Pattern Recognition 137 (2023) 109251

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

RGB-D salient object ranking based on depth stack and truth stack for complex indoor scenes

Jingzheng Deng[a], Jinxia Zhang[a,b,*], Zewen Hu[a], Liantao Wang[c], Jiacheng Jiang[a], Xinchao Zhu[a], Xinyi Chen[a], Yin Yuan[a], Chao Wang[a]

[a] The Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China
[b] Southeast University Shenzhen Research Institute, Shenzhen 518057, China
[c] The College of Internet of Things Engineering, Hohai University, Changzhou 213022, China

ARTICLE INFO

Article history:
Received 7 May 2022
Revised 2 November 2022
Accepted 13 December 2022
Available online 16 December 2022

Keywords:
Complex scenes
RGB-D
Salient object ranking
Indoor
Depth

ABSTRACT

RGB-D salient object detection has achieved a great development in recent years due to its extensive applications. Previous studies mainly focus on simple scene images with one single object. These models usually become overwhelmed by complex scenes with multiple objects. Moreover, these methods model salient object detection as a binary segmentation problem. However, psychology studies show that humans shift their visual attention from one object to another and rank salient objects, especially in complex indoor scenes. Following the psychological studies, we propose to rank salient objects in RGB-D images of complex indoor scenes. Due to the lack of such data, we first construct a RGB-D salient object ranking dataset containing complex indoor scenes with multiple objects. The saliency ranking of different objects is defined based on the order that an observer notices these objects. The final salient object ranking result is an average across the saliency rankings of 13 observers. This RGB-D salient object ranking dataset is also analyzed with current mainstream RGB-D salient object detection dataset for comparison. Since location information provided by depth images can help to determine the saliency ranking of objects, we further propose an end-to-end network exploiting depth stack and ground truth stack to predict the order of salient objects in complex scenes. The quantitative and qualitative comparisons demonstrate the effectiveness of the proposed method.

1. Introduction

Salient object detection can be used as a pre-processing technique for many vision-related applications such as semantic segmentation [1], foreground map evaluation [2,3], visual tracking [4], image parsing [5], image captioning [6] and person re-identification [7,8]. Therefore, research in salient object detection, which has attracted the interest of many researchers, has grown extensively in recent years [9–11]. Salient object detection has a strong correlation with the object's location. Moreover, the depth map, which can provide contour and location information of the object, is a vital aid in determining the saliency of the object. Therefore, RGB-D salient object detection gains more and more attention from researchers.

Most RGB-D salient object detection datasets contain simple outdoor scene images with a single object. However, the real-world scene is usually complex and contains multiple objects. This limits the application capacity of RGB-D salient object detection models to real-world vision tasks to some extent. Furthermore, psychology studies show that humans have the ability to shift their visual attention from one object to another. This ability can help humans to deal with complex scenes with multiple objects and rank salient objects accordingly. Besides, the judgment of saliency is usually subjective. Observers judge the saliency of different objects both differently and similarly. Thus, it is difficult to evaluate the saliency of objects in complex scenes in a straightforward way [12]. However, the existing RGB-D salient object detection task is treated as a binarization problem, which does not match the real visual perception of humans. Following the psychological studies, this paper introduces salient object ranking into the RGB-D saliency detection domain. Due to the lack of such a dataset, a complex indoor saliency ranking dataset with multiple objects is constructed in this paper. Based on the NYU Depth-v2 dataset [13], we invited 13 annotators to label the objects considered salient based on the

[*] Corresponding author at: The Key Laboratory of Measurement and Control of CSE, Ministry of Education, School of Automation, Southeast University, Nanjing 210096, China.
E-mail address: jinxiazhang@seu.edu.cn (J. Zhang).

https://doi.org/10.1016/j.patcog.2022.109251
0031-3203/© 2022 Elsevier Ltd. All rights reserved.

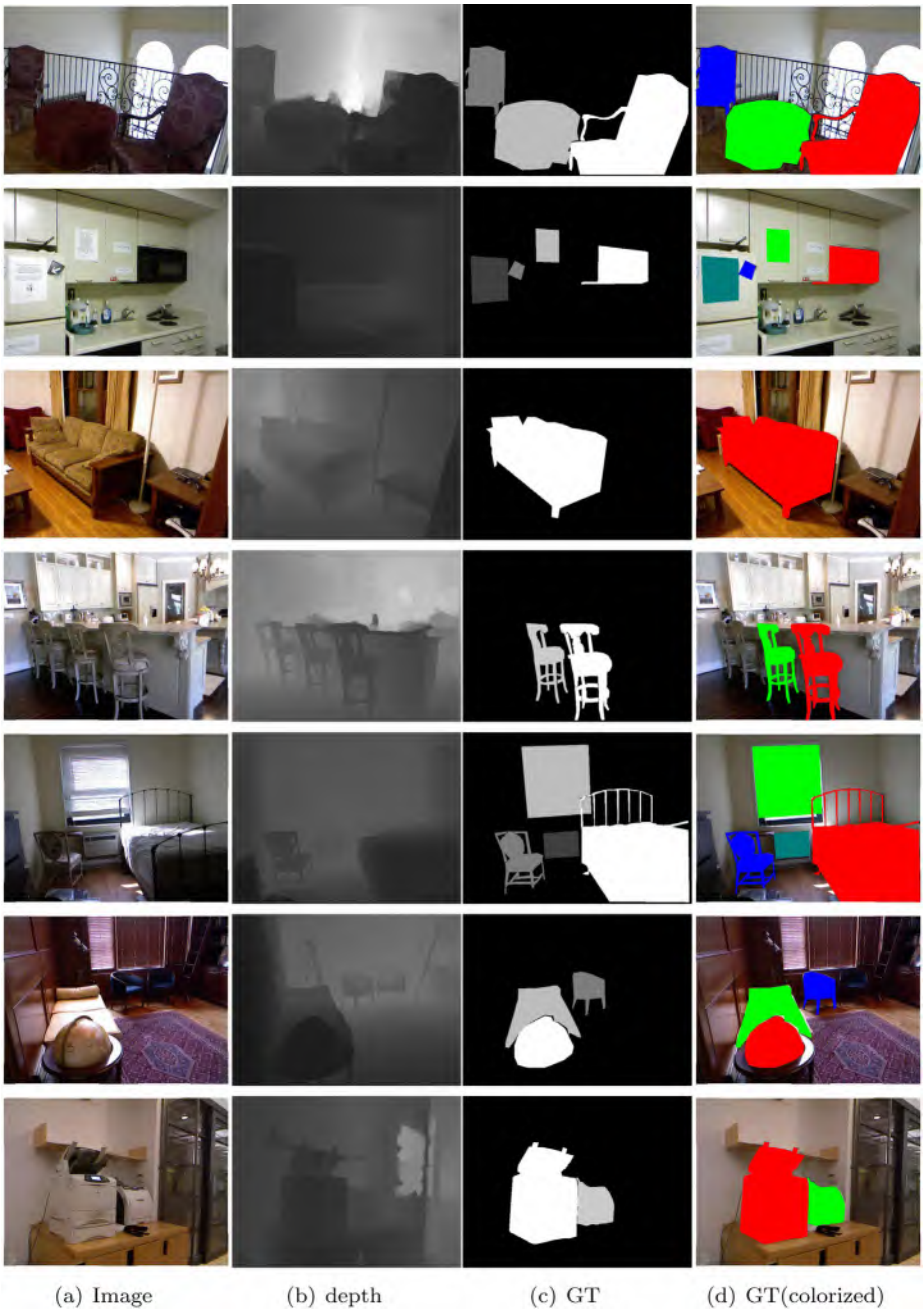


Fig. 1. (a) Image from our dataset, (b) corresponding depth map, (c) corresponding ground truth(GT) for saliency rank and (d) corresponding GT for saliency rank (colorised, red represents for rank 1, green represents for rank 2, blue represents for rank 3 and cyan represents for rank 4+). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

order they noticed the objects for each image. The final salient object ranking result is an average across the saliency ranks of these observers. We further analyze the proposed RGB-D salient object ranking dataset with the mainstream RGB-D salient object detection dataset for comparison.

