Multi-scale Residual Interaction for RGB-D Salient Object Detection

Mingjun Hu[1], Xiaoqin Zhang[1], and Li Zhao[1]

[1]College of Computer Science and Artificial Intelligence, Wenzhou University, China
humingjun98@gmail.com, zhangxiaoqinnan@gmail.com, lizhao@wzu.edu.cn

Abstract. RGB-D salient object detection (SOD) is used to detect the most attractive object in the scene. There is a problem in front of the existing RGB-D SOD task: how to integrate the different context information between the RGB and depth map effectively. In this work, we propose the Siamese Residual Interactive Refinement Network (SiamRIR) equipped with the encoder and decoder to handle the above problem. Concretely, we adopt the Siamese Network shared parameters to encode two modalities and fuse them during decoding phase. Then, we design the Multi-scale Residual Interavtive Refinement Block (RIRB) which contains Residual Interactive Module (RIM) and Residual Refinement Module (RRM). This block utilizes the multi-type cues to fuse and refine features, where RIM takes interaction between modalities to integrate the complementary regions with residual manner, and RRM refines features during fusion phase by incorporating spatial detail context with multi-scale manner. Extensive experiments on five benchmarks demonstrate that our method outperforms the state-of-the-art RGB-D SOD methods both quantitatively and qualitatively.

Keywords: RGB-D salient object detect · Multi-scale interactive · Siamese Network.

1 Introduction

Salient object detection (SOD) aims to segment the most attractive object from the scene [1–3]. As a pre-processing task, it plays an important role in computer vision tasks, such as semantic segmentation [4–6], object detection [7, 8], person re-identification [9, 10], and object tracking [11, 12]. In the past years, various SOD methods have been proposed and achieved promising performances with only take RGB as the input [13–18], but may suffer from challenges when on the indistinguishable and complex scenarios. Alternatively, we can obtain some complementary informations from depth maps. In fact, owing to the popularity of depth sensing technologies and the importance of the depth information, RGB-D SOD has attracted the attention of researchers, and various RGB-D SOD methods have been designed to detect the salient object from the RGB image and corresponding depth maps [19–24]. Traditional RGB-D SOD methods adopted the image priors with hand-crafted feature to detect the saliency object

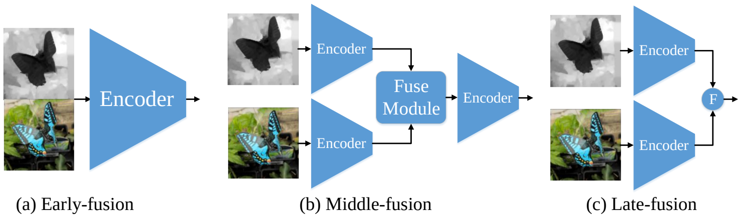


Fig. 1. The architecture of three fusion manners. Early-fusion methods directly concatenate two modalities. Middle-fusion methods utilize two-stream architecture to fuse two modalities. Late-fusion methods fuse features as a post-processing step.

in scenes, including contrast [19], shape [25] and so on. However, the hand-crafted features cannot be represented well to the complex scenario, which limited the performance of these methods. Recently, benefiting from the progress of Convolutional Neural Networks (CNNs) and the representation ability of features, several CNN-based RGB-D SOD approaches were proposed [26, 27]. Depending on merging the RGB and depth map features in different stages, these methods can be divided into three categories: early-fusion, middle-fusion and late-fusion, the architectures of these manners are shown in Fig. 1. Early-fusion schema merged the RGB and depth to a four-channels input and fed it into network directly [28, 29]. While middle-fusion methods usually designed a two-branch architecture network to fused the features [27, 23]. Late-fusion [30] methods extracted the features of RGB and depth map separately, and fuse these features as a post-processing step.

Though above middle-fusion methods have achieved promising performance, there is still a problem in front of them, which is how to integrate the different context information between the RGB and depth map effectively. To this end, we propose the Siamese Residual Interactive Refinement Network (SiamRIR) with residual manner to fuse two modalities and refine the features by incorporating the cues from encoder. Specifically, we adopt the Siamese Network as the encoder since it contains less parameters. Then, the Context Enhancement Module (CEM) is proposed to utilize the multi-scale features to improve the global context information. After that, we design a Residual Interactive Refinement Block (RIRB) to model the interactions during two modalities, which contains Residual Interactive Module (RIM) and Residual Refinement Module (RRM). The RIM takes interaction between the RGB and depth features to integrate the complementary regions, and RRM incorporates the spatial details which extracted by the encoder to refine the features.

In summary, the contributions of this paper are as follows:

- We propose a Siamese Residual Interactive Refinement Network (SiamRIR), which considers the different context cues. Our SiamRIR explores the complementary regions by taking interaction between modalities and refines the features by incorporating the spatial detail context.

- We design Multi-Scale Residual Interactive Refinement Block (RIRB) as decoder with residual manner to fuse the multi-type context. In this process, the complementary regions between two modalities are utilized by RIM, and the spatial details are explored by RRM with multi-scale manner, which can refine the features during the fusion phase, therefore the performance of SiamRIR is improved.
- We conduct extensive quantitative and qualitative evaluations on five RGB-D SOD benchmarks, which illustrates that SiamRIR outperforms previous state-of-the-art RGB-D SOD approaches.

