Computational Visual Media
https://doi.org/10.1007/s41095-020-0200-x Vol. 7, No. 1, March 2021, 115–125

Research Article

WGI-Net: A weighted group integration network for RGB-D salient object detection

Yanliang Ge[1,†], **Cong Zhang**[1,†], **Kang Wang**[1], **Ziqi Liu**[1], **and Hongbo Bi**[1] (✉)

Abstract Salient object detection is used as a preprocess in many computer vision tasks (such as salient object segmentation, video salient object detection, etc.). When performing salient object detection, depth information can provide clues to the location of target objects, so effective fusion of RGB and depth feature information is important. In this paper, we propose a new feature information aggregation approach, weighted group integration (WGI), to effectively integrate RGB and depth feature information. We use a dual-branch structure to slice the input RGB image and depth map separately and then merge the results separately by concatenation. As grouped features may lose global information about the target object, we also make use of the idea of residual learning, taking the features captured by the original fusion method as supplementary information to ensure both accuracy and completeness of the fused information. Experiments on five datasets show that our model performs better than typical existing approaches for four evaluation metrics.

Keywords weighted group; depth information; RGB-D information; salient object detection; deep learning

1 Introduction

In recent years, salient object detection (SOD) has attracted widespread interest; it aims to distinguish the most visually obvious objects or regions in a given image. Salient object detection uses computers to imitate human visual mechanisms to detect and distinguish salient objects in given images. SOD has been applied to many fields, including content-based image editing [1–4], image and video compression [5], object segmentation and recognition [6–10], visual tracking [11–13], image retrieval [14, 15], etc. Due to their powerful ability to extract information, SOD [16, 17] and other related tasks (e.g., video salient object detection [18, 19], co-saliency detection [20, 21], light field salient object detection [22–24], etc.) are often used as preprocesses in visual tasks. Most early SOD approaches considered a single RGB image or a set of them. As depth cameras (such as Kinect, RealSense, etc.) began to be applied to computer vision, combining the use of depth information for salient object detection, namely RGB-D SOD, becomes a topic of interest.

Depth cues can supply additional information about appearance, so it is useful to fuse depth information into salient object detection. A model incorporating depth information is able to identify target objects in given images more quickly and accurately.

In recent years, more and more researchers have considered RGB-D SOD [25–27] as a way to improve salient object detection. Existing RGB-D SOD methods mostly fused depth input in one of 3 stages: fusion at an early stage [28–31], fusion at a middle stage [32–35], or fusion at a late stage [36–38]. Early stage fusion directly fuses the input, both RGB and depth features, into one channel to extract information. In Ref. [28], Peng et al. proposed a multi-stage RGB-D SOD algorithm that combines depth cues and appearance features in a coupled manner. As a result of the distribution gap between the two inputs, it is not easy to fit the data in one model.

[†] Yanliang Ge and Cong Zhang contributed equally to this article.

[1] School of Electrical Information Engineering, Northeast Petroleum University, Daqing 163000, China. E-mail: Y. Ge, 15804593399@139.com; C. Zhang, congzhang98@126.com; K. Wang, kangwangwww@163.com; Z. Liu, zicky2015@sina.com; H. Bi, bhbdq@126.com (✉).

Manuscript received: 2020-08-19; accepted: 2020-11-19

Some methods fuse depth features in a middle-stage, they first extract RGB features at each level and then combine them with depth features to generate saliency maps. For example, in Ref. [32], Feng et al. proposed a method that utilizes RGB-D saliency features to obtain angular spread directions. Fusing the input at a late stage firstly determines salient RGB and depth information in two channels, and then utilizes pixel-wise summation or multiplication to fuse the RGB and depth saliencies. For example, Cheng et al. [38] proposed a method that exploits visual saliency cues in color and depth spaces to compute the saliency map.

Since depth information can help to locate salient objects in an image, in this article, we present a weighting strategy to obtain more accurate depth feature cues. Furthermore, to exploit both RGB and depth information, we propose a novel feature integration method, weighted group integration (WGI), that can well employ each category of information. See Fig. 1. The first row shows that our model is able to accurately detect salient objects in complex scenes. The second row shows that, although the depth map is noisy, the predicted saliency map from our method is still close to the ground truth. Extensive experiments demonstrate that the proposed method achieves comparable results to other state-of-art models on five public benchmarks.

In summary, our main contributions are

1. A novel feature fusion method, WGI, which can effectively integrate RGB features and depth features to accurately distinguish salient objects in given images. It shows significant performance improvements over existing feature fusion modules like DRB.
2. A series of experiments on five popular datasets to verify the effectiveness and efficiency of the proposed approach.

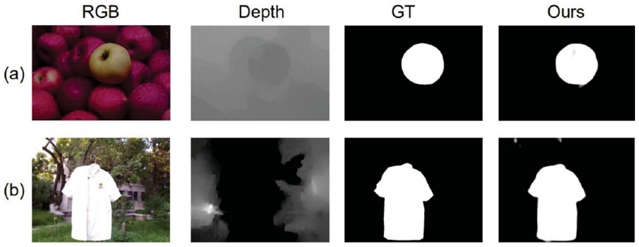

Fig. 1 Saliency maps from our model: (a) with a complex background, (b) with noisy depths.