Some example images and corresponding truths can be seen in Fig. 1. The first column of this figure is RGB images of the dataset, and it can be seen that it contains indoor scenes with multiple objects. Moreover, the background of the images in the proposed dataset is messy and complex. These characteristics of this dataset fit the realistic application scenario of salient object ranking. The second column contains depth images of the dataset, and the third column is the ground truths for salient object ranking. The fourth column contains colorized ground truths where red represents rank 1 (the most salient object), green represents rank 2 (the second salient object), blue represents rank 3 (the third salient object) and cyan represents for rank 4+. The number of salient objects in our dataset is not fixed and there may be one, two, three or more salient objects. Objects with a saliency rank greater than three are labeled as rank 4+. Not fixing the number of salient objects is more in line with the logic of salient object ranking. From the second, third, and fourth columns, we can see that objects with different saliency ranking levels usually have different depth values. Generally, objects with a higher level salient rank have smaller depth values. This shows that the depth map is significantly helpful in predicting the salient object ranking. And there is a correlation between the saliency ranking level of the object and the depth value.

To fully exploit the depth information, we propose an end-to-end learning network to rank salient objects based on the depth stack and ground truth stack. The depth stack consists of sub-depth maps for four different depth intervals. We treat the sub-depth maps as depth stack. Similarly, we separate the ground truth into four different sub-ground truths as a ground truth stack. Based on these stacks, four different coarse saliency prediction maps are generated. And then, a saliency re-fusion module is proposed to combine them to generate the final prediction map for salient object ranking.

The main contributions of the work include: Firstly, we propose a new research problem to rank salient objects in the RGB-D saliency detection field. This research problem is inspired by the visual perception of humans, who shift their attention from one object to another. Secondly, we construct an RGB-D salient object ranking dataset that contains complex indoor images with multiple objects. We analyze the dataset in-depth and compare it with salient object detection datasets. Thirdly, we propose an end-to-end learning network that fully uses depth information based on depth stack and ground truth stack to perform RGB-D salient object ranking task. Experimental comparisons demonstrate the effectiveness of the proposed method.

2. Background

2.1. RGB-D saliency detection datasets

RGB-D salient object detection is a study to locate the most salient objects for a given scene using RGB maps and depth maps [14]. With the development of depth sensors, depth maps with rich location information have become easier to acquire. This has led to significant advances in RGB-D salient object detection [15].

Since depth information can provide rich location and contour information in the feature extraction process, more and more researchers are now interested in RGB-D salient object detection. However, the existing scenarios for RGB-D salient object detection are relatively simple. Most existing RGB-D datasets collect images which have a prominent object or a relatively clean background such as STERE [16], NLPR [17], NJUD [18], DUT-RGBD [19] and SIP [20]. The sample images of these datasets are shown in the first five columns of Fig. 2. It can be seen that most of the scenes in the previous dataset contain only one salient object or person, and the background is relatively simple. This is a great difference from the actual application scenes. Real-world applications often encounter more complex situations, such as occlusions, appearance changes, and low illumination, which may degrade the performance of salient object detection.

Some datasets also contain complex scenes with multiple objects, such as GIT [21] and DES [22] shown in the sixth and seventh columns of Fig. 2. However, the number of images for these datasets is generally small. GIT dataset contains 80 images, and the DES dataset contains 135 images. These datasets are mostly around one hundred images, which cannot meet the training requirements for the deep learning network [23]. All these previous RGB-D datasets above treat salient object detection as a two-class classification task. However, humans shift attention from one object to another in a scene, and different people perceive saliency differently. Therefore, this paper constructs an RGB-D salient object ranking dataset. The constructed dataset contains complex images with multiple objects, shown in the last column of Fig. 2. The background, object scale, lighting condition, and appearance variance of the images in our dataset are more complex than in other datasets.

2.2. RGB-D salient object detection

The most distinctive part of RGB-D salient object detection is the fusion of the RGB maps and the depth maps. There are generally three types of fusion, i.e., early fusion, late fusion, and multi-scale fusion. Early fusion is implemented by directly overlaying the RGB channels and the depth channels or fusing the low-level features computed based on RGB or depth channels [24,25]. Late fusion is carried out by overlaying high-level features from two parallel streams or by fusing two coarse saliency maps from two parallel streams [26–28]. Multi-scale fusion is to first fuse RGB and depth features from different layers and then integrate these features into the decoder network [29,30].

Current RGB-D salient object detection treats this task as a binarized prediction task. This approach is not in line with the visual perception of humans. In real life, visual attention will shift from one object to another. Meanwhile, people have subjectivity and inconsistency about the saliency of various objects in a complex scene. Based on the above psychological observations, we introduce the salient object ranking task to the RGB-D field. We propose an end-to-end learning network that leverages depth information based on depth stack and ground truth stack to perform the RGB-D salient object ranking task.

2.3. Salient object ranking

Salient object ranking is a new task proposed in the RGB field in recent years to rank the objects in a scene based on the saliency levels. Islam first presented this task in 2018 [31] based on the Pascal-S [32] dataset. The annotation information of Pascal-S contains only the number of times each instance has been labeled, which is the number of the 12 labellers who considered the instance to be a salient object. Islam used this annotation information as a basis for saliency ranking. However, there is no information of attention shift in this dataset. In 2020, Siris [33] constructed an RGB salient object ranking dataset based on an image segmentation dataset and an eye fixation prediction dataset. Since this dataset does not contain information about majority voting, the results may be not generalizable and would be biased towards a particular annotator. In contrast, the annotations of our dataset contain not only the number of times one instance labeled as salient object but also the order in which the annotators noticed the object. In this way, the ground truth of our dataset contains information about attention shifts, not just the saliency based on the number of annotations. Such a dataset is more consistent with the human eye's pattern of recognizing salient objects in multi-object scenes. Moreover, both Islam and Siris rank salient objects in the RGB field, while we introduce the salient object ranking task to the RGB-D field to better exploiting the depth information rather than only RGB images.

In the work [31], Islam proposed a deep network that used the ground truth stacks according to different saliency levels to generate salient object ranking maps [31]. We also exploit the ground truth stack in the proposed method. But our method and the method of the work differ a lot. Firstly, instead of redundantly stacking 12 sub-GTs, our ground truth stack can be explicitly divided into four sub-GTs. Secondly, prior work does not take full advantage of the depth information. In contrast, the depth stack module is proposed for synergy between depth stack and ground truth stack in our method. Depth information can provide objects' contour and location information, which is a vital aid in determining the saliency ranking of the objects. Meanwhile, depth sensors are now developing rapidly, and it is becoming easier to obtain depth information.

3D vision has attracted great interest of research community. And we expect RGB-D salient object ranking in this dataset to con-

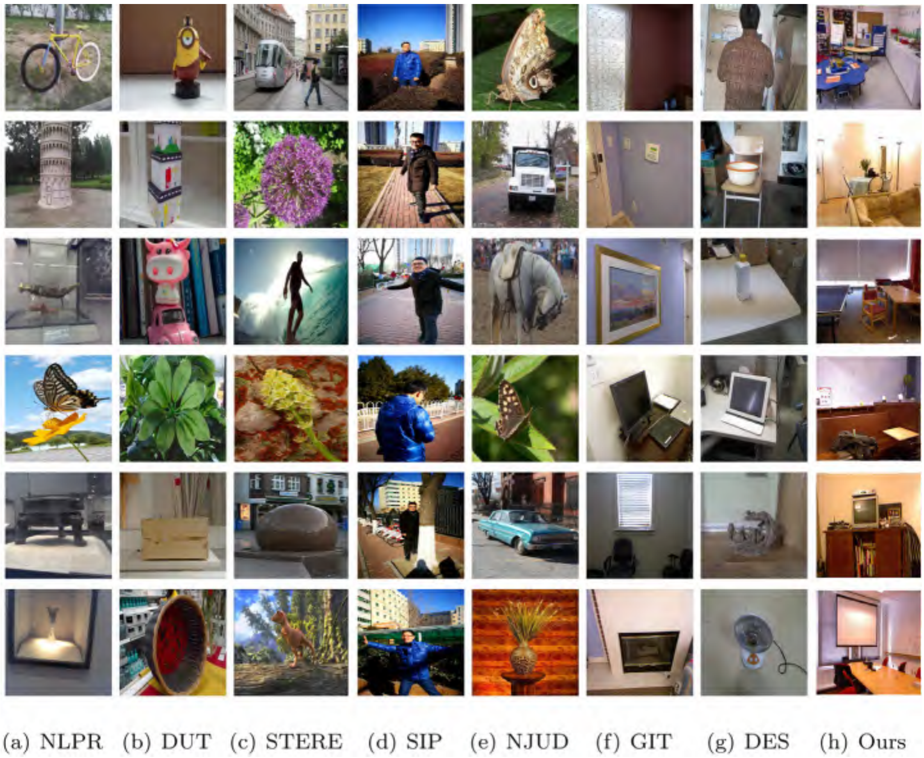


Fig. 2. This is a visual comparison between our RGB-D salient object ranking dataset and the current mainstream RGB-D salient object detection datasets. The background, object scale, lighting condition and appearance variance of the images in our dataset are more complex than other datasets, where most of the objects are single and the background lighting conditions are simple and easy to distinguish.

tribute to other subsequent tasks in 3D vision. Thus, this paper introduces the salient object ranking task into the RGB-D field.

