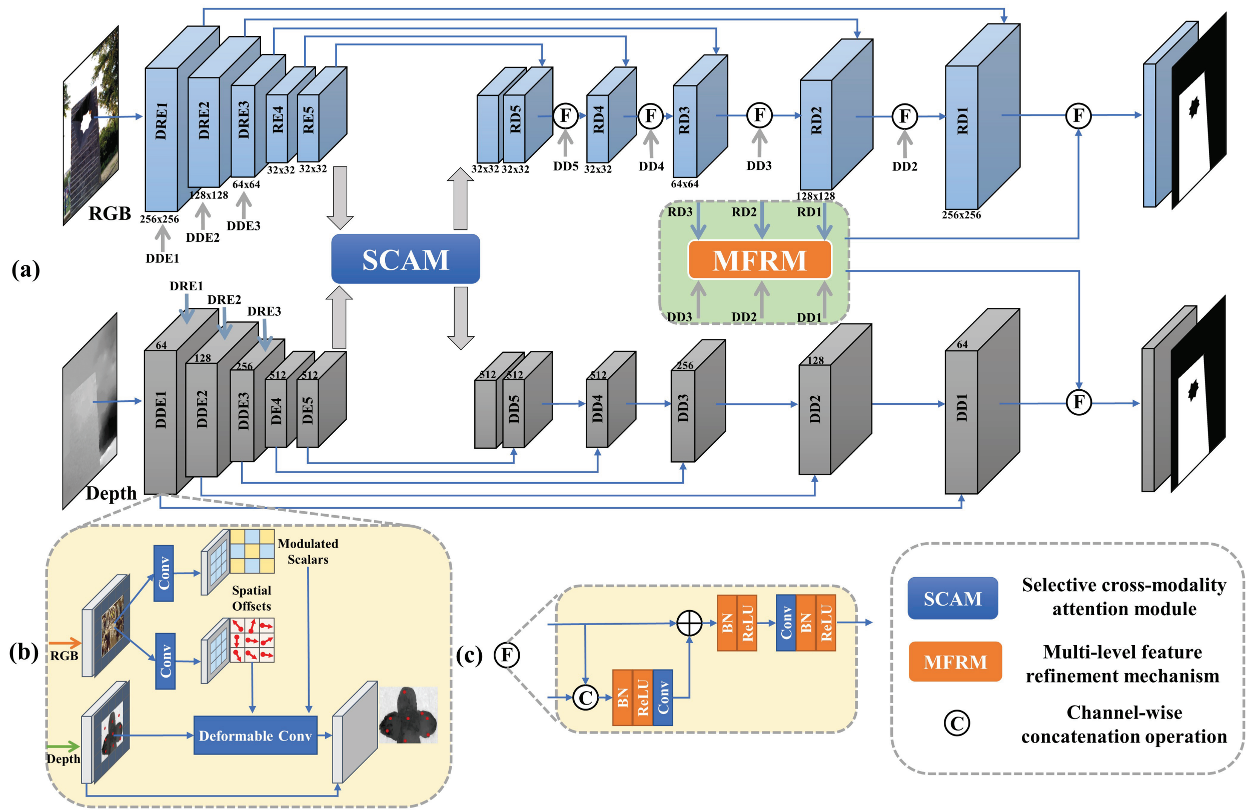
Graphical Abstract

AMDFNet: Adaptive Multi-level Deformable Fusion Network for RGB-D Saliency Detection

Fei Li,Jiangbin Zheng,Yuan-fang Zhang,Nian Liu,Wenjing Jia



Highlights

AMDFNet: Adaptive Multi-level Deformable Fusion Network for RGB-D Saliency Detection

Fei Li,Jiangbin Zheng,Yuan-fang Zhang,Nian Liu,Wenjing Jia

- We propose a selective cross-modality attention module that adaptively integrates the information from both modes to reduce the fusion ambiguity caused by unreliable inputs and maximally retain the realistic details.
- The proposed cross-modality deformable module can extract additional cues from another branch to adaptively alter the sampling locations and cover the irregular boundaries of the salient objects.
- The multi-level feature refinement mechanism is able to fuse cross-modality features in multiple scales and incredibly aggregate those unique cues from small size features.

AMDFNet: Adaptive Multi-level Deformable Fusion Network for RGB-D Saliency Detection

Fei Li[a,1], Jiangbin Zheng[a,b,*], Yuan-fang Zhang[b,c], Nian Liu[d] and Wenjing Jia[c]

[a] *School of Software and Microelectronics, Northwestern Polytechnical University, P. R. China*
[b] *School of Computer Science and Engineering, Northwestern Polytechnical University, P. R. China*
[c] *Faculty of Engineering and IT, University of Technology Sydney, Australia*
[d] *Mohamed bin Zayed University of Artificial Intelligence, United Arab Emirates*

ARTICLE INFO

Keywords:
salient object detection
multi-modality fusion
cross-modality deformable convolution
RGB-D

ABSTRACT

Effective exploration of useful contextual information in multi-modal images is an essential task in salient object detection. Nevertheless, the existing methods based on the early-fusion or the late-fusion schemes cannot address this problem as they are unable to effectively resolve the distribution gap and information loss. In this paper, we propose an adaptive multi-level deformable fusion network (AMDFNet) to exploit the cross-modality information. We use a cross-modality deformable convolution module to dynamically adjust the boundaries of salient objects by exploring the extra input from another modality. This enables incorporating the existing features and propagating more contexts so as to strengthen the model's ability to perceiving scenes. To accurately refine the predicted maps, a multi-scaled feature refinement module is proposed to enhance the intermediate features with multi-level prediction in the decoder part. Furthermore, we introduce a selective cross-modality attention module in the fusion process to exploit the attention mechanism. This module captures dense long-range cross-modality dependencies from a multi-modal hierarchical feature's perspective. This strategy enables the network to select more informative details and suppress the contamination caused by the negative depth maps. Experimental results on eight benchmark datasets demonstrate the effectiveness of the components in our proposed model, as well as the overall saliency model.

1. Introduction

In salient objection detection (SOD), the main objective is to extract the most predominant objects from a natural scene. It has been an essential function in computer vision since SOD has many useful applications, including image/video compression [18, 27], object segmentation and recognition [68, 67, 44, 23], content-based image editing [52, 55], informative common object discovery [63, 64], and image retrieval [47]. Many SOD methods are based on the assumption that the inputs are RGB images [40, 54, 57, 53, 66] or video sequences [56, 25].

With the advancement of the depth cameras such as Microsoft Kinect and time-of-flight sensors [20], the SOD based on the RGB-D ('D' means the depth images) offers new opportunities, where the depth images provide complementary cues that are not available in the RGB images. Such cues are game-changers in challenging SOD scenarios, *e.g.,* cluttered background or salient objects that have similar appearance with the background, as shown in Fig. (1). Compared with the SOD using RGB images, the depth information, if available, supplies geometric cues that are otherwise invisible in color space. This significantly enhances the final predicted maps and has motivated the extensive recent research activities on RGB-D based salient object detection.

In the existing research, several studies [9, 10, 8] have investigated designing hand-crafted features with domain-specific knowledge, such as the tendency of humans to focus on the center objects for saliency detection. However, using hand-crafted features lacks generalization ability and hence is not applicable to other scenes, mainly due to missing high-level representations.

Figure 1: Several low-quality depth samples obtained from the existing RGB-D SOD benchmarks. The first row shows the RGB images and the second row their depth samples.)

To address the generalization issue, relevant investigations have been proposed using convolution neural networks (CNNs) to learn the representative features. Several studies [2, 46] have also attempted to overcome the limitation caused by missing high-level representations by incorporating the depth information effectively.

Although in many SOD research works, the strategies for cross-modality fusion have been investigated, the following issues still exist. First of all, the main challenge for the existing SOD methods is the lack of sufficient high-quality depth datasets for training the backbone networks and extracting the critical features. Secondly, the need for large datasets is due to the sophisticated architecture of the networks [2, 3] with many parameters. These issues have undermined fea-

*Corresponding author
lflovelife@mail.nwpu.edu.cn (F. Li); zhengjb@nwpu.edu.cn (J. Zheng); zyf.robinzhang@gmail.com (Y. Zhang); liunian228@gmail.com (N. Liu); Wenjing.Jia@uts.edu.au (W. Jia)
ORCID(s): 0000-0001-7511-2910 (F. Li)

ture extraction and led to sub-optimal solutions. Moreover, the existing RGB-D benchmarks are collected by different laboratories who have used different metrics for choosing and labeling the images. This results into some low-quality depth images being included which contribute little or even negatively to the training. These low-quality samples may further affect the accuracy of the final saliency detection, especially if the adopted method indiscriminately integrates the RGB and depth information. The fusing strategy and capturing sufficient cross-modality complementary information also play critical roles in RGB-D SOD. The selective fusion scheme is adopted in the fusing process to prevent the contamination caused by unreliable depth information and effectively integrate the multi-modal information. Therefore, it is essential to address the negative impact of the low-quality depth images and select reliable and accurate information in the fusion process.

The existing works have explored different contributions between the early- [41, 21, 33, 46] and late-fusion [51]. Specifically, the early-fusion schemes take both RGB and depth data as inputs and process them in a unified mode. However, such a fusion strategy ignores the distribution gap and different feature characters in both modalities. It is also not easy for one model to fit both modalities. By comparison, the late-fusion strategy means that the data of both modalities are handled in two separate processing branches to produce the corresponding saliency maps. Both maps are then designed through a concentration operation. Nevertheless, the major issue with this scheme is the inner supervision between both modalities. The rich cross-modality cues are also compressed and lost in the two separate branches.

Both of the fusion strategies mentioned above result in the learning process being trapped in a local optimum, where it becomes biased towards the RGB information. This is due to the channel concatenation degrading the learning outcomes, where the final prediction is dominated by the RGB features without incorporating the contribution of the cross-modality informative feature. To enhance the fusion process of the depth maps, several works [2, 3, 4, 19] proposed middle-fusion strategies to conduct intermediate independent features by two-stream CNNs. Such a network is then used to simultaneously extract independent hierarchical features from the RGB and depth images. Both features are then integrated to eliminate the distribution gap. This scheme further introduces rich cross-modality features with well-designed intermediate processing actions. Hence, the desired fusion method can consider different properties in both modalities and adaptively alter the contribution of both modalities in the final prediction results.

To address the abovementioned issues, we revisit the fusion process of cross-modality complementary and propose a novel adaptive multi-level deformable fusion network (AMDFNet) for the RGB-D SOD. Our approach comprises of the adaptive adjustment of the salient objects' boundaries in both modalities. We further optimize the fusion process of RGB and depth information based on a selective cross-modality attention mechanism.

