [30] $S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	STERE [45] $S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	DES [10] $S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	NLPR [46] $S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	SSD [73] $S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$	SIP [18] $S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\xi\uparrow$	$M\downarrow$
LHM[14] [46]	.514	.632	.724	.205	.562	.683	.771	.172	.562	.511	.653	.114	.630	.622	.766	.108	.566	.568	.717	.195	.511	.574	.716	.184
ACSD[14] [30]	.699	.711	.803	.202	.692	.669	.806	.200	.728	.756	.850	.169	.673	.607	.780	.179	.675	.682	.785	.203	.732	.763	.838	.172
LBE[16] [19]	.695	.748	.803	.153	.660	.633	.787	.250	.703	.788	.890	.081	.762	.745	.855	.081	.621	.619	.736	.278	.727	.751	.853	.200
DCMC[16] [12]	.686	.715	.799	.172	.731	.740	.819	.148	.707	.666	.773	.111	.724	.648	.793	.117	.704	.711	.786	.169	.683	.618	.743	.186
SE[16] [24]	.664	.748	.813	.169	.708	.755	.846	.143	.741	.741	.856	.090	.756	.713	.847	.091	.675	.710	.800	.165	.628	.661	.771	.164
MDSF[17] [56]	.748	.775	.838	.157	.728	.719	.809	.176	.741	.746	.851	.122	.805	.793	.885	.095	.673	.703	.779	.192	.717	.698	.798	.167
CDCP[17] [74]	.669	.621	.741	.180	.713	.664	.786	.149	.709	.631	.811	.115	.669	.621	.741	.180	.603	.535	.700	.214	.595	.505	.721	.224
DTM[20] [11]	.706	.716	.799	.190	.747	.743	.837	.168	.752	.697	.858	.123	.733	.677	.833	.145	.677	.651	.773	.199	.690	.659	.778	.203
DF[17] [50]	.763	.804	.864	.141	.757	.757	.847	.141	.752	.766	.870	.093	.802	.778	.880	.085	.747	.735	.828	.142	.653	.657	.759	.185
CTMF[18] [25]	.849	.845	.913	.085	.848	.831	.912	.086	.863	.844	.932	.055	.860	.825	.929	.056	.776	.729	.865	.099	.716	.694	.829	.139
PCF[18] [4]	.877	.872	.924	.059	.875	.860	.925	.064	.842	.804	.893	.049	.874	.841	.925	.044	.841	.807	.894	.062	.842	.838	.901	.071
AFNet[19] [57]	.772	.775	.853	.100	.825	.823	.887	.075	.770	.729	.881	.068	.799	.771	.879	.058	.714	.687	.807	.118	.720	.712	.819	.118
CPFP[19] [66]	.878	.877	.923	.053	.879	.874	.925	.051	.872	.846	.923	.038	.888	.867	.932	.036	.807	.766	.852	.082	.850	.851	.903	.064
MMCI[19] [6]	.859	.853	.915	.079	.873	.863	.927	.068	.848	.822	.928	.065	.856	.815	.913	.059	.813	.781	.882	.082	.833	.818	.897	.086