3. RGB-D saliency ranking dataset

3.1. Definition of RGB-D salient object ranking

The goal of the RGB-D salient object ranking task is to identify and rank the salient objects in the scene based on the RGB and depth information. The input of this task is an RGB image and a depth map. The output of this task is a two-dimentional gray map, as is shown below:

$$\mathbf{S} = Model(I_{RGB}, I_{depth}) \tag{1}$$

where S denotes the output prediction map, $Model$ denotes a method, and I_{RGB} and I_{depth} denote the input RGB map and depth map respectively.

The different values of pixels in the prediction map S represent different saliency levels, with larger values representing higher saliency levels. For example, the value of 255 represents rank 1, while the value of 0 represents the background. The rank order of one salient instance in prediction map is obtained by averaging the saliency scores of different pixels within that instance mask [31]:

$$Rank\Big(\mathbf{S}(\delta)\Big) = \frac{\sum_{i=1}^{\rho_\delta} \mathbf{S}_\delta\big(x_i, y_i\big)}{\rho_\delta}, \tag{2}$$

where δ represents a particular instance of the predicted saliency map ($\mathbf{S}$), ρ_δ denotes total number of pixels the instance δ contains, and $\mathbf{S}_\delta(x_i, y_i)$ refers to saliency score for the pixel (x_i, y_i) inside the instance δ.

The proposed task helps to model the visual attention and attention shift of humans in 3D scenes and investigate the association between depth information and salient object ranking. This task does not consider saliency as a binarization task, which is more in line with the visual perception of attention shift when observing objects in complex scenes. Moreover, RGB-D salient object ranking task can be employed as a preprocessing precedure for subsequent tasks in 3D vision, such as image segmentation, object detection and so on. Depth sensors are now developing rapidly, and it is becoming easier to obtain depth information. Thus 3D vision has attracted the great interest of research community. We expect RGB-D salient object ranking in this dataset to contribute to other subsequent tasks in 3D vision.

3.2. Data collection and truth generation

Previous RGB-D salient object detection task is treated as a binary classification problem. And the RGB-D models usually detect salient objects in simple scenes with one prominent object. In this paper, we construct a RGB-D salient object ranking dataset containing the indoor complex scenes with multiple objects. We construct the RGB-D salient object ranking dataset based on the NYU Depth-v2 [13] dataset, which contains 1449 indoor complex RGB images with multiple objects, high-quality depth maps and the ground truths for instance segmentation. The proposed RGB-D salient object ranking dataset is marked as RGBD NYU-rank.

We let 13 annotators label different objects in each image according to the order they noticed the object. In the labeling process, we do not limit the number of objects they can label. So a particular annotator may think that there are many salient objects in a scene when annotating an image, or may only annotate a few salient objects, or even think there are no salient objects in the image.

We first analyze the annotation data to figure out the proper number of the rank for salient objects. An instance is considered as a salient object if it is annotated as saliency more than six times out of 13 annotators according to the majority voting. We then counted the number of salient objects for each image. The num-

Table 1
Count and percentage of images corresponding to different numbers of salient objects in NYU dataset.

Salient object	1	2	3	4	5	6	7+	Total
Images	178	447	626	104	72	18	4	1449
Distribution (%)	0.123	0.308	0.432	0.070	0.049	0.012	0.003	1

Table 2
Comparison of our dataset with existing mainstream RGB-D salient object detection datasets and RGB-D salient object ranking datasets. SOD represents salient object detection and SOR represents salient object ranking. MV represents majority voting information, AS represents attention shift information and DI represents depth information.

Dataset	Size	Object	Types	Task	MV	AS	DI
STERE [16]	1000	Single	Outdoor	SOD	Yes	No	Yes
GIT [21]	80	Multiple	Home environment	SOD	Yes	No	Yes
DES [22]	135	Single	Complex indoor	SOD	Yes	No	Yes
NLPR [17]	1000	Multiple	Simple indoor/Outdoor	SOD	Yes	No	Yes
NJUD [20]	1985	Single	Moive/Internet/Photo	SOD	Yes	No	Yes
DUT-RGBD [19]	1200	Multiple	Simple indoor/Outdoor	SOD	Yes	No	Yes
SIP [20]	929	Multiple	Person in wild	SOD	Yes	No	Yes
Pascal-S [31]	850	Multiple	Indoor/Outdoor	SOR	Yes	No	No
ASR [33]	11,500	Multiple	Indoor/Outdoor	SOR	No	Yes	No
NYU-rank	1449	Multiple	Complex indoor	SOR	Yes	Yes	Yes

ber of images with different numbers of salient objects and corresponding proportions in the dataset are listed in Table.1. From the table, we can see that the images which contain 1, 2 or 3 salient objects cover most of the dataset, and the total proportion reaches 86.3%. According to these characteristics, we set four rank levels, rank 1, rank 2, rank 3 and rank 4+. It should be noted that the number of salient objects in the image is arbitrary. Depending on the majority voting results, there can be one or more salient objects, not a fixed number of four salient objects. In this way, the number of salient objects in the dataset are variable for each image.

First of all, we consider the objects in the scene that are labeled more than 6 times as salient objects according to the principle of majority voting. Then we rank these salient objects in order of saliency. To model the attention shift mechanism of humans, different annotated salient objects are assigned different scores according to the labeling order. The score for the first labeled salient object by each annotator is 1 point, and the score decreases by 10% for the next annotated salient object. That is, the scores for the second and third annotated saliency objects are 0.9 and 0.8 points separately. The first labeled instance gets most of the visual attention during annotation. This is consistent with the attention mechanism in our daily lives, where people tend to catch the most attractive things first, followed by the less attractive ones [33].

To consider the overall saliency rank behavior of different annotators, the annotation scores of the same instances from different annotators are summed up. Then the summed scores of different instances are sorted within each image. The instances considered salient by majority voting are first selected as salient objects, then the scores of these salient objects are used to rank them. In each image, the instance with the highest annotation score is labeled as rank 1, the instance with the second highest annotation score is labeled as rank 2, and so on for four rank levels. In particular, those with scores outside of the top three are labeled as rank 4+.

3.3. Data analysis

We first compare the proposed RGBD NYU-rank dataset with other RGB-D salient object detection datasets, concluded in Table 2. The visual comparisons between different datasets are shown in Fig. 2. According to Fig. 2 and Table 2, The STERE [16], NLPR [17], DUT-RGBD [19] and SIP [20] datasets have about one thousand images, but they contain primarily single object and simple background.

These datasets all contain distinct foreground objects, while the background is relatively clean and easy to distinguish, which is very different from the actual scene. Meanwhile, our dataset contains indoor, complex scenes closer to the actual scenes. In particular, although NLPR [17] and DUT-RGBD [19] contain multiple objects, there are very few of these images that contain multiple objects. In the NLPR dataset, only six of the first hundred images have ground truth containing multiple salient objects, while the remaining 94 have only one salient object. Similarly, three of the first hundred images in the DUT-RGBD contain multiple salient objects, while the remaining 97 images contain only one salient object. In contrast, 88 of the first hundred images in our dataset have multiple salient objects in the ground truth. Our dataset has far more images with multiple salient objects than the other two datasets.

The DES [22] and GIT [21] have complex indoor scenes, but the number of images is small. Specifically, DES dataset contains 135 images and GIT contains 80 images. The proposed RGBD NYU-rank dataset contains 1449 indoor complex scenes with multiple objects. Most importantly, all the above datasets treat saliency detection as the binary classification problem, which does not match the real visual perception of humans. The proposed RGBD NYU-rank dataset treats saliency detection as the salient object ranking problem, which is consistent with the attention shift mechanism of humans.

