Postprint: Channel-Weight-Based Sequential Refinement RGB-D Salient Object Detection Network

Authors: Bian Huajun, Wang Huajun, Zhao Hewei

Date: 2022-04-07T15:01:57+00:00

Abstract

A novel network framework (SR-Net) for RGB-D salient object detection is proposed. To effectively integrate the complementarity of multimodal features, depth feature extraction is employed as an independent branch, the Convolutional Block Attention Module (CBAM) is adopted for depth feature enhancement, and the complementary information between the enhanced depth features and RGB features is integrated. To eliminate feature redundancy and reduce the interference of background noise on prediction results, a sequential refinement network is designed in the upsampling network, which obtains primary global features by integrating the complementarity of multi-level and multi-scale features, and employs a channel-weight-based Primary Global Feature Weight Matrix Acquisition Module (PFW) to obtain the weight matrix of primary global features; secondly, the obtained weight matrix is utilized to refine features at each level to suppress interference caused by background noise; finally, to better optimize the entire network, a novel loss function is proposed. Experimental results on four public datasets demonstrate that the model outperforms nine state-of-the-art methods in recent years across different evaluation metrics, achieving excellent performance.

Full Text

Preamble

Vol. 39 No. 8
Application Research of Computers
ChinaXiv Cooperative Journal

Sequential Refined RGB-D Salient Object Detection Network Based on Channel Weight

Bian Huajun, Wang Huajun, Zhao Hewei
(School of Network Security, Chengdu University of Technology, Chengdu 610059, China)

Abstract: This paper proposes a novel network framework for RGB-D salient object detection called SR-Net. To effectively integrate the complementarity of multi-modal features, depth feature extraction is treated as an independent branch, and the Convolutional Block Attention Module (CBAM) is employed for depth feature enhancement, followed by the integration of complementary information between enhanced depth features and RGB features. To remove feature redundancy and reduce background noise interference in prediction results, a sequential refining network is designed in the upsampling network. This network first obtains primary global features by integrating complementary multi-level and multi-scale features, then uses a channel-weight-based Primary Global Feature Weight Matrix Acquisition Module (PFW) to obtain the weight matrix of primary global features. Subsequently, the obtained weight matrix refines features at each level to suppress interference from background noise. Finally, a new loss function is proposed to better optimize the entire network. Experimental results on four public datasets demonstrate that the proposed model outperforms nine state-of-the-art methods across different evaluation metrics, achieving superior performance.

Keywords: salient object detection; RGB-D; channel weight; sequential refinement

0 Introduction

RGB-D salient object detection (RGB-D SOD) aims to identify the most attractive regions from a pair of RGB and depth images. Over the past decade, salient object detection (SOD) has attracted significant attention due to its wide application as a preprocessing step in image segmentation [**?**], image editing [**?**], and video analysis [**?**]. Traditional SOD methods primarily rely on hand-crafted low-level features [**?**, **?**, **?**, **?**], but they struggle to achieve satisfactory results in complex backgrounds due to their limited ability to capture semantic information of salient objects. Recently, with the rapid development of deep learning, researchers have begun applying convolutional neural networks (CNNs) to RGB-D SOD and have achieved promising results. Li et al. [**?**] first built a saliency model based on multi-scale features using deep neural networks. Wu et al. [**?**] proposed a Cascaded Partial Decoder (CPD) model that integrates deeper features from the backbone network to obtain an initial saliency map, which is then refined through a holistic attention module to produce the final saliency map. Liu et al. [**?**] argued that the backbone network extracts multi-level features from shallow to deep layers to generate coarse saliency maps that locate

salient objects but lose contour details. In their DRCNN-Net, they employed DRCNN to render salient objects from deep to shallow layers, where low-level side outputs could generate salient objects at multiple scales with the help of deep side outputs, original depth cues, and coarse saliency maps, thereby preserving more contour details. Wu et al. [**?**] proposed in MCMF-Net a method to detect salient object boundaries from corresponding geometric information using depth data, rather than simply extracting salient object features from depth data. However, as research progresses, two major challenges remain to be addressed: (1) how to effectively integrate the complementarity of multi-modal, multi-scale, and multi-level features; and (2) how to effectively suppress interference from complex background noise and remove redundant information in features.