TANet[19] [5]	.878	.874	.925	.060	.871	.861	.923	.060	.858	.827	.910	.046	.886	.863	.941	.041	.839	.810	.897	.063	.835	.830	.895	.075
DMRA[19] [48]	.886	.886	.927	.051	.886	.886	.938	.047	.900	.888	.943	.030	.899	.879	.947	.031	.857	.844	.906	.058	.806	.821	.875	.085
cmSalGAN[20] [29]	.903	.896	.940	.046	.900	.894	.936	.050	.913	.899	.943	.028	.922	.907	.957	.027	.791	.735	.867	.086	.865	.864	.906	.064
ASIFNet[20] [32]	.889	.888	.927	.047	.878	.878	.927	.049	.934	.935	.974	.019	.906	.888	.944	.030	.857	.834	.884	.056	.857	.859	.896	.061
ICNet[20] [34]	.894	.891	.926	.052	.903	.898	.942	.045	.920	.913	.960	.027	.923	.908	.952	.028	.848	.841	.902	.064	.854	.857	.903	.069
A2dele[20] [49]	.871	.874	.916	.051	.878	.879	.928	.044	.886	.872	.920	.029	.898	.882	.944	.029	.802	.776	.861	.070	.828	.833	.889	.070
JLDCF[20] [20]	.903	.903	.944	.043	.905	.901	.946	.042	.929	.919	.968	.022	.925	.916	**.962**	.022	.830	.795	.885	.068	.879	.885	.923	.051
S^2MA[20] [38]	.894	.889	.930	.053	.890	.882	.932	.051	.941	.935	.973	.021	.915	.902	.953	.030	.868	.848	.909	.052	.872	.877	.919	.057
UCNet[20] [63]	.897	.895	.936	.043	.903	.899	.944	.039	.933	.930	.976	.018	.920	.903	.956	.025	.865	.854	.907	.049	.875	.879	.919	.051
SSF[20] [65]	.899	.896	.935	.043	.893	.890	.936	.044	.904	.884	.941	.026	.914	.896	.953	.026	.845	.824	.897	.058	.876	.882	.922	.052
Cas-GNN[20] [43]	.911	.903	.933	.035	.899	.901	.930	.039	.905	.906	.947	.028	.919	.904	.947	.028	.872	.862	.915	.047	.875	.879	.919	.051
CMMS[20] [33]	.900	.897	.936	.044	.895	.893	.939	.043	.937	.930	.976	.018	.915	.896	.949	.027	**.874**	.864	**.922**	.046	.872	.877	.911	.058
CoNet[20] [28]	.895	.893	.937	.046	**.908**	.905	**.949**	.040	.909	.896	.945	.028	.908	.887	.945	.031	.853	.840	.915	.059	.858	.867	.913	.063
DANet[20] [68]	.899	.910	.935	.045	.901	.892	.937	.043	.924	.928	.968	.023	.915	.916	.953	.028	.864	.866	.914	.050	.875	.892	.918	.054
PGAR[20] [9]	.909	.907	.940	.042	.907	.898	.939	.041	.913	.902	.945	.026	**.930**	.916	.961	.024	.865	.838	.898	.057	.876	.876	.915	.055
D^3Net[21] [18]	.900	.900	.950	.041	.899	.891	.938	.046	.898	.885	.946	.031	.912	.897	.953	.030	.857	.834	.910	.058	.860	.861	.909	.063
SP-Net (Ours)	**.925**	**.935**	**.954**	**.028**	.907	**.915**	.944	**.037**	**.945**	**.950**	**.980**	**.014**	.927	**.925**	.959	**.021**	.871	**.883**	.915	**.044**	**.894**	**.916**	**.930**	.043

