Chapter 9
RGB-D Salient Object Detection: A Review

Tongwei Ren, Ao Zhang

Abstract Salient object detection focuses on extracting attractive objects from the scene, which serves as a foundation of various vision tasks. Benefiting from the progresses in acquisition devices, depth cue is convenient to obtain, and is used in salient object detection on RGB-D images by combining with color cue. In this chapter, we comprehensively review the advances in RGB-D salient object detection. We firstly introduce the task and key concepts in RGB-D salient object detection. Then, we briefly review the evolution of salient object detection technology, especially those on RGB images, since many RGB-D salient object detection methods derive from the existing RGB ones. Next, we present the typical RGB-D salient object detection methods, evaluate their performance on public datasets, and summarize their issues. Finally, we discuss some open problems and suggestions for future research.

1.1 Introduction

When introducing salient object detection from a cognitive perspective, it refers to finding objects, which can attract more attention than the surrounding regions when human visual system perceives the scene. The task of salient object detection in computer vision is inspired by early tasks which try to simulate human attention [17, 13], a concept has been studied in cog-nitive psychology for many years [27]. Because of the complexity of human visual system, the criterion of judging whether an object is salient cannot be explicitly listed with a couple of simple standards. There are a bunch of factors that can influence the judgement of salient objects, for example, salient objects are context dependent. The change of scene or even the change of location in the same scene may cause a difference in the saliency rank of objects. Both local contrast and global contrast with other objects in the same context should be taken into consideration.

When introducing the salient object from a precise and computational perspective, it refers to segmenting the entire objects, which are the most attention-grabbing compared to surrounding regions, rather than only parts of the objects [2]. Referring to some popular salient object detection dataset construction [33, 4, 1, 5], the concrete way of judging whether an object is salient, is to let a couple of annotators to choose the most attention-grabbing object in the scene. Fig. 1.1 shows an example of salient object detection.

Saliency analysis technology mainly includes fixation prediction and salient object detection. Different from fixation prediction, salient object detection aims to extract the entire attractive objects rather than presenting the gaze points by highlighting a few spots on heat maps, which is more useful to serve as a foundation of various vision tasks, such as object detection, information retrieval and video analysis.

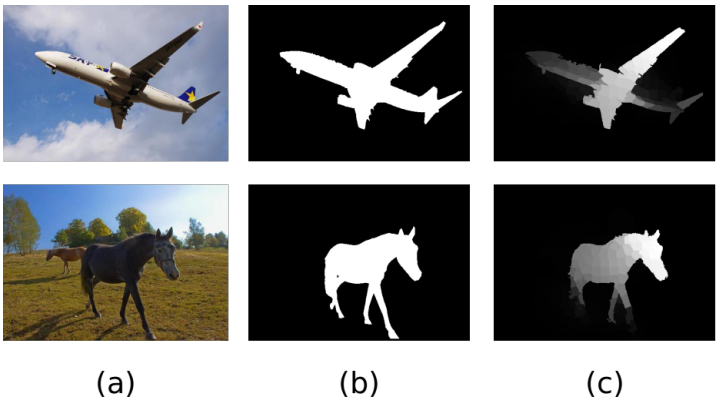

Fig. 1.1 Examples of salient object detection. (a) Original images. (b) Groundtruths of salient object detection. (c) Saliency maps. The saliency maps are generated by [11].

In recent years, benefiting from the progresses of acquisition devices, depth cue can be conveniently obtained by depth cameras and binocular cameras, and its potentiality in salient object detection is explored. In reality, human visual system perceives both color and depth information from the scene, and uses them in distinguishing salient objects together. Depth cue helps to distinguish salient objects from the background, especially when the objects have complex structure or texture. Fig. 1.2 shows comparison between saliency maps using color cue and saliency maps using both color cue and depth cue. Thus, it is useful to combine depth cue with color cue in salient object detection on RGB-D images.

However, due to the performance limitation of current acquisition devices, the depth maps are usually of low quality, low resolution and accuracy in particular, which brings serious noises and even misleads into salient object detection. How to handle the low quality of depth maps in salient object

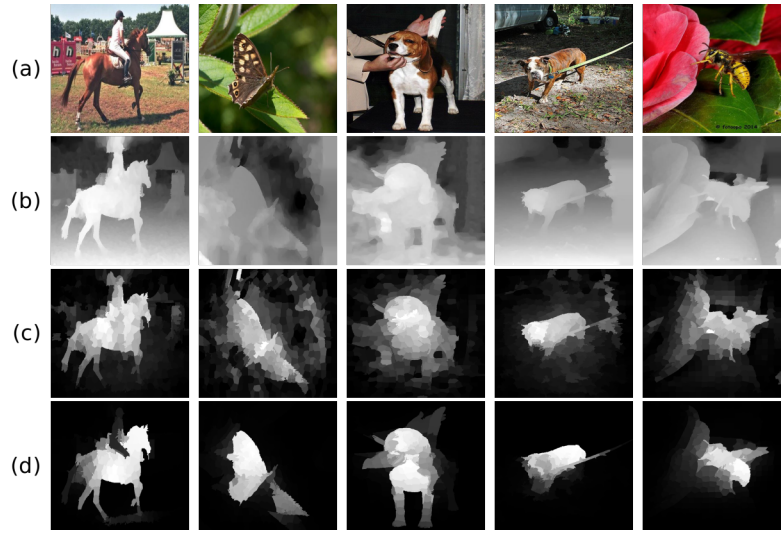

Fig. 1.2 Examples of comparison between saliency maps using only color cue and saliency maps using both color cue and depth cue. (a) Original images. (b) Depth maps. (c) Saliency maps using only color cue. (d) Saliency maps using both color cue and depth cue. The saliency maps are generated by [11].

detection has not yet been solved. Moreover, color cue and depth cue play complementary roles in salient object detection, but they conflict with each other sometimes. How to combine color cue and depth cue while handling their inconsistency still needs further investigation.