In our approach, instead of indiscriminately integrating multi-modal information from RGB and depth maps, we devise a selective cross-modality attention module (SCAM). The SCAM captures the long-range dependencies from a multi-level cross-modality perspective. The obtained attention associations, along with the existing local and multi-scale features in the other modality, facilitate the fusion process by highlighting the salient objects. Inspired by the Non-local (NL) operation [59], the SCAM also supplies extra complementary cues on more important contextual features that should be emphasized in propagating the features. This improves the accuracy of locating salient object boundaries.

To further enhance the independent hierarchical features simultaneously from both views, we also introduce a novel feature refinement scheme. Here, we first design a cross-modality deformable convolution module (CDCM) based on the standard deformable convolution operation [12]. This module adjusts the boundaries of the salient objects in both modes to prevent contamination caused by unreliable depth maps. The CDCM also emphasizes the salient regions and object boundaries. As shown in Fig. (1), several depth samples lost the details of salient objects because of the cluttered background. This may result in low-quality features being extracted by both feature extraction branches. The CDCM extracts accurate geometric boundaries of the salient objects using both modalities to regulate the negative samples' training by emphasizing the geometric boundaries. This significantly reduces the negative impact of these samples. Specifically, another modality feature provides offsets to adjust the filter boundaries, hence resulting in the convolution block to emphasize the image content, with the nodes on the foreground having support for covering the whole target object. In contrast, other nodes in the background are ignored to better focus on the salient target.

Moreover, we employ a multi-level feature refinement mechanism (MFRM) to improve the integration of different levels of hierarchical features in the decoding stage. Different modalities are not equally informative or beneficial to the final segmented map. This is because some images or depth information are affected by imperfect alignment or direct concatenation. Besides, it is challenging to compensate the details of modalities explicitly or implicitly within a single resolution scale. To address this issue, we introduce the MFRM to further improve the performance of the precision maps from various feature levels in both modalities. In the MFRM module, the depth features provide the learning offset and the modulated scalar for the image features, whereas the image features provide the corresponding coefficients for the depth branch. By introducing the deformable convolution operation, the network decoder block adaptively adjusts the reference image and supporting information at the feature level without warping and blurring, which are usually caused by direct concatenation.

The main contributions of this work are summarized as follows: 1) This paper proposes a selective cross-modality attention module that adaptively integrates the information from both modes, reducing the fusion ambiguity caused by unreliable inputs and maximally reserving realistic details. 2) The proposed cross-modality deformable convolution module can extract additional cues from another branch to adaptively alter the sampling locations and cover the irregular boundaries of the salient objects. 3) The multi-level feature refinement mechanism aims to fuse cross-modality features in the multi-scale terms, incredibly aggregating some unique cues from small size features.

2. Related Work

In this section, we review the salient object detection models for RGB and RGB-D images with a focus on deep learning based methods.

2.1. Saliency Detection on RGB-D Images

The conventional methods for RGB-D SOD predict high-quality saliency maps via hand-crafted features based on image characteristics such as contrast and shape. Niu *et al.* [35] introduced the disparity contrast and domain knowledge into stereoscopic photography for measuring the stereo saliency. Several other SOD studies relying on hand-crafted features were also extended for RGB-D SOD, *e.g.,* based on contrast [8, 11, 36], boundary prior [9, 29, 50], or compactness [10]. Since the above methods heavily rely on hand-crafted heuristic features, they often have limited generalizability to more complex scenarios.

Furthermore, in the existing methods, domain knowledge priors induced by both 2D images and RGB-D cues have not been exploited. This is often addressed by the CNN-based methods. Such methods outperform the traditional methods because of their enhanced representativeness. Most of the recent advances in SOD [38, 31, 15] are based on CNNs.

The depth maps also supply extra details that are invisible in RGB images. Emerging deep learning-based approaches have also been adopted and become a mainstream approach in RGB-D SOD. Qu *et al.* adopted an early fusion strategy to handle hand-crafted RGB and depth features together as inputs to the CNN. Besides, early fusion schemes in [15, 21, 33] formulated four-channel inputs, treating the depth map as the 4^{th} channel of the corresponding RGB images as the CNN inputs. Unlike the early fusion for an extra channel, the middle fusion strategy is adopted in [2, 3, 4, 19] to fuse intermediate depth and RGB features. Specifically, Chen *et al.* [2] proposed a complementarity-aware fusion module to obtain cross-modality and cross-level features. Besides, Wang *et al.* [51] used a switch map to adaptively fuse the RGB images with depth saliency maps. Chen *et al.* [6] introduced the depth map enhancement module to improve the salient object performance.

2.2. Self-Attention to Cross-Modality Attention

Vaswani *et al.* [48] proposed a self-attention network for language learning. In their proposed network, they first calculated the attention weight between the query and each key in a set of key-value pairs. Then, they aggregated the values through a weighted sum with the attention weights as the final output. Motivated by various approaches, Wang *et al.* [59] then proposed the NL model for learning self-attention in computer vision. Nam *et al.* [34] also proposed a dual attention model to learn multi-modal attention. Wan *et al.* [49] extracted three-modality attention for a code retrieval task.

In RGB-D SOD, standard self-attention cannot meet the requirement, and cross-modality attention influence should be considered. In this paper, we propose a fusion scheme to accurately extract multi-scale cross-modality attention from both modality views in this work.

2.3. Deformable Convolutional Network

A deformable convolution network [12, 69] adaptively determines the object scales or receptive field sizes without being affected by the fixed structures of the convolution kernels. Dai *et al.* [12] proposed deformable convolutional networks (DCNs), where additional offsets were learned to allow the network to obtain information from its regular local neighborhood. This improved the capability of the regular convolutions. Based on the DCNs, Zhu *et al.* [69] then proposed the modulation deformable convolution network, which introduced an additional modulated scale to enable the adaptive scale to control the learned offsets.

Deformable convolutions are widely used in various image processing applications, such as semantic segmentation [12], video super-resolution [58], object detection [7], SOD [17, 30] and video SOD [5].

3. Methodology

Here, we propose a novel cross-modality fusion model for the RGB-D images to improve the SOD performance. We first briefly review the deformable convolution networks and then design a cross-modality deformable convolution module (CDCM). We then devise a multi-level feature refinement mechanism (MFRM) which integrates cross-modality features from coarse features to fine features. We then propose a selective cross-modality attention module (SCAM) for fusing informative and complementary details using multi-scale features extracted in the pyramid non-local block. Finally, we describe the implementation details of the proposed RGB-D SOD system and the corresponding hybrid loss function.

3.1. Modulation Deformable Convolutional Network

It is generally challenging to extract the desired cross-modality features in SOD using the RGB-D data. The CNNs of the cascaded standard convolution layers are also limited by the fixed geometric structure of the standard convolution filters. Therefore, they are often unable to adaptively fuse useful features in both modalities. Since salient objects generally have arbitrary sizes and compositions, especially in their depth maps, the regular-gridded sampling filters impose feature extraction from the rectangular regions. This results in lower-quality features and hence degrades the SOD performance.

The primary motivation for adopting the modulation deformable convolutional networks (DCNV2) is to lead the

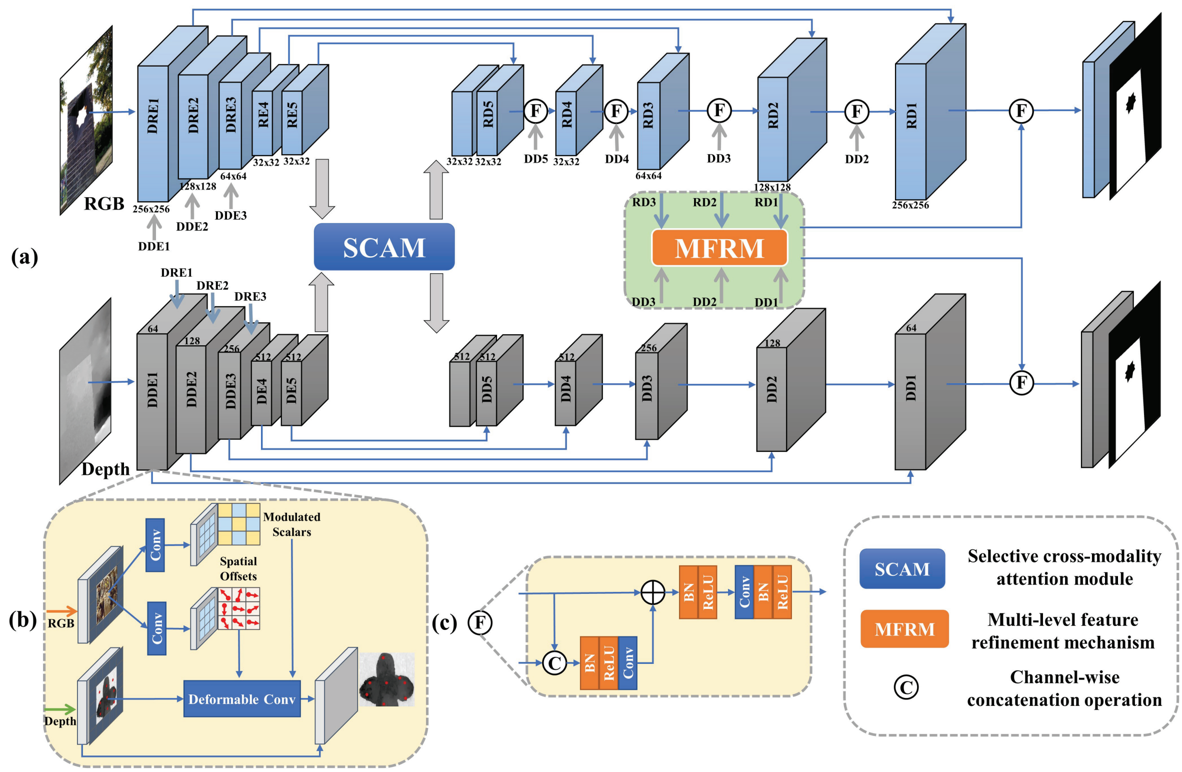


Figure 2: The network architecture of the proposed RGB-D saliency detection network. (a) Overview of our propose network architecture. The whole network is a two-steam CNN architecture, which consists of a RGB and a depth branch. $\mathbf{DRE_i}$ and $\mathbf{DDE_i}$ $(i = 1, 2, 3)$ denote the features generated by the beginning three layers with **cross-modality deformable convolution module** at encoding stage of both branches respectively, and $\mathbf{RE_i}$ and $\mathbf{DE_i}$ $(i = 4, 5)$ are the features generated from normal convolutional blocks. The $\mathbf{RD_i}$ and $\mathbf{DD_i}$ $(i = 5, 4, \cdots, 1)$ represent the features of both decoder stages. (b) The architecture of the cross-modality deformable convolution module (CDCM). (c) Details of the feature fusion operation.