2 Related work

2.1 Traditional

Traditional RGB-D saliency models usually rely on hand-crafted features to distinguish salient objects in given images. Existing widely-used hand-crafted features including contrast [28, 38, 39], compactness [39, 40], center-surround difference [41, 42], center or boundary prior [43, 44], background enclosure [32], and various fused saliency measures [29]. In Ref. [45], Niu et al. proposed a pioneering model for RGB-D SOD that applied disparity contrast and domain information into stereoscopic photography to measure stereo saliency. In Ref. [32], Feng et al. proposed a hand-crafted feature, local background enclosure (LBE) feature, that can directly assess salient structure from depths. LBE features distinguish the background from target objects or candidate regions. In Ref. [39], to reduce the influence of poor depth maps on saliency detection, Cong et al. turned the input into a graph and applied depth information to graph construction. They proposed a new method that utilizes RGB and depth features to compute a compactness saliency map. However, this hand-crafted feature has limitations, such as difficulty in providing high-level semantic information, slow and imprecise extraction of information, and poor generalizability in complex scenarios.

2.2 Deep learning based

To overcome the limitations of hand-crafted features, and benefit from the powerful information extraction capability of deep learning, recent works have applied convolutional neural networks (CNN) to RGB-D saliency detection. This improves the expressiveness of models and improves detection performance [25, 46–53]. Shigematsu et al. [33] proposed a pioneering method, BED, which applies deep-learning to RGB-D based SOD models. To obtain background enclosure features and depth contrast in given images, BED extracted ten hand-crafted depth features based on super-pixels. These features were then fed into a CNN to fuse them with RGB features to give superpixel saliency values. In Ref. [30], Qu et al. designed a method that firstly generated RGB and depth feature vectors for superpixels or patches, and then fused these vectors in the CNN to generate saliency values, ultimately utilizing a Laplacian function to obtain the predicted maps. More recently, Han et al. [47]

designed an end-to-end model that extracted features from both RGB images and depth maps, and used a fully connected layer to obtain the final saliency map.

3 Our approach

In this paper, we propose an integration network which fuses RGB and depth feature information for RGB-D salient object detection. In this section, we describe our proposed model, WGI-Net, for RGB-D salient object detection. We also explain the advantage of weighting depth information and clarify how to weight the depth information in detail. Finally, we expound the proposed feature fusion module, WGI that aggregates the depth features and RGB features to distinguish the salient objects in given images.

3.1 Overall network architecture

To explain our network for RGB-D based saliency detection, Fig. 2 depicts an example backbone with two branches (an RGB branch and a depth branch), each having a hierarchy of five levels. The RGB branch is utilized to obtain the main feature information, including low-level features (color, location, texture, etc.), high-level features (semantic information), and contextual features. The depth branch is used to capture depth cues from the image to help accurately and completely detect the salient objects. To better fuse the depth information with the RGB information, we present a novel feature fusion module, WGI. We employ element-wise addition to integrate the output of each WGI module, F_i^{fused} ($i = 1, \cdots, 5$). Finally, we feed the summed values into the FRU (see Ref. [34]) to obtain more detailed and accurate saliency maps.

3.2 Weighted depth information

Both RGB and depth information are significant for salient object detection and other segmentation. Specifically, the depth information can provide powerful cues to locate and distinguish salient objects in an image. It is difficult to accurately detect and distinguish salient objects in images only by appearance features when the background environment is complex or the color contrast between the foreground and background is low.

To the best of our knowledge, most existing models only consider depth information but not weighted depth information. In this paper, we apply weights