In this chapter, we comprehensively review the advances in RGB-D salient object detection, and the rest of the chapter is organized as follows. In Section 1.2, we briefly review the evaluation of salient object detection, especially those on RGB images, since many RGB-D salient object detection methods derive from the existing RGB ones. In Section 1.3, we present the typical RGB-D salient object detection methods, evaluate their performance on public datasets, and summarize their issues. In Section 1.6, we discuss some open problems and suggestions for future research.

1.2 Salient object detection evolution

In the past decades, a great progress has been made in salient object detection on RGB images. A large number of RGB salient object detection methods are proposed, and they achieve significant performance. These methods explore the effectiveness of color cue in salient object detection, while providing the useful inspiration for depth cue in RGB-D salient object detection. The incipient RGB salient object detection methods are mainly based on

the handcrafted features of global or local contrast, while there are many corresponding RGB-D methods [8, 9, 14, 15, 16, 22, 18, 25, 28, 31]. These methods perform well on the images which have simple and high-contrast salient objects and background, but easily suffer from many problems on complex images, such as incomplete objects. To improve the completeness of the detected salient objects, graph-based models are used to propagate the saliency among adjacent and similar regions, which can effectively enhance the missing parts in the salient objects while suppressing the residual saliency on the background. Graph based methods also inspire some RGB-D salient object detection methods [22, 11]. Recently, deep learning based methods show their remarkable abilities in salient object detection, including deep neural networks, multi-context deep networks, multi-scale deep networks, symmetrical networks and weakly-supervised deep networks [12, 23, 3].

Beyond extracting salient objects from a single image, co-saliency detection focuses on detecting common salient objects from several related images [10, 26, 7, 6]. By exploring the inter-image correspondence among images, co-saliency can extract the salient objects with similar appearances from multiple images effectively. Compared to RGB-D salient object detection, the multiple images used in co-saliency detection have the same modality, *i.e.*, color cue, but not different ones. Moreover, co-saliency detection requires that the objects should be salient in all the images, but the objects are usually only in color cue or depth cue in RGB-D salient object detection. Fig. 1.3 shows an example of co-saliency object detection. Recently, some research works combine co-saliency detection and RGB-D salient object detection, and extract common salient objects from multiple RGB-D images.

Fig. 1.3 Examples of co-saliency object detection. (a) Image series. (b) Saliency maps. (c) Groundtruths of co-saliency object detection. The saliency maps are generated by [6].

Video salient object detection aims to extract salient objects from video sequences [29, 30]. From a certain perspective, video salient object detection can be treated as a special co-saliency detection, in which all the adjacent video frames contain the common salient objects with similar appearances.

Fig. 1.4 shows an example of video salient object detection. Nevertheless, video salient object detection is usually conducted in a different way. In one aspect, the adjacent video frames are similar in both objects and background. And it follows that inter-frame analysis can provide little additional information compared to single frame analysis. From another perspective, the motion cue that can be estimated from the adjacent frames usually plays a key role in salient object detection, because the moving objects are easy to attract human attention. The exploration [29, 30] of motion cue has some similar characteristics to that of depth cue, for example, the estimated object motion is usually inaccurate and the detection results on color cue and motion cue conflict each other sometimes. Thus, the studies on video salient object detection, especially on the fusion of color cue and motion cue, may provide useful inspiration to RGB-D salient object detection.

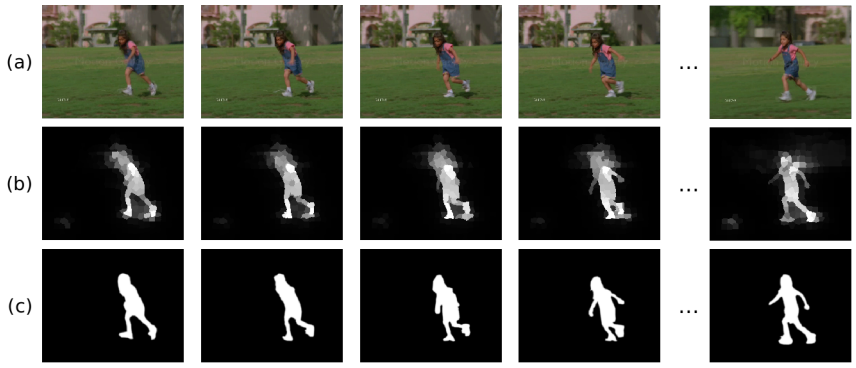

Fig. 1.4 Examples of video salient object detection. (a) Video frames. (b) Saliency maps. (c) Groundtruths of video salient object detection. The saliency maps are generated by [30].

1.3 RGB-D Salient object detection

Based on the numbers of modalities and images used in salient object detection, RGB-D salient object detection can be roughly classified into three categories: depth based salient object detection, depth and color based salient objet detection and RGB-D co-saliency detection.