Figure 5. PR curves (1st row) and F-measure curves (2nd row) on NJU2K [30], STERE [45], DES [10], NLPR [46], and SIP [18].

[56], and CDCP [74]) by a large margin on all six datasets. Besides, our method outperforms all of the comparison state-of-the-art methods and obtains the best performance in terms of four evaluation metrics on NJU2K, DES, and SIP datasets. Moreover, it is worth noting that our model obtains better performance 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 *SP-Net* obtains promising performance in locating salient object(s) in a given scene. In addition, we show the PR curves [2] and F-measure curves in Fig. 5. For a clear view, we provide the results of 14 RGB-D saliency detection methods, including 13 SOTA models with complete saliency maps. As observed, the superiority of our model is more visible on these reported datasets.
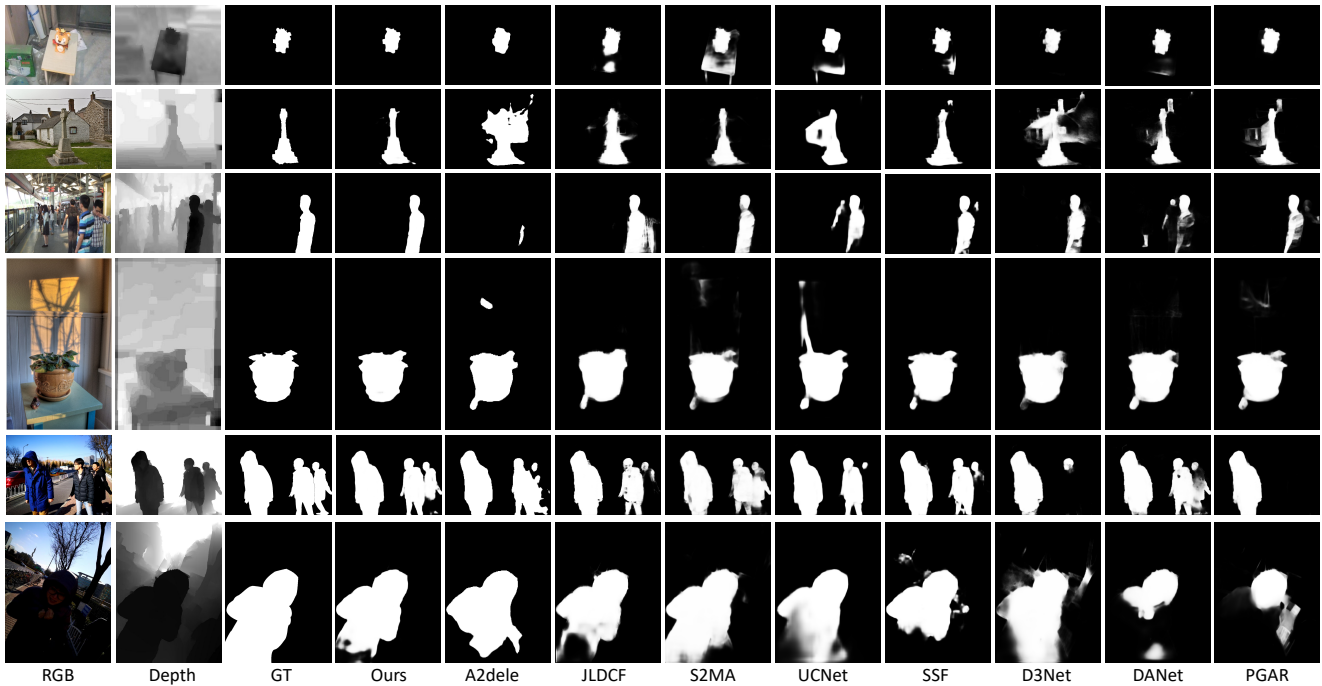
Figure 6. Visual comparisons of our method and eight state-of-the-art methods (including A2dele [49], JLDCF [20], S2MA [38], UCNet [63], SSF [65], D3Net [18], DANet [68] , and PGAR [9].

Qualitative Evaluation: Fig. 6 shows several representative samples of results comparing our model with eight top state-of-the-art methods. The first row shows a scene with a small object. Our method, A2dele, PGAR, and D3Net can accurately detect the salient object, while JLDCF, S2MA, SSF, and UCNet predict some non-object regions. In the 2nd and 3rd rows, we show two examples when the scene is with complex backgrounds. From the comparison results, it can be observed that our method and S2MA produce reliable results, while other RGB-D saliency detection models fail to locate the object or confuse the background as a salient object. In the 4th row, the comparison methods (except D3Net) locates a non-salient and small object. In the 5th row, we show an example with multiple salient objects, where it is challenging to accurately locate all salient objects. Our method locates all salient objects and segments them more accurately, generating sharper edges compared to other approaches. We show an example under low-light conditions in the last row. Some approaches fail to detect the entire extent of the salient object. Our model can produce promising results by suppressing background distractors to boost the saliency detection performance.

4.3. Ablation Study

To verify the relative contribution of different components of our model, we carry out ablation studies by removing or replacing them from our full model.

(A) **Effectiveness of CIM**. Since the proposed CIM is

Table 2. Quantitative evaluation for ablation studies.