2 Related Work

2.1 RGB-D Salient Object Detection

RGB-D salient object detection aims to detect the object in a scene that the human would be most interesting in. Peng [31] proposed a single-stream architecture to fuse the RGB and depth directly as inputs to predict the saliency maps. Song [22] adopted the multi-scale fusion to combine low-level, mid-level and high-level feature to calculate the saliency maps. Liu [32] fed the RGB-D four-channels into network to generate multiple level features. Then a depth recurrent network is applied to render salient object outline from deep to shallow hierarchically and progressively. The fusion strategy of these methods is called early-fusion, which merge the RGB and depth as single input of network. In contrast to early-fusion, middle-fusion can make full fusion of RGB and depth, thus there are many methods apply this strategy. Liu [33] proposed a two-stream network to fuse features from different level by directly adding the features. Zhang [34] designed a bilateral attention network with a complementary attention mechanism to better utilize salient informations in foreground and background. Huang [35] considered corresponding semantic information to distinguish the informative and non-informative regions in the input RGBD images. Chen [36] utilized RGB images to predict a hint map, then used hint map to enhance the depth map, this approach resolved the low-quality issue of depth maps. Wang [37] added an adaptive weights to depth maps to reduce the negative effects of unreliable depth maps.

Different from above approaches, our method considers the multi-type cues during fusion stage including the different context between two modalities, and spatial detail context from the features extracted by encoder.

2.2 Siamese Network

In order to reduce the number of parameter of the model, several works utilized Siamese network to extract the features of RGB and depth. Siamese network was proposed by Bromley [38] for hand-written signature verification. In their paper, two uniform networks were designed to deal with different signatures, where these two networks shared the same parameter. During the learning process,

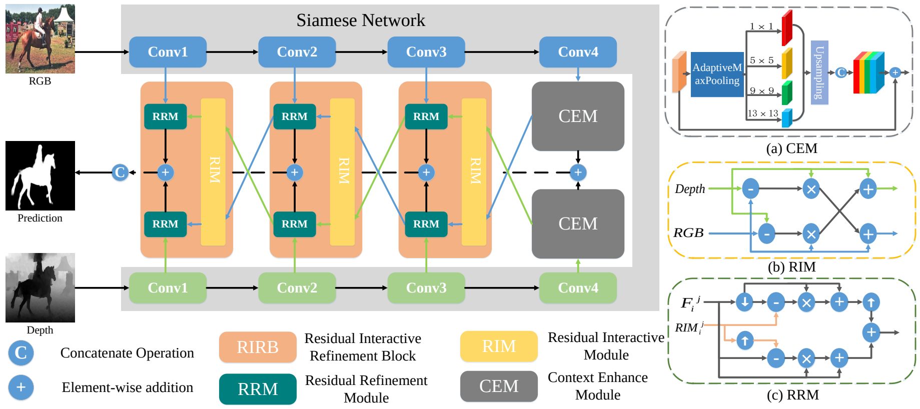


Fig. 2. The overall architecture of Siamese Residual Interactive Refinement Network (SiamRIR).

the features were constrained by the distance measure. Since it is suitable for calculating the distance from two similar inputs, Siamese network was further applied in sundry tasks where the inputs were similar, such as Chan [39] proposed a novel siamese implicit region proposal network with compound attention for visual tracking, and Fan [40] designed a multi-stage tracking framework, which consists of a sequence of RPNs cascaded from deep high-level to shallow low-level layers in a Siamese network. Recently, several works adopted Siamese network to salient object detection. For example, Zhao [41] proposed a lightweight and real-time model with a simple initialization strategy, which can make full use of the pre-trained model on ImageNet [42] to extract the discriminative features and utilized the depth to guide the fusion between RGB and depth.

Different from above RGB-D SOD methods, in this work, the Siamese network is applied to take advantage of the complementary regions between RGB and depth, rather than measuring distance. Specifically, we concatenate the RGB and depth along the batch dimension and fed it into the network [43], after that we fuse the features during decoding stage to achieve the interactive between RGB and depth in stead of measuring distance.

3 Approach

3.1 Architecture Overview

The architecture of the proposed framework is illustrated in Fig. 2. We utilize two backbones to extract the features from two modalities, where the parameters in these two backbones are shared (*i.e.*, Siamese Network). To be concise, we define the features of RGB branch in the encoder as $F_i^R (i \in \{1, 2, 3, 4\})$ and the features of depth map in the encoder as $F_i^D (i \in \{1, 2, 3, 4\})$. Then the features are fed into Context Enhance Module (CEM), to enhance the global context of the features by different receptive fields. After that, we design a Residual Interactive Refinement Block (RIRB) to decode features from the CEM. Specifically, the features are first input to Residual Interactive Module (RIM) to explore the complementary regions by taking interactive between the RGB and depth map with residual manner. Then, Residual Refinement Module (RRM) takes the features from RIM and encoder as inputs to refine the features with integrating the spatial information from the encoder. Finally, in order to reduce the parameters we adopt the element-wise addition to fuse the output of RRMs in the one RIRB and the output of CEMs, then concatenate these features as the prediction. In the following contents, we will describe the details of each components in the architecture.