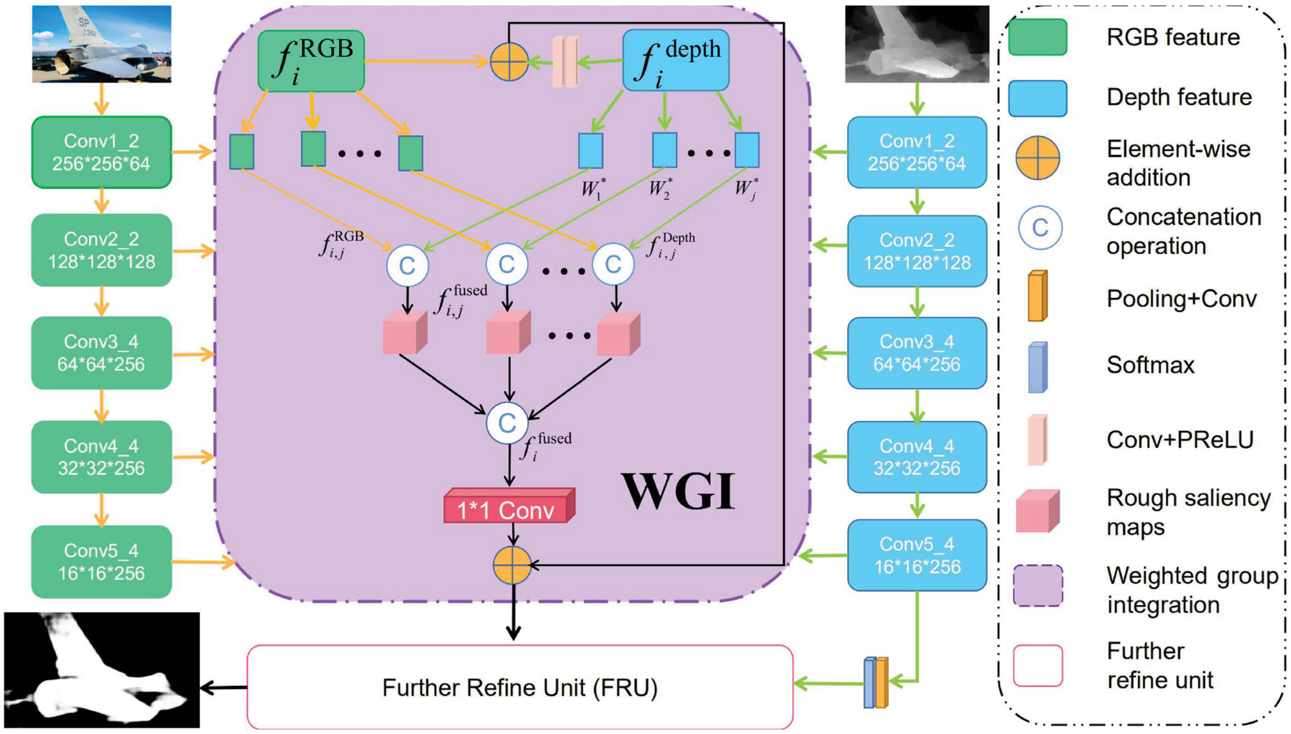

Fig. 2 An overview of our proposed model. f_i^{RGB} and f_i^{depth} ($i = 1, \cdots, 5$) represent RGB and depth features at each level, respectively. W_j^* ($j = 1, \cdots, 8$) are the weights for each block of depth information. $f_{i,j}^{\text{RGB}}$ and $f_{i,j}^{\text{depth}}$ represent features for the j^{th} part of the i^{th} level of RGB and depth information, respectively.

to depth information to obtain more accurate saliency maps. As shown in Fig. 2, we record the RGB feature and depth information of each layer as f_i^{RGB} and f_i^{depth} $(i = 1, \cdots, 5)$, respectively. We then divide both f_i^{RGB} and f_i^{depth} into 8 parts in each level, namely, $f_{i,j}^{\mathrm{RGB}}$ and $f_{i,j}^{\mathrm{depth}}$ $(j = 1, \cdots, 8)$.

To obtain the depth residual feature we firstly feed $f_{i,j}^{\mathrm{depth}}$ into a 3×3 convolutional layer, and compute:

$$f_{i,j}^{\mathrm{depth}'} = \mathrm{Conv}_3(f_{i,j}^{\mathrm{depth}}) \tag{1}$$

where $\mathrm{Conv}_3(.)$ represents a convolutional layer with a kernel size of 3.

This depth residual feature can provide cues that are ignored in the process of forwarding extracted information. Then, we divide $f_{i,j}^{\mathrm{depth}'}$ into two parts, and feed them into two branches. In one branch, we feed the $f_{i,j}^{\mathrm{depth}'}$ into a series of weight layers composed of a Pooling+Conv layer and a Softmax layer to capture more detailed and accurate information:

$$f_{i,j}^{\mathrm{depth}''} = S(\mathrm{AvgPooling} * f_{i,j}^{\mathrm{depth}'}) \tag{2}$$

where $S(.)$ denotes the softmax function and $*$ represents the convolution operation. In the other branch, we do nothing with $f_{i,j}^{\mathrm{depth}'}$. Finally, we utilize element-wise multiplication to fuse $f_{i,j}^{\mathrm{depth}'}$ and $f_{i,j}^{\mathrm{depth}''}$ to obtain complete information:

$$f_{i,j}^{\mathrm{Depth}} = f_{i,j}^{\mathrm{depth}'} \times f_{i,j}^{\mathrm{depth}'} \tag{3}$$

where × denotes element-wise multiplication. The weighted depth information can provide more accurate detailed information, i.e., $f_{i,j}^{\mathrm{Depth}}$ is more complementary to the RGB information.

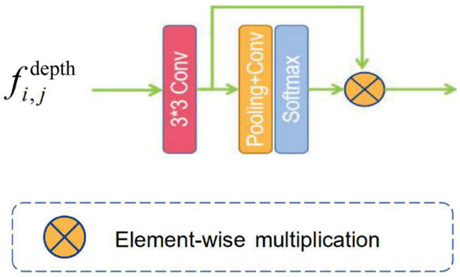

Fig. 3 The process of weighting depth information.

3.3 Weighted group integration module

In order to effectively utilize feature information from the image, we introduce a new feature fusion method, WGI. In this module, we divide the RGB information and depth information into 8 parts, respectively. Then, we use a concatenation operation to fuse the RGB feature and the depth feature for each part to obtain that part's saliency map. Next, we again utilize a concatenation operation to integrate the predicted maps to collect all useful information. Details of the WGI module are as follows.