To address these two challenges, this paper proposes a Sequential Refined RGB-D Salient Object Detection Network based on channel weight (SR-Net). Specifically, in SR-Net, we employ the attention mechanism-based CBAM (Convolutional Block Attention Module) to enhance depth features and effectively integrate the complementarity of multi-modal features. We also design a sequential refining network that first obtains primary global features by integrating complementary multi-level and multi-scale features (as shown in Figure 2), then uses the channel-weight-based Primary Global Feature Weight Matrix Acquisition Module (PFW) to obtain the weight matrix of primary global features and remove redundant information, and finally uses the obtained weight matrix to refine features at each level to suppress background noise interference.

RGB images contain color and texture information of salient objects, while depth images provide structural and spatial layout information, making their features complementary. Unlike approaches that merely treat depth images as a supplement to RGB images [**?**], we adopt two independent ResNet-50 backbone branches in the downsampling network to extract depth and RGB features separately. The extracted depth features are enhanced using the attention mechanism-based CBAM module, and the enhanced depth features are then integrated with RGB features to effectively fuse complementary multi-modal information.

Background noise can severely affect the final salient object prediction results. We aim to remove redundant information from features at each level during upsampling and use primary global features to refine each level's features, thereby emphasizing and enhancing important information. To this end, in the upsampling network, we first integrate complementary multi-level and multi-scale features to combine texture information from low-level features and semantic information from high-level features, obtaining primary global features (see Figure 2). These primary global features are then fed into the PFW module to remove redundant information and obtain a weight matrix (as shown in Figure 2), which is used to refine features at each level and reduce background noise interference.

Unlike previous methods that simply integrate complementary multi-modal fea-

tures, our approach first uses two ResNet-50 backbone branches to extract RGB and depth features separately, with CBAM modules enhancing depth features to effectively integrate multi-modal complementarity. Furthermore, to remove redundant information and reduce background noise interference, we design a sequential refining network in the upsampling network and propose the channel-weight-based PFW module to remove redundancy from primary global features and obtain a weight matrix for refining features at each level. As shown in Figure 2, our model produces salient object predictions with clear edges (as in the first row of Figure 1) and complete structures (as in the second and third rows of Figure 1). In summary, our contributions are:

1) We employ an attention mechanism-based CBAM module for depth feature enhancement. Unlike previous works that treat depth features merely as a supplement to RGB features, we use an independent ResNet-50 backbone branch for depth feature extraction.

2) We design a sequential refining network that first obtains primary global features by integrating multi-level and multi-scale features, then uses the weight matrix of primary global features to refine features at each level to remove redundant information.

3) We design a Primary Global Feature Weight Matrix Acquisition Module (PFW) that uses an attention mechanism to remove feature redundancy from primary global features and obtain a corresponding weight matrix for refining features at each level.

4) To better optimize our network, we propose a new loss function. Experimental results demonstrate that under the optimization of this new loss function, our SR-Net achieves excellent performance across four public datasets.

---

1 Overall Model Architecture

As shown in Figure 2, we propose the Sequential Refined RGB-D Salient Object Detection Network based on channel weight (SR-Net). The architecture comprises two independent ResNet-50 feature extraction backbone branches, a primary global feature acquisition branch, and an upsampling feature refinement branch based on channel weight and primary global features. Specifically, in the ResNet-50 feature extraction backbone branches, conv$i$ represents the feature extraction backbone at each layer. The extracted depth features undergo enhancement through the Depth Feature Enhancement Module (CBAM), after which the enhanced features are fused with RGB features extracted by the backbone network to obtain multi-modal integrated features for the upsampling network. In the primary global feature acquisition branch, the multi-modal integrated features first pass through the Global Contextual Module (GCM) and upsampling operations for contextual information integration and upsampling.

Subsequently, complementary multi-level and multi-scale features are integrated to obtain primary global features.