SOD network for locating adaptive neighborhoods for each pixel position in the intermediate feature maps. The pixels in the current position and the corresponding details from another branch enhance these cross-modality features in the RGB or depth modality.

The DCNV2 [69] adjusts offsets in perceiving the input features and further modulates the amplitudes of the input feature from different spatial samples. Therefore, the DCNV2 can vary the spatial distribution and the relative influence of its samples. Specifically, the offset dynamically extends the size of the receptive field to obtain the desired convolutional region. The learning modulation mechanism also provides the network module with an extra degree of freedom to adjust its spatial support regions.

Compared with the standard convolution layer, the DCNV2 emphasizes the irregularity and variety of the object structures. This is because DCNV2 changes the sampling location of the convolution kernels by adding the offsets and modulated scalars. Moreover, both coefficients are adaptive and can highlight the significant boundaries, and hence suppress the unnecessary regions extracted by the standard convolution rectangular filter. The DCNV2 then adaptively expands the receptive field for the object according to its size. The dynamic receptive fields further ensure that the feature map of the object responds to the target and removes those unnecessary regions without informative details.

In the DCNV2, images for post Δp_k and Δm_k are the learning offset and the modulation scalar for the k-th location, respectively, *i.e.,* K is the number of locations within the convolution grid. A 3×3 kernel is defined with $K = 9$ and $p_k \in \{(-1, -1), (-1, 0), \cdots, (1, 1)\}$ which denotes a 3×3 convolutional kernel with a dilation of 1. Besides, the modulation scalar Δm_k is in $[0, 1]$. Both coefficients are obtained via a 1×1 convolution layer applied over the same input feature map x as shown in Fig. (2)-(b). Hence, the modulated deformable convolution can be written as:

$$y(p) = \sum_{k=1}^{K} w_k \cdot x(p + p_k + \Delta p_k) \cdot \Delta m_k. \tag{1}$$

The output has $3K$ channels, where the first $2K$ channels correspond to the learned offsets Δp_k, and the remaining K channels are fed into a sigmoid layer to obtain the modulation scalars Δm_k. The learning offsets Δp_k are usually fractional, and hence bilinear interpolation [12] is adopted to ensure an integer value. The initial values of Δp_k and Δm_k are 0 and 0.5, respectively.

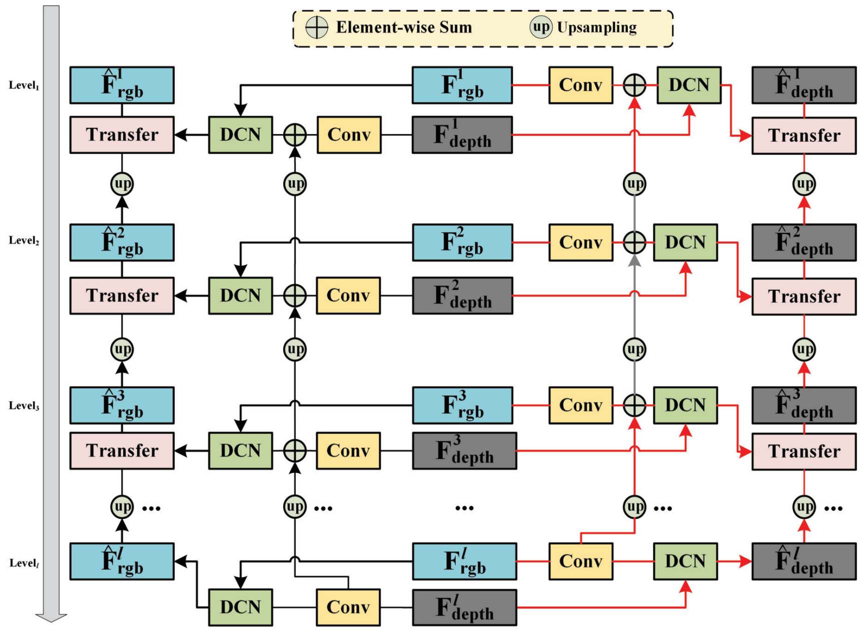


Figure 3: The details of our proposed multi-level feature refinement mechanism (MFRM). The black and red lines denote the image and the corresponding depth processing branch, respectively.

3.1.1. Cross-modality Deformable Convolution Module

As demonstrated in Fig. (1), there are several low-quality depth images in these widely used RGB-D SOD datasets. If we only regard the two processing branches without necessary treatments, these negative samples will affect the final prediction map. Moreover, it is challenging for conventional feature extractors (*e.g.,* VGG or ResNet) to extract the desired features in the separate stream for RGB and depth maps. The considerable distribution gap between the data in both modalities data worsens the issue.

To address this issue, we adopt the deformable progressive extraction strategy to adaptively extract the cross-modality details. Based on the DCNV2, we propose the cross-modality deformable convolution module (CDCM) as shown in Fig. (2)-(b), which receives the features of another branch to produce the modulated scalars and offsets. The offsets and scalars learned by the depth maps provide the accurate position of the salient objects for the image branch. This is because the depth images effectively locate the boundary of the significant objects. The geometric transformation ability enables the feature extractor to obtain more accurate boundary information. Nevertheless, the image details also provide offsets and modulated scalars for depth information, ensuring that the complementary details contain the saliency regions so as to reduce the negative effect caused by the background and non-salient objects.

Here, we employ CDCM to guide the cross-modality feature extraction, which can dynamically adjust the receptive field to focus on the object body in the saliency boundaries together. In our design, we replace the traditional convolution layer with the module at the first three encoder blocks (*i.e.*, $\mathbf{DRE_i}$ and $\mathbf{DDE_i}$ $i \in \{1, 2, 3\}$).

We consider the additional features consisting of the RGB and depth information F^r and F^d, where $(\cdot)^r$ and $(\cdot)^d$ indicate whether the parameter serves in the RGB image or depth branch. We further assume that both features can predict the desired values of Δp_k and Δm_k adopted in DCNV2 [69] for other branches. This enables the supply of more accurate information through learnable offsets and modulated scalars.

The detailed processing can be expressed as:

$$F^r(p) = \sum_{k=1}^{K} w_k^r \cdot F^r(p + p_k + \Delta p_k^d) \cdot \Delta m_k^d \tag{2}$$

and

$$F^d(p) = \sum^{K} w_k^d \cdot F^d(p + p_k + \Delta p_k^r) \cdot \Delta m_k^r, \tag{3}$$

where

$$\begin{aligned} \Delta p^d &= Conv(F^d) \\ \Delta m^d &= Conv(F^d) \\ \Delta p^r &= Conv(F^r) \\ \Delta m^r &= Conv(F^r). \end{aligned}$$

Here, the module receives F^r and F^d as its inputs and then extracts the enhanced cross-modality features $\hat{F}^r$ and $\hat{F}^d$ as:

$$\hat{F}^r = CDCM(F^r, F^d) + F^r \tag{4}$$

and

$$\hat{F}^d = CDCM(F^d, F^r) + F^r. \tag{5}$$

Using this module, the cluttered background and unclear salient object get highlighted using the information from the

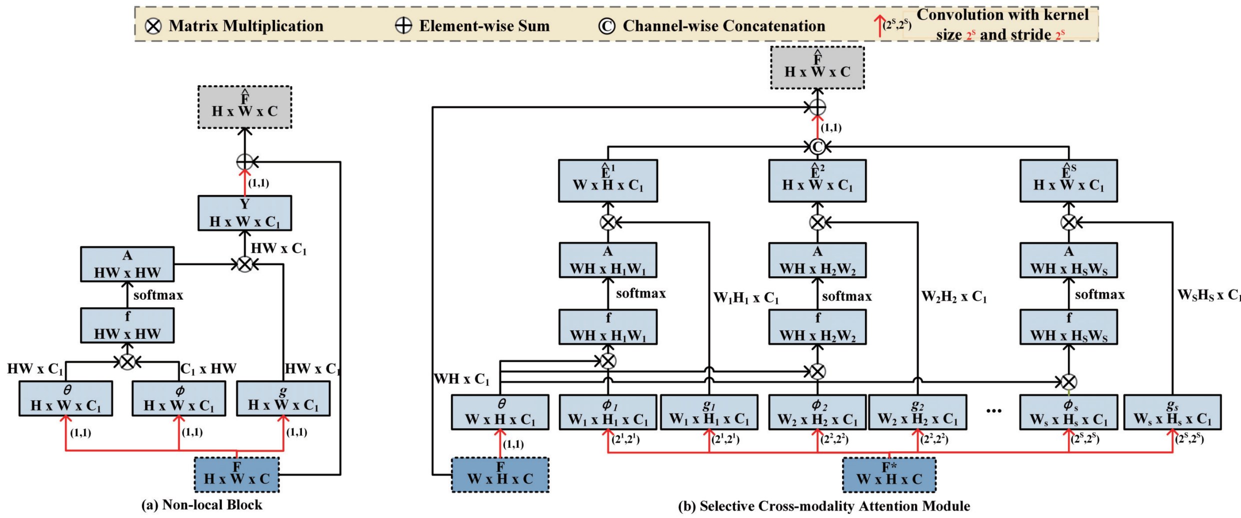


Figure 4: The architecture of the prior non-local block (a) and the proposed Selective Cross-Modality Attention Module (SCAM) (b). In SCAM, input features F and additional features F^* are the output from the RGB and depth encoder streams respectively. ϕ_s and g_s are computed by multi-scale feature in F^*, while θ transformed by F is shared in all scales. Besides, the SCAM is symmetrical and we denote the depth and RGB features as F and F^*, respectively.

other branch. The irregular object structures can then be accurately sampled. These adaptively-learned parameters then adjust the boundary of the receptive field to recover more critical details and remove the regions with irrelevant background.

3.1.2. Semantic Feature Refinement

In multi-modality feature fusion, it is essential to prevent the contamination introduced by unreliable depth maps. To achieve this goal, we design a multi-level feature refinement mechanism (MFRM), as demonstrated in Fig. (3), to combine the inner cues existing in features with different sizes. This leads to a more primitive visual context covering different scales and shapes of the non-rigid salient objects. The proposed MFRM is a symmetrical structure consisting of two paths, *i.e.*, RGB and depth streams. The MFRM aggregates the features with different scales in both modalities. This reduces the interference of different modalities of the single-sized features.

Here, we obtain features $[F^1_{rgb}, F^2_{rgb}, F^3_{rgb}]$ and $[F^1_{depth}, F^2_{depth}, F^3_{depth}]$ from the image decoder module ($\mathbf{RD_3}$-$\mathbf{RD_1}$) and the depth decoder module ($\mathbf{DD_3}$-$\mathbf{DD_1}$), respectively. We then employ a 3×3 Conv layer to obtain the sampling position offsets Δp and controlling scalar Δm from F^l_{rgb} or F^l_{depth}. Besides, the DCN receives the learning parameters and original feature F^l_{rgb} or F^l_{depth}. This means the intermediate scaled features $\hat{F}^l_{rgb}$ and $\hat{F}^l_{depth}$ can extract different cross-modality cues and cover more details.