Instead of fusing RGB and depth feature information using convolution layers, we seek alternative methods with powerful feature integration ability, while maintaining a similar or less computational load. In particular, we replace RGB and depth feature information with smaller groups of feature information blocks, while at the same time, the previous fusion information is connected in a similar residual style. As shown in the purple box in Fig. 2, the WGI module contains two branches, an RGB branch and a depth branch. In the RGB branch, we evenly divide the obtained RGB feature of each level into 8 sub-information blocks, $f_{i,j}^{\mathrm{RGB}}$ $(i = 1, \cdots, 5, j = 1, \cdots, 8)$. Each block of the RGB branch has the same number of channels, 1/8 of the input image. In the depth information branch, we perform the same operation on the input depth map as in the RGB branch, dividing the input depth map of each layer into 8 parts, $f_{i,j}^{\mathrm{depth}}$. We also perform the operation given in Section 3.2 on the 8 blocks to obtain more instructive depth information $(f_{i,j}^{\mathrm{Depth}})$.

Then, we separately merge each obtained RGB information block with the corresponding depth information block by concatenation on channel dimension:

$$f_{i,j}^{\mathrm{fused}} = \mathrm{Concat}(f_{i,j}^{\mathrm{RGB}}, f_{i,j}^{\mathrm{Depth}}) \tag{4}$$

where $\mathrm{Concat}(.)$ represents concatenation on channel dimension and $f_{i,j}^{\mathrm{fused}}$ denotes the fused feature information of each block. We then concatenate the 8 saliency prediction maps obtained from the previous step:

$$f_i^{\mathrm{fused}} = \mathrm{Concat}(f_{i,1}^{\mathrm{fused}}, \cdots, f_{i,8}^{\mathrm{fused}}) \tag{5}$$

where f_i^{fused} is the saliency prediction map generated by the 8 $f_{i,j}^{\mathrm{fused}}$ in this layer. Since concatenation changes the number of channels in the result, we perform a 1×1 convolution operation on the obtained prediction map to ensure it has the same size as the input map for each level. Thus, the reshaped fused information can be written as

$$f_i^{\mathrm{next}} = \mathrm{Conv}_1(f_i^{\mathrm{fused}}) \tag{6}$$

where $\mathrm{Conv}_1(.)$ represents a convolution layer with kernel size 1.

In order to ensure the completeness and accuracy of the information, the WGI module utilizes the information obtained by the original fusion method as residual information to correct the predicted saliency maps, allowing it to highly accurately distinguish the salient objects:

$$F_i^{\text{fused}} = f_i^{\text{next}} + f_i^{\text{res}} \tag{7}$$

where F_i^{fused} $(i = 1, \cdots, 5)$ denotes the output of the WGI module, and f_i^{res} $(i = 1, \cdots, 5)$ denotes the original fused information of each level.

In this module, we use segmentation and fusion to slice the depth information and RGB information extracted from each layer and then fuse them separately. This method is conducive to the use of global information and can more effectively fuse the two types of information.

4 Experiments

4.1 Datasets

The following datasets were chosen for evaluation. Ju et al. [41] proposed a dataset, NJUD, for detecting salient objects or pixels in given images. The dataset consists of 2000 images with mask labels. Its stereo images were taken with a Fuji W3 camera. Images were collected from the Internet, 3D movies, and photos. Because of the labeling differences between 2D images and 3D environments, the labels were all provided by Nvidia 3D vision, to ensure accuracy of mask labeling. The RGBD135 dataset [38] comprises 135 indoor images with manually marked labels. The images are taken by Kinect and the resolution of each image is 640 × 480. To address the problem of strong complementarity between RGB and depth, Peng et al. [28] proposed a benchmark, NLPR, for RGB-D salient object detection. It contains 1000 natural images and manually matched ground truths. Zhang et al. [54] presented a dataset, LFSD, based on general salient object segmentation and saliency detection on light fields. The dataset comprises 3 parts, including outdoor scenes, indoor scenes, and corresponding ground truths. DUT-RGBD [34] presented by Piao et al. is composed of 1200 pairs of images taken by a Lytro camera. Most images have complex backgrounds so are suitable for evaluating the effectiveness of our proposed model.

4.2 Evaluation metrics

In this paper, we utilize four common measures to evaluate the quality of predicted saliency maps against the ground truth: MAE [59], F-measure [60], S-measure [61], and E-measure [62]. MAE evaluates the mean absolute error between saliency maps S and corresponding ground-truth G over all image pixels:

$$\text{MAE} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} \left| S(x, y) - G(x, y) \right| \tag{8}$$

where H and W denote the height and the width of the image, respectively. F-measure computes the weighted harmonic mean between precision P and recall R of binarized saliency maps, defined as

$$F_\beta = \frac{(\beta^2 + 1) PR}{\beta^2 P + R} \tag{9}$$

where β^2 is generally set to 0.3 to emphasize precision. $F_{\max}$ is maximum F-measure. S-measure computes the structural similarity of the object-aware S_{o} and the region-aware S_{r} comparing the non-binary saliency map and the ground truth.

$$S = \alpha S_{\text{o}} + (1 - \alpha) S_{\text{r}} \tag{10}$$

Following previous work [61], α is set to 0.5. E-measure utilizes both local and global pixels to obtain local pixel matching information and image-pixel statistics. Unlike S-measure, E-measure evaluates binary maps:

$$E = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_{\text{FM}}(x, y) \tag{11}$$

where ϕ_{FM} is the enhanced alignment matrix. We consider maximum E-measure, $E_{\max}$.