3.2 Context Enhance Module (CEM)

The global context is useful for SOD method to detect the object. Therefore we design Context Enhance Module (CEM) to enhance the global context of the features from the encoder. Specifically, we fed the output of Conv4 (*i.e.*, F_4^R, F_4^D) into CEM, then we utilize four adaptive max pooling with different sizes to acquire four feature maps with different receptive fields,

$$F_i = AdaptiveMaxPooling_i(F_4^j) \tag{1}$$

where F_i denote the output of adaptive max pooling, $AdaptiveMaxPooling_i$ is the adaptive max pooling with different size ($i = 1, 5, 9, 13$). F_4^j is the input of CEM ($j \in \{R, D\}$). After that, we use convolution layer and ReLu to reduce the numbers of channels for the features to one fourth of F_4^j and up-sample these features to the size of F_4^j,

$$\overline{F}_i^j = \uparrow (ReLU(Conv(F_i^j))) \tag{2}$$

where $ReLU(*)$ is the ReLU function, $Conv(*)$ is the convolution layer and $\uparrow$ is the bilinear upsample operation. Finally, we concat these features and add the input of CEM to it to generate the features which contain the global context,

$$CEM^j = Cat(\overline{F}_1^j, \overline{F}_5^j, \overline{F}_9^j, \overline{F}_{13}^j) + F_4^j \tag{3}$$

where $Cat(*)$ denotes the concatenate operation. The details of this module are shown in Fig. 2 (a).

3.3 Residual Interactive Refinement Block (RIRB)

As shown in Fig. 2, the Residual Interactive Refinement Block (RIRB) contains two components, *e.g.*, Residual Interactive Module (RIM) and Residual Refinement Module (RRM). We embed RIRBs (*e.g.*, $RIRB_1, RIRB_2, RIRB_3$ represents the first RIRB from left to right) to decode the features, which can achieve the interactions of two modalities. Specifically, we adopt RIM to calculate the complementary regions between the features of two modalities with residual manner. Then, RRM is applied to incorporate the spatial informations obtained from the encoder to refine the features. In the following context, we will describe the details of RIM and RRM.

Residual Interactive Module (RIM) In the training process, we utilize the same ground-truth to supervise the prediction of RGB and corresponding depth map, the predictions of these two modalities contains consistency. However, the informations contained in the RGB and the corresponding depth map are different (*e.g.*, RGB contains semantic informations, depth map contains spatial depth informations), this will lead the network to locate the different regions. In order to explore the complementary regions between the features of two modalities during decoding phase for adjusting the prediction, we design the Residual Interactive Module (RIM) with residual manner to interact with the features of two modalities. For the RIM_3, as is shown in Fig. 2 (b), the blue and green lines represent the feature extracted from RGB (*i.e.*, CEM^R), the feature extracted from depth map (*i.e.*, CEM^D), respectively. We firstly adopt the convolution layer to reduce the channels of features to 256, Sigmoid and BatchNormalization are used to map the value of features to the range of 0 to 1. Then we subtract the CEM^R from the CEM^D to obtain the complementary regions between these two features,

$$Com^R = ReLU(Bn(Conv(CEM^D))) \tag{4}$$

$$- ReLU(Bn(Conv(CEM^R))) \tag{5}$$

$$Com^D = ReLU(Bn(Conv(CEM^R))) \tag{6}$$

$$- ReLU(Bn(Conv(CEM^D))) \tag{7}$$

where $Bn(*)$ represents the BatchNormalize layer. Then we multiply the F^R and F^D by the corresponding $Com^j (j \in \{R, D\})$, the channels of Com^j also be reduced to 256 by convolution layer, which can assign the weights to the complementary regions between two modalities.

$$Weight^R = Com^R * Conv(CEM^D) \tag{8}$$

$$Weight^D = Com^D * Conv(CEM^R) \tag{9}$$

$$RIM^R = Weight^R + CEM^D \tag{10}$$

$$RIM^D = Weight^D + CEM^R \tag{11}$$

Finally, we add the $Weight^R$ and the F^R together, add the $Weight^D$ and the F^D together. The next RIMs (*e.g.*, RIM_2^j, RIM_1^j) takes the outputs of the previous RIRB as the inputs. Thus, the final prediction of the saliency map can be adjusted step by step.

Residual Refinement Module (RRM) Since, there are more semantic context than detailed context in the high-level features, the details in the the RIM^R

and RIM^D are unsatisfactory. Hence, we design the Residual Refinement Module (RRM) to refine the features by incorporating the features from the encoding, which contains the spatial detail information. The RRM takes the output of the RIM and features from the encoder as the inputs, which are extracted from different modalities. It is worth mentioning that, we adopt the multi-scale manner to explore the detail context contained in different scale features. The architecture of RRM is shown in Fig. 2 (c). Firstly, since the channels of F_i^j and RIM_i^j are different, we utilize the convolution layer to decrease the channels of F_i^j to 256. After that, we down-sample the F_i^j to the size of RIM_i^j and up-sample the RIM_i^j to the size of F_i^j, respectively.

$$\hat{F}_i^j =\downarrow Conv(F_i^j) \tag{12}$$

$$\hat{R}IM_i^j =\uparrow RIM_i^j \tag{13}$$

Then we subtract the $\hat{F}_i^j$ from the RIM_i^j and $\hat{R}IM_i^j$ from F_i^j respectively, to locate the redundant regions by incorporating the features from the encoding between these two modalities.