	NJU2K [30]		STERE [45]		DES [10]		NLPR [46]		SSD [73]		SIP [18]	
	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
Ours	**.925**	**.028**	**.907**	**.037**	**.945**	**.014**	**.927**	**.021**	**.871**	**.044**	**.894**	**.043**
A1	.916	.034	.898	.042	.939	.016	.926	.022	.869	.047	.892	.044
A2	.921	.031	.895	.042	.938	.016	.925	.022	.865	.051	.896	.042
A3	.919	.032	.895	.043	.938	.016	.929	.020	.864	.049	.887	.048
A4	.924	.029	.903	.038	.930	.019	.927	.023	.867	.049	.888	.046
B1	.918	.034	.901	.041	.939	.017	.922	.024	.858	.050	.885	.048
B2	.924	.029	.900	.041	.941	.015	.926	.022	.864	.049	.893	.044
B3	.921	.031	.903	.039	.938	.016	.925	.022	.863	.050	.891	.045
C	.913	.037	.900	.047	.935	.019	.922	.025	.861	.055	.880	.051

used to fuse cross-modal features and learn their shared representation, we utilize a direct concatenation strategy instead of the CIM. Specifically, the two features f_m^R and f_m^D (as shown in 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 evaluation as "A1" in Table 2. From the comparison results, it can be seen 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 boosting the saliency detection performance. Besides, there are two parts in CIM, *i.e.*, cross-modal feature enhancement and adaptive feature fusion. Thus, to evaluate the contribution of each part, we denote the CIM with only cross-modal feature enhancement or adaptive feature fusion as "A2" and "A3", respectively. When comparing the two independent parts with the full version of the CIM, we can see that the effectiveness of the
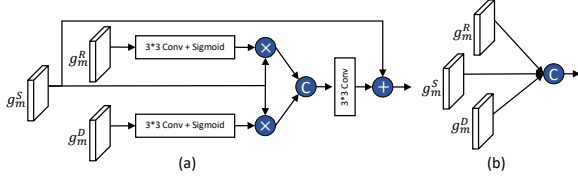
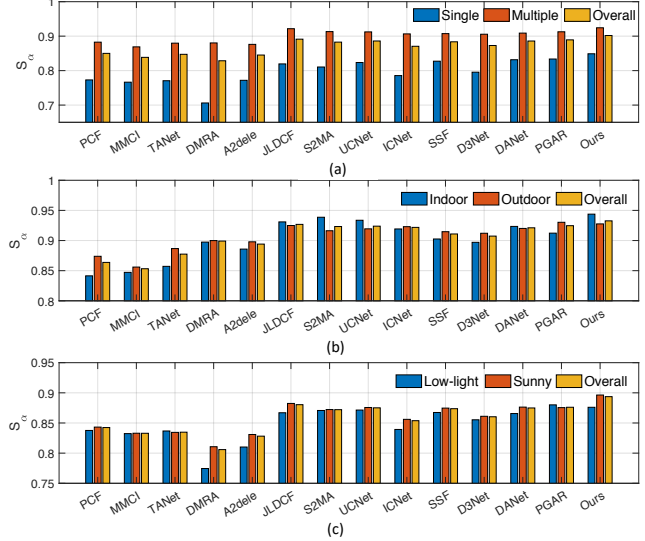Figure 7. Comparison of MFA module with other fusion strategies.

Figure 8. Attribute-based evaluation *w.r.t.* (a) number of salient objects (*i.e.*, single vs. multiple), (b) indoor vs. outdoor environments, and (c) light conditions (low-light vs. sunny).

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 delete this propagation in the CIM, denoted as "A4". The comparison results between "A4" and CIM show that this propagation strategy improves the saliency detection performance.

(B) **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 delete this module, denoted as "B1". Besides, we consider comparing two other feature fusion strategies with our MFA. As shown in Fig. 7, one is the cross-modal feature enhancement fusion; the other is a simple concatenation strategy. The comparison experiments for the two strategies are denoted "B2" and "B3". As shown in Table 2, Comparing "B1" and our full model, the results demonstrate the effectiveness of integrating the features learned into the shared decoder. Comparing "B2" and "B3" with our full model, we can see that the MFA module outperforms both of the other fusion strategies.