4.3 Training details

We implemented our method using the Pytorch toolbox and utilized an NVIDIA 1080 Ti GPU for acceleration. As training dataset we used the same one as DMRA [34], with input maps set to 256×256. Other experimental parameters, momentum and weight decay, were set to 0.9 and 0.0005, respectively. In addition, the learning rate was 10^{-10}, and the batch size was 2.

4.4 Comparison

We made a detailed comparison between our method and seven other state-of-the-art frameworks for SOD based on RGB-D: CPFP [25], MMCI [35], TAN [55], DMRA [34], A2dele [56], ASIF [57], and D3Net [58]. To fully compare our proposed method, WGI-Net with these existing approaches, we re-evaluated these models using available source code or directly used saliency maps provided by their authors.

4.4.1 Quantitative comparision

Detailed comparative results of experiments based on the above four metrics are listed in Table 1. As it can be seen, our framework achieves good performance. Our proposed approach performs better than all other approaches across all four metrics on most datasets. Specifically, in terms of $F_{\max}$ and $E_{\max}$, WGI-Net achieves the best performance across all datasets. For the NJUD dataset, our model achieves the best performance on all four evaluation metrics. Compared with the second ranked models, D3Net, $F_{\max}$ score, $E_{\max}$ score, and S score are higher by 0.006, 0.002, and 0.001 respectively, while MAE is lower by 0.005. In the LFSD dataset, compared with the second place, A2dele, our F score, E score, and S score are higher by 0.022, 0.016, and 0.015, respectively.

4.4.2 Visual comparison

Figure 4 provides sample saliency maps predicted by the proposed method and several other algorithms. It intuitively illustrates the outstanding ability of our method to highlight correct salient object regions. Specifically, as shown in the 1st and 2nd rows of Fig. 4, the saliency maps provided by our method are closer to the ground-truth. Our method can detect the edges of objects more completely and accurately, while the maps output by other models lose certain items. For the LFSD dataset, for example, our results have no holes or extra parts. For the NJUD dataset, our saliency maps are more similar to the ground-truth: e.g., ours clearly detects the flags on the car, while others only detect part or none. For the NLPR dataset, our method accurately distinguishes salient objects in the foreground, while other methods detect incomplete objects or extraneous objects as salient objects. The 1st and 2nd rows of the RGB135 dataset show results for small and large objects; our method is able to provide accurate results in the cases. In summary, our model is able to handle various complex situations and provide highly accurate saliency maps.

4.5 Ablation study

To verify the effectiveness of our WGI-Net, an ablation experiment was conducted comparing just the backbone with additionally using WGI. Our backbone is DMRA [34] following the identical implementation setup. We conducted experiments on two datasets, RGBD135 and DUT-RGBD.

Table 1 Performance comparison to seven state-of-the-art architectures, for five datasets. Maximum F-measure $F_{\max}$, maximum E-measure $E_{\max}$, S-measure S, and MAE are utilized to assess performance. ↑ and ↓ indicate the higher the score the better, and the lower the better, respectively. "$*$" indicates that the author has not provided corresponding saliency maps. The top three ranking results in each row are indicated in red, green, and blue, respectively

		2019	2019	2019	2019	2020	2020	2020	
		CVPR	PR	ICCV	ICCV	CVPR	CVPR	TNNLS	Ours
		CPFP [25]	MMCI [35]	TAN [55]	DMRA [34]	A2dele [56]	ASIF [57]	D3Net [58]	
NJUD [41]	$F_{\max}$ ↑	0.799	0.853	0.874	0.889	0.873	0.888	0.889	0.895
	$E_{\max}$ ↑	0.835	0.915	0.925	0.927	0.916	0.927	0.932	0.934
	S ↑	0.798	0.859	0.878	0.880	0.869	0.889	0.895	0.896
	MAE ↓	0.079	0.079	0.060	0.053	0.051	0.047	0.051	0.046
LFSD [54]	$F_{\max}$ ↑	0.825	0.771	0.796	0.841	0.836	0.824	0.819	0.858
	$E_{\max}$ ↑	0.871	0.839	0.847	0.886	0.880	0.860	0.864	0.896
	S ↑	0.828	0.787	0.801	0.823	0.837	0.823	0.832	0.852
	MAE ↓	0.088	0.132	0.111	0.087	0.074	0.090	0.099	0.076
DUT-RGBD [34]	$F_{\max}$ ↑	0.795	0.767	0.790	0.889	0.892	0.821	0.786	0.903
	$E_{\max}$ ↑	0.859	0.859	0.861	0.927	0.930	0.876	0.857	0.937
	S ↑	0.818	0.791	0.808	0.869	0.885	0.838	0.814	0.893
	MAE ↓	0.076	0.113	0.093	0.057	0.042	0.073	0.086	0.047
NLPR [28]	$F_{\max}$ ↑	0.868	0.815	0.863	0.883	0.880	0.888	0.886	0.890
	$E_{\max}$ ↑	0.932	0.913	0.941	0.940	0.945	0.944	0.946	0.947
	S ↑	0.888	0.856	0.886	0.890	0.896	0.906	0.906	0.905
	MAE ↓	0.036	0.059	0.041	0.035	0.028	0.030	0.034	0.031
RGBD135 [38]	$F_{\max}$ ↑	0.845	0.822	0.827	0.869	0.867	$*$	0.882	0.889
	$E_{\max}$ ↑	0.923	0.928	0.910	0.933	0.923	$*$	0.939	0.942
	S ↑	0.874	0.848	0.858	0.878	0.885	$*$	0.906	0.901
	MAE ↓	0.037	0.065	0.046	0.035	0.028	$*$	0.030	0.030