To ensure the training flexibility, we sum the l-th learning parameters with the upper value in $(l+1)$-th level, processed by one ×2 upsampling operation. Hence, the Δp and Δm for RGB and depth in different spatial level are defined as follows:

$$\Delta p^l_{rgb} = Conv(F^l_{depth}) + (\Delta p^{l+1}_{rgb})^{up\times 2} \tag{6}$$

$$\Delta p^l_{depth} = Conv(F^l_{rgb}) + (\Delta p^{l+1}_{depth})^{up\times 2} \tag{7}$$

$$\Delta m^l_{rgb} = Conv(F^l_{depth}) + (\Delta m^{l+1}_{rgb})^{up\times 2} \tag{8}$$

$$\Delta m^l_{depth} = Conv(F^l_{rgb}) + (\Delta m^{l+1}_{depth})^{up\times 2} \tag{9}$$

where $Conv$ represents a 1×1 convolution layers and l indicates the spatial level.

Based on Eq. (6) to Eq. (9), the enhanced features $\hat{F}^l_{rgb}$ and $\hat{F}^l_{depth}$ are handled with the input parameters Δm^l and Δp^l. It is then concentrated with the upper one $\hat{F}^{l+1}$ as:

$$\hat{F}^l_{rgb} = T(DCN(F^l_{rgb}, \Delta p^l_{rgb}, \Delta m^l_{rgb}), (\hat{F}^{l+1}_{rgb})^{up\times 2}), \tag{10}$$

and

$$\hat{F}^l_{depth} = T(DCN(F^l_{depth}, \Delta p^l_{depth}, \Delta m^l_{depth}), (\hat{F}^{l+1}_{depth})^{up\times 2}), \tag{11}$$

where $(\cdot)^{up\times 2}$ denotes the up-sampling operation by a factor of 2, T represents a transfer module and consists of a concentration operation and a 1×1 convolution layer to reduce the channel dimension. The outputs $\hat{F}^1_{rgb}$ and $\hat{F}^1_{depth}$ denote the enhanced features for RGB and depth stream, respectively. Here l is set to 3.

3.2. Selective Cross-modality Attention Module

The existing approaches [3, 4, 19] that adopted the middle-fusion strategy have treated the intermediate features of both modalities equally. However, considering that there is complementarity due to the inconsistency of the cross-modality RGB-D data (*e.g.*, contamination from unreliable depth maps), direct integration of the cross-modality information may introduce negative results. Hence, it is essential yet challenging to capture the pertinent details of the feature fusion process, especially the depth image.

To address the uncertainty issue of the fusing features, we propose an information selection module SCAM. The SCAM strengthens the important features containing helpful complementary information using an attention strategy. The proposed SCAM aims to capture the long-range dependencies existing between the multi-level RGB and depth features.

A non-local (NL) [59] structure is proposed to exploit the channel and spatial relationship between all pixels. As demonstrated in Fig. (4)-(a), $X \in \mathbb{R}^{H\times W\times C}$ denotes the input feature activation map, where H, W, C refer to the height, weight and channel, respectively. The enhanced feature representation $\mathbf{Z}$ is defined as:

$$\mathbf{Z} = \mathcal{T}\left(\frac{1}{\mathcal{D}(\mathbf{F})}\mathcal{M}(\mathbf{F})\mathcal{G}(\mathbf{F})\right) + \mathbf{F}, \tag{12}$$

where $\mathcal{M}(\mathbf{F}) \in \mathbb{R}^{HW\times HW}$ is the self-similarity matrix, and $\mathcal{G}(\mathbf{F}) \in \mathbb{R}^{HW\times C_1}$ denotes the channel transformation operation responsible for deducing the channel dimension from C to C_1. In general, C_1 is set as $C/2$ to reduce the computational cost. Besides, $\mathcal{D}(\mathbf{F})$ produces a diagonal matrix for normalization purposes. Here, we adopt the $Softmax$ operation to normalize the intermediate features. Furthermore, $\mathcal{T}(\cdot)$ reproduces the enhanced feature back into its original channel dimension. Specifically, $\mathcal{T}(\cdot)$ applies a 1×1 Conv layer to recover the feature from C_1- to $C-$dimension.

The correlation matrix $\mathcal{M}$ and $\mathcal{G}$ are defined as:

$$\begin{aligned}\mathcal{M}(\mathbf{F}) &= \exp\left(\mathcal{F}_{emb}\left(\mathbf{F}, \mathbf{W}_\theta\right)\mathcal{F}_{emb}\left(\mathbf{F}, \mathbf{W}_\phi\right)^{\mathrm{T}}\right)\\ \mathcal{G}(\mathbf{F}) &= \mathcal{F}_{emb}\left(\mathbf{F}, \mathbf{W}_g\right)\end{aligned} \tag{13}$$

where $\mathcal{F}_{emb}(\mathbf{F}, \mathbf{W})$ is implemented using a 3×3 Conv layer of parameters W (*i.e.,* W_θ, W_ϕ and $W_g \in \mathbb{R}^{C\times C_1}$ are the embedding weights). In $\mathcal{M}(\mathbf{F})$, each element $f_{i,j}$ denotes the affinity between the i-th and j-th spatial locations in $\mathbf{F}$.

By exploiting the long-range dependencies of the image pixel or region in both modalities, we create an attention map for each branch. The attention map indicates the extent of information contribution from another one.

Nevertheless, there exist two limitations. First, the computational complexity and memory usage of the correlation matrix increase quadratically with the increase of the size of the input features. The second limitation is that the direct processing of the single-sized features might not fully exploit the hidden cues and unable to obtain optimal predictions. These challenge the utilization of a selective cross-modality attention module for the large feature.

To address the computational complexity issue and establish the cross-modality attention association, we propose the SCAM to exploit the mutual attention in both modalities. To do this, the SCAM computes the selective attention map at the multi-level feature level. Here, we take the RGB features as the target source, and the depth features as the reference. In other words, we establish the attention association between the original RGB features and corresponding depth features in multi-size.

Specifically, taking the enhancement of the RGB features $\hat{\mathbf{F}}_r$ as an instance. The $\hat{\mathbf{F}}_r$ denotes the feature by the concentration of embedding depth features $\hat{\mathbf{E}}_d^s$ as shown in Fig. (4)-b. Here, we take the input consisting of $F_r \in \mathbb{R}^{H\times W\times C}$ and the depth features $F_d^s \in \mathbb{R}^{H\times W\times C}$ to create the attention relationships among multi-scale features. The self-similarity matrix $\mathcal{M}(\mathbf{F})$ and transformation operation $\mathcal{G}(\mathbf{F})$ in the s-th level are defined as:

$$\begin{aligned}\mathcal{M}(\mathbf{F_r^s}) &= \exp\left(\mathcal{F}_{emb}\left(\mathbf{F}_d^s, \mathbf{W}_\theta^s\right)\mathcal{F}_{emb}\left(\mathbf{F}_d^s, \mathbf{W}_\phi^s\right)^{\mathrm{T}}\right)\\ \mathcal{G}(\mathbf{F_r^s}) &= \mathcal{F}_{emb}\left(\mathbf{F}_d^s, \mathbf{W}_g^s\right)\end{aligned} \tag{14}$$

The kernel size and stride of the convolutional layer for the depth feature in the s-th scale are set to 2^s, whereas the values in the image features are set to 1. Because the proposed module employs downsampling depth features to compute the weights $\mathbf{W}_\theta$ and $\mathbf{W}_\phi$, the rows in both weights are reduced to $HW/4^s$. This significantly reduces the computational complexity of obtaining the self-similarity matrix.

Furthermore, the enhanced embedding features $\mathbf{E}^s$ is obtained as:

$$\hat{\mathbf{E}}^s = \frac{1}{\mathcal{D}(\mathbf{F}^s)}\mathcal{M}(\mathbf{F}^s)\mathcal{G}(\mathbf{F}^s) \quad (s \in \{1, \cdots s\}) \tag{15}$$

The embedded features are concatenated together, followed by a 1×1 convolution layer to reproduce its channel from $sC1$ to C. Therefore, the final output in both branches processed by the SCAM are:

$$\hat{\mathbf{F}}_{rgb} = \mathcal{T}\left(\left[\hat{\mathbf{E}}_{rgb}^1, \cdots, \hat{\mathbf{E}}_{rgb}^s\right], \mathbf{W}_\psi\right) + \mathbf{F}_{rgb} \tag{16}$$

and

$$\hat{\mathbf{F}}_{depth} = \mathcal{T}\left(\left[\hat{\mathbf{E}}_{depth}^1, \cdots, \hat{\mathbf{E}}_{depth}^s\right], \mathbf{W}_\psi\right) + \mathbf{F}_{depth} \tag{17}$$

Here, we concentrate the enhanced feature representation $\hat{\mathbf{E}}^s$ by a concentration operation $[\cdot]$, and $\mathcal{T}(\cdot,\cdot)$ denotes a 1×1 convolution layer with weight $\mathbf{W}_\psi \in \mathbb{R}^{sC1\times C}$. This is reasonable for restoring the features to their original dimensions. In our experiments, we set $S = 3$.

Compared with the standard NL block adopted in SOD [31], the proposed SCAM significantly reduces the computational complexity and further improves feature aggregation capability from multi-scale and cross-modality aspects. Furthermore, the SCAM captures the long-range dependencies from a cross-modality and multi-scale perceptive, where $\hat{\mathbf{E}}_d^s$ exploits the depth information to generate a spatial weight for the RGB feature, and $\hat{\mathbf{E}}_r^s$ refines the depth information by using the spatial weight generated from the RGB feature.

3.3. RGB-D Saliency Detection Network

As shown in Fig. (2), we propose a symmetrical two-stream encoder-decoder architecture for RGB-D SOD based on the proposed SCAM and deformable feature fusion strategy.

Figure 5: Qualitative comparison of the proposed approach with some state-of-the-art RGB-D SOD methods. (a) RGB images. (b Depth map. (c) GT. (d) Ours. (e) A2dele[38]. (f) S^2MA[31]. (g) D3Net[15]. (h) DMRA[37]

Here, we denote the output features of the RGB branch in the encoder blocks as $\mathbf{DRE_i}(i = 1, 2, 3)$ and $\mathbf{RE_i}(i = 4, 5)$, and the features of the depth branch in the decoder block as $\mathbf{RD_i}(i = 1, 2, \cdots, 5)$. The structure of the depth branch is analogous to the RGB branch.