In the upsampling feature refinement branch using primary global features, the obtained primary global features first pass through the attention weight mechanism-based Primary Global Feature Weight Matrix Acquisition Module (PFW) to remove redundant information and generate a corresponding weight matrix (as shown as 'W' in Figure 2). This weight matrix is then used to refine features at each level. Finally, refined multi-level and multi-scale features are integrated to obtain the final salient object prediction result. To better optimize the proposed channel-weight-based sequential refinement network, we perform upsampling at different levels of the network to obtain saliency prediction maps at each level and compute sub-loss functions. Specifically, different weights are assigned to sub-loss functions at each level based on their impact on the final prediction (as shown as '$*$' in Figure 2). The detailed network architecture is described below.

1.1 Depth Feature Enhancement Module (CBAM)

To effectively integrate the complementarity of RGB and depth features, previous works often employ simple fusion methods such as concatenation, element-wise multiplication, addition, or treat depth features merely as a supplement to RGB features without deeply considering that direct simple fusion may introduce redundancy and noise due to intrinsic modality differences and depth feature redundancy. Inspired by [**?**], we construct a depth feature enhancement module using channel attention and spatial global attention mechanisms to enhance depth features. As shown in Figure 3, the input feature map F_{input} undergoes max-pooling and avg-pooling to obtain channel-wise weights. After ratio transformation to extract global channel information and element-wise addition, we obtain the channel attention-based feature map F_{CA}. The specific computation process is as follows:

$$f_1 = \delta(conv_{c\to c/ratio}(maxpool(F_{input}))) \tag{1}$$

$$f_2 = \delta(conv_{c\to c/ratio}(avgpool(F_{input}))) \tag{2}$$

$$F_{CA} = \sigma(conv_{c\to c}([f_1, f_2])) \tag{3}$$

where *maxpool* and *avgpool* represent global max pooling and global average pooling, respectively, $conv_{i\to j}$ denotes a 1×1 convolution that transforms the number of channels from i to j, *ratio* represents the scaling factor, δ denotes the ReLU activation function, f_1 and f_2 represent intermediate transition variables, and F_{CA} denotes the feature map refined by the channel attention mechanism.

Subsequently, F_{CA} undergoes spatial-based max pooling and avg pooling to obtain spatial weights regarding salient objects. These are then concatenated,

and a 7×7 convolution transforms the channel number to 1, yielding the spatial attention-based feature map F_{SA}. The specific computation process is as follows:

$$F_{SA} = \sigma(conv_{2\to1}([maxpool(F_{CA}), avgpool(F_{CA})]))$$

where F_{CA} represents the feature map refined by the channel attention mechanism, *maxpool* and *avgpool* denote spatial global max pooling and global average pooling, respectively, and F_{SA} represents the feature map refined by the global attention mechanism.

1.2 Primary Feature Acquisition

As shown in Figure 1, the depth-enhanced features are element-wise added with RGB features extracted by the backbone network to integrate contextual information, which is then fed into the Global Contextual Module (GCM) for comprehensive contextual information processing to obtain features S_i. Since multi-level and multi-scale features contain complementary information about salient objects, effectively integrating these features yields primary global features that contain more essential information about salient targets. When used for attention mechanism-based global information weight acquisition, the resulting weights exhibit higher confidence. Based on this idea and considering that features at different levels have different scales, we first upsample each level's features to the same size ($88 \times 88 \times 32$). The specific upsampling ($up * n$) computation is as follows:

$$S_i = Relu(BN(conv_{3\times3}(S_i^*))) \tag{4}$$
$$S_i^* = upsample_n(S_i) \tag{5}$$

where S_i^* represents the feature after redundancy removal by the Global Contextual Module (GCM), $upsample_n$ denotes n-times upsampling operation, *Relu* represents the ReLU activation function, and S_i denotes the output feature after upsampling. Finally, the upsampled features at each level undergo element-wise multiplication to obtain the primary global feature. The specific computation process is as follows:

$$F_{prd1} = S_{u1} \otimes S_{u2} \otimes S_{u3} \otimes S_{u4} \otimes S_{u5}$$

where S_{ui} represents the output feature after upsampling, $conv_{3\times3}$ denotes 3×3 convolution, BN represents batch normalization, $\otimes$ denotes element-wise multiplication, and F_{prd1} represents the obtained primary global feature.