1.3.1 Depth based salient object detection

Depth based salient object detection aims to explore the effectiveness of depth cue in salient object detection directly and independently, *i.e.*, extracting salient objects from depth maps without considering color cue.

Based on the assumption that depth is an intrinsic part of biological vision, Ouerhani et al. [21] investigated the power of depth in saliency analysis, and pointed out that depth cue is beneficial in predicting human gazes. Ju et al. [15, 16] proposed the first depth based salient object detection method with the assumption that salient objects stand out from their surroundings in depth. The method is based on anisotropic center-surround difference, and refines its results by integrating the 3D spatial prior. However, they used fixed weights to combine depth contrast from different directions to predict pixel level saliency, which might lead to low quality on some specific directions of the saliency map. There is also another disadvantage that the area chosen to generate depth contrast in each direction for a single pixel was fixed, which may lead to a vague saliency map under some condition, especially when the salient object takes up a big portion of the whole image.

In order to detect salient objects easier and more accurate, Sheng et al. [24] enhanced the depth comparison between salient objects and the background instead of extracting features from depth maps directly, based on the fact that contrast between pixels in many depth maps is not obvious due to various view points used to capture depth maps.

Depth cue is simpler than color cue in saliency analysis because it only contains one channel rather than three. However, it suffers from the problems of low quality, which tends to hamper the accurate salient object detection. Moreover, the depth maps of natural images are usually connected, which prevents segmenting the salient objects from the background without the assistance of color cue [31].

1.3.2 Depth and color based salient object detection

As compared to only using depth cue, it is a common and better solution to combine depth cue and color cue in salient object detection. Early works usually directly treat the depth cue as a complement channel of color cue [14] or mix the features from depth cue with those from color, luminance and texture [8], which ignores the differences among different modalities in saliency representation.

To study whether and how depth information influences visual saliency, Lang et al. [18] built a 3D eye fixation dataset using Kinect to study the power of depth in attention prediction. They drew a set of conclusions based on their observations, including: (i) Humans are likely to draw fixation on area with closer depth. (ii) The majority of fixation consists of only a few interesting

objects both in 2D and 3D. (iii) There is a non-linear relationship between depth and saliency and the relationship is different under different scenes with different depth ranges. (iv) The incorporation of depth cue will cause a huge difference between fixation distribution of 2D version and fixation distribution of 3D version, especially in complex scenes. The additional depth information led to an increased difference of fixation distribution between 2D and 3D version, especially when there are multiple salient stimuli located in different depth planes. Based on the above observations, they integrated depth into 2D methods as a probabilistic prior and found that the predictive power could be increased by 6% to 7%. However, they combine depth prior through simple summation or multiplication, which are not efficient enough and even suffer when there are conflicts between color cue and depth cue.

Based on the observations that there are obvious depth gaps between salient objects and background and some domain knowledge in stereoscopic photography, Niu et al. [20] proposed to compute the saliency based on the global disparity contrast, and leverage domain knowledge of stereoscopic photography in salient object detection. However, there are drawbacks that the fact that they considered the depth cue as the fourth channel of color cue that ignores the differences among different modalities in saliency representation, and there are some certain salient objects whose depth comparison between background are consistent rather than abrupt which is conflicted with their basic assumption.

Peng et al. [22] built a RGB-D dataset using Kinect and combined depth and existing 2D models for improvement. They proposed a multi-level saliency map combination method. For low level saliency maps, a multi-contextual feature combining local, global and background contrast to measure pixel-wise saliency is employed. The feature performs a fixed, passive measurement of depth contrast. For mid level saliency maps, a graph based propagation method is adopted, which are helpful in reducing the saliency value in the background area. Notably, most of the contrast based methods without further optimization would suffer from the problem of high saliency in the background, while graph based methods show a better performance on this problem. For high level saliency maps, some spatial priors are incorperated. Because of the fact that most of the salient objects are always laid on the central area of the scene, spatial priors could contribute to eliminating some interference from background objects with high contrast for color cue or depth cue. Finally, they combine three levels' saliency maps by adding the first two levels' saliency maps and then multiplying high level saliency maps. Despite the delicate process of multi-contextual features in low level and diverse feature extraction in different levels, the combination method consists of simply summation and multiplication, which cannot make an effective combination of different saliency maps.

To eliminate the regions with high depth contrast in the background, Feng et al. [9] computed a local background enclosure feature, then applied the priors on depth, spatial, and background, and refined the boundaries of salient

objects with Grabcut segmentation. There are several improvements compared to Ju et al. [15, 16] on how to take advantage of depth cue, including: (i) Incorporation of angular could be considered as a kind of contrast with adaptive weights which ameliorated the problem brought by fixed weights of contrast in different directions in [15, 16]. (ii) The area of contrast for each pixel was reduced compared to Ju et al. [15, 16], which only drew attention to distinguishing salient objects from local background.

Guo et al. [11] further proposed a salient object detection method based on saliency evolution, which generated the accurate but incomplete salient objects by fusing the saliency analysis results on color cue and depth cue, and refined the saliency maps by propagating saliency among adjacent and similar regions in super-pixel level. The main contribution of Guo et al. [11] was that they proposed an effective method to combine color cue and depth cue. To be more specific, the saliency evolution strategy implemented with a single-layer cellular automata can reduce the high saliency regions in the background and improve the completeness of salient objects. However, if some parts of the salient object are very thin compared to the main part, like a tentacle of an alien, the final saliency map would be vague in these thin parts, due to the fact that evolution strategy tends to assign higher saliency value when most of its surrounding area has high saliency value, while the surrounding of the thin parts do not have high saliency value.