We employ the CDCM at the beginning convolution blocks in both branches, (*i.e.*, $\mathbf{DRE_1}$-$\mathbf{DRE_3}$ and $\mathbf{DDE_1}$-$\mathbf{DDE_3}$), to handle the geometric variations and process the cross-modality cues, especially in the depth maps. Supervised by these cross-modality details, both encoder branches can extract more valuable low-level features. For the details, we replace the last Conv layer with a cross-modality deformable convolution module (CDCM) to enable these blocks to receive and losslessly process the geometric information. Taking the first image encoder block $\mathbf{DRE_1}$ as an instance, the last regular 3×3 Conv layer is then replaced by a 3×3 CDCM. (*i.e.,* Conv(3,3) → ReLU → Conv(3,3) → ReLU → CDCM(3,3), where (3,3) represents the kernel size).

We then obtain the features from the RGB and depth branches in the CNN and perform the proposed SCAM to obtain the cross-modality attention. The global contexts for both views are then propagated.

The decoder blocks of the two branches progressively integrate multi-scale features. We first apply 512 channels to the convolution layers at $\mathbf{RD_5}$ and $\mathbf{DD_5}$ to receive the enhanced features from the SCAM. Following the UNet[43] architecture, we then to progressively skip-connect the corresponding encoder features (*e.g.,* $\mathbf{RE_1}$-$\mathbf{RD_5}$ and $\mathbf{DE_1}$-$\mathbf{DD_5}$).

To further improve the performance of the final saliency map, we then apply the cross-stream fusion operation Ⓕ to fuse the image features and the corresponding depth features with a cascaded residual module as shown in Fig. (2)-(c).

We also employ the MFRM at the final decoder blocks $\mathbf{RD_1}$ and $\mathbf{DD_1}$ to refine the final saliency map. The RGB features $\left[F_r^1, F_r^2, F_r^3\right]$ and the depth feature vector $\left[F_d^1, F_d^2, F_d^3\right]$ are obtained from $\mathbf{RD_3}$-$\mathbf{RD_1}$ and $\mathbf{DD_3}$-$\mathbf{DE_1}$, respectively. The enhanced feature is propagated forward in both branches, and we employ the operation Ⓕ to concentrate the feature in the current module with the previous one. To ensure that the dimension of the final prediction is the same as the input, we adopt a 3×3 convolution layer with one channel on the last decoder feature map. We also use the sigmoid activation function to obtain the final saliency map for both streams. Each convolution layer in our decoder has a 3×3 kernel and is followed by a BN [22] layer and the ReLU activation.

3.4. Loss Function

As for the training loss of both streams, we consider a hybrid loss function between the predicted saliency maps and the ground truth mask. We also use in-depth supervision for each decoder module, where we first apply a 3×3 Conv layer with the sigmoid activation function on each decoder feature map to generate a saliency map compute their loss. We then set up a scale aggregation architecture for each side-output branch that densely accumulates the features from the largest scale 256×256 in $\mathbf{RD_1}$ and $\mathbf{DD_1}$ to the smallest scale 32×32 in $\mathbf{RD_5}$ and $\mathbf{DD_5}$. The aggregation of the features from each scale is then used to estimate the saliency maps and supervised by the ground-truth saliency maps.

Our hybrid loss is defined as the summation of the intermediate and final saliency result losses as:

$$\mathcal{L} = \sum_{k=1}^{K} (\alpha_k \ell_r^{(k)} + \beta_k \ell_d^{(k)}), k \in \{1, 2, \cdots, 5\}, \tag{18}$$

where $\ell_r^{(k)}$ denotes the loss of the k-th side output in the RGB branch, $\ell_d^{(k)}$ is the loss of the k-th side output in the $depth$ stream, and K denotes the total number of the outputs. Moreover, α_k and β_k are the weight of each loss in both branches.

To obtain high-quality region segmentation and clear boundaries, the hybrid loss $\ell^{(k)}$ for each scaled prediction is defined as:

$$\ell^{(k)} = \ell_{bce}^{(k)} + \ell_{ssim}^{(k)} + \ell_{edge}^{(k)}, \tag{19}$$

where $\ell_{bce}^{(k)}$, $\ell_{ssim}^{(k)}$ and $\ell_{edge}^{(k)}$ denote the BCE loss [1], SSIM loss [60] and Edge loss, respectively. Hence, we supervise these multi-scale predicated saliency maps in both streams using a hybrid loss. Here, we consider BCE loss in $\ell_{bce}^{(k)}$ as follows:

$$\begin{aligned} \ell_{bce}^{k} = -\sum_{i,j} & [G_k[i,j]) \log(S_k[i,j]) \\ & + (1 - G_k[i,j]) \log(1 - S_k[i,j])], \end{aligned} \tag{20}$$

where $G_k[i,j]$ and $S_k[i,j]$ denote the values at the location (i,j) of the ground truth map G_k and the corresponding estimated saliency map S_k, respectively.

For the edge-preserving loss $\ell_{edge}^{(k)}$, we compute the difference between the extracted edge information S_k^e of the side-output saliency map S_k and the corresponding boundary G_k^e of the ground-truth saliency map G_k as:

$$\begin{aligned} \ell_{edge}^{k} = -\sum_{i,j} & (G_k^e[i,j]) \log(S_k^e[i,j]) \\ & + (1 - G_k^e[i,j]) \log(1 - S_k^e[i,j])], \end{aligned} \tag{21}$$

where $G_k^e[i,j]$ and $S_k^e[i,j]$ denote the values at the location (i,j) of the obtained edge details from the ground truth map G_k and the corresponding estimated saliency map S_k, respectively. Both edge map prediction G_k^e and S_k^e are obtained using the Canny edge detector.

Besides, the SSIM strengthens the saliency boundary's supervision, as illustrated in [40]. Therefore, we employ the SSIM loss as a key component in the joint loss function, which is defined as:

$$\ell_{\text{ssim}}^{k} = 1 - \frac{1}{M} \sum_{j=1}^{M} \frac{\left(2\mu_{x_j}\mu_{y_j} + C_1\right)\left(2\sigma_{x_j y_j} + C_2\right)}{\left(\mu_{x_j}^2 + \mu_{y_j}^2 + C_1\right)\left(\sigma_{x_j}^2 + \sigma_{y_j}^2 + C_2\right)} \tag{22}$$

Here, the estimated map S^k and the ground truth map G^k are divided into M patches using a sliding window of 11×11 with a stride of 1. We then obtain the patches for both maps $\{x_1, \cdots, x_M\}$ and $\{y_1, \cdots, y_M\}$, respectively. In the above, μ_{x_j}, μ_{y_j}, σ_{x_j} and σ_{y_j} are the mean and standard deviation of patches x_j and y_j, where $j \in \{1, \cdots, M\}$. Furthermore, σ_{x_j} and σ_{y_j} are their covariance, while C_1 and C_2 are constant used to avoid division by zero.

Table 1
Quantitative performance comparison of our proposed model with several other state-of-the-art RGB-D saliency models on eight benchmark datasets in terms of four evaluation metrics. (Figures highlighted in red indicate the best performance).

Dataset	Metrics	ACSD [24]	LBE [16]	DCMC [11]	SE [42]	DF [45]	CTMF [19]	MMCI [4]	PCFN [2]	TAN [3]	CPFP [65]	DMRA [37]	D3Net [15]	A2dele [39]	S2MA [31]	Ours
NJU2K [24]	S_m ↑	0.699	0.695	0.685	0.644	0.763	0.849	0.858	0.877	0.878	0.879	0.886	0.895	0.892	0.894	0.902
	max-F ↑	0.711	0.748	0.715	0.748	0.804	0.845	0.852	0.872	0.874	0.877	0.886	0.889	0.888	0.889	0.902
	E_ξ ↑	0.803	0.803	0.799	0.813	0.864	0.913	0.915	0.924	0.925	0.926	0.927	0.932	0.930	0.929	0.940
	MAE ↓	0.202	0.153	0.172	0.169	0.141	0.085	0.085	0.059	0.060	0.053	0.051	0.051	0.053	0.054	0.044
NLPR [36]	S_m ↑	0.673	0.762	0.724	0.756	0.802	0.860	0.856	0.874	0.886	0.888	0.894	0.911	0.890	0.915	0.923
	max-F ↑	0.607	0.745	0.648	0.713	0.778	0.825	0.815	0.841	0.863	0.867	0.888	0.896	0.875	0.902	0.907
	E_ξ ↑	0.780	0.855	0.793	0.847	0.880	0.929	0.913	0.925	0.941	0.932	0.944	0.953	0.937	0.953	0.956
	MAE ↓	0.179	0.081	0.117	0.091	0.085	0.056	0.059	0.044	0.041	0.036	0.036	0.030	0.030	0.030	0.026
STERE [35]	S_m ↑	0.692	0.660	0.731	0.708	0.757	0.848	0.873	0.875	0.871	0.879	0.886	0.886	0.879	0.890	0.896
	max-F ↑	0.669	0.633	0.740	0.755	0.757	0.831	0.863	0.860	0.861	0.874	0.886	0.886	0.879	0.882	0.888
	E_ξ ↑	0.806	0.787	0.819	0.846	0.847	0.912	0.927	0.925	0.923	0.925	0.938	0.938	0.928	0.932	0.933
	MAE ↓	0.200	0.250	0.176	0.148	0.141	0.086	0.068	0.064	0.060	0.051	0.047	0.047	0.044	0.051	0.047
RGBD135 [8]	S_m ↑	0.728	0.703	0.707	0.741	0.752	0.863	0.848	0.842	0.858	0.872	0.900	0.897	0.883	0.941	0.939
	max-F ↑	0.756	0.788	0.666	0.726	0.766	0.844	0.822	0.804	0.827	0.846	0.888	0.884	0.873	0.935	0.937
	E_ξ ↑	0.850	0.890	0.773	0.856	0.870	0.932	0.928	0.893	0.910	0.923	0.943	0.945	0.920	0.973	0.978
	MAE ↓	0.169	0.208	0.111	0.090	0.093	0.055	0.065	0.049	0.046	0.038	0.030	0.031	0.030	0.021	0.019
SSD100 [26]	S_m ↑	0.675	0.621	0.704	0.675	0.747	0.776	0.813	0.841	0.839	0.807	0.857	0.857	0.803	0.868	0.877
	max-F ↑	0.682	0.619	0.711	0.710	0.735	0.729	0.781	0.807	0.810	0.766	0.844	0.834	0.776	0.848	0.859
	E_ξ ↑	0.785	0.736	0.786	0.800	0.828	0.865	0.882	0.894	0.897	0.852	0.906	0.911	0.861	0.906	0.922
	MAE ↓	0.203	0.278	0.169	0.165	0.142	0.099	0.082	0.062	0.063	0.082	0.058	0.059	0.070	0.052	0.047
LFSD [28]	S_m ↑	0.727	0.729	0.746	0.692	0.783	0.788	0.779	0.786	0.794	0.820	0.839	0.824	0.826	0.829	0.843
	max-F ↑	0.763	0.722	0.813	0.786	0.813	0.787	0.767	0.775	0.792	0.821	0.797	0.815	0.828	0.831	0.842
	E_ξ ↑	0.829	0.797	0.856	0.832	0.857	0.857	0.831	0.827	0.840	0.864	0.846	0.856	0.867	0.865	0.878
	MAE ↓	0.195	0.214	0.155	0.174	0.146	0.127	0.139	0.119	0.118	0.095	0.083	0.106	0.084	0.102	0.090
DUT-RGBD [62]	S_m ↑	0.361	0.568	0.659	0.499	0.736	0.831	0.791	0.801	0.808	0.818	0.889	0.824	0.885	0.903	0.907
	max-F ↑	0.247	0.625	0.723	0.411	0.740	0.823	0.767	0.771	0.790	0.795	0.898	0.815	0.891	0.900	0.904
	E_ξ ↑	0.590	0.734	0.800	0.654	0.823	0.899	0.859	0.856	0.861	0.859	0.933	0.856	0.930	0.937	0.941
	MAE ↓	0.332	0.174	0.280	0.243	0.144	0.097	0.113	0.100	0.093	0.076	0.048	0.073	0.043	0.043	0.043
SIP [15]	S_m ↑	0.732	0.727	0.683	0.628	0.653	0.720	0.716	0.833	0.835	0.850	0.806	0.860	0.870	0.872	0.877
	max-F ↑	0.763	0.751	0.618	0.661	0.465	0.702	0.608	0.771	0.803	0.821	0.811	0.861	0.865	0.877	0.880
	E_ξ ↑	0.614	0.651	0.598	0.592	0.565	0.793	0.704	0.845	0.870	0.870	0.875	0.909	0.910	0.918	0.917
	MAE ↓	0.172	0.200	0.186	0.164	0.165	0.118	0.139	0.086	0.075	0.064	0.085	0.063	0.063	0.058	0.053
ReDWeb-S [32]	S_m ↑	-	0.637	0.427	0.435	0.595	0.641	0.660	0.655	0.656	0.685	0.592	0.688	0.705	0.710	0.719
	max-F ↑	-	0.629	0.348	0.393	0.579	0.607	0.641	0.627	0.623	0.645	0.579	0.669	0.685	0.694	0.706
	E_ξ ↑	-	0.730	0.549	0.587	0.683	0.739	0.754	0.743	0.741	0.744	0.712	0.765	0.772	0.779	0.783
	MAE ↓	-	0.253	0.313	0.283	0.233	0.204	0.176	0.166	0.165	0.142	0.188	0.149	0.145	0.140	0.141