$$Redu_{i\ down}^R = \hat{F}_i^D - RIM_i^R \tag{14}$$

$$Redu_{i\ down}^D = \hat{F}_i^R - RIM_i^D \tag{15}$$

$$Redu_{i\ up}^D = F_i^D - \hat{R}IM_i^R \tag{16}$$

$$Redu_{i\ up}^R = F_i^R - \hat{R}IM_i^D \tag{17}$$

$$\tag{18}$$

where $Redu_{i\,down}^j$ and $Redu_{i\,up}^j$ represent the redundant regions which subtract from different scales, then we multiply the $Redu_{i\,down}^j$ by $\hat{F}_i^j$ and add the $\hat{F}_i^j$. Meanwhile, multiply $Redu_{i\,up}^j$ by $\hat{F}_i^j$ and add the F_i^j.

$$RRM_i^R =\uparrow (Redu_{i\ down}^R * \hat{F}_i^D + \hat{F}_i^D) \tag{19}$$

$$+ Redu_{i\ up}^R * F_i^D + F_i^D \tag{20}$$

$$RRM_i^D =\uparrow (Redu_{i\ down}^D * \hat{F}_i^R + \hat{F}_i^R) \tag{21}$$

$$+ Redu_{i\ up}^D * F_i^R + F_i^R \tag{22}$$

At last, we add the features of these two different scales as the output of RRM_i^j.

3.4 Decoder Network

We integrate the outputs of CEM and RIRB as the prediction of the proposed method. Concretely, we sum the results of RRM in the same RIRB and sum the outputs of CEM, respectively. Then, we upsample the outputs of CEM and RIRBs to the size of ground-truth,

$$RIRB_i =\uparrow (RRM_i^R + RRM_i^D) \tag{23}$$

$$\overline{CEM} =\uparrow (CEM^R + CEM^D) \tag{24}$$

Notably, in the RIRBs and the features integration stage, we utilize the pixel-wise addition to integrate the features of two modalities, which can reduce the parameters of network and avoid the network modeling bias for a modalities. Inspired by [44] we concat $RIRB_i$ and $\overline{CEM}$ to retain the various levels of contexts, then convolution layer is adopted to reduce the channels of feature to 1 as the final prediction of out method. Besides, the $RIRB_i$ and $\overline{CEM}$ is also used to be the coarse maps,

$$S^f = Conv(Cat(RIRB_i, \overline{CEM})) \tag{25}$$

$$S_i^a = RIRB_i \tag{26}$$

where $Cat(*)$ represents the concatenate operation, S^f is the final prediction map of our method and S_i^a denotes the coarse maps to coarsely locate the objects. The S^f and S_i^a suffer from the same supervision operation to make sure these maps are consistent.

3.5 Implementation Details

Loss function In salient object detection fields, binary cross-entropy loss is the classical loss function to calculate the relation between the ground truth and the predicted saliency map, which is defined as:

$$\begin{aligned} \ell = -\frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} & [G_{ij} \log (S_{ij}) \\ & + (1 - G_{ij}) \log (1 - S_{ij})] \end{aligned} \tag{27}$$

where H, W indicate the height and weight of the image respectively, G_{ij} denotes the ground truth of the pixel (i, j) and S_{ij} denotes the predicted saliency map of the pixel (i, j). In order to coarsely detect the results, we utilize the auxiliary loss ℓ_{aux}^i $(i = 1, 2, 3, 4)$ at decoder stages. Specifically, we apply 3×3 convolution layer to decrease the channel of the feature maps to 1. After that, these maps are fed into bilinear interpolation to up-sample the feature maps to ground truth size. The total loss function ℓ_{total} is formulated as:

$$\ell_{\text{total}} = \ell_f \left(S^f, G\right) + \sum_{i=1}^{4} \lambda_i \ell_{aux}^i \left(S_i^a, G\right) \tag{28}$$

where S^f is the final predicted result, λ_i indicates the weight of different loss we set λ_i as $\{1, 1, 1, 1\}$, S_i^a represents the coarse maps predicted.

Network Training We apply Pytorch as our training platform. In the training process, we resize the input image to 384×384, then randomly crop a patch with the size of 240×240. Noticeably, we convert the depth map into three channel by simple gray color mapping. The module parameters are optimized by Adam optimization algorithm, with the batch size of 16, the momentum parameter 0.9 and the weight decay to $5e^{-4}$. We set the learning rate to $1e^{-4}$ and stop training after 48 epochs.