Wang et al. [28] proposed a multistage salient object detection method, which generated color cue and depth cue based saliency maps, weighted them with depth bias and 3D spatial prior, and fused all the saliency maps by multi-layer cellular automata. Different from Guo et al. [11] which utilized a single-layer cellular automata on the multiplication of different saliency maps, they use a multi-layer cellular automata to fuse all saliency maps directly, which shows a superiority in performance.

Song et al. [25] exploited different features on multiple levels and generated several multi-scale saliency maps by performing a discriminative saliency fusion on hundreds of corresponding regional saliency results. To be more specific, the discriminative saliency fusion employed a random forest regressor to find the most discriminative ones, which would be used in generating multi-scale saliency maps. Different from many other proposed fusion methods that use weighted summation or multiplication, the discriminative fusion is non-linear which will not suffer when the amount of salient results exceed one hundred. Based on several generated multi-scale saliency maps, a further fusion is needed to generate a final saliency map. Bootstrap learning was employed to combine these saliency maps, which performed salient objects segmentation at the same time. Evidently, the segmentation contributed to both reducing the saliency value in the background and refining the boundary of saliency objects.

In recent years, similar to that in many other vision tasks, deep learning shows its power in salient object detection. However, recent deep learning methods mainly pay their attention to color cue, while there are few of them

taking advantage of both color cue and depth cue. In the following part, we introduce two RGB-D salient object detection methods which are deep learning based.

Qu et al. [23] designed a Convolutional Neural Network(CNN) to fuse different low level saliency cues into hierarchical features for automatic detection of salient objects. They adopted the well-designed saliency feature vectors as the input instead of directly feeding raw images to the network, which could take advantage of the knowledge in the previous advances in salient object detection and reduced learning ambiguity to detect salient object more effectively. Moreover, it integrates Laplacian propagation with the learned CNN to extract a spatially consistent saliency map. Thanks to the superiority of CNN in fusing different feature vectors, the performance is improved compared to other non-deep learning based methods, but they ignored the strong power of CNN in feature extraction.

Han et al. [12] transferred the structure of the RGB-based deep neural network to be applicable for depth cue, and fused the deep representations of both color and depth views automatically to obtain the final saliency map. Different from Qu et al. [23], CNN is used in all of stages including feature extraction and feature fusion.

Chen et al. [3] designed a complementarity-aware fusion module and explored the complement across all levels in order to obtain sufficient fusion results. There is a difference between Han et al. [12] and Chen et al. [3] that Han et al. [12] combined depth cue and color cue after feature extraction, Chen et al. [3] fused two cues from the beginning of the feature extraction and performed fusion in every stage of the process.

1.3.3 RGB-D co-saliency detection

RGB-D co-saliency detection aims to further explore the inter-image correspondence and to perform better in salient object detection.

Fu et al. [10] utilized the depth cue to enhance identification of similar foreground objects via a proposed RGB-D co-saliency map, as well as to improve detection of object-like regions and provide depth-based local features for region comparison. Moreover, they formulated co-segmentation in a fully-connected graph structure together with mutual exclusion constraints to deal with the images where the common object appears more than or less than once.

Song et al. [26] proposed a RGB-D co-saliency method via bagging-based clustering, which generates the saliency maps on single images, clusters them into weak co-saliency maps, and integrates the weak co-saliency maps adaptively into the final saliency map based on a clustering quality criterion.

Cong et al. [7] proposed an iterative RGB-D co-saliency method, which utilizes the existing single saliency maps as the initialization, and generates the final RGB-D co-saliency map by using a refinement-cycle model.

Another method proposed by Cong et al. [6] utilized the depth cue to enhance identification of co-saliency. It calculated the intra saliency maps on each single image and the inter saliency maps based on the multi-constraint feature matching, refined the saliency maps with cross label propagation, and integrated all the original and optimized saliency maps to the final co-saliency result.

1.4 Evaluation

1.4.1 Datasets

There are many datasets for RGB salient object detection, such as MSRA10K [5] and XPIE [32], but the datasets for RGB-D salient object detection are quite scarce.

For depth and color based salient object detection, also for depth based salient object detection, there are two existing datasets: RGBD1000 [22] and NJU2000 [16]. Specifically, RGBD1000 dataset consists of 1000 RGB-D images with the maximum resolution of 640×640, which are captured by Kinect. RGBD1000 also provides two versions of depth cues, including raw depth map and smoothed depth map. Fig. 1.5 shows an overview of RGBD1000. NJU2000 dataset consists of 2000 RGB-D images with the maximum resolution of 600×600, whose depth cues are generated by a depth estimation algorithm. Fig. 1.6 shows an overview of NJU2000.

For RGB-D co-saliency detection, there are two typical datasets: RGBD Coseg183 [10] and RGBD Cosal150 [6]. Specifically, RGBD Coseg183 dataset consists of 183 RGB-D images captured by Kinect, which are divided into 16 groups and each group contains 12 to 36 images, and the maximum resolution of the images is 640×480; RGBD Cosal150 dataset consists of 150 RGB-D images with the estimated depth cues, which are divided into 21 groups and each group contains 2 to 20 images, and the maximum resolution of the images is 600×600. Fig. 1.7 shows an overview of Coseg183 [10] and RGBD Cosal150 [6].