(C) **Effectiveness of Modality-specific Decoder**. We delete the two modality-specific decoders, and the evaluation is as shown in "C" of Table 2. It can be observed that the performance will degrade without using the two parts. This indicates the effectiveness of the modality-specific decoder, which can provide supervision signals to ensure that modality-specific properties can be learned.

4.4. Attribute-based Evaluation

There are several challenging factors that affect the performance of RGB-D saliency detection models, such as the number of salient objects, indoor or outdoor environment, light conditions, and so on. Thus, it is interesting to evaluate the saliency detection performance under different conditions, to show the strengths and weaknesses of state-of-the-art models in handling these challenges. (1) Single vs. multiple objects. In this evaluation, we construct a hybrid dataset with 1,229 images collected from the NLPR [46] and SIP [18] datasets. The comparison results using S_α are shown in Fig. 8 (a). As can be observed, it is easier to detect a single salient object than multiple. Besides, our model outperforms other state-of-the-art methods in locating single and multiple objects. (2) Indoor vs. outdoor. DES [10]

and NLPR [46] include indoor and outdoor scenes, we thus construct a hybrid dataset collected from the two datasets. The comparison results are shown in Fig. 8 (b). As can be observed, many models struggle more to detect salient objects in indoor scenes than outdoor scenes, while JLDCF, S2MA, UCNet, ICNet, SSF, DANet, and our model perform a little better in outdoor scenes. (3) Light conditions. We carry out this evaluation on the SIP dataset [18], and the data is grouped into two categories, *i.e.*, sunny and low-light. The comparison results are shown in Fig. 8 (c). As can be seen, all models struggle more to detect salient objects in low-light conditions, confirming that low-light negatively impacts SOD performance.

5. Conclusion

In this paper, we have proposed a novel *SP-Net* for RGB-D saliency detection. Different from most existing works, which mainly focus on learning shared representations, our model not only explores the shared cross-modal information but also compensates modality-specific characteristics to boost the saliency detection performance. Besides, the proposed CIM can propagate information across modalities and layers, while our MFA module can provide specific properties to the shared decoder to enhance the complementary multi-modal information. Quantitative and qualitative evaluations conducted on six challenging benchmark datasets demonstrate the superiority of our *SP-Net* over other existing RGB-D saliency detection approaches. In the future, we can apply our model to the light field saliency detection task [21, 40].

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604. IEEE, 2009. 5, 6

[2] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE TIP*, 24(12):5706–5722, 2015. 6

[3] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML*, pages 129–136, 2009. 3

[4] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *CVPR*, pages 3051–3060, 2018. 1, 2, 5, 6

[5] Hao Chen and Youfu Li. Three-stream attention-aware network for RGB-D salient object detection. *IEEE TIP*, 28(6):2825–2835, 2019. 1, 2, 5, 6

[6] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition*, 86:376–385, 2019. 1, 2, 5, 6

[7] Hao Chen, You-Fu Li, and Dan Su. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In *IEEE IROS*, pages 6821–6826. IEEE, 2018. 2

[8] Ming-Ming Chen and Deng-Ping Fan. Structure-measure: A new way to evaluate foreground maps. *IJCV*, 129:2622–2638, 2021. 5, 6

[9] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In *ECCV*. Springer, 2020. 5, 6, 7

[10] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *ICIMCS*, pages 23–27, 2014. 5, 6, 7, 8

[11] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE TCYB*, 2019. 5, 6

[12] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *SPL*, 23(6):819–823, 2016. 5, 6

[13] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013. 2

[14] Changxing Ding and Dacheng Tao. Robust face recognition via multimodal deep face representation. *IEEE TMM*, 17(11):2049–2058, 2015. 3

[15] Yu Ding, Zhi Liu, Mengke Huang, Ran Shi, and Xiangyang Wang. Depth-aware saliency detection using convolutional neural networks. *Journal of Visual Communication and Image Representation*, 61:1–9, 2019. 1