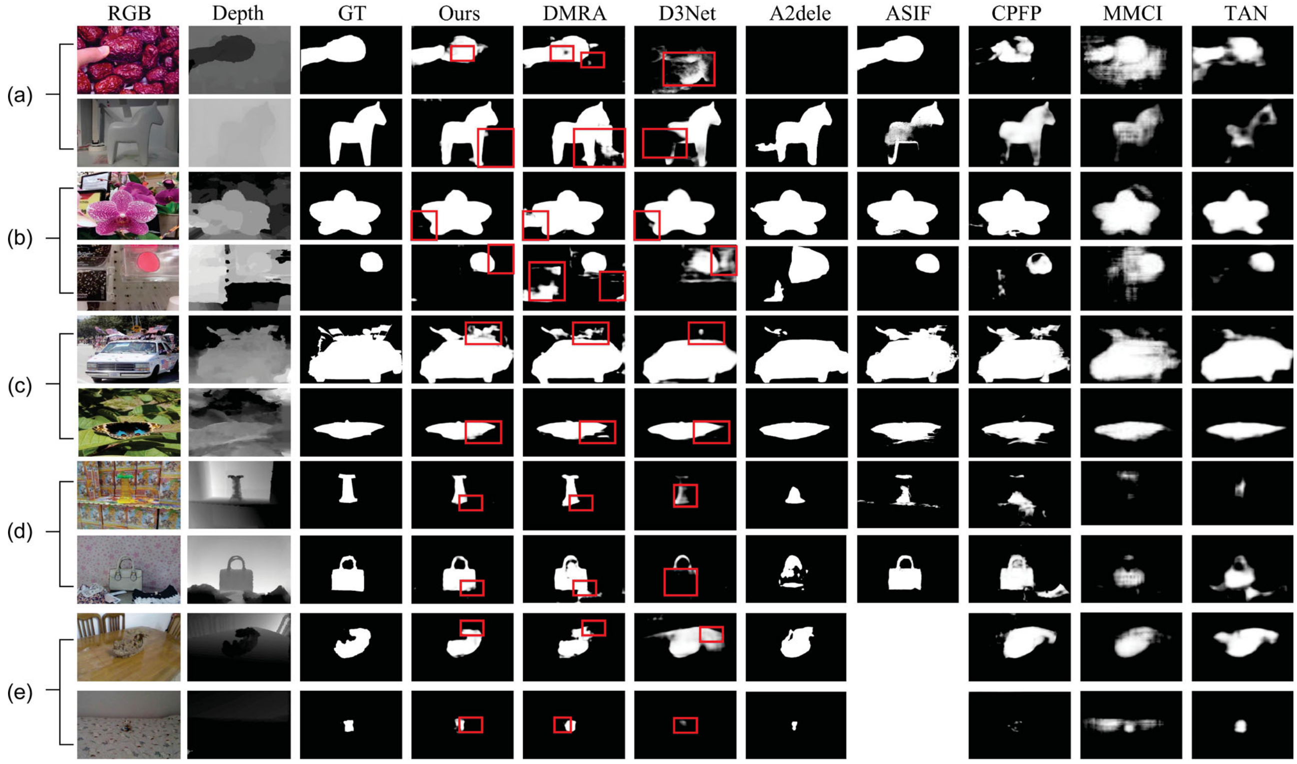

Fig. 4 Visual comparison of results from our method and other existing state-of-the-art algorithms on five public datasets. First two columns input images and the depth maps, respectively. 3rd column: ground truth saliency. 4th column: saliency map from our model. Areas in red boxes highlight advantages of our algorithm. Remaining columns: output from other method. Rows: (a) DUT-RGBD, (b) LFSD, (c) NJUD, (d) NLPR, (e) RGBD135. Empty cells indicate that the author did not provide corresponding saliency maps.

Experimental results are listed in Table 2. Compared to the baseline with no WGI module, the performance of our approach is improved. Specifically, for the RGBD135 dataset, $F_{\max}$, $E_{\max}$, and S score increase by 0.020, 0.009, and 0.023 respectively, while MAE decreases by 0.005. For the DUT-RGBD dataset, $F_{\max}$, $E_{\max}$, and S score increase by 0.004, 0.010, and 0.024 respectively, while MAE score by 0.010.

As shown in Fig. 5, the saliency maps from our method are closer to the ground-truth. Unlike the backbone (DMRA), our method is able to eliminate background interference and accurately detect salient objects against complex backgrounds.