1.4.2 Metrics

The evaluation of RGB-D salient object detection performance uses the same metrics as other salient object detection tasks. By comparing the generated saliency map to the manually labeled groundtruth, several evaluation metrics

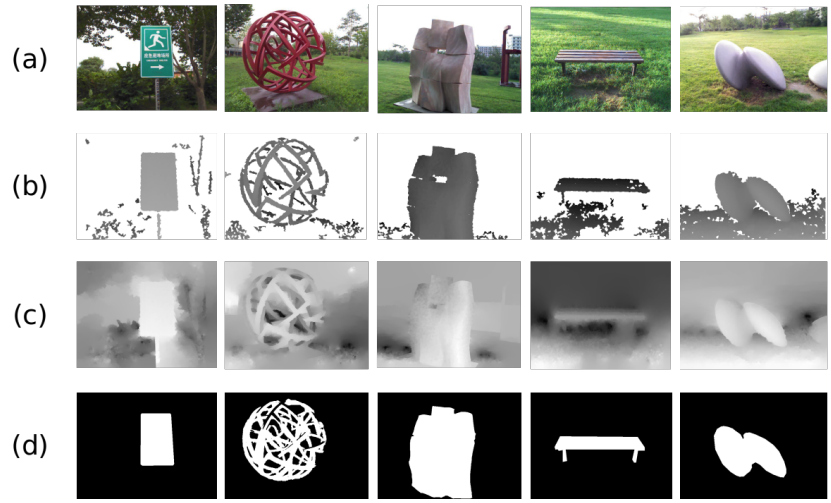

Fig. 1.5 Overview of RGB1000. (a) Original images. (b) Raw depth maps. (c) Smoothed depth maps. (d) Groundtruths of salient object detection.

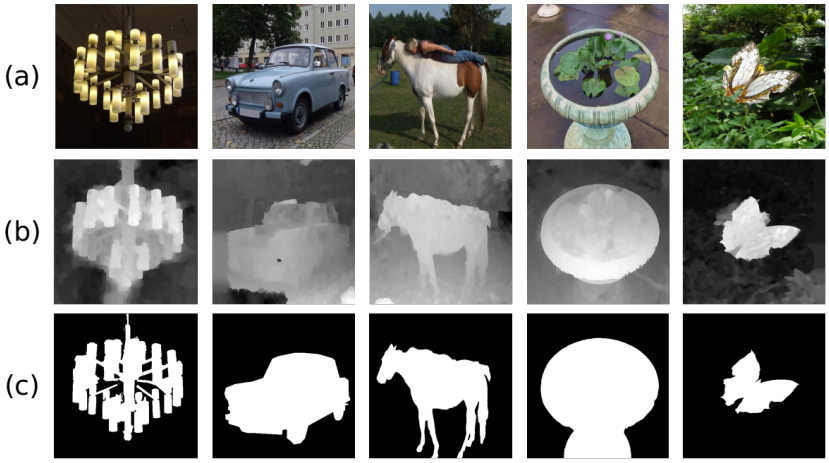

Fig. 1.6 Overview of NJU2000. (a) Original images. (b) Depth maps. (c) Groundtruths of salient object detection.

can be calculated for quantitative evaluation, including Area Under the Curve (AUC), F-measure and Mean Absolute Error (MAE). Specifically, AUC metric calculates the area under Receiver Operating Characteristic (ROC) curve, which is better if larger. F-measure calculates a weighted harmonic mean of precision P and recall R, which is defined as follows:

$$F_\beta = \frac{(1+\beta^2)P \times R}{\beta^2 \times P + R}, \tag{1.1}$$

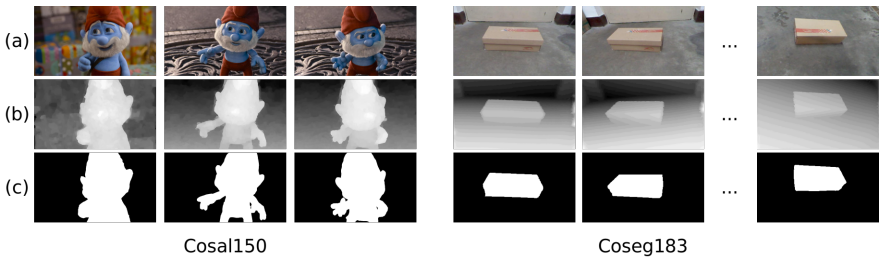

Fig. 1.7 Overview of NJU1000. (a) Original images. (b) Depth maps. (c) Groundtruths of co-saliency object detection.

where β^2 is usually set to 0.3 to emphasize the precision. A larger F_β score means better performance.

Weighted F-measure calculates the F-measure with weighted precision P^w and recall R^w, which is defined as follows:

$$F_\beta^w = \frac{(1+\beta^2)P^w \times R^w}{\beta^2 \times P^w + R^w}. \tag{1.2}$$

Specially, it will be lower than normal F-measure. The specific calculation of weighted precision P^w and recall R^w can be reffered in [19].

MAE is calculated based on the difference between the salient object detection result S and the groundtruth G, which is defined as follows:

$$MAE = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} |S(i,j) - G(i,j)|, \tag{1.3}$$

where w and h are the width and height of the image. A smaller MAE score means the better performance.