4. Experiments

4.1. Benchmark Datasets and Evaluation Metrics

In this work, we conduct experiments on nine widely used RGB-D SOD datasets, including NJU2K [24] (1985 RGB-D images), NLPR [36] (1000 RGB-D images), RGBD135 [8] (135 RGB-D images), STERE [35] (1000 RGB-D images), LFSD [28] (100 RGB-D images), SSD [26] (80 RGB-D images), DUT-RGBD[37] (1200 RGB-D images), SIP [15] (929 RGB-D images) and ReDWeb-S [32] (3600 RGB-D images). For fair comparisons, we perform the same training as described in [37, 39], which contains 800 samples from the DUT-RGBD dataset, 1485 samples from NJU2K and 700 samples from NLPR for training. The remaining images and the other five datasets are used for testing to evaluate the performance.

To avoid over-fitting, we adopt the following data augmentation. First, we resize the training images, and the corresponding depth maps to 288×288 pixels and then randomly crop 256×256 regions to train the network. We also use random horizontal flipping. To match the channel dimension between depth and RGB images to fit the network input layer, we further replicate each depth map to three channels. Besides, each image and the three-channel depth map are subtracted by their mean pixel values before being considered as the inputs to the whole network.

Following the recent work [15, 31], we adopt the maximum F-measure (max-F), Structure-measure (S_m), Enhanced-alignment measure (E_ξ) and Mean Absolute Error (MAE) for quantitative evaluations. Specifically, max-F is the weighted harmonic mean of precision and recall, and it is a comprehensive measure indicating the performance. Further, S_m [13] score measures the difference between the saliency map and ground truth, and the larger of the score, the higher the performance. Also, E_ξ [14] is a reasonable measure to capture both global statistics and local pixel matching information of the saliency maps. The MAE score further measures the difference between the continuous saliency map and the ground truth. The smaller the value of the MAE, the smaller the gap, indicating a higher performance.

4.2. Implementation Details

We implement the proposed network by using the PyTorch package and two NVIDIA 1080 Ti GPUs for computing acceleration. The stochastic gradient descent (SGD) with the momentum algorithm is adopted to optimize our network with a total of 40,000 iterations. The weight decay, momentum and batch size are set to 1e-4, 0.9 and 8, respectively. The initial learning rate is set to 0.01 and divided by 10 at the $15,000^{th}$ and the $30,000^{th}$ iterations.

4.3. Comparisons with State-of-the-art Methods

We compare our method with 14 state-of-the-art RGB-D SOD methods (including four classical traditional non-deep models, *i.e.* ACSD [24], LBE [16], DCMC [11], and SE [42], and ten learning-based models, *i.e.* DF [45], CTMF [19], MMCI [4], PCFN [2], TAN [3], CPFP [65], DMRA [37], D3Net [15], A2dele [39] and S^2MA [31]. We use the released codes and default hyper-parameters as provided by the corresponding authors to reproduce the final saliency maps.

1) **Qualitative Evaluation:** To illustrate the advantages of the proposed method, we provide several visual examples of different methods. As shown in Fig. (5), the proposed method can obtain better experimental results with precise saliency location, clean background, complete structure, and sharp boundaries. Moreover, it is efficient in various challenging scenarios, such as low contrast, complicated scene, background disturbance, and unreliable depth maps. To be specific:

(a) Our model handles the disturbance of a similar appearance between the salient object and the background. For example, in the eighth image, the robot's arms and legs are similar to the background, and the whole scene has low contrast. The existing methods are unable to address this challenging case very well as their results ignore these almost identical regions with the background. By contrast, our method shows a competitive advantage in terms of completeness, sharpness, and accuracy. Specifically, AMDFNet highlights the robot and its entire limbs using the depth maps.

(b) Our model can produce robust results even in the cases where the available depth information is inaccurate or blurred (*e.g.,* the second and fifth images). This indicates the robustness of the SCAM. In these challenging scenarios, because of the negative effect caused by unreliable depth maps, the existing methods are unable to locate the accurate boundaries of the salient objects. The proposed method, however, utilizes the cross-modal complementary information and suppresses the impact of unreliable depth maps.

(c) Our model produces a complete structure and sharp boundaries in the results. For example, in the third and fourth images, the irregular shape of the purple flower is accurately and entirely detected by the existing methods, such as A2dele [38], and S^2MA [31] and the unnecessary background (*e.g.*, the red flower at the right of the third image and purple petals at the right of the fourth image) are wrongly retained. By contrast, our method obtains complete and accurate boundaries and has an improved ability to process complex scenarios.

In summary, the experimental results indicate that our model accurately localizes the salient objects and segments them precisely, whereas the existing models are disturbed in the complex scenes.

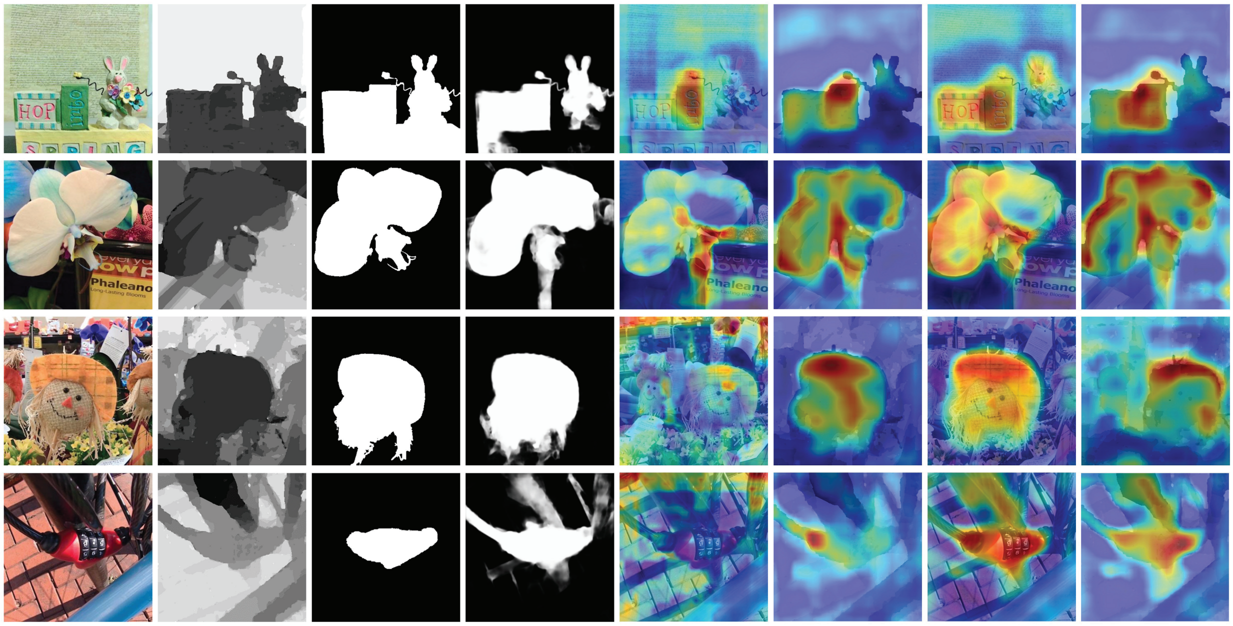
2) **Quantitative Evaluation:** For a more intuitive comparison of performance, here we obtain the quantitative metrics including max-F, S_m, E_ξ, and MAE score in Tab. (1). It can be seen that our proposed method outperforms almost all of the existing methods on all datasets, except for the LFSD and RGND135. On these two datasets, our model also achieves a performance comparable to the best existing methods.

Furthermore, AMDFNet outperforms all other methods by a notable margin on the DUT-RGB, SIP and ReDWeb datasets, containing more challenging scenarios. The experimental results further indicate that our modifications integrate informational cues in both modalities and transfer the qualified depth knowledge to facilitate a more accurate final saliency prediction.

4.4. Ablation Study

To verify the effectiveness of each key component in our proposed network, including CDCM, SCAM and MFRM, we conduct ablation studies on NJU2K, NLPR, RGBD135 and LFSD datasets. The basic model with the standard fusion decoder modules is regarded as the baseline model to guarantee the fairness of the ablation experiments. Tab. (2) validates all components in our proposed system based on four widely used benchmarks and the above four metrics.