Table 2 Ablation study of our proposed model on RGBD135 and DUT-RGBD datasets. (Ours represents Baseline+WGI)

Ablation study	RGBD135			
	$F_{\max}\uparrow$	$E_{\max}\uparrow$	$S\uparrow$	MAE $\downarrow$
Baseline	0.869	0.933	0.878	0.035
Ours	0.889	0.942	0.901	0.030
Ablation study	**DUT-RGBD**			
	$F_{\max}\uparrow$	$E_{\max}\uparrow$	$S\uparrow$	MAE $\downarrow$
Baseline	0.889	0.927	0.869	0.057
Ours	0.903	0.937	0.893	0.047

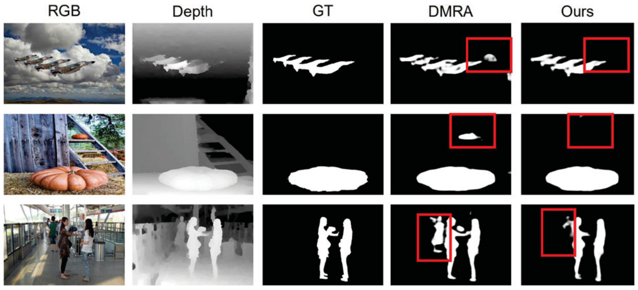

Fig. 5 Visual comparison of results from our method and the backbone (DMRA). Areas in red boxes highlight improvements in saliency maps produced by our algorithm.

5 Conclusions

In this paper, we have proposed a simple but efficient fusion approach, WGI, to make effective use of RGB feature information and depth feature information. The extracted RGB and depth features are sliced into 8 parts, and then concatenation is used to fuse the features of each block to more effectively integrate the two kinds of feature information. We also apply a series of weight layers to the depth information to obtain more accurate cues about the locations of the salient objects. Experiments on five datasets verify that our method performs better than current work for different evaluation metrics. Although our approach can accurately detect salient objects in complex environments through weighted group integration, it requires a large amount of calculation time. Therefore, in future we will focus on improving the fusion of information to make more effective use of feature information.

Acknowledgements

This work was supported by the NEPU Natural Science Foundation under Grants Nos. 2017PY ZL-05, 2018QNL-51, JY_CX_CX06 2018, JY_CX_JG06 2018, and JY_CX_14_2020.

References

[1] Marchesotti, L.; Cifarelli, C.; Csurka, G. A framework for visual saliency detection with applications to image thumbnailing. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 2232–2239, 2009.

[2] Ding, Y.; Xiao, J.; Yu. J. Importance filtering for image retargeting. In: Proceedings of the Conference on Computer Vision and Pattern Recognition, 89–96, 2011.

[3] Goferman, S.; Zelnik-Manor, L.; Tal, A. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 34, No. 10, 1915–1926, 2012.

[4] Wang, W.; Shen, J. Deep cropping via attention box prediction and aesthetics assessment. In: Proceedings of the IEEE International Conference on Computer Vision, 2205–2213, 2017.

[5] Guo, C. L.; Zhang, L. M. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE Transactions on Image Processing* Vol. 19, No. 1, 185–198, 2010.

[6] Rutishauser, U.; Walther, D.; Koch, C.; Perona, P. Is bottom–up attention useful for object recognition? In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, II, 2004.

[7] Han, J.; Ngan, K. N.; Li, M. J.; Zhang, H. J. Unsupervised extraction of visual attention objects in color images. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 16, No. 1, 141–145, 2006.

[8] Liu, Z.; Shi, R.; Shen, L. Q.; Xue, Y. Z.; Ngan, K. N., Zhang, Z. Y. Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut. *IEEE Transactions on Multimedia* Vol. 14, No. 4, 1275–1289, 2012.

[9] Jerripothula, K. R.; Cai, J. F.; Yuan, J. S. Image co-segmentation via saliency co-fusion. *IEEE Transactions on Multimedia* Vol. 18, No. 9, 1896–1909, 2016.

[10] Ye, L. W.; Liu, Z.; Li, L. N.; Shen, L. Q.; Bai, C.; Wang, Y. Salient object segmentation via effective integration of saliency and objectness. *IEEE Transactions on Multimedia* Vol. 19, No. 8, 1742–1756, 2017.

[11] Borji, A.; Frintrop, S.; Sihite, D. N.; Itti, L. Adaptive object tracking by learning background context. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 23–30, 2012.

[12] Hong, S.; You, T.; Kwak, S.; Han, B. Online tracking by learning discriminative saliency map with convolutional neural network. In: Proceedings of the International Conference on Machine Learning, 597–606, 2015.

[13] Lee, H.; Kim, D. Salient region-based online object tracking. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1170–1177, 2018.

[14] Gao, Y.; Wang, M.; Tao, D. C.; Ji, R. R.; Dai, Q. H. 3-D object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing* Vol. 21, No. 9, 4290–4303, 2012.

[15] He, J.; Feng, J.; Liu, X.; Cheng, T.; Lin, T.; Chung, H.; Chang, S. Mobile product search with bag of hash bits and boundary reranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3005–3012, 2012.

[16] Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P. H. Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3203–3212, 2017.

[17] Wang, W.; Zhao, S.; Shen, J.; Hoi, S. C.; Borji, A. Salient object detection with pyramid attention and salient edges. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1448–1457, 2019.

[18] Fan, D.; Wang, W.; Cheng, M.; Shen, J. Shifting more attention to video salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 8554–8564, 2019.

[19] Bi, H.; Lu, D.; Li, N.; Yang, L.; Guan, H. Multi-level model for video saliency detection. In: Proceedings of the IEEE International Conference on Image Processing, 4654–4658, 2019.