1.4.3 Comparison analysis

We compared the performance of typical RGB-D salient object detection methods. All the results are provided by the authors or generated by their source codes. For depth based salient object detection, we compared Ju et al. [16] and Sheng et al. [24]; for depth and color based salient object detection, we compared Lang et al. [18], Niu et al. [20], Peng et al. [22], Guo et al. [11], Qu et al. [23], and Chen et al. [3]; for RGB-D co-saliency detection, we compared Song et al. [26] and Cong et al. [6].

Table 1.1 to 1.3 show the performance of the compared methods in depth based salient object detection, depth and color based salient object detection, and RGB-D co-saliency detection, respectively. We can see that:

(i) As shown in Table 1.1, Sheng et al. [24] is slightly better than Ju et al. [16] in all three metrics. A possible explanation would be discussed as follows. They both employed depth cue as the basic cue to generate salieny maps. However, Ju et al. [16] only emphasised the depth contrast on the origin depth map, and they fixed the weights of depth contrast from different directions rather than using adaptive weights, which may lead to low quality on some specific direction of the saliency map. Ju et al. [16] also used the weighted summation of the biggest contrast values among a relatively large area from different directions in depth maps to calculate pixel level salient value, which would lead to a vague saliency map, especially when the salient object takes up a big portion of the whole image. As shown in Fig. 1.1, prediction of small salient object is relatively more accurate than that of big salient object. By contrast, Sheng et al. [24] developed a new preprocessing method to enhance the depth contrast on depth maps and then used the preprocessed depth map to generate saliency maps.

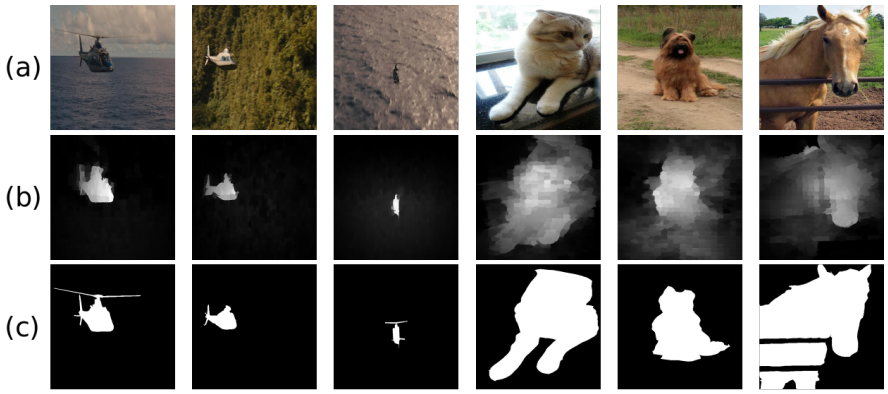

Fig. 1.8 Examples of Ju et at. [16] (a) Original images. (b) Saliency maps. (c) Groundtruths of salient object detection.

(ii) By comparing methods of Lang et al. [18], Niu et al. [20], Peng et al. [22] and Guo et al. [11] in Table 1.2, which used both color cue and depth cue without deep learning modules, we find that Guo et al. [11] outperform other methods and Peng et al. [22] take the second place. A possible explanation would be discussed as follows. Lang et al. [18] and Niu et al. [20] combined color cue and depth cue simply by adding or multiplying saliency maps generated with different cues. Similarly, Peng et al. [22] calculated the final saliency map by adding the first two levels' saliency maps and multiplying the third level's saliency maps.

In spite of the similarity between the fusion methods of Peng et al. [22], Lang et al. [18] and Niu et al. [20], Peng et al. [22] incorporated different levels' depth contrast, *e.g.*, local contrast, global contrast, and background contrast. Notably, Peng et al. [22] also employed graph based method to gen-

erate saliency maps of the second level, which contributed to reducing high saliency maps in the background. All the above works of Peng et al. [22] helped to generate saliency maps with high quality. By contrast, Guo et al. [11] proposed a new method to combine depth cue and color cue in salient object detection. They generated saliency maps using color cue and saliency maps using depth separately, which are both of low quality. After multiplying two saliency maps, Guo et al. [11] conducted a refinement step by employing a single layer cellular automaton that boosted the final performance. Fig. 1.9 shows a comparison between the above methods. To conclude, simply calculating summation and multiplication are not efficient ways to fuse different saliency maps. There is still a demand for exploiting other efficient fusing strategies.