First, we choose the basic network that removes the multi-level feature refinement module (MFRM), removes the cross-modality deformable convolution module (CDCM), and replaces the selective cross-modality attention module (SCAM) with the standard channel and spatial attention operation [61] as the baseline (denoted as "B"). From the Tab. (2), compar-



(a) (b) (c) (d) (e) (f) (g) (h)

Figure 6: Visualization of the output from SCAM. (a) RGB image. (b) Depth maps. (c) GT. (d) Predicted saliency maps. (e) and (f) Heat-maps of RGB and depth channel (without SCAM). (g) and (h) Heat-maps of RGB and depth channel (with SCAM).

Table 2
Ablation study of module verification on NJU2K, NLPR, RGBD135 and LFSD dataset. The best results on each dataset are highlighted in **boldface**.

Settings				NJUD-test [24]				NLPR-test [36]				RGBD135 [8]				LFSD [28]			
B	B^S	MF	C	S_m	max-F	E_ξ	MAE	S_m	max-F	E_ξ	MAE	S_m	max-F	E_ξ	MAE	S_m	max-F	E_ξ	MAE
✓				0.865	0.852	0.902	0.072	0.897	0.873	0.941	0.039	0.875	0.834	0.927	0.046	0.786	0.775	0.836	0.131
	✓			0.893	0.887	0.928	0.056	0.915	0.896	0.952	0.032	0.933	0.924	0.970	0.024	0.821	0.824	0.854	0.105
	✓	✓		0.897	0.892	0.933	0.052	0.923	0.909	0.957	0.028	0.939	0.932	0.972	0.023	0.838	0.846	0.873	0.097
	✓	✓	✓	**0.902**	**0.902**	**0.940**	**0.044**	**0.923**	**0.907**	**0.956**	**0.026**	**0.939**	**0.937**	**0.978**	**0.019**	**0.843**	**0.842**	**0.878**	**0.090**

ing the "B" with the "B^S", we replace the standard attention operation by the selective cross-modality attention module (denoted as 'B^S') which improves the baseline by about $3 \sim 4$ points in terms of the maximum F-measure in the NJU2K dataset. Our proposed SCAM aims to adaptively select the informative and vital details in depth to solve two issues: (1) how to effectively remove the adverse effects from the low-quality depth input. (2) how to provide complementary information to support cross-modality fusion. The experimental results prove that adding the cross-modality attention module can significantly improve the SOD performance.

By adding the multi-level feature refinement module in the last feature decoding block (denoted as '$B^S + MF$'), the F-measure increases to 0.902 on the NJU2K dataset which is comparable with the state-of-the-art methods. Furthermore, the performance is significantly enhanced after adding the CDCM at the first three encoder blocks (denoted as '$B^S + MF + C$'), which yields the best performance with F-measure and MAE percentage gains of 5.0% and 2.8%, respectively compared with the original baseline on the NJU2K dataset. The MFRM applies the advantages of multi-scale feature and cross-modality deformable operation. This effectively captures the global context in multi-scale features and determine the salient object fully and resolve the challenging ambiguity in the SOD with a similar appearance and a cluttered background. The experiments on the other three datasets, *i.e.,* NLPR, RGBD135 and LFSD, also show the effectiveness of the proposed components significantly.

Selective Cross-modality Attention Mechanism (SCAM) To thoroughly understand the selective cross-modality attention mechanism, we visualize several feature maps and their corresponding heat-maps in Fig. (6). Taking the RGB output produced by SCAM as an example, the module learns the cross-modality complementarity from a cross-modality perspective and prevent unreliable depth maps. As shown in Fig. (6), the model with SCAM accurately locates the salient object positions, and the focus covers the whole object (*e.g.,* the first and second images). In case of a cluttered background or where the depth input is not ideal, the third image contains several cans, and the foreground has a similar

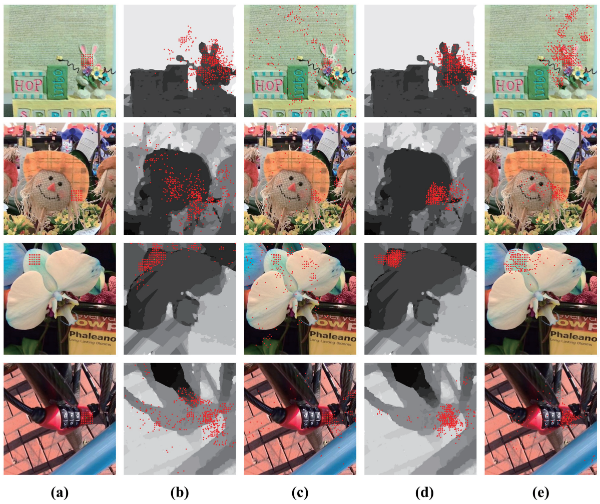


Figure 7: Visualization of the sampling locations in RGB and depth stream employed in the original convolution, modulated convolution network (DCNv2) and cross-modality deformable convolution module (CDCM). The green dots in each image represent the activation units and the red dots are sampling locations. (a) Standard convolution. (b-c) DCN in depth and RGB stream. (d-e) CDCM in depth and RGB stream.

appearance to the background. This results in an unclear attention map in the heat-map produced by the baseline ('B'). By adding the SCAM, our model maintains more structural information of the desired mode and successfully suppresses most background noise.

To verify the effectiveness of SCAM in memory reduction, we design an ablation study to analyze the required computational resources in terms of floating-point operations (FLOPs), memory consumption and parameters. The results are shown in Tab. (3). Specifically, all experimental results are obtained by testing methods on a 256×256 input sample. We compare our method with SCAM against the original non-local block. The original non-local operation dramatically increases memory consumption since it requires computing a large correlation matrix. In contrast, the additional memory requirement of the proposed SCAM (1.251Gb) is 22.5% less than (1.621Gb) the standard non-local operation. This means that our method can reduce the required memory in the training process, and our method allows larger training batch size or bigger image size under the same GPU memory.

In summary, the designed SCAM strengthens the feature from a cross-modality perspective and prevents contamination caused by unreliable depth maps. Furthermore, the computing and memory consumption significantly decreased compared with the relevant structure.

Cross-modality Deformable Convolution Module (CDCM) To better understand the behavior of CDCM, we visualize the sampling location [69], which contributes significantly to the final network prediction. Specifically, we analyze the visual support regions in both feature encoder modules (*i.e.,* RGB and depth streams). First, we employ standard convolution layer in $\mathbf{DRE_i}$ and $\mathbf{DDE_i}$ ($i = 1, 2, 3$) as baseline. Besides, the three 3×3 standard convolutions layers inserted in the above blocks are replaced by DCNv2 [69] and the sampling locations of this operation are shown in Fig. (7)-(e) and (f). In comparison, we employ CDCM in corresponding convolution blocks, and the sampling results are illustrated in Fig. (7)-(e) and (f).

Table 3
Ablation study of efficiency in terms of floating point operations (FLOPs) and memory consumption.

Non-Local Module Type	FLOPs	Memory	#Params
NLB [59]	142.27G	1.614Gb	1.949M
Ours	140.83G	1.251Gb	1.311M

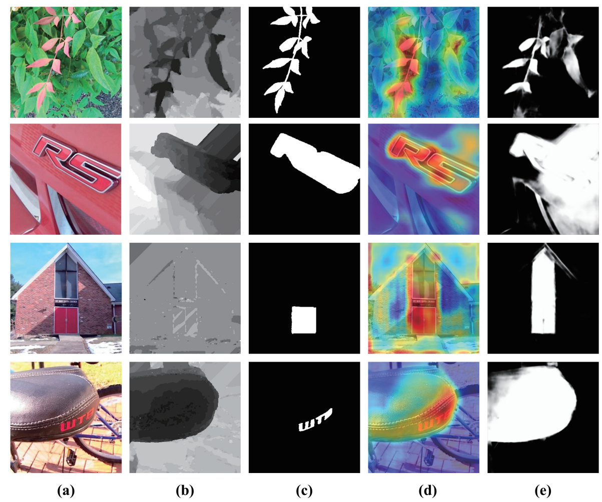


Figure 8: Failure examples. (a) RGB images. (b) Depth maps. (c) GT. (d) Heat maps. (e) Our results.

As shown in Fig. (7), the spatial support of the DCNv2 expands the sampling distribution and enlarges the receptive field by deformable filters significantly. The network's ability to model geometric transformation is considerably enhanced, and the spatial support adapts much more to image content, with nodes on the foreground having support covering the whole salient object. In contrast, nodes on the background have expanded support that encompasses greater context. However, the range of spatial support may be inexact, *i.e.,* the boundary splitting salient regions and background could not be detected, and salient regions contain irrelevant areas.

To regulate the sampling distribution and make full use of cross-modal cues, the CDCM receives extra information from another modal to guide the filter training and enhance the network's feature extraction ability. Based on these visible results, we observed that these adaptive sampling location produced by the CDCM highly emphasises the salient object boundaries and dramatically suppresses the interference of background information.

4.5. Failure Cases

To further promote the SOD, Fig. (8) shows several failure cases produced by our AMDFNet. As it shows in this figure, our approach had troubles to recognize the accurate boundaries of the salient objects in these examples. Further investigating the typical characteristics of the failure cases, we can identify two factors that contribute to the low quality of the predicted maps. First, the conflict of a salient object between the depth maps and the RGB images leads to false alarms. Although our SCAM reduces the adverse effects resulted from the depth maps and the heat-maps, it is challenging to suppress the contamination for these cases. Secondly, the combination of the salient object and background region significantly interferes with the results. For the cases where the spatial distance between the objects is small, especially when the salient object is embedded in other non-salient objects in the background (*e.g.*, the red door is located in a house and the letters are printed on the seats), the depth maps cannot provide the exact location details. This results in incorrect SOD by the algorithm.

5. Conclusion

In this paper, we have proposed a selective cross-modality attention module to capture the dense attention among various features maps in both modalities. The proposed module enables selecting informative regions and suppressing the impact of unreliable depth maps. We have also developed a multi-level feature refinement mechanism to adaptively strengthen those maps of different scales and refine the features from the multi-scale and cross-modality perspectives. Both the embedded selective attention module and densely cooperative refinement strategy have been empirically proved to be effective for exploiting the cross-modality complementarity. Our next challenge is to improve the quality of the depth maps. The work presented in this paper lays the groundwork for future therapeutic research. The multi-modal feature fusion method further provides new insights into other challenging visual tasks, *e.g.*, RGB-D image enhancement and multi-source image fusion.

Acknowledgment

This work was sponsored by Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (CX201959) and Synergy Innovation Foundation of the University and Enterprise for Graduate Students in Northwestern Polytechnical University (XQ201910). This work was also supported in part by the National Natural Science Foundation of China under Grant 61972321.