[16] Andreas Eitel, Jost Tobias Springenberg, Luciano Spinello, Martin Riedmiller, and Wolfram Burgard. Multimodal deep learning for robust rgb-d object recognition. In *IROS*, pages 681–687. IEEE, 2015. 3

[17] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018. 5, 6

[18] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE TNNLS*, 32(5):2075–2089, 2021. 2, 5, 6, 7, 8

[19] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for RGB-D salient object detection. In *CVPR*, pages 2343–2350, 2016. 2, 5, 6

[20] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, Qijun Zhao, Jianbing Shen, and Ce Zhu. Siamese network for RGB-D salient object detection and beyond. *IEEE TPAMI*, 2021. 5, 6, 7

[21] Keren Fu, Yao Jiang, Ge-Peng Ji, Tao Zhou, Qijun Zhao, and Deng-Ping Fan. Light field salient object detection: A review and benchmark. *arXiv preprint arXiv:2010.04968*, 2020. 8

[22] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 2020. 3, 5

[23] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *JMLR*, 12:2211–2268, 2011. 3

[24] Jingfan Guo, Tongwei Ren, and Jia Bei. Salient object detection for RGB-D image via saliency evolution. In *ICME*, pages 1–6. IEEE, 2016. 1, 2, 5, 6

[25] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE TCYB*, 48(11):3171–3183, 2017. 1, 2, 5, 6

[26] Junlin Hu, Jiwen Lu, and Yap-Peng Tan. Sharable and individual multi-view metric learning. *IEEE TPAMI*, 40(9):2281–2288, 2017. 2, 3

[27] Wei Ji, Jingjing Li, Shuang Yu, Miao Zhang, Yongri Piao, Shunyu Yao, Qi Bi, Kai Ma, Yefeng Zheng, Huchuan Lu, et al. Calibrated RGB-D salient object detection. In *CVPR*, pages 9471–9481, 2021. 2

[28] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *ECCV*, 2020. 5, 6

[29] Bo Jiang, Zitai Zhou, Xiao Wang, Jin Tang, and Bin Luo. cmsalgan: RGB-D salient object detection with cross-view generative adversarial networks. *IEEE TMM*, 2020. 5, 6

[30] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119. IEEE, 2014. 2, 5, 6, 7

[31] Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. Depth matters: Influence of depth cues on visual saliency. In *ECCV*, pages 101–115. Springer, 2012. 2

[32] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu Fu, Guopu Zhu, Dingwen Zhang, and Qingming Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE TCYB*, 2020. 5, 6

[33] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In *ECCV*. Springer, 2020. 2, 5, 6

[34] Gongyang Li, Zhi Liu, and Haibin Ling. Icnet: Information conversion network for RGB-D based salient object detection. *IEEE TIP*, 29:4873–4884, 2020. 5, 6

[35] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for RGB-D salient object detection. In *ECCV*. Springer, 2020. 2

[36] Fangfang Liang, Lijuan Duan, Wei Ma, Yuanhua Qiao, Zhi Cai, and Laiyun Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. *Neurocomputing*, 275:2227–2238, 2018. 2

[37] Di Liu, Yaosi Hu, Kao Zhang, and Zhenzhong Chen. Two-stream refinement network for RGB-D saliency detection. In *ICIP*, pages 3925–3929. IEEE, 2019. 1, 2

[38] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for RGB-D saliency detection. In *CVPR*, 2020. 5, 6, 7

[39] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, 2021. 1

[40] Nian Liu, Wangbo Zhao, Dingwen Zhang, Junwei Han, and Shao Ling. Light field saliency detection with dual local graph learning and reciprocative guidance. In *ICCV*, 2021. 8

[41] Zhengyi Liu, Song Shi, Quntao Duan, Wei Zhang, and Peng Zhao. Salient object detection for RGB-D image by single stream recurrent convolution neural network. *Neurocomputing*, 363:46–57, 2019. 1

[42] Yan Lu, Yue Wu, Bin Liu, Tianzhu Zhang, Baopu Li, Qi Chu, and Nenghai Yu. Cross-modality person re-identification with shared-specific feature transfer. In *CVPR*, pages 13379–13389, 2020. 2, 3