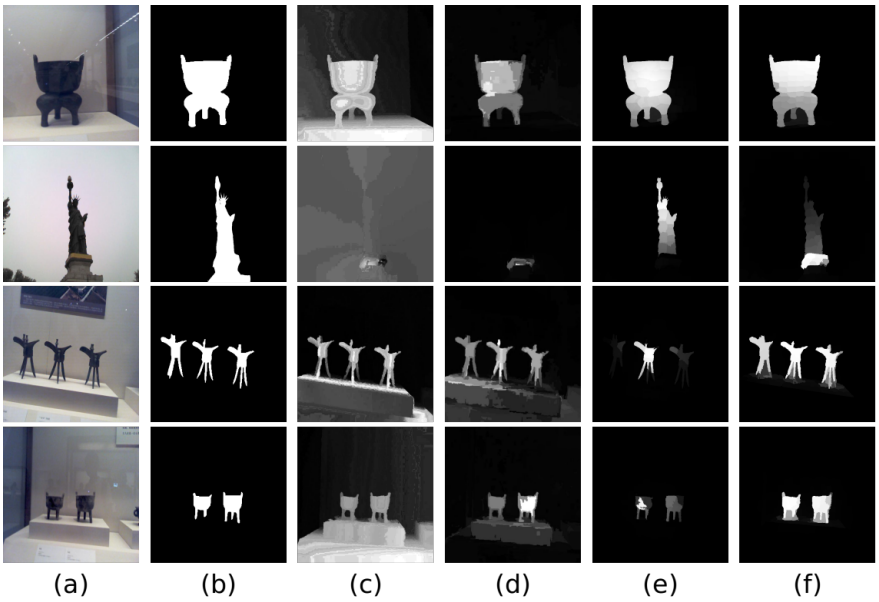

Fig. 1.9 Examples of saliency maps using both color cue and depth cue without deep learning modules. (a) Original images. (b) Groundtruths. (c) Results of Lang et al. [18] (d) Results of Niu et al. [20]. (e) Results of Peng et al. [22]. (f) Results of Guo et al. [11].

(iii) By comparing two deep learning based methods, Qu et al. [23] and Chen et al. [3], we find that Chen's method is better than Qu's method. A possible explanation would be discussed as follows. Qu et al. [23] only used the deep learning module to fuse two saliency maps generated independently with depth cue and color cue. By contrast, Qu et al. [23] employed Convolutional Neural Network (CNN) both to extract features from RGB images and depth

maps and fuse saliency maps, which utilized the power of CNN in feature extraction. Thus, Qu et al. [23] can make a better performance.

(iv) Table 1.2 shows that the deep learning based methods, *e.g.* Qu et al. [23] and Chen et al. [3], outperform other methods, which shows the power of deep learning in saliency feature representation.

(v) By comparing Table 1.1 and 1.2, the depth based methods are not inferior to many methods based on color and depth. It shows that the effective combination of color cue and depth cue is not yet achieved. Simply multiplying or adding saliency maps generated with different cues are not efficient.

(vi) By comparing Table 1.1 and 1.3, the performance of RGB-D co-saliency detection is better than that on single images. It shows that the analysis of inter-image correspondence is beneficial to salient object detection.

Table 1.1 Evaluation of different depth based salient object detection methods on RGBD1000 and NJU2000 datasets.

	RGBD1000			NJU2000		
	AUC	F_β	MAE	AUC	F_β	MAE
Ju et al. [16]	0.92	0.67	0.16	0.93	0.75	0.19
Sheng et al. [24]	0.95	0.68	0.15	0.95	0.78	0.16

Table 1.2 Evaluation of different depth and color based salient object detection methods on RGBD1000 and NJU2000 datasets.

	RGBD1000				NJU2000			
	AUC	F_β	F_β^w	MAE	AUC	F_β	F_β^w	MAE
Lang et al. [18]			0.16	0.33			0.31	0.29
Niu et al. [20]	0.80	0.47	0.23	0.18	0.81	0.61	0.35	0.22
Peng et al. [22]			0.46	0.11			0.34	0.21
Guo et al. [11]		0.55	0.55	0.10		0.43	0.60	0.20
Qu et al. [23]	0.88	0.64		0.12	0.83	0.64		0.20
Chen et al. [3]		0.82				0.83		

Table 1.3 Evaluation of different RGB-D co-saliency detection methods on RGBD Coseg183 and RGBD Cosal150 datasets.

	RGBD Cosal183		
	AUC	F_β	MAE
Song et al. [26]	0.97	0.83	0.05
Cong et al. [6]	0.96	0.84	0.14

1.5 Discussion

By analyzing all above methods, we summarize three main points related to the effect of depth cue in salient object detection, which may give some inspiration for future RGB-D salient object detection models' design:

The first point is about feature extraction. In the past few years, there are mainly two ways to extract features in depth maps, including various contrast based methods and deep learning based methods. It should be noted that graph based methods are not ways to extract features. They are used to make refinement or generate final saliency maps. For contrast based methods, a bunch of different contrasts are developed to make a better performance, while there are relatively less deep learning based methods paying attention to depth feature extraction.

The second point is about saliency map fusion. With the incorporation of depth cue, there is often a need to fuse several candidate saliency maps, or some intermediate results. The amount of saliency maps or intermediate results required to fuse are quite different in various proposed models from two to three hundred. Especially when the amount is as high as three hundred, the effectiveness of fusion strategy will matter a lot for the final results. The simplest strategies are weighted summation and point-wise multiplication, while there are many other more effective ones, like evolution based fusion [11], multi-layer cellular based fusion [28], random forest regressor selection based fusion [25], bootstrap based fusion [25] and deep learning based fusion [12, 23, 3].

The third point is about refinement of saliency maps, which includes two aspects: eliminate saliency in the background and make better segmentation in the foreground. Most of contrast based methods without further refinement will suffer from high saliency in the background, due to the fact that there are many objects in the background that have strong contrast with surrounding areas for either color cue or depth cue. To avoid the high saliency in the background, graph based methods are proposed which propagate saliency based on some specific seed points instead of generating saliency value directly on the whole image or depth map. For the second aspect, there is often an incompleteness of salient objects or vagueness in some specific areas, because many parts are not obviously distinct from the background or big enough to be detected by some proposed models. In this condition, refinement like using Grabcut [9] and bootstrap based segmentation [25] can help to make a better segmentation of foreground objects.