We also compare the proposed RGBD NYU-rank dataset with existing salient object ranking datasets in Table 2. Both Pascal-S [32] dataset and ASR [33] dataset do not contain depth information maps. In contrast, the proposed dataset provides depth maps. Among them, the Ground Truth of Pascal-S [32] dataset contains the number of times 12 annotators annotate each instance. Therefore, the annotation information can be used to find the salient objects according to the principle of majority voting. The number of times labelled for one instance is also used to determine the saliency level of the instances. But Pascal-S does not contain information about attention shift. The salient object ranking task is primarily designed to model the human visual attention mechanism. The human visual attention mechanism allows humans to process visual information and respond quickly to overwhelming visual input: the human visual attention mechanism allows humans to focus on the most attractive information first, and then shift to focus on the less attractive information. This phenomenon is called attention shift. Adding attention shift information makes salient object ranking task more consistent with the visual attention mechanism of the human eye. Without attention shift information, we can only assign the same score to different objects in the same scene. This does not correspond to the actual phenomenon of the human eye when observing multiple objects simultaneously.

The ASR dataset uses eye fixation annotation to score all instances in the image. This dataset considers the attention shifts, but this information is only collected via one annotator. Since the dataset does not contain information about majority voting, the generated ground truth is not generalizable and may be biased towards a particular annotator. In contrast, our dataset can obtain salient objects and saliency ranking levels based on the order of attention shifts of multiple annotators during labelling.

We further conduct the data analysis to disclose the characteristics of the proposed dataset by comparing the proposed dataset with four other mainstream RGB-D salient object detection datasets. We first calculate the number of salient categories in the dataset. There are 895 categories in the original NYU Depth-v2 [13] dataset. 208 categories appear in our ground truth, which means 208 out of 895 categories are salient.

We also compute the number of times each specific category is labeled as the salient object in our dataset, seen in Fig. 3(a). The horizontal coordinate indicates the different categories and the vertical coordinate indicates the number of instances of the category. The most common category is picture, which appears 493 times. The second common category is chair, which appears 353 times. And the third common category is the pillow, which appears 181 times, as seen in Fig. 3(b). The horizontal coordinate indicates the different categories and the vertical coordinate indicates the number of instances of the category. It can be seen that the dataset contains indoor objects with a large number of semantic categories, which indicates that it is an indoor scene dataset.

The sizes of different salient objects are analyzed in Fig. 3(c). The horizontal coordinate indicates the scale of the salient object in the whole image, calculated as the pixel number of the salient object divided by the pixel number of the entire image. The vertical coordinates indicate the number of instances with different scales. To facilitate comparison between different datasets, we normalize the number of instances with each scale to [0, 100]. From the data, we can see that the peaks of the curves for other datasets are around 0.2 to 0.4, while the curve peak of the proposed dataset is around 0.1. The scales of salient objects in other datasets are much larger than ours, and most of the salient objects in our dataset only occupy a proportion of ten percent of the image. The above observation indicates that the proposed dataset is more complex.

We also analyze the depth information of the dataset, shown in Fig. 3(d). The horizontal coordinate represents the pixel's depth value, and the vertical coordinate represents the proportion of the pixels in the entire dataset. We count the number of pixels for all the regions labeled as salient objects for each depth value. It can be seen that the most salient points concentrate in locations with smaller depth values. This indicates a correlation between the value of the depth and the saliency of objects. Besides the proposed dataset, a large amount of data from the NLPR [17] and the SIP [20] datasets also have a lot of salient points with smaller depth values, which also demonstrate the correlation between saliency and depth.

The light and saturation situation contrast are also calculated, as shown in Fig. 3(e) and (f). The horizontal coordinate represents the pixel's light or saturation value, and the vertical coordinate represents the proportion of the pixels in the entire dataset. Our dataset has a smoother result for both light and saturation situation contrast, which indicates a wider and more diverse distribution of light conditions and color information in the dataset. This shows that our dataset has a more complex lighting situation and diverse color information.

4. Proposed network architecture

According to the observation that there is usually a strong correlation between salient objects and depth information, in this paper, we propose an end-to-end network by exploiting the depth stack and the ground truth stack to solve the problem of salient object ranking in RGB-D complex indoor scenes.

We propose a Depth Stack Module (DSM) and a Saliency Map Re-fusing Module (SMRM) to fully exploit the information of the depth stack and the ground truth stack. The whole network is described in Fig. 4. A backbone network is first used to extract initial features.

Next, the DSM module is proposed to make use of each depth interval which provides depth information at specific locations and produce corresponding coarse saliency prediction map.

Finally, we propose an SMRM module to integrate different coarse saliency prediction maps obtained based on different DSM modules. SMRM module utilizes the information of different rank levels and helps improve the network's effectiveness in determining the saliency ranking level. This module parses the information of different rank levels through multiple convolution operations to obtain the final prediction map.

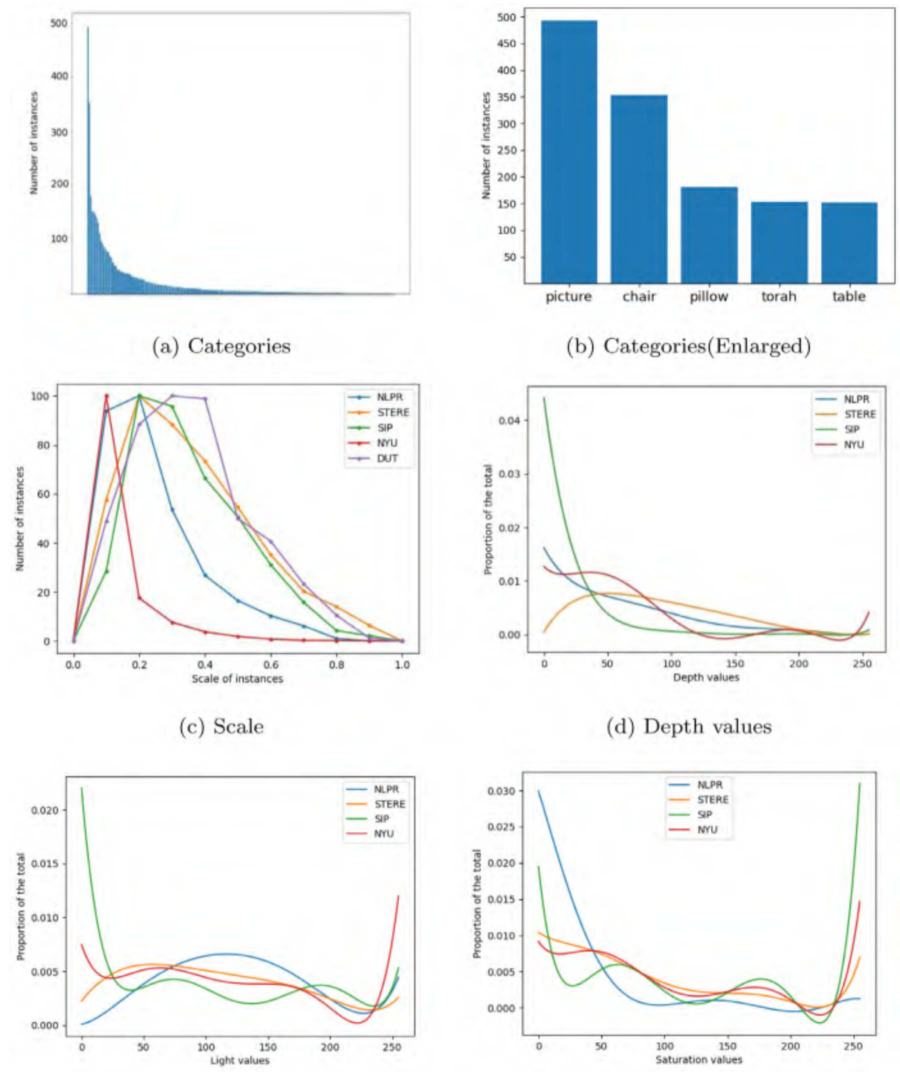
4.1. Generation of depth stack and ground truth stack

As shown in Fig. 5. A depth stack including depth interval 1, depth interval 2, depth interval 3 and depth interval 4+ is generated based on the depth map. Similarly, based on the ground truth, a ground truth stack is generated, which includes a sub-ground truth map containing the most salient object (rank 1), a sub-ground truth map containing the two most salient objects (rank 1 and rank 2), a sub ground truth map containing the first three salient objects (rank 1, rank 2 and rank 3) and a sub ground truth map containing all salient objects (rank 1, rank 2, rank 3 and rank 4+).