1.3 Primary Global Feature Weight Matrix Acquisition Module (PFW)

As shown in Figure 4, the primary global feature acquisition branch effectively integrates complementary multi-level and multi-scale features to obtain the primary global feature F_{prd1}. Since global features contain more important features of salient objects, using them to guide the refinement of features at each level can remove redundant information, automatically select and enhance important features, and reduce background noise interference. Based on this idea, we propose the Primary Global Feature Weight Acquisition Module (PFW), detailed as follows:

First, the primary global feature F_{prd1} obtained from the primary global feature acquisition branch undergoes a downsampling judgment based on the network level it will refine. Notably, considering that upsampling introduces more noise compared to downsampling, we choose to downsample F_{prd1} rather than upsample smaller-scale features when unifying feature sizes. The specific downsampling judgment formula is:

$$F'_{prd1} = \begin{cases} \text{interpolate}(F_{prd1}), & \text{if } size(F_{prd1}) \neq size(S_i) \\ F_{prd1}, & \text{otherwise} \end{cases}$$

where S_i and F_{prd1} represent the features at each level and the primary global feature, respectively, $size$ denotes feature size, interpolate represents bilinear interpolation-based downsampling, and F'_{prd1} denotes the output after the downsampling judgment.

Then, the output F'_{prd1} undergoes spatial-level global average pooling. Notably, we employ spatial global average pooling rather than spatial global max pooling because we believe max pooling exhibits specificity and instability, where weights from single channels could significantly impact the final overall weight distribution. Spatial global average pooling ensures greater robustness and accuracy for the entire network.

Finally, the feature after global average pooling sequentially passes through a 3×3 convolution and sigmoid activation function to generate the final spatial weight matrix of the primary global feature for guiding subsequent feature refinement. The specific computation process is as follows:

$$W_i = \sigma(conv_{3\times 3}(avgpool(F'_{prd1})))$$

where σ represents the sigmoid activation function, F'_{prd1} denotes the output after the downsampling judgment, $avgpool$ represents spatial global average pooling, and W_i denotes the spatial weight matrix of the primary global feature.

1.4 Feature Refinement Network

As shown in Figure 2, since primary global features contain more information about salient objects, using them to guide the refinement of features at each network level can remove redundant information and automatically select and enhance key information. Therefore, we multiply the obtained primary global feature spatial weights with features at each level to obtain refined features guided by primary global features. Subsequently, we integrate the refined features at each level in a top-down sequential manner to effectively combine complementary multi-level and multi-scale features and obtain the final salient object prediction result. The specific feature refinement process is as follows:

$$S_{ri} = S_i \otimes W_i$$

where S_i represents the feature after redundancy removal by the Global Contextual Module (GCM), $\otimes$ denotes element-wise multiplication, W_i represents the spatial weight of the primary global feature, and S_{ri} denotes the output result at each level after refinement by primary global features.

Furthermore, since the refined features at each level contain different information, to integrate complementary multi-level and multi-scale features, we perform element-wise multiplication or concatenation of features at each level in a top-down manner. For clarity, we instantiate the inputs as S_{r4} and S_{r5}. The specific computation process is:

$$S_{45} = BN(conv_3(S_{r4} \otimes S_{r5}))$$

where S_{r4} and S_{r5} represent refined features at each level, and S_{45} denotes the feature obtained after integrating the complementarity of these two levels.

Finally, the feature fused from multi-level and multi-scale information (S_{45}) is upsampled to the same size as the Ground Truth (GT) map (352×352). Considering that direct upsampling may lose details and introduce noise, we employ a simple yet effective Feature Size Conversion Module (FSC). Specifically, FSC first uses a 1×1 convolution to change the feature channel number, then adopts a residual network to upsample the input feature map, improving information flow and preventing gradient vanishing and degradation caused by network depth. The specific computation process is as follows:

$$f_3 = Relu(conv_5(f_4)) \tag{6}$$

$$Result = Relu(BN(conv_3(f_4))) + f_3 \tag{7}$$

where *Result* represents the final output of the upsampling network, *Relu* denotes the ReLU activation function, f_3 represents an intermediate transition

variable, $conv_5$ denotes the upsampling operation using convolutional layers of different sizes in the residual network, and $Result$ represents the final prediction result of the model.