1.6 Conclusion

In this chapter, we comprehensively reviewed the advances in RGB-D salient object detection, including depth based salient object detection, depth and

color based salient object detection and RGB-D co-saliency. We first introduced the evolution of salient object detection, and analyzed the relationship between RGB-D salient object detection and salient object detection on other media, *e.g.*, RGB images, multiple images for co-saliency detection and videos. Furthermore, we presented the typical methods of these three categories, and evaluated their performance on four public datasets.

Though many RGB-D salient object detection methods have been proposed, there are still many unsolved issues. The low quality of depth maps may influence the performance of RGB-D salient image detection methods. How to enhance depth maps or improve the robustness to depth noise will be a critical issue for RGB-D salient object detection. Moreover, compared to the datasets for RGB salient object detection, the datasets for RGB-D salient object detection is scarce and their sizes are smaller. It will be significant to construct a large-scale datasets for RGB-D salient object detection.

1.7 Acknowledgements

This work is supported by National Science Foundation of China (61202320, 61321491) and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

1. Borji, A.: What is a salient object? a dataset and a baseline model for salient object detection. IEEE Transactions on Image Processing **24**(2), 742–756 (2014)
2. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE transactions on image processing **24**(12), 5706–5722 (2015)
3. Chen, H., Li, Y.: Progressively complementarity-aware fusion network for rgb-d salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3051–3060 (2018)
4. Cheng, M.M., Mitra, N.J., Huang, X., Hu, S.M.: Salientshape: Group saliency in image collections. The Visual Computer **30**(4), 443–453 (2014)
5. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(3), 569–582 (2014)
6. Cong, R., Lei, J., Fu, H., Huang, Q., Cao, X., Hou, C.: Co-saliency detection for rgbd images based on multi-constraint feature matching and cross label propagation. IEEE Transactions on Image Processing **27**(2), 568–579 (2017)
7. Cong, R., Lei, J., Fu, H., Lin, W., Huang, Q., Cao, X., Hou, C.: An iterative co-saliency framework for rgbd images. IEEE transactions on cybernetics **49**(1), 233–246 (2017)
8. Fang, Y., Wang, J., Narwaria, M., Le Callet, P., Lin, W.: Saliency detection for stereoscopic images. IEEE Transactions on Image Processing **23**(6), 2625–2636 (2014)
9. 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, pp. 2343–2350 (2016)
10. Fu, H., Xu, D., Lin, S., Liu, J.: Object-based rgbd image co-segmentation with mutex constraint. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4428–4436 (2015)
11. Guo, J., Ren, T., Bei, J.: Salient object detection for rgb-d image via saliency evolution. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2016)
12. Han, J., Chen, H., Liu, N., Yan, C., Li, X.: Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics **48**(11), 3171–3183 (2017)
13. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis & Machine Intelligence **20**(11), 1254–1259 (1998)
14. Jeong, S., Ban, S.W., Lee, M.: Stereo saliency map considering affective factors and selective motion analysis in a dynamic environment. Neural networks **21**(10), 1420–1430 (2008)
15. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 1115–1119. IEEE (2014)
16. Ju, R., Liu, Y., Ren, T., Ge, L., Wu, G.: Depth-aware salient object detection using anisotropic center-surround difference. Signal Processing: Image Communication **38**, 115–126 (2015)
17. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. In: Matters of intelligence, pp. 115–141. Springer (1987)
18. Lang, C., Nguyen, T.V., Katti, H., Yadati, K., Kankanhalli, M., Yan, S.: Depth matters: Influence of depth cues on visual saliency. In: European conference on computer vision, pp. 101–115. Springer (2012)
19. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 248–255 (2014)

20. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 454–461. IEEE (2012)
21. Ouerhani, N., Hugli, H.: Computing visual attention from scene depth. In: Proceedings 15th International Conference on Pattern Recognition. ICPR-2000, vol. 1, pp. 375–378. IEEE (2000)
22. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: Rgbd salient object detection: a benchmark and algorithms. In: European conference on computer vision, pp. 92–109. Springer (2014)
23. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. IEEE Transactions on Image Processing **26**(5), 2274–2285 (2017)
24. Sheng, H., Liu, X., Zhang, S.: Saliency analysis based on depth contrast increased. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1347–1351. IEEE (2016)
25. Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., Ren, T.: Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Transactions on Image Processing **26**(9), 4204–4216 (2017)
26. Song, H., Liu, Z., Xie, Y., Wu, L., Huang, M.: Rgbd co-saliency detection via bagging-based clustering. IEEE Signal Processing Letters **23**(12), 1722–1726 (2016)
27. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive psychology **12**(1), 97–136 (1980)
28. Wang, A., Wang, M.: Rgb-d salient object detection via minimum barrier distance transform and saliency fusion. IEEE Signal Processing Letters **24**(5), 663–667 (2017)
29. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3395–3402 (2015)
30. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE transactions on pattern analysis and machine intelligence **40**(1), 20–33 (2017)
31. Wang, Y., Ren, T., hua Zhong, S., Liu, Y., Wu, G.: Adaptive saliency cuts. Multimedia Tools and Applications **77**, 22213–22230 (2018)
32. Xia, C., Li, J., Chen, X., Zheng, A., Zhang, Y.: What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4142–4150 (2017)
33. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162 (2013)