Different depth intervals are used in the DSM module separately to extract the information of objects at different depths. The DSM module distinguishes image features in different depth intervals and facilitates the effective use of location information in the saliency ranking process. By comparing the coarse saliency prediction map with the corresponding ground truth interval, the proposed network with the depth stack module and ground truth stack is able to generate saliency prediction maps of different saliency rank levels: the coarse saliency prediction map with the most salient objects, the coarse saliency prediction map with the two most salient objects, the coarse saliency prediction map with the three most salient objects and the coarse saliency prediction map with all salient objects. These modules facilitate the integration of depth and RGB information and extract the location information from the depth map to rank salient objects. Note that in this process, if there are less than 4 salient objects in the scene, the sub-GTs of the higher rank level in the ground truth stack remain the same as those of the lower rank level. The depth stack works the same way.

4.2. Backbone network

BBSNet [34] is used as the backbone network to fuse the RGB map and the depth map to provide the initial image feature. Three sets of low-level features and three sets of high-level features are generated. It incorporates each stage of the feature extraction stream of the depth map into the RGB feature extraction stream of



(a) Categories

(b) Categories(Enlarged)

(c) Scale

(d) Depth values

(e) Luminance

(f) saturation

Fig. 3. (a) The distribution of categories in the dataset, (b) The distribution of categories with a relatively large number, (c) Comparison of the scale of salient objects in images, (d) Comparison of the depth values of salient objects, (e) Comparison of the luminance, (f) Comparison of the saturation.

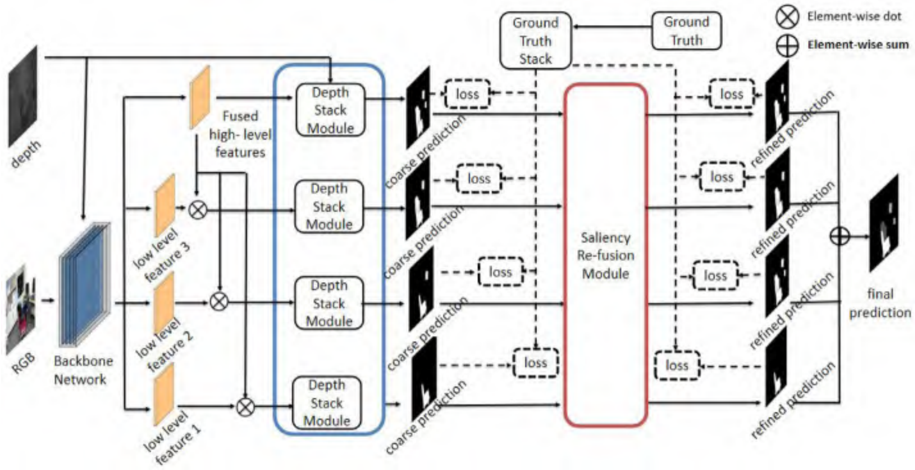
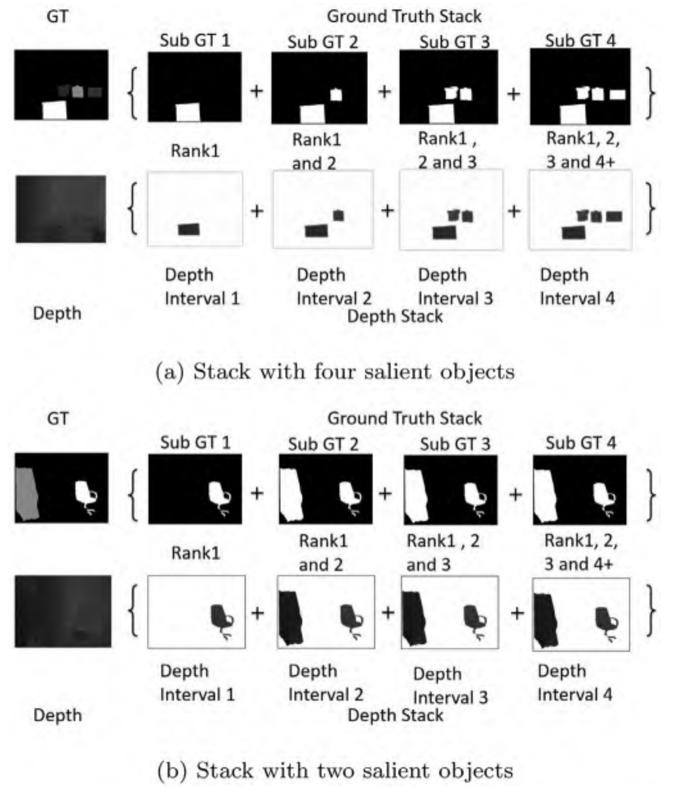


Fig. 4. Architecture overview.



(a) Stack with four salient objects

(b) Stack with two salient objects

Fig. 5. Depth stack and GT stack. (a) Depth stack and GT stack with four salient objects, (b) Depth stack and GT stack with two salient objects.

the corresponding stage by depth enhancement operations. Therefore, different stages of RGB and depth fused image features are generated in that network, including three sets of high-level features and three sets of low-level features. The generated three sets of high-level features are fused to generate the final high-level feature. The fused final high-level feature was used to generate coarse prediction map for all salient objects. Furthermore, the fused final high-level feature is used as an attention feature to guide the low-level features. The new attention features are combined with the low-level features at each input. We use upper low-level features as input features when generating a higher rank level coarse prediction map. Lower low-level image features are input when generating a lower rank level coarse prediction map. These operations allow for better utilization of image features at each stage and more efficient generation of coarse saliency prediction maps at different rank levels.

4.3. Depth stack module

As shown in Fig. 4, the DSM module appears four times in the network framework. Each depth interval is inputted into a DSM module. Four coarse prediction maps are generated for four rank levels based on the backbone features and four different depth intervals. The detailed description of the DSM module is presented in Fig. 6. The yellow module represents the convolution operation, and the blue module represents the upsampling operation.

First, we convert the depth map into four different depth intervals using the method from Sun's study [35]. According to the different depth values of different pixel points in the depth map, we can get a histogram with a depth value of [0, 255] in the horizontal coordinate and the number of occurrences of that pixel value in the vertical coordinate. The depth values of individual pixel points in an instance tend to be relatively close to each other, so multiple peaks appear in the histogram. The vicinity of each peak may be the depth interval of pixel points of one or more instances. We take depth values of the vicinity of each peak to form an interval, then take the location of the pixels in this interval to form a sub-map. These sub-maps are often composed of specific instances, and the depth values of the instances are closely related to the saliency rank of the instances. We build sub-maps in lower branches using depth intervals with smaller pixel values. Therefore, the instances in the depth interval with smaller depth values tend to have a higher saliency rank, which facilitates the determination of saliency levels.

Then we combine the initial image features $\{f_i^{bbs}; i = 1, 2, \ldots, 96\}$ obtained from the backbone network and the binarized depth interval f^{depth} with dot product:

$$f_i^{fuse} = f_i^{bbs} \odot f^{depth}, \tag{3}$$

So we get new 96-dimensional image features $\{f_i^{fuse}; i = 1, 2, \ldots, 96\}$ which are highlighted by the depth interval. Then the newly generated features and the original features are combined by concatenation. Thus, both the depth highlighted features and the original features are taken into account:

$$f^{con} = \mathbf{O}_c(f^{fuse}, f^{depth}), \tag{4}$$

where $\mathbf{O}_c$ denotes the concatenation operation. The concatenated features have 192 channels. Then two 2 × 2 convolution operations are adopted to change the channels of the features to 64. After each convolutional layer, there are a Relu layer and a batch normalization layer:

$$f^{concise} = Conv\big(Conv(f^{con})\big), \tag{5}$$

Finally, we used two upsampling operations and a 1 × 1 convolution operation to obtain a coarse prediction maps:

$$CS = Conv\Big(\mathbf{F}_{UP}\big(\mathbf{F}_{UP}(f_i^{concise})\big)\Big), \tag{6}$$

where $\mathbf{F}_{UP}$ denotes the up-sampling operation and $\{CS\}$ is the coarse prediction map. After four depth stack modules, four coarse saliency prediction maps are generated, which can be denoted as CS1, CS2, CS3 and CS4 separately.

In this way, the features of objects with different depths are utilized individually based on depth intervals for better salient object ranking.

4.4. Saliency map re-fusion module

We propose a Saliency Map Re-fusion Module (SMRM) to integrate different coarse saliency prediction maps generated from different DSM modules. This module parses the information of different coarse saliency prediction maps for different rank levels through multiple convolution operations to obtain the final saliency prediction map.