1.5 Loss Function

To better train the entire network, we propose a new loss function. Experiments show that under the optimization of this new loss function, the model converges to an optimal point, producing final salient object predictions with more complete structures and clearer edges. The specific loss function composition is described below.

As shown in Figure 2, the primary global feature, feature refinement branch output, and final salient object prediction result are upsampled to the same size as the Ground Truth map. The specific upsampling process has been detailed in Equation (5) of Section 1.2. We then compute loss functions for the upsampled features separately, building upon the binary cross-entropy loss function. The binary cross-entropy loss function is calculated as:

$$loss = G\log(S) + (1-G)\log(1-S)$$

where G represents the Ground Truth map and S represents the prediction result map. Smaller values indicate that the final prediction is closer to the Ground Truth.

To better align the loss function with the actual operation of the model, we assign different weights to loss functions at different fusion nodes to emphasize the varying degrees of influence that predictions at each network level have on the final salient object prediction result. The specific loss function formula is as follows:

$$loss = 0.1 \cdot loss_1 + 0.3 \cdot loss_2 + 0.5 \cdot loss_3$$

where $loss_1$, $loss_2$, and $loss_3$ represent the loss functions computed at different fusion nodes of the upsampling network, and $loss$ represents the overall loss function for the entire model's final prediction output.

---

2 Experiments and Results Analysis

This section first introduces the four public datasets, five evaluation metrics, and experimental details used in this paper. We then compare our proposed method with nine state-of-the-art models from recent years. Finally, a series of ablation experiments demonstrate the effectiveness of the proposed methods and modules.

2.1 Datasets

To comprehensively validate the effectiveness of the SR-Net model, we conduct experiments on four public datasets: SIP [**?**], NJUD [**?**], NLPR [**?**], and LFSD [**?**]. SIP [**?**] contains 929 high-resolution person images captured by Huawei Meta10, focusing on real-world human subjects. NJUD [**?**] comprises 1,985 images collected from the Internet and 3D movies. NLPR [**?**] contains 1,000 RGB-D images with pixel-level Ground Truth maps, where depth images are captured by Kinect under various lighting conditions and scenes, and images may contain multiple salient objects. LFSD [**?**] includes 100 indoor and outdoor images with a resolution of 360×360 captured by a Lytro light field camera.

2.2 Evaluation Metrics

To quantitatively evaluate our proposed model, we introduce the Precision-Recall curve (PR curve) and five evaluation metrics: $F_{\max}$, F_{ω}^{β}, E_m, and MAE. The PR curve is generated from a series of precision-recall pairs; curves closer to (1,1) indicate higher model prediction accuracy. Precision (P) and Recall (R) are calculated as:

$$P = \frac{|S \cap G|}{|S|}, \quad R = \frac{|S \cap G|}{|G|}$$

where G represents the Ground Truth map and S is the binarized mask of the prediction result based on a threshold. Since precision and recall can sometimes be contradictory, comprehensive consideration is necessary. The most common method is F_β, which is the weighted harmonic mean of precision and recall, defined as:

$$F_\beta = \frac{(1+\beta^2) \times P \times R}{\beta^2 \times P + R}$$

where P and R represent precision and recall, respectively, and β represents the weight. Following the recommendation in [**?**], we set β to 0.3 to emphasize precision.

MAE represents the average pixel-level error between the prediction result and Ground Truth. Smaller values indicate higher prediction accuracy. The specific calculation formula is:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x,y) - G(x,y)|$$

where S represents the model's prediction result, G represents the Ground Truth map, and H and W represent the height and width of the prediction map, respectively.