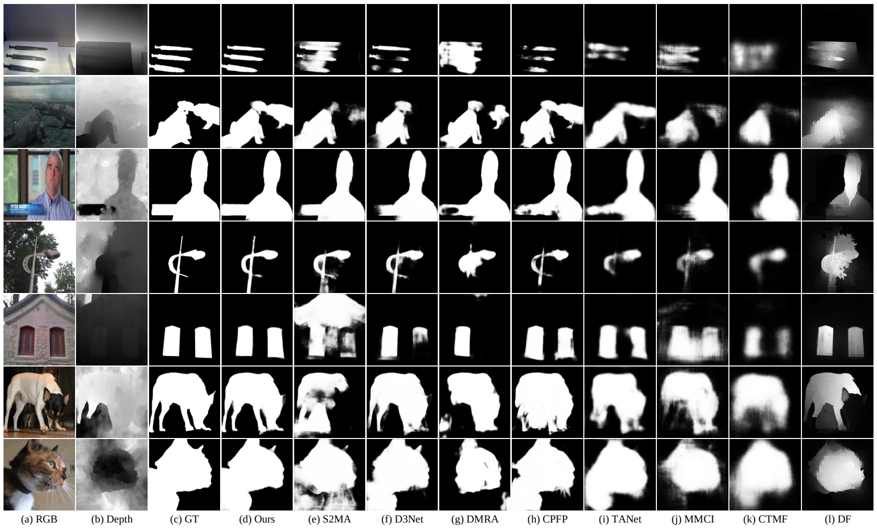


Fig. 3. Visual comparisons of the state-of-the-art RGB-D SOD methods and our method. As shown in this figure, the saliency maps generated by our method are closer to ground-truth than others, especially when the color of object is similar to background and complex background (*e.g.*, the chameleon in the second row, the street lights in the fourth row).

4 Experiments

4.1 Datasets

NJU2000 [45] consists of 1985 RGB-D image pairs, which is collected from the internet, movies and photographs. NLPR [31] contains 1000 RGB-D image pairs respectively, with diverse scenarios collected by Kinect. Following [23], we select the 1500 image pairs from NJU2000 and 700 image pairs from NLPR as the training set. STERE [46] contains 1000 samples collected from internet and the depth maps are produced by sift flow algorithm [47]. SSD is a small-scale but high-resolution dataset with 80 image pairs picked up from movies. The last dataset is SIP [29], which contains 929 high-quality person images.

4.2 Evaluation Metrics

To quantitatively compare our method with other methods, we adopt five widely-used metrics, *i.e.*, S-measuer (S_α), maximum F-measure (F_β^{max}), maximum E-measure (E_ϕ^{max}) and Mean Absolute Error (M).

Table 1. Quantitative results of the state-of-the-art method and the proposed method. The best and second scores are marked with red and blue, respectively. ↑ / ↓ for a metric means higher/lower value is better.

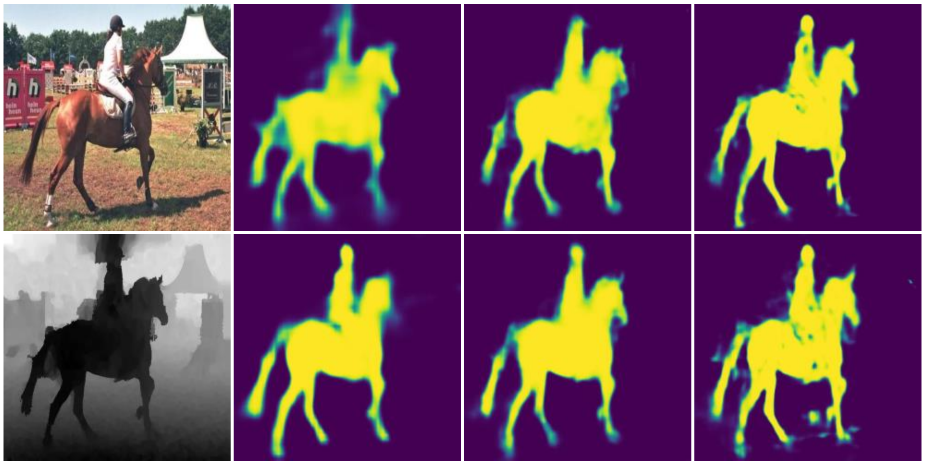
Dataset	Metrics	PCF (2018)	TANet (2019)	CPFP (2019)	DMRA (2019)	D3Net (2020)	MCINet (2021)	ASIF (2021)	FANet (2022)	FCMNet (2022)	SiamRIR (Ours)
NJU2K	S_{α} ↑	0.877	0.878	0.879	0.886	0.895	0.900	0.889	0.899	0.901	0.912
	M ↓	0.059	0.06	0.053	0.051	0.051	0.050	0.047	0.044	0.044	0.039
	E_{ϕ}^{max} ↑	0.924	0.925	0.926	0.927	0.932	0.920	0.921	0.914	0.929	0.948
	F_{β}^{max} ↑	0.872	0.874	0.877	0.886	0.889	0.873	0.900	0.892	0.907	0.911
STERE	S_{α} ↑	0.875	0.871	0.879	0.886	0.891	0.901	0.869	0.881	0.899	0.908
	M ↓	0.064	0.06	0.051	0.047	0.054	0.042	0.050	0.047	0.043	0.039
	E_{ϕ}^{max} ↑	0.925	0.923	0.925	0.938	0.93	0.929	0.926	0.908	0.939	0.943
	F_{β}^{max} ↑	0.86	0.861	0.874	0.886	0.881	0.872	0.894	0.863	0.904	0.899
NLPR	S_{α} ↑	0.874	0.886	0.888	0.899	0.906	0.917	0.884	0.913	0.916	0.926
	M ↓	0.044	0.041	0.036	0.031	0.034	0.027	0.050	0.026	0.024	0.023
	E_{ϕ}^{max} ↑	0.925	0.941	0.932	0.947	0.946	0.947	0.926	0.951	0.949	0.958
	F_{β}^{max} ↑	0.841	0.863	0.867	0.879	0.885	0.890	0.894	0.885	0.908	0.912
SIP	S_{α} ↑	0.842	0.835	0.85	0.806	0.864	0.867	0.373	-	0.858	0.873
	M ↓	0.071	0.075	0.064	0.085	0.063	0.056	0.269	-	0.062	0.054
	E_{ϕ}^{max} ↑	0.901	0.895	0.903	0.875	0.910	0.909	0.552	-	0.912	0.914
	F_{β}^{max} ↑	0.838	0.83	0.851	0.821	0.862	0.840	0.250	-	0.881	0.871
SSD	S_{α} ↑	0.841	0.84	0.807	0.857	0.858	0.860	0.849	-	0.855	0.877
	M ↓	0.062	0.063	0.082	0.058	0.059	0.052	0.059	-	0.055	0.043
	E_{ϕ}^{max} ↑	0.892	0.897	0.852	0.906	0.910	0.901	0.888	-	0.903	0.916
	F_{β}^{max} ↑	0.804	0.81	0.766	0.844	0.834	0.820	0.846	-	0.860	0.851