[20] Fu, H. Z.; Cao, X. C.; Tu, Z. W. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing* Vol. 22, No. 10, 3766–3778, 2013.

[21] Zhang, K.; Li, T.; Shen, S.; Liu, B.; Chen, J.; Liu, Q. Adaptive graph convolutional network with attention graph clustering for co-saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9050–9059, 2020.

[22] Wang, A. Z.; Wang, M. H.; Li, X. Y.; Mi, Z. T.; Zhou, H. A two-stage Bayesian integration framework for salient object detection on light field. *Neural Processing Letters* Vol. 46, No. 3, 1083–1094, 2017.

[23] Zhang, M.; Ji, W.; Piao, Y. R.; Li, J. J.; Zhang, Y.; Xu, S.; Lu, H. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing* Vol. 29, 6276–6287, 2020.

[24] Zhang, J.; Wang, X. Light field salient object detection via hybrid priors. In: Proceedings of the International Conference on Multimedia Modeling, 361–372, 2020.

[25] Zhao, J.; Cao, Y.; Fan, D.; Cheng, M.; Li, X.; Zhang, L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3927–3936, 2019.

[26] Zhang, J.; Fan, D. P.; Dai, Y.; Anwar, S.; Saleh, F. S.; Zhang, T.; Barnes, N. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8582–8591, 2020.

[27] 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. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3049–3059, 2020.

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

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

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

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

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

[33] Shigematsu, R.; Feng, D.; You, S.; Barnes, N. Learning RGB-D salient object detection using background enclosure, depth contrast, and top–down features. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2749–2757, 2017.

[34] Piao, Y.; Ji, W.; Li, J.; Zhang, M.; Lu, H. Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision, 7254–7263, 2019.

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

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

[37] Fan, X.; Liu, Z.; Sun, G. Salient region detection for stereoscopic images. In: Proceedings of the 19th International Conference on Digital Signal Processing, 454–458, 2014.

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

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

[40] 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* Vol. 50, No. 8, 3627–3639, 2019.

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

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

[43] Wang, A. Z.; Wang, M. H. RGB-D salient object detection via minimum barrier distance transform and saliency fusion. *IEEE Signal Processing Letters* Vol. 24, No. 5, 663–667, 2017.

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

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

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

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

[48] Huang, P.; Shen, C.; Hsiao, H. RGBD salient object detection using spatially coherent deep learning framework. In: Proceedings of the IEEE 23rd International Conference on Digital Signal Processing, 1–5, 2018.

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

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

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

[52] Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F.; Aliakbarian, S.; Barnes, N. Uncertainty inspired RGB-D saliency detection. *arXiv preprint* arXiv:2009.03075, 2020.

[53] Zhang, M.; Zhang, Y.; Piao, Y. R.; Hu, B. Q.; Lu, H. C. Feature reintegration over differential treatment: A top–down and adaptive fusion network for RGB-D salient object detection. In: Proceedings of the 28th ACM International Conference on Multimedia, 4107–4115, 2020.

[54] Zhang, J.; Wang, M.; Lin, L.; Yang, X.; Gao, J.; Rui, Y. Saliency detection on light field. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 13, No. 3, 1–22, 2017.

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

[56] Piao, Y.; Rong, Z.; Zhang, M.; Ren, W.; H. Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 9060–9069, 2020.

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

[58] Fan, D. P.; Lin, Z.; Zhang, Z.; Zhu, M. L.; Cheng, M. M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems* doi: 10.1109/TNNLS.2020.2996406, 2020.

[59] Borji, A.; Sihite, D. N.; Itti, L. Salient object detection: A benchmark. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7573.* Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 414–429, 2012.

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

[61] Fan, D.; Cheng, M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, 4558–4567, 2017.

[62] Fan, D.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. *arXiv preprint* arXiv:1805.10421, 2018.

Yanliang Ge received his bachelor degree in communications engineering in 2002 from Northeast Petroleum University, Daqing, China. He received his master degree in 2008 from Northeast Petroleum University in oil and gas information and control engineering. Currently he is an associate professor in the School of Electrical Information Engineering in Northeast Petroleum University. His main research interests concern digital watermarking, signal processing, and digital video processing.

Cong Zhang is pursuing her master degree at Northeast Petroleum University. Her current research interests include camouflaged object detection, RGB-D salient object detection, and deep learning.

Kang Wang is pursuing his master degree at Northeast Petroleum University. His current research interests include co-saliency detection, camouflaged object detection, RGB-D salient object detection, and deep learning.

Ziqi Liu is pursuing her master degree at Northeast Petroleum University. Her current research interests include RGB-D salient object detection, camouflaged object detection, RGB salient object detection, and deep learning.

Hongbo Bi received his bachelor and master degrees in communications engineering from Northeast Petroleum University in 2001 and 2004, respectively. He received his Ph.D. degree in 2013 from Beijing University of Posts and Telecommunications and worked as a postdoctoral fellow in Harbin Engineering University in 2014–2017. He also worked as a visiting scholar in the University of Waterloo, Canada in 2014–2015. Currently, he is an associate professor in the School of Electrical Information Engineering in Northeast Petroleum University. His main research interests focus on salient object detection, camouflaged object detection, compressive sensing, deep learning, digital watermarking, and signal processing.