2.3 Experimental Details

Following [**?**, **?**], we select 1,485 and 700 images from NJUD and NLPR datasets as training sets, respectively. The remaining images, along with SIP and LFSD datasets, are used for testing. We use ResNet-50 as the backbone network and optimize the entire network using the Adam algorithm. Training is performed on a single NVIDIA GeForce RTX 2080Ti GPU with a batch size of 8. The initial learning rate is set to $1e^{-4}$ and is reduced to 0.1 times its original value every 60 epochs. The entire network stops training at 200 epochs, and the best model is saved for testing. All experiments are implemented on the PyTorch platform.

2.4 Comparison with State-of-the-Art Models

In this section, we compare our proposed SRNet with nine state-of-the-art models [**?**] from both qualitative and quantitative perspectives. For fair comparison, we either reproduce experimental results using the authors' released source code (e.g., [**?**]) or directly use the salient object predictions provided by the authors.

2.4.1 Qualitative Analysis

1) As shown in Figure 5, we randomly select six advanced models from the nine comparison methods for qualitative analysis with SR-Net. In the first row of Figure 5, when detecting hands and held objects, many detection methods such as CoNet [**?**], BiANet [**?**], CMWNet [**?**], and D3Net [**?**] fail to obtain accurate salient objects, and their results contain significant noise. While cmSalGAN [**?**] detects the general contour of salient objects, it lacks many edge details. In contrast, our model accurately detects both hands and held objects with clearer edges, which is also demonstrated in the second row of images.

2) Our proposed SRNet can accurately detect salient objects in multi-object scenarios. As shown in the third row of Figure 5, due to multiple objects in the image, some detection methods including BBS-Net [**?**], CoNet [**?**], BiANet [**?**], CMWNet [**?**], and D3Net [**?**] fail to accurately detect the main salient object, and their results contain varying degrees of noise. In contrast, our model can precisely identify the salient object among multiple objects and effectively reduce noise interference, as also shown in the fourth row of Figure 5.

3) Our proposed model can detect salient objects in complex backgrounds. As shown in the sixth row of Figure 5, due to the complex background behind the car, some detection models fail to capture the complete car contour, such as CoNet [**?**] and BiANet [**?**]. Although CMWNet [**?**] and D3Net [**?**] obtain the general car contour, they introduce considerable noise, making the results appear cluttered. In contrast, our model can completely detect the car while effectively reducing background noise interference, fully demonstrating its capability to handle complex background problems.

2.4.2 Quantitative Analysis To more intuitively demonstrate the effectiveness of our proposed model, we compare it with nine state-of-the-art methods using five evaluation metrics and PR curves, as shown in Table 1 and Figure 6. Specifically, our proposed model achieves the highest precision-recall rates on three public datasets (SIP, NJUD, NLPR) and ranks second on the LFSD dataset. Furthermore, Table 1 presents quantitative evaluations across five metrics, clearly showing that our model outperforms recent state-of-the-art methods on all five metrics for SIP and NLPR datasets. Compared with the most recent cmSalGAN (TMM21) [**?**], our model leads by a large margin across four datasets. For example, on the SIP dataset, SRNet improves over cmSalGAN by 2.6% and 3.9% on $F_{\max}$ and F_{ω}^{β} metrics, respectively, while reducing MAE by 17%. This fully proves that our model achieves superior performance compared to the latest cmSalGAN [**?**] model. Finally, compared with the relatively best method among the nine comparison methods, BBS-Net, our SRNet still achieves outstanding performance. Specifically, SRNet outperforms BBS-Net on all five metrics for SIP and NLPR datasets, and only slightly underperforms BBS-Net on some metrics (e.g., MAE) for NJUD and LFSD datasets, fully demonstrating that our model maintains clear advantages over the best competing method.

2.5 Ablation Experiments

This section conducts ablation experiments to validate the effectiveness of the sequential refining network, PFW module, and loss function designed in SR-Net.

1) To verify the effectiveness of our proposed sequential refining network, we visualize the three fusion nodes in Figure 2 ($Predict_1$, $Predict_2$, $Result$). The visualization results in Figure 7 show that as the sequential refining network progresses, salient objects in the image gradually become more complete under the guidance of primary global features, with most background noise filtered out. To further demonstrate effectiveness, we also perform quantitative analysis on the three fusion node outputs. As shown in Table 2, we measure five evaluation metrics for the three fusion nodes across three datasets. The results clearly show that the quality of salient object detection results improves progressively through the sequential refining network. Therefore, both visual and quantitative perspectives perfectly validate the effectiveness of our proposed sequential refining network.