[43] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for RGB-D salient object detection. In *ECCV*. Springer, 2020. 5, 6

[44] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *ICML*, 2011. 3

[45] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461. IEEE, 2012. 5, 6, 7

[46] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *ECCV*, pages 92–109. Springer, 2014. 1, 2, 5, 6, 7, 8

[47] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, pages 733–740. IEEE, 2012. 5, 6

[48] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 2, 5, 6

[49] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *CVPR*, 2020. 5, 6, 7

[50] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. RGBD salient object detection via deep fusion. *IEEE TIP*, 26(5):2274–2285, 2017. 2, 5, 6

[51] Konstantinos Rapantzikos, Yannis Avrithis, and Stefanos Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In *CVPR*, pages 1454–1461, 2009. 1

[52] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for RGB-D saliency detection. In *CVPRW*, pages 25–32, 2015. 1, 2

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015. 2, 3

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 3, 5

[55] Wataru Shimoda and Keiji Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In *ECCV*, pages 218–234. Springer, 2016. 1

[56] Hangke Song, Zhi Liu, Huan Du, Guangling Sun, Olivier Le Meur, and Tongwei Ren. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE TIP*, 26(9):4204–4216, 2017. 1, 5, 6

[57] Ningning Wang and Xiaojin Gong. Adaptive fusion for RGB-D salient object detection. *IEEE Access*, 7:55277–55284, 2019. 1, 5, 6

[58] Wenguan Wang, Jianbing Shen, Ruigang Yang, and Fatih Porikli. Saliency-aware video object segmentation. *IEEE TPAMI*, 40(1):20–33, 2017. 1

[59] Jun Wei, Shuhui Wang, and Qingming Huang. F3Net: Fusion, feedback and focus for salient object detection. *AAAI*, 2019. 5

[60] Martha White, Xinhua Zhang, Dale Schuurmans, and Yaoliang Yu. Convex multi-view subspace learning. In *NIPS*, pages 1673–1681, 2012. 3

[61] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019. 4

[62] Changqing Zhang, Qinghua Hu, Huazhu Fu, Pengfei Zhu, and Xiaochun Cao. Latent multi-view subspace clustering. In *CVPR*, pages 4279–4287, 2017. 3

[63] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Saleh, Sadegh Aliakbarian, and Nick Barnes. Uncertainty inspired RGB-D saliency detection. *IEEE TPAMI*, 2021. 2, 5, 6, 7

[64] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Xin Yu, Yiran Zhong, Nick Barnes, and Ling Shao. RGB-D saliency detection via cascaded mutual information minimization. In *ICCV*, 2021. 1

[65] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for RGB-D saliency detection. In *CVPR*, 2020. 5, 6, 7

[66] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *CVPR*, pages 3927–3936, 2019. 1, 2, 5, 6

[67] Rui Zhao, Wanli Oyang, and Xiaogang Wang. Person re-identification by saliency learning. *IEEE TPAMI*, 39(2):356–370, 2016. 1

[68] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *ECCV*. Springer, 2020. 5, 6, 7

[69] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. *Computational Visual Media*, pages 1–33, 2021. 1

[70] Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal MR image synthesis. *IEEE TMI*, 39(9):2772–2781, 2020. 2

[71] Tao Zhou, Changqing Zhang, Xi Peng, Harish Bhaskar, and Jie Yang. Dual shared-specific multiview subspace clustering. *IEEE TCYB*, 50(8):3517–3530, 2019. 2

[72] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *ICME*, pages 199–204, 2019. 1, 2

[73] Chunbiao Zhu and Ge Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *ICCVW*, pages 3008–3014, 2017. 5, 6, 7

[74] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *ICCVW*, pages 1509–1515, 2017. 2, 5, 6

[75] Jun-Yan Zhu, Jiajun Wu, Yan Xu, Eric Chang, and Zhuowen Tu. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE TPAMI*, 37(4):862–875, 2014. 1