4.3 Comparisons with State-of-the-art Methods

We compare our method with other state-of-the-art RGB-D SOD methods, including PCF[48], TANet [24], CPFP [23], DMRA [49], D3Net [29], MCINet [50], ASIF [51], FANet [52], FCMNet [53].

Quantitative Evaluation Table 1 illustrates the quantitative evaluation result in terms of four metrics on five datasets. As shown in Table 1, on the NJU2K and NLPR dataset, our method achieves the best performance in terms of the four metrics, on the STERE, SIP and SSD dataset, our method achieves the best performance in terms of three metrics (*i.e.*, S_{α}, M and E_{ϕ}^{max}). In addition, our method achieves competitive results for the SIP dataset (*i.e.*, our approach ranks second in terms of the four metrics). This indicates the effectiveness of the proposed approach.

Qualitative Evaluation In order to make the comparisons more intuitive, we further provide qualitative results. The visual results of our method and other state-of-the-art are shown in Fig. 3. From Fig. 3, we can find that the saliency maps generated by DF, CTMF and MMCI is not clear (*i.e.*, the edges of the salient object are slightly blurred). In addition, the results of DMRA and S2MA is not accurate (*e.g.*, in the fifth row, the saliency map of DMRA only



(a) RGB & Depth (b) $Input^j$ (c) RIM_1^j (d) RRM_1^j

Fig. 4. The features of the RIRBs. From this figure we can observe that the complementary regions of the features are recovered by RIM and the details of the features are refined by RRM.

detect one window and S2MA segment the wall as salient object). Moreover, the CPFP and D3MA are failure when the object is similar to background (*e.g.*, D3Net only detect one chameleon in the second, CPFP regard the background as a part of the dog since the color of background is similar to the head of the dog). In contrast to the above methods, the saliency maps predicted by our method are more accurate and clear, especially under the similar and complex background. Besides, when the quality of depth map is poor our method can also detect the object accurately (*e.g.*, the sword in the first row). Overall, the comparisons of quantitative and qualitative illustrate that our method achieve better performance than other state-of-the-art approaches and is less influenced by background and the quality of depth maps.

4.4 Ablation Study

To demonstrate the effectiveness of the components in our method, we perform ablation study on NJU2K, STERE and NLPR, as shown in Tab 2. We first explore the validity of concat the features from different level decoder as the prediction. Then, we investigate the effect of RIRB on model performance, which contains the RIM and RRM. We will make detailed analysis of these factors in following parts.

Effectiveness of multi-scale features fusion In Tab 2, baseline represents that we do not concat the features from various level decoder as the predictions, S_f represents we adopt the S_f as the predictions. By observing the first two rows

Table 2. Ablation studies with different components. '✓' means adding the corresponding component. ↑ / ↓ for a metric means higher/lower value is better.

Model	baseline	✓	✓	✓	✓	✓
	l_f		✓	✓	✓	✓
	RRM			✓	✓	✓
	RIM				✓	✓
	CEM					✓
NJU2K	S_α ↑	0.881	0.892	0.896	0.908	**0.912**
	M ↓	0.057	0.050	0.048	0.041	**0.039**
	E_ϕ^{max} ↑	0.927	0.937	0.935	0.944	**0.948**
	F_β^{max} ↑	0.872	0.888	0.890	0.903	**0.911**
STERE	S_α ↑	0.876	0.887	0.892	0.906	**0.908**
	$M \downarrow$ ↓	0.060	0.052	0.049	0.041	**0.039**
	E_ϕ^{max} ↑	0.919	0.926	0.931	0.941	**0.943**
	F_β^{max} ↑	0.853	0.873	0.877	0.897	**0.899**
NLPR	S_α ↑	0.897	0.908	0.916	0.920	**0.926**
	$M \downarrow$ ↓	0.033	0.030	0.027	0.025	**0.023**
	E_ϕ^{max} ↑	0.941	0.944	0.952	0.954	**0.958**
	F_β^{max} ↑	0.867	0.886	0.894	0.905	**0.912**

of Tab 2 we can find that the performance in terms of four metrics have declined, demonstrates that supervise the S_f is useful to improve the performance of network since S_f contains the multi-scale context informations.