2) As described in Section 1.3, we first obtain primary global features that contain abundant main features of salient objects. When used as guidance features, they can refine and enhance important features in the guided features while removing redundant information. Therefore, we propose the PFW module to remove redundancy and obtain the weight matrix of primary global features for guiding fusion. To prove the effectiveness of the PFW module, we remove it from Figure 2 (denoted as SRNet1), where primary global features are obtained only through element-wise multiplication of multi-level features without subsequent redundancy removal and

weight matrix acquisition via PFW. The ablation results in Table 3 show that the comparison model without PFW underperforms SR-Net across three datasets by an average of 1-2 percentage points, fully proving the effectiveness of the PFW module proposed in SR-Net for obtaining primary global feature weights.

3) As described in Section 1.5, to better train the entire network, we design a new loss function that assigns different weights to different fusion nodes to emphasize their varying influence on the final loss. To verify the effectiveness of our loss function, we modify it by computing loss only for the final salient object prediction result with weight 1, without computing losses for intermediate fusion nodes. The specific formula becomes $loss = loss_3$. The ablation results (denoted as SRNet2) in Table 3 show that under our designed loss function, our model outperforms SRNet2 across all metrics on three datasets by 1-2 percentage points, fully proving that our new loss function enables more accurate salient object predictions.

2.6 Failure Cases

To facilitate future research in this field, this section introduces some failure cases from our experiments and provides insights for these cases, as shown in Figure 8.

1) **Depth map misguidance**: In the first row of Figure 8, the depth image primarily highlights the first toy without emphasizing subsequent toys, causing our model and CoNet [**?**] to detect only the first toy as the salient object and fail to recognize subsequent toys. The second row of images further demonstrates this issue.

2) **Interference from backgrounds with similar color contrast to salient objects**: In the third row of Figure 8, the sculpture in the RGB image has very similar color to the background toys. Even though the depth map emphasizes only the sculpture, the interference from background colors with similar contrast in the RGB image causes Ours, cmSalGAN [**?**], and CoNet [**?**] to include background noise in their detection results.

---

3 Conclusion

This paper proposes a novel network framework for RGB-D salient object detection (SR-Net). To effectively integrate the complementarity of multi-modal features, depth feature extraction is treated as an independent branch, and the depth feature module CBAM is used for depth feature enhancement to integrate complementary information between enhanced depth features and RGB features. Secondly, to remove feature redundancy and reduce background noise interference in prediction results, a sequential refining network is designed in the upsampling network. This network integrates multi-level and multi-scale

features to obtain primary global features, uses the weight matrix obtained through the PFW module to refine features at each level, and finally proposes a new loss function. Experimental results on four public datasets demonstrate that the model outperforms nine state-of-the-art methods across different evaluation metrics.

References

[1] Wang Wenguan, Shen Jianbing, Yang Ruigang, et al. Saliency-aware video object segmentation [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, pp. 20-33.

[2] Cheng Mingming, Mitra N J, Huang Xing, et al. Repfinder: Finding approximately repeated scene elements for image editing [C]// ACM Transactions on Graphics, 2010: 83: 1-83: 8.

[3] Fan Dengping, Wang Ww, Cheng Mingming, et al. Shifting more attention to video salient object detection [C]// The IEEE Conference on Computer Vision and Pattern Recognition, 2019: 8554-8564.

[4] Borji A. and Itti L. State-of-the-art in visual attention modeling [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2012: 185-207.

[5] Borji A. Saliency prediction in the deep learning era: Successes and limitations [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019: 679-700.

[6] Guo Jingfang, Ren Tongwei, Bei Jia, et al. Salient object detection in RGB-D image based on saliency fusion and propagation [C]// Proceedings of the International Conference on Internet Multimedia Computing and Service (ICIMCS), 2015: 1-5.

[7] Woo S, Park J, and Lee J Y. Cbam: Convolutional block attention module [C]// Proceedings of the European conference on computer vision (ECCV), 2018: 3-19.