As shown in Fig. 7, we take the coarse saliency prediction maps for different rank levels as four input feature channels $\{CS_i; i = 1, 2, 3, 4\}$ and perform an concatenate operation, where $\mathbf{C}$ denotes concatenate operation:

$$CS = \mathbf{C}(CS_1, CS_2, CS_3, CS_4) \tag{7}$$

These features are further fused through a 2 × 2 convolution operation and two 3 × 3 convolution operations:

$$F = \mathbf{T}(CS), \tag{8}$$

where $\mathbf{T}$ represents a series of convolution operations: a 2 × 2 convolution and two 3 × 3 convolutions. After each convolutional layer, there are a Relu layer and a batch normalization layer.

Then, we use a 1 × 1 convolution to generate feature map of four channels. Each channel represents a refined saliency prediction map of different rank level. This stage generates four different

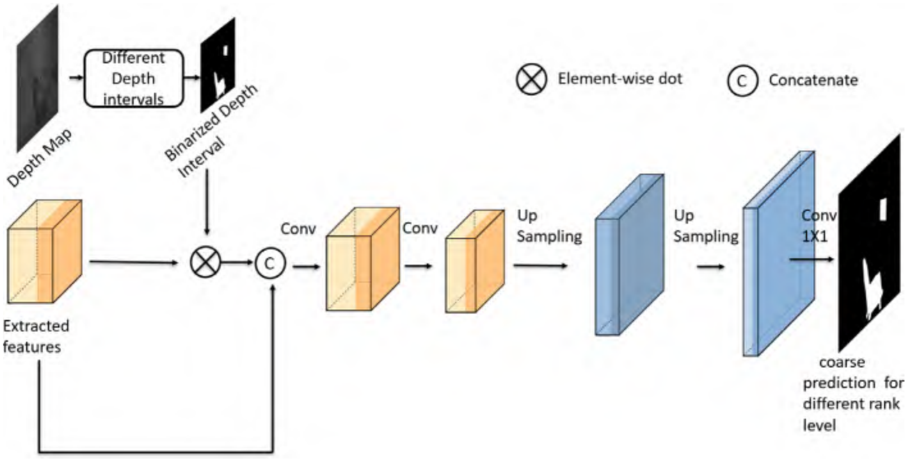


Fig. 6. Graphical representation of depth stack module.

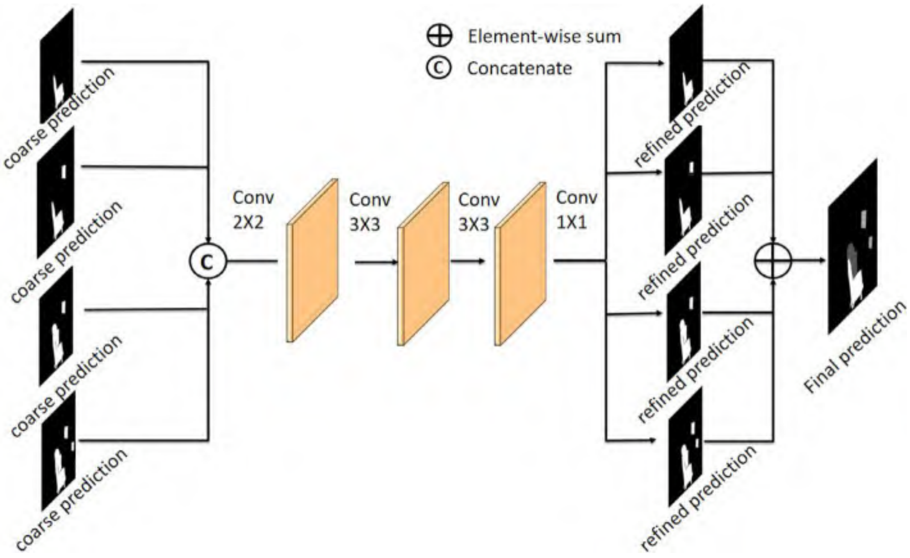


Fig. 7. Graphical representation of saliency map re-fusion module.

refined saliency maps, $\{RS_i; i = 1, 2, 3, 4\}$.

$$RS_i = Conv(F), \tag{9}$$

Finally, we add up these four refined saliency maps to get the final prediction map:

$$S = RS_1 + RS_2 + RS_3 + RS_4 \tag{10}$$

where S denotes the final saliency prediction map.

Through summation and multiple convolutions, SMRM module fuses the information of different coarse saliency maps for different rank levels, which helps to improve the performance of salient object ranking.

4.5. Loss

Inspired by the previous paper of Islam [31], besides the loss between the final saliency prediction map and the ground truth, we also exploit the synergy between the depth stack and the ground truth stack to accomplish the task more efficiently.

It can be seen that the depth map can be divided into multiple regions based on depth information, and these regions have a firm consistency with the ground truth stack of saliency ranking, which is exploited to construct the SMRM module.

According to the ground truth map, a ground truth stack containing four sub-ground truths is generated. We take the instance of rank 1 (the most salient instance) to form a binarized sub-ground truth. Similarly, we take instances of rank 1 and rank 2 (the two most salient instances) to form the second binarized sub-ground truth. And three instances of rank 1, rank 2 and rank 3 are combined to form the third binarized sub-ground truth. And the forth binarized sub-ground truth includes all salient objects.

By calculating the loss between the coarse saliency prediction and the corresponding sub ground truth, four different coarse saliency prediction maps with different saliency rank levels can be generated: a coarse saliency prediction map with the most salient objects, a coarse saliency prediction map with the two most salient objects, a coarse saliency prediction map with the first three salient objects and a coarse saliency prediction map with all salient objects. The proposed coarse saliency loss $\mathcal{L}_{\mathcal{C}}$ is as follows:

$$\mathcal{L}_{\mathcal{C}} = \alpha(\ell_{ce}(CS_1, G_1) + \ell_{ce}(CS_2, G_2) + \ell_{ce}(CS_3, G_3) + \ell_{ce}(CS_4, G_4)), \tag{11}$$

In the above equation, CS1, CS2, CS3 and CS4 represent the coarse prediction maps of different rank levels after four DSM modules, respectively. And G1, G2, G3 and G4 represent the binarized sub ground truths of different rank levels. ℓ_{ce} represents the widely used binary cross entropy loss. And we also calculate the loss $\mathcal{L}_{\mathcal{R}}$ between generated refined prediction map corresponding to each coarse prediciton map and sub ground truth.

$$\mathcal{L}_{\mathcal{R}} = \alpha(\ell_{ce}(RS_1, G_1) + \ell_{ce}(RS_2, G_2) + \ell_{ce}(RS_3, G_3) + \ell_{ce}(RS_4, G_4)), \tag{12}$$

RS1, RS2, RS3 and RS4 represent the refined prediction maps of different rank levels in SMRM modules, respectively. And G1, G2, G3 and G4 represent the binarized sub ground truths of different rank levels. ℓ_{ce} represents the widly used binary cross entropy loss.

Our model is divided into four rank levels, extracting four different image features. Using the four image features of different rank levels generated before, we calculate the loss by corresponding these features to the ground truth of different rank levels in the saliency ranking. The previous image features are upsampled separately to generate four binarized saliency maps compared with ground truth at different rank levels. As a result, we obtain the saliency prediction maps for different rank levels.

The final saliency loss is the weighted combination of the coarse saliency loss and the refined saliency loss.

$$\mathcal{L} = \alpha(\mathcal{L}_{\mathcal{C}}) + \beta(\mathcal{L}_{\mathcal{R}}), \tag{13}$$

In the above equation, α and β are the weight parameters. In this paper, we will simply set α as 1/3 and set β to be 1. Since this work is for Salient object Ranking for Complex and Indoor scenes, the proposed network is abbreviated as SRCINet for simplicity.

5. Experiments

5.1. Implementation details

Necessary image augmentations are exploited, such as random rotation, random crop and random flip to avoid potential overfitting. The input image is resized to have a resolution of 640 × 480. Our model is implemented by the Pytorch framework and trained on TITAN RTX GPU. We set the mini-batch size to be 10. The total number of epochs for each training is 200. The Adam optimizer is employed with the learning rate of 10^{-4}.

5.2. Datasets

The proposed database, i.e. NYU rank, is used to compare the performances of different models. The proposed RGBD NYU-rank dataset contains RGB images, depth images and truth images for salient object ranking. We randomly divide this dataset into a training set of 1160 images and a test set of 289 images.