Effectiveness of RIRB In our method, the RIRB consists of two components which are RRM and RIM, the RRM can explore the complementary regions by interacting the features of two modalities with residual manner and we adopt the RIM to refine the features during decoding phase by incorporating the spatial detail context from the encoder. We can find from Tab 2 that, after embed the RIM and RRM, the performance of our method is improved. Besides, the results of different variants to SiamRIR are listed in Tab 3, where $R_i (i \in \{1, 2, 3\}$ represents there are different number of RIRB in the SiamRIR. From this table we can find that, as the number of RIRB increases the performance of SiamRIR gradually improves. This study demonstrates the effectiveness of the RIRB, which takes the features of two modalities and features from encoder interaction.

In order to illustrate the RIRB intuitively, we provide the visualization results of $RIRB_1$ which are shown in Fig. 4, where $Input^j$, RIM_1^j and RRM_1^j represent the input of the $RIRB_1$, the output of the RIM_1^j, the output of the RRM_1^j, respectively (*e.g.*, $j = R$ where features in the first row, $j = D$ where features in the second row). By observing the first row in Fig. 4 (b) we can find that, after the RIM_1^R the complementary regions of rider are located, then after the RRM_1^j the details of the rider and the head of horse are refined. The similar phenomenon

Table 3. Ablation studies with RIRB. '✓' means adding the corresponding component. ↑ / ↓ for a metric means higher/lower value is better.

Model			NJU2K				STERE				NLPR			
R_1	R_2	R_3	$S_\alpha \uparrow$	$M \downarrow$	$E_\phi^{max} \uparrow$	$F_\beta^{max} \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$E_\phi^{max} \uparrow$	$F_\beta^{max} \uparrow$	$S_\alpha \uparrow$	$M \downarrow$	$E_\phi^{max} \uparrow$	$F_\beta^{max} \uparrow$
✓			0.891	0.048	0.933	0.887	0.893	0.048	0.93	0.878	0.909	0.029	0.946	0.888
	✓		0.904	0.043	0.938	0.902	0.903	0.041	0.937	0.892	0.925	0.025	0.954	0.911
		✓	**0.912**	**0.039**	**0.948**	**0.911**	**0.908**	**0.039**	**0.943**	**0.899**	**0.926**	**0.023**	**0.958**	**0.912**

can also be observed in the second row. Therefore, we can concluded that the features of two modalities can be adjusted (*i.e.*, the complementary regions are recovered and the spatial details are refined) by RIRBs with taking interaction of two modalities and spatial context, which can improve the performance of the proposed method.

Effectiveness of CEM Since there are different size of objects in the scene, the global context informations are important for our method to detect the salient object. Therefor, we design the CEM to enhance the global context in the features from high-level encoder. From the Tab 2 we can observe that, without the CEM, the performance of our method in three test datasets are decreased, which verifies the benefit of the CEM and also verifies the importance of global context cues to the SOD task.

5 Conclusion

In this work, we propose a novel RGB-D SOD method Siamese Residual Interactive Refinement Network (SiamRIR). In order to utilize the different context information during fusion stage effectively, we design a Multi-scale Residual Interactive Refinement Block (RIRB) with residual manner to interact the saliency maps of two modalities and the spatial detail information extracted by the encoder, which can explore the complementary regions and refine the features during decoding phase. And then, the Context Enhance Module (CEM) is proposed to improve the global context information. Extensive experiments illustrate that SiamRIR outperforms the state-of-the-art methods on RGB-D SOD task in terms of quantitative and qualitative.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China [grant nos. 61922064, U2033210, 62101387] and Zhejiang Xinmiao Talents Program [grant nos. 2022R429B046].

References

1. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
2. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE transactions on Pattern Analysis and Machine Intelligence **37**(3), 569–582 (2014)
3. Borji, A., Cheng, M.M., Hou, Q., Jiang, H., Li, J.: Salient object detection: A survey. Computational visual media **5**(2), 117–150 (2019)
4. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.: Understanding convolution for semantic segmentation pp. 1451–1460 (2018)
5. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation pp. 1520–1528 (2015)
6. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation pp. 3431–3440 (2015)
7. Chen, Z.M., Jin, X., Zhao, B.R., Zhang, X., Guo, Y.: Hce: hierarchical context embedding for region-based object detection. IEEE Transactions on Image Processing **30**, 6917–6929 (2021)
8. Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y.: Relation networks for object detection pp. 3588–3597 (2018)
9. Liu, W., Chang, X., Chen, L., Phung, D., Zhang, X., Yang, Y., Hauptmann, A.G.: Pair-based uncertainty and diversity promoting early active learning for person reidentification. ACM Transactions on Intelligent Systems and Technology (TIST) **11**(2), 1–15 (2020)
10. Zheng, Z., Zheng, L., Yang, Y.: A discriminatively learned cnn embedding for person reidentification. ACM Transactions on Multimedia Computing, Communications, and Applications **14**(1), 1–20 (2017)
11. Mahadevan, V., Vasconcelos, N.: Biologically inspired object tracking using center-surround saliency mechanisms. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(3), 541–554 (2012)
12. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. Acm computing surveys **38**(4), 13–es (2006)
13. Wang, W., Zhao, S., Shen, J., Hoi, S.C., Borji, A.: Salient object detection with pyramid attention and salient edges pp. 1448–1457 (2019)
14. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections pp. 3203–3212 (2017)
15. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection pp. 234–250 (2018)
16. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection pp. 714–722 (2018)
17. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach pp. 2083–2090 (2013)
18. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection pp. 8779–8788 (2019)
19. Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X.: Depth enhanced saliency detection method pp. 23–27 (2014)
20. Ren, J., Gong, X., Yu, L., Zhou, W., Ying Yang, M.: Exploiting global priors for rgb-d saliency detection pp. 25–32 (2015)