[8] Fan Dengping, Lin Zheng, Zhang Jiaxing. Rethinking RGB-D salient object detection: Models, datasets, and large-scale benchmarks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020: 463-473.

[9] Ran Ju, Ge Ling, Ge Wenjing, T. Ren, et al. Depth saliency based on anisotropic center-surround difference [C]// International Conference on Image Processing, 2014: 1115-1119.

[10] Hou Wenpeng, Li Bing, Wei Huaxiong, et al. RGBD salient object detection: A benchmark and algorithms [C]// Proceedings of the European Conference on Computer Vision (ECCV), 2014: 92-109.

[11] Li Nianyi, Ye Jingwei, Yu Ji, et al. Saliency detection on light field [C]// The IEEE Conference on Computer Vision and Pattern Recognition, 2014: 2806-2813.

[12] Piao Yongrui, Ji Wei, Li Jingjing, Zhang et al. Depth-induced multiscale recurrent attention network for saliency detection [C]// The IEEE International Conference on Computer Vision, 2019: 7254-7263.

[13] Chen Hao, Li Youfu. Progressively complementarity-aware fusion network for RGB-D salient object detection [C]// The IEEE Conference on Computer Vision and Pattern Recognition, 2018: 3051-3060.

[14] Jiang Bing, Zhou Zhi, Wang Xing, et al. cmSalGAN: RGB-D Salient Object Detection With Cross-View Generative Adversarial Networks [J]. IEEE Transactions on Multimedia, 2021: 1343-1353.

[15] Zhai Yinjie, Fan Dengping, Yang Jufeng. Bifurcated backbone strategy for RGB-D salient object detection [J]. IEEE Transactions on Image Processing, 2021: 8727-8742.

[16] Ji Wei, Li Jingjing, and Zhang Miao. Accurate RGB-D Salient Object Detection via Collaborative Learning [C]// Proceedings of the European Conference on Computer Vision (ECCV), 2020: 52-69.

[17] Zhao Xiaoqi, Zhang Lihe, Pang Youwei, et al. A Single Stream Network for Robust and Real-Time RGB-D Salient Object Detection [C]// Proceedings of the European Conference on Computer Vision (ECCV), 2020: 646-662.

[18] Zhang Zhao, Lin Zheng, Xu Jun. Bilateral attention network for rgb-d salient object detection [J]. IEEE Transactions on Image Processing, 2021: 1949-1961.

[19] Li Gongyang, Liu Zhi, Ye Linwei, et al. Cross-Modal Weighting Network for RGB-D Salient Object Detection [C]// Proceedings of the European Conference on Computer Vision (ECCV), 2020: 665-681.

[20] Fan Dengping, Lin Zheng, Zhang Jiaxiang, et al. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks [J]. IEEE Transactions on Neural Networks and Learning Systems, 2020.

[21] Zhao Jiaxing, Cao Yang, Fan Dengping, et al. Contrast prior and fluid pyramid integration for RGBD salient object detection [C]// Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2019: 3927-3936.

[22] Li, Guanbing, and Yu Yizhou. Visual saliency based on multiscale deep features [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 5455-5463.

[23] Wu Zhe, Su Li, and Huang Qingming. Cascaded partial decoder for fast and accurate salient object detection [C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

[24] Liu Zhenyi, Shi Song, Zhao Peng, et al. Salient object detection for RGB-D image by single stream recurrent convolution neural network [C]// Neurocomputing, 2019: 46-57.

[25] Wu Junwei, Zhou Wujie, Luo Ting, et al. Multiscale multilevel context and multimodal fusion for RGB-D salient object detection [C]// Signal Processing, 2021.

[26] Wang Haocong, Zhang Songlong, Peng Li. Salient region detection based on fusion of boundary information and color features [J]. Computer engineering and Application, 2019, 55 (3): 179-183.

[27] Zhai Jiyou, Tu Lizhong, Zhuang yan. Significance detection of boundary a priori and adaptive region merging [J]. Computer engineering and Application, 2018, 54 (6): 178-182.

Note: Figure translations are in progress. See original paper for figures.