5.3. Evaluation metrics

Two metrics are employed to measure RGB-D Saliency ranking performance, including Salient Object Ranking (SOR) and mean absolute error (MAE). SOR metric is used to assess the saliency ranking accuracy of different prediction maps. It is expressed as Spearman's Rank-Order correlation between the predicted rank order of the salient objects and the ground truth. Correlation coefficients are during the interval $[-1, 1]$ from the absolutely wrong prediction to the perfectly positive correlation [31]. If the predicted ranking series is the same as the actual ranking series, it is strongly correlated, and the Spearman coefficient is 1. If the predicted ranking series is the opposite of the actual ranking series, the coefficient is -1. In this paper, all correlation coefficients are normalized to be in the range of [0,1] for a better linear formulation.

The MAE is used to measure the pixel-level difference between the predicted map and the ground truth by averaging the absolute value of the difference over all pixels. This metric is used to evaluate the accuracy of the generated saliency maps. This metric is helpful for both salient object ranking and salient object detection [33].

Table 3
Quantitative analysis of different models. The backbone network without the proposed DSM module and the proposed SMRM module is directly inputted into a simple convolution operation to obtain the results as baseline. ↓ (↑) means the higher(lower) the better.

Method	MAE↓	SOR↑
DMRA [36]	0.191	0.627
D3Net [37]	0.114	0.716
SP-Net [10]	0.110	0.715
Baseline (A)	0.161	0.679
Baseline + DSM (B)	0.122	0.689
Baseline + SMRM(C)	0.112	0.719
SRCINet (D)	**0.108**	**0.732**

5.4. Quantitative analysis

Previous saliency ranking studies have been conducted on RGB images, and our study is the first RGB-D saliency ranking work. Besides the proposed SRCINet, we propose another three methods by adding none or part modules to the backbone network for comparison. The first method is to obtain the prediction map by simple convolution operations based on the image features extracted by the backbone. This method is used as our baseline. The second method is to add the DSM module to the baseline. In this method, we add up the four coarse prediction maps and divide them by four as the final prediction map. The third method is to add the SMRM module to the baseline. The fourth method is the proposed SRCInet, where both the DSM module and SMRM module are added to the baseline.

All the experiments are trained for 200 epochs. The epoch with the best MAE is taken as the final experimental result. Table 3 shows the quantitative comparison of different models. It can be seen from the MAE metrics that the addition of both the DSM module and the SMRM module improves the prediction performance. The best MAE result is obtained when both the DSM module and the SMRM module are added simultaneously. The experimental results show that the proposed SRCINet has the best performance with the lowest MAE. As seen from the SOR metrics, the addition of only the DSM module results in a slight decrease in the SOR metrics. When only the SMRM module is added, the SOR metric slightly increases. Moreover, when both modules are added at the same time, the SOR metric gets a considerable improvement. This indicates that the two modules we designed can enhance the prediction of salient object ranking when interacting with each other. From the modeling perspective, the DSM module improves the MAE metric but reduces the SOR metrics when used alone. At the same time, the SRMR module alone improves the MAE metric and slightly improves the SOR metric. The MAE and SOR metrics are greatly improved when these two modules are used together. This reflects the effectiveness of the proposed model and validates our idea that depth stack and ground truth stack can help to improve the RGB-D saliency ranking performance.

We have compared the proposed method with three more RGB-D salient object detection methods: SPNnet, D3Net and DMRA. As it can be seen in Table 3, the performance of the proposed model SRCINet are the best comparing with other three RGB-D salient object detection models based on both MAE metric and SOR metric.

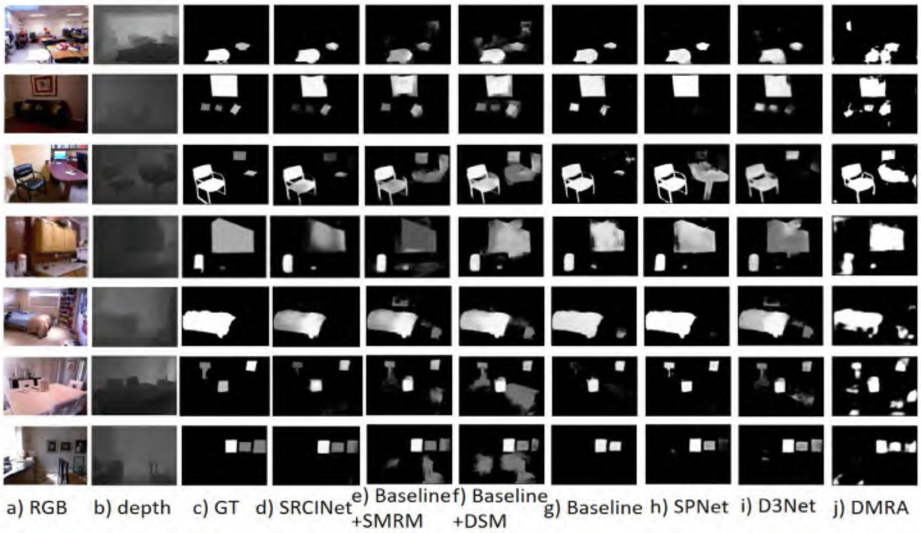


Fig. 8. Visual comparison of different models, (a) RGB images, (b) depth images, (c) ground truth, (d) The proposed SRCINet, (e) Baseline + SMRM, (f) Baseline + DSM, (g) Baseline, (h) SPNet [10], (i) D3Net [37], (j) DMRA [36].

5.5. Qualitative analysis

Fig. 8 shows the visual comparison of the results based on different models. The first column of the figure is the RGB image, the second column is the depth map, and the third column is the ground truth. The fourth column shows the prediction results of the complete SRCINet model. The fifth column shows the prediction results of the baseline added SMRM module, the sixth column shows the prediction results of the baseline added DSM module, and the seventh column shows the effect of the baseline.

It can be seen that when only the baseline model is used, the saliency ranking of the prediction results is not satisfactory. The low rank level objects are often ignored. For example, in the seventh column of the third row, only the contour of the object instance of rank one can be predicted, but not the objects of rank 2 and rank 3. Also, the network is not able to correctly discern the saliency level of the object. When only the DSM module is added, the segmentation of the objects with low saliency rank is improved, but some cluttered lines appear in the images. Its saliency ranking prediction is also unsatisfactory. For example, in the sixth column of the second row, all rank-level objects can be predicted, but the contour of the objects is not precise enough.

The segmentation of objects is improved by adding only the SRMR module. And the saliency ranking of the model has been improved. However, some background regions are incorrectly predicted as salient regions. For example, in the fifth column of the third row, all three rank-level objects can be predicted, and the predicted saliency levels are basically correct. However, a large number of background pixel points in the figure are predicted to be salient.

After adding both the DSM module and SRMR module, the model segmentation is significantly improved. And the saliency ranking results have improved considerably. For example, in the fourth column of the third row, the contours of all three rank-level objects can be predicted more accurately, and the three predicted objects are correctly ranked in terms of saliency.

We also visually compare the proposed method with three RGB-D salient object detection models. It can be seen that SPNet can predict the saliency level relatively accurately, but there is a problem of predicting some pixels in the background as salient. D3Net can predict the contour information for accuracy, but failed to predict the saliency level correctly in some cases. For example, in the ninth column of the second row, the saliency level prediction deviates widely. DMRA can detect salient object, but does not

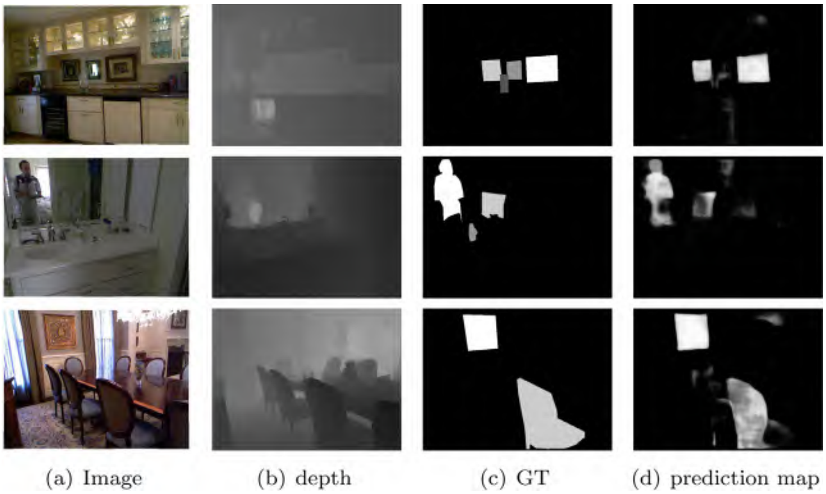


Fig. 9. Failure cases. (a) Image, (b) corresponding depth map, (c) ground truth(GT), (d) prediction map.

predict the saliency level. The contours of the objects were not correctly predicted, while the saliency levels of different objects could not be discerned in the prediction maps.