21. Cong, R., Lei, J., Fu, H., Hou, J., Huang, Q., Kwong, S.: Going from rgb to rgbd saliency: A depth-guided transformation model. IEEE Transactions on Cybernetics **50**(8), 3627–3639 (2019)
22. Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., Ren, T.: Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Transactions on Image Processing **26**(9), 4204–4216 (2017)
23. Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L.: Contrast prior and fluid pyramid integration for rgbd salient object detection pp. 3927–3936 (2019)
24. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. IEEE Transactions on Image Processing **28**(6), 2825–2835 (2019)
25. Ciptadi, A., Hermans, T., Rehg, J.M.: An in depth view of saliency (2013)
26. Zhao, S., Chen, M., Wang, P., Cao, Y., Zhang, P., Yang, X.: Rgb-d salient object detection via deep fusion of semantics and details. Computer Animation and Virtual Worlds **31**(4-5), e1954 (2020)
27. Chen, H., Li, Y., Su, D.: 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)
28. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. IEEE Transactions on Image Processing **26**(5), 2274–2285 (2017)
29. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems **32**(5), 2075–2089 (2020)
30. Chen, H., Deng, Y., Li, Y., Hung, T.Y., Lin, G.: Rgbd salient object detection via disentangled cross-modal fusion. IEEE Transactions on Image Processing **29**, 8407–8416 (2020)
31. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: a benchmark and algorithms pp. 92–109 (2014)
32. Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P.: Salient object detection for rgb-d image by single stream recurrent convolution neural network. Neurocomputing **363**, 46–57 (2019)
33. Liu, D., Hu, Y., Zhang, K., Chen, Z.: Two-stream refinement network for rgb-d saliency detection pp. 3925–3929 (2019)
34. Zhang, Z., Lin, Z., Xu, J., Jin, W.D., Lu, S.P., Fan, D.P.: Bilateral attention network for rgb-d salient object detection. IEEE Transactions on Image Processing **30**, 1949–1961 (2021)
35. Huang, N., Luo, Y., Zhang, Q., Han, J.: Discriminative unimodal feature selection and fusion for rgb-d salient object detection. Pattern Recognition **122**, 108359 (2022)
36. Chen, Q., Fu, K., Liu, Z., Chen, G., Du, H., Qiu, B., Shao, L.: Ef-net: A novel enhancement and fusion network for rgb-d saliency detection. Pattern Recognition **112**, 107740 (2021)
37. Wang, J., Chen, S., Lv, X., Xu, X., Hu, X.: Guided residual network for rgb-d salient object detection with efficient depth feature learning. The Visual Computer **38**(5), 1803–1814 (2022)
38. Bromley, J., Bentz, J.W., Bottou, L., Guyon, I., LeCun, Y., Moore, C., Säckinger, E., Shah, R.: Signature verification using a “siamese” time delay neural network. International Journal of Pattern Recognition and Artificial Intelligence **7**(04), 669–688 (1993)

39. Chan, S., Tao, J., Zhou, X., Bai, C., Zhang, X.: Siamese implicit region proposal network with compound attention for visual tracking. IEEE Transactions on Image Processing **31**, 1882–1894 (2022)
40. Fan, H., Ling, H.: Siamese cascaded region proposal networks for real-time visual tracking pp. 7952–7961 (2019)
41. Zhao, X., Zhang, L., Pang, Y., Lu, H., Zhang, L.: A single stream network for robust and real-time rgb-d salient object detection pp. 646–662 (2020)
42. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database pp. 248–255 (2009)
43. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection pp. 3052–3062 (2020)
44. Zhang, P., Liu, W., Zeng, Y., Lei, Y., Lu, H.: Looking for the detail and context devils: High-resolution salient object detection. IEEE Transactions on Image Processing **30**, 3204–3216 (2021)
45. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference pp. 1115–1119 (2014)
46. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis pp. 454–461 (2012)
47. Liu, C., Yuen, J., Torralba, A.: Sift flow: Dense correspondence across scenes and its applications. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(5), 978–994 (2010)
48. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection pp. 3051–3060 (2018)
49. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection pp. 7254–7263 (2019)
50. Huang, Z., Chen, H.X., Zhou, T., Yang, Y.Z., Liu, B.Y.: Multi-level cross-modal interaction network for rgb-d salient object detection. Neurocomputing **452**, 200–211 (2021)
51. Li, C., Cong, R., Kwong, S., Hou, J., Fu, H., Zhu, G., Zhang, D., Huang, Q.: Asif-net: Attention steered interweave fusion network for rgb-d salient object detection. IEEE Transactions on Cybernetics **51**(1), 88–100 (2020)
52. Zhou, X., Wen, H., Shi, R., Yin, H., Zhang, J., Yan, C.: Fanet: Feature aggregation network for rgbd saliency detection. Signal Processing: Image Communication **102**, 116591 (2022)
53. Jin, X., Guo, C., He, Z., Xu, J., Wang, Y., Su, Y.: Fcmnet: Frequency-aware cross-modality attention networks for rgb-d salient object detection. Neurocomputing **491**, 414–425 (2022)