CRediT authorship contribution statement

Fei Li: Conceptualization of this study, Methodology, Writing - Original draft preparation. **Jiangbin Zheng:** Methodology, Writing - Original draft preparation. **Yuan-fang Zhang:** Data curation, Writing - Original draft preparation. **Nian Liu:** Methodology. **Wenjing Jia:** Writing - Original draft preparation.

References

[1] Boer, P.D., Kroese, D.P., Mannor, S., Rubinstein, R., 2005. A tutorial on the cross-entropy method. Annals of Operations Research 134, 19–67.

[2] Chen, H., Li, Y., 2018. Progressively complementarity-aware fusion network for rgb-d salient object detection. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 3051–3060.

[3] Chen, H., Li., Y., 2019. Three-stream attention-aware network for rgb-d salient object detection. IEEE Transactions on Image Processing 28, 2825–2835.

[4] Chen, H., Li, Y., Su, D., 2019a. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recognit. 86, 376–385.

[5] Chen, J.R., Song, H., Zhang, K., Liu, B., Liu, Q., 2021a. Video saliency prediction using enhanced spatiotemporal alignment network. Pattern Recognit. 109, 107615.

[6] Chen, Q., Fu, K., Liu, Z., Chen, G., Du, H., Qiu, B., Shao, L., 2021b. Ef-net: A novel enhancement and fusion network for rgb-d saliency detection. Pattern Recognit. 112, 107740.

[7] Chen, Y., Yang, T., Zhang, X., Meng, G., Xiao, X., Sun, J., 2019b. Detnas: Backbone search for object detection, in: NeurIPS.

[8] Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X., 2014. Depth enhanced saliency detection method, in: ICIMCS '14.

[9] Cong, R., Lei, J., Fu, H., Hou, J., Huang, Q., Kwong, S., 2020. Going from rgb to rgbd saliency: A depth-guided transformation model. IEEE Transactions on Cybernetics 50, 3627–3639.

[10] Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C., 2016a. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Processing Letters 23, 819–823.

[11] Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C., 2016b. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Processing Letters 23, 819–823.

[12] Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y., 2017. Deformable convolutional networks. 2017 IEEE International Conference on Computer Vision (ICCV) , 764–773.

[13] Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A., 2017. Structure-measure: A new way to evaluate foreground maps. 2017 IEEE International Conference on Computer Vision (ICCV) , 4558–4567.

[14] Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A., 2018. Enhanced-alignment measure for binary foreground map evaluation. ArXiv abs/1805.10421.

[15] Fan, D.P., Lin, Z., Zhao, J., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.M., 2020. Rethinking rgb-d salient object detection: Models, datasets, and large-scale benchmarks. IEEE transactions on neural networks and learning systems PP.

[16] Feng, D., Barnes, N., You, S., McCarthy, C., 2016. Local background enclosure for rgb-d salient object detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2343–2350.

[17] Feng, G., Bo, H., Sun, J., Zhang, L., Lu, H., 2020. Cacnet: Salient object detection via context aggregation and contrast embedding. Neurocomputing 403, 33–44.

[18] Guo, C., Zhang, L., 2010. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Transactions on Image Processing 19, 185–198.

[19] Han, J., Chen, H., Liu, N., Yan, C., Li, X., 2018. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics 48 11, 3171–3183.

[20] Han, J., Shao, L., Xu, D., Shotton, J., 2013. Enhanced computer vision with microsoft kinect sensor: A review. IEEE Transactions on Cybernetics 43, 1318–1334.

[21] Huang, P.S., Shen, C.H., Hsiao, H.F., 2018. Rgbd salient object detection using spatially coherent deep learning framework. 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP) , 1–5.

[22] Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: ICML.

[23] Jerripothula, K.R., Cai, J., Yuan, J., 2016. Image co-segmentation via saliency co-fusion. IEEE Transactions on Multimedia 18, 1896–1909.
[24] Ju, R., Ge, L., Geng, W., Ren, T., Wu, G., 2014. Depth saliency based on anisotropic center-surround difference. 2014 IEEE International Conference on Image Processing (ICIP) , 1115–1119.
[25] Lai, Q., Wang, W., Sun, H., Shen, J., 2020. Video saliency prediction using spatiotemporal residual attentive networks. IEEE Transactions on Image Processing 29, 1113–1126.
[26] Li, G., Zhu, C., 2017. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. 2017 IEEE International Conference on Computer Vision Workshops (ICCVW) , 3008–3014.
[27] Li, J., Fu, B., Liu, Z., 2019. Panchromatic image compression based on improved post-transform for space optical remote sensors. Signal Process. 159, 72–88.
[28] Li, N., Ye, J., Ji, Y., Ling, H., Yu, J., 2014. Saliency detection on light field. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1605–1616.
[29] Liang, F., Duan, L., Ma, W., Qiao, Y., Cai, Z., Qing, L., 2018. Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing 275, 2227–2238.
[30] Liu, J., Wang, H., Yan, C., Yuan, M., Su, Y., 2020a. Soda^2:salient object detection with structure-adaptive scale-adaptive receptive field. IEEE Access 8, 204160–204172.
[31] Liu, N., Zhang, N., Han, J., 2020b. Learning selective self-mutual attention for rgb-d saliency detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 13753–13762.
[32] Liu, N., Zhang, N., Shao, L., Han, J., 2020c. Learning selective mutual attention and contrast for rgb-d saliency detection. ArXiv abs/2010.05537.
[33] Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P., 2019. Salient object detection for rgb-d image by single stream recurrent convolution neural network. Neurocomputing 363, 46–57.
[34] Nam, H., Ha, J.W., Kim, J., 2017. Dual attention networks for multimodal reasoning and matching. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) , 2156–2164.
[35] Niu, Y., Geng, Y., Li, X., Liu, F., 2012. Leveraging stereopsis for saliency analysis. 2012 IEEE Conference on Computer Vision and Pattern Recognition , 454–461.
[36] Peng, H., Li, B., Xiong, W., Hu, W., Ji, R., 2014. Rgbd salient object detection: A benchmark and algorithms, in: ECCV.
[37] Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H., 2019. Depth-induced multi-scale recurrent attention network for saliency detection. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) , 7253–7262.
[38] Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H., 2020a. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 9057–9066.
[39] Piao, Y., Rong, Z., Zhang, M., Ren, W., Lu, H., 2020b. A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 9057–9066.
[40] Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jägersand, M., 2019. Basnet: Boundary-aware salient object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 7471–7481.
[41] Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q., 2017. Rgbd salient object detection via deep fusion. IEEE Transactions on Image Processing 26, 2274–2285.
[42] Quo, J., Ren, T., Bei, J., 2016. Salient object detection for rgb-d image via saliency evolution. 2016 IEEE International Conference on Multimedia and Expo (ICME) , 1–6.
[43] Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation, in: MICCAI.
[44] Rutishauser, U., Walther, D.B., Koch, C., Perona, P., 2004. Is bottom-up attention useful for object recognition?, in: CVPR 2004.
[45] Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. CoRR abs/1409.1556.
[46] Song, H., Liu, Z., Du, H., Sun, G., Meur, O.L., Ren, T., 2017. Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Transactions on Image Processing 26, 4204–4216.
[47] Thomas, S.S., Gupta, S., Subramanian, V., 2019. Context driven optimized perceptual video summarization and retrieval. IEEE Transactions on Circuits and Systems for Video Technology 29, 3132–3145.
[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need, in: NIPS.
[49] Wan, Y., Shu, J., Sui, Y., Xu, G., Zhao, Z., Wu, J., Yu, P.S., 2019. Multi-modal attention network learning for semantic source code retrieval. 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE) , 13–25.
[50] Wang, A., Wang, M., 2017. Rgb-d salient object detection via minimum barrier distance transform and saliency fusion. IEEE Signal Processing Letters 24, 663–667.
[51] Wang, N., Gong, X., 2019. Adaptive fusion for rgb-d salient object detection. IEEE Access 7, 55277–55284.
[52] Wang, W., Shen, J., 2017. Deep cropping via attention box prediction and aesthetics assessment. 2017 IEEE International Conference on Computer Vision (ICCV) , 2205–2213.
[53] Wang, W., Shen, J., Cheng, M.M., Shao, L., 2019a. An iterative and cooperative top-down and bottom-up inference network for salient object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 5961–5970.
[54] Wang, W., Shen, J., Dong, X., Borji, A., Yang, R., 2020. Inferring salient objects from human fixations. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 1913–1927.
[55] Wang, W., Shen, J., Ling, H., 2019b. A deep network solution for attention and aesthetics aware photo cropping. IEEE Transactions on Pattern Analysis and Machine Intelligence 41, 1531–1544.
[56] Wang, W., Shen, J., Xie, J., Cheng, M.M., Ling, H., Borji, A., 2021. Revisiting video saliency prediction in the deep learning era. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 220–237.
[57] Wang, W., Zhao, S., Shen, J., Hoi, S., Borji, A., 2019c. Salient object detection with pyramid attention and salient edges. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 1448–1457.
[58] Wang, X., Chan, K.C.K., Yu, K., Dong, C., Loy, C.C., 2019d. Edvr: Video restoration with enhanced deformable convolutional networks. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) , 1954–1963.
[59] Wang, X., Girshick, R.B., Gupta, A., He, K., 2018. Non-local neural networks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition , 7794–7803.
[60] Wang, Z., Simoncelli, E.P., Bovik, A., 2003. Multiscale structural similarity for image quality assessment. The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003 2, 1398–1402 Vol.2.
[61] Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: ECCV.
[62] Zagoruyko, S., Lerer, A., Lin, T.Y., Pinheiro, P.O., Gross, S., Chintala, S., Dollár, P., 2016. A multipath network for object detection, in: BMVC.
[63] Zhang, D., Han, J., Li, C., Wang, J., Li, X., 2016. Detection of co-salient objects by looking deep and wide. International Journal of Computer Vision 120, 215–232.
[64] Zhang, D., Meng, D., Han, J., 2017. Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 865–878.
[65] Zhao, J., Cao, Y., Fan, D.P., Cheng, M.M., yi Li, X., Zhang, L., 2019. Contrast prior and fluid pyramid integration for rgbd salient object detection. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 3922–3931.
[66] Zhao, T., Wu, X., 2019. Pyramid feature attention network for saliency detection. 2019 IEEE/CVF Conference on Computer Vision and

Pattern Recognition (CVPR) , 3080–3089.

[67] Zhou, T., Fu, H., Gong, C., Shen, J., Shao, L., Porikli, F., 2020a. Multi-mutual consistency induced transfer subspace learning for human motion segmentation. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 10274–10283.

[68] Zhou, W., Lv, Y., Lei, J., Yu, L., 2020b. Global and local-contrast guides content-aware fusion for rgb-d saliency prediction. IEEE Transactions on Systems, Man, and Cybernetics , 1–9.

[69] Zhu, X., Hu, H., Lin, S., Dai, J., 2019. Deformable convnets v2: More deformable, better results. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) , 9300–9308.