5.6. Failure cases

There are still some failure cases in the results of our experiments as shown in Fig. 9. When small objects appear in a scene with multiple salient objects, the model sometimes does not predict the contours of small objects well. For example, in the fourth column of the first row and second row, objects with small objects are not correctly predicted. Further improvements can be made in the optimization of small salient object detection. Another kind of failure case is that the prediction maps sometimes contain noise in the background. For example, in the fourth column of the third row, there is some noise in the background of the prediction map. This shortcoming can be solved by introducing instance segmentation. Subsequent works can introduce instance segmentation into RGB-D salient object ranking to further improve the prediction results.

6. Conclusion

The current RGB-D salient object detection is treated as a binarized segmentation task. Besides, indoor and complex scenes with multiple objects are usually uncommon in the current RGB-D saliency detection dataset. This paper introduces the salient object ranking task into the RGB-D field. Since the lack of such a dataset, we reconstruct an RGBD NYU-rank dataset for salient object ranking tasks. The dataset contains indoor and complex scenes. We also propose a novel end-to-end neural network for salient object ranking using the synergistic features of depth stacks and ground truth stacks. It exploits the location and contour information in the depth map to compensate for the missing information in RGB images and perform the saliency ranking task more effectively. The experiments demonstrate the effectiveness of our proposed neural network. It is proved that the proposed DSM module and SMRM module can help to improve the effectiveness of the salient object ranking performance.

3D vision has attracted the great interest of the research community, and we expect that the proposed RGB-D salient object ranking task can contribute to other subsequent tasks of 3D vision. In the future, further attention can be paid to improving salient object ranking results for small and low rank-level objects in the dataset. Also, an attempt can be made to introduce instance segmentation into salient object ranking tasks.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

This work was supported by the National Key R&D Program of China (2018YFB1500800), Guangdong Basic and Applied Basic Research Foundation (2022A1515011435), ZhiShan Scholar Program of Southeast University and the Fundamental Research Funds for the Central Universities, Natural Science Foundation of Jiangsu Province (Grant BK20201160), Science and Technology Project of State Grid Corporation of China (Intelligent operation and maintenance technology of distributed photovoltaic system SGTJDK00DYJS2000148).We thank the Big Data Computing Center of Southeast University for providing the facility support on the numerical calculations in this paper.

References

[1] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, STC: a simple to complex framework for weakly-supervised semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2016) 2314–2320.
[2] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: a new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4548–4557.
[3] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, arXiv preprint arXiv:1805.10421(2018).
[4] V. Mahadevan, N. Vasconcelos, Saliency-based discriminant tracking, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2009, pp. 1007–1013.
[5] B. Lai, X. Gong, Saliency guided dictionary learning for weakly-supervised image parsing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3630–3639.
[6] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, arXiv (2015-02-10) https://arxiv.org/abs/1502.03044 v3
[7] R. Zhao, W. Ouyang, X. Wang, Person re-identification by salience matching, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2528–2535.
[8] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3586–3593.
[9] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4019–4028.
[10] T. Zhou, H. Fu, G. Chen, Y. Zhou, D.-P. Fan, L. Shao, Specificity-preserving RGB-D saliency detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4681–4691.
[11] W. Ji, J. Li, S. Yu, M. Zhang, Y. Piao, S. Yao, Q. Bi, K. Ma, Y. Zheng, H. Lu, et al., Calibrated RGB-D salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 9471–9481.
[12] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, N. Barnes, Uncertainty inspired RGB-D saliency detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (9) (2021) 5761–5779.
[13] N. Silberman, D. Hoiem, P. Kohli, R. Fergus, Indoor segmentation and support inference from RGBD images, in: European Conference on Computer Vision, Springer, 2012, pp. 746–760.
[14] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, A. Borji, Salient objects in clutter: bringing salient object detection to the foreground, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 186–202.
[15] H. Chen, Y. Li, Y. Deng, G. Lin, CNN-based RGB-D salient object detection: learn, select, and fuse, Int. J. Comput. Vis. 129 (7) (2021) 2076–2096.
[16] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 454–461.
[17] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[18] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1115–1119.
[19] Y. Piao, X. Li, M. Zhang, J. Yu, H. Lu, Saliency detection via depth-induced cellular automata on light field, IEEE Trans. Image Process. 29 (2019) 1879–1889.
[20] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2075–2089.
[21] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[22] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: Proceedings of International Conference on Internet Multimedia Computing and Service, 2014, pp. 23–27.
[23] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: a survey, Comput. Vis. Media 7 (1) (2021) 37–69.
[24] J. Ren, X. Gong, L. Yu, W. Zhou, M. Ying Yang, Exploiting global priors for RGB-D saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 25–32.
[25] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[26] K. Desingh, K.M. Krishna, D. Rajan, C. Jawahar, Depth really matters: improving visual salient region detection with depth, in: BMVC, 2013, pp. 1–11.
[27] J. Han, H. Chen, N. Liu, C. Yan, X. Li, CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybern. 48 (11) (2017) 3171–3183.

[28] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.

[29] N. Huang, Y. Luo, Q. Zhang, J. Han, Discriminative unimodal feature selection and fusion for RGB-D salient object detection, Pattern Recognit. 122 (2022) 108359.

[30] G. Feng, J. Meng, L. Zhang, H. Lu, Encoder deep interleaved network with multi-scale aggregation for RGB-D salient object detection, Pattern Recognit. 128 (2022) 108666.

[31] M.A. Islam, M. Kalash, N.D. Bruce, Revisiting salient object detection: simultaneous detection, ranking, and subitizing of multiple salient objects, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7142–7150.

[32] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.

[33] A. Siris, J. Jiao, G.K. Tam, X. Xie, R.W. Lau, Inferring attention shift ranks of objects for image saliency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12133–12143.

[34] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, L. Shao, BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network, in: European Conference on Computer Vision, Springer, 2020, pp. 275–292.

[35] P. Sun, W. Zhang, H. Wang, S. Li, X. Li, Deep RGB-D saliency detection with depth-sensitive attention and automatic multi-modal fusion, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 1407–1417.

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

[37] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2020) 2075–2089.

Jingzheng Deng received the B.S. degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2020. He is currently pursuing the Master's degree in Artificial Intelligence and Pattern Recognition in Southeast University. His research interests include saliency detection and computer vision.

Jinxia Zhang received the B.S. degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2009 and the PhD degree in the Department of Computer Science and Engineering, Nanjing University of Science and Technology, China in 2015. She was a visiting scholar in Visual Attention Lab at Brigham and Women's Hospital and Harvard Medical School from 2012 to 2014. She is currently an associate professor and PhD supervisor in the School of Automation, Southeast University. Her research interests include saliency detection, knowledge transfer, computer vision and machine learning.

Zewen Hu received the B.E. degree in automation from Donghua University. He is currently pursuing the Ph.D. degree with the Key Laboratory of Measurement and Control of Complex Engineering Systems, Ministry of Education, Southeast University. His research interests include robot automatic path planning and constant force control.

Liantao Wang received the B.S. degree in mechanical engineering and the Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, China, in 2004 and 2015, respectively. From 2012 to 2014, he was a Visiting Scholar with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently an Assistant Professor with the College of Internet of Things Engineering, Hohai University, Changzhou, China. His current research interest includes weakly supervised learning methods for object classification and localization.

Jiacheng Jiang received the B.S. degree in Automation from Jiangnan University, in 2020, and is currently working toward the M.S. degree in Electronic and Information Engineering at Southeast University. His current research interests include image classification and image segmentation. His recent work has focused on image classification and defect inspection.

Zhu Xinchao received the B.S. degree in automation from Southeast University, Nanjing, in China, in 2021. He is currently pursuing the M.S. degree in Artificial Intelligence and Pattern Recognition in Southeast University. His research interests include salient object ranking and instance segmentation.

Xinyi Chen received B.E. degree in automation from Chongqing University, Chongqing, China, in 2021. She is currently working towards the M.E. degree in Control Science and Engineering from Southeast University, Nanjing. Her current research interests include defect detection and deep learning.

Yin Yuan received her bachelor's degree in automation from Sichuan University, Chengdu, in China, in 2021. She is currently pursuing the Master's degree in Artificial Intelligence and Pattern Recognition in Southeast University. Her research interests include defect detection and learning with noisy labels.