Salient Object Detection: A Short Review

Sakshi Aggarwal[1],Deepika Tyagi[2],Neha Tyagi[3],Priya Singh[4]
{staff.sakshi.aggarwal@iitrpr.ac.in[1],deepika.tyagi@jssaten.ac.in[2],neha.tyagi@jssaten.ac.in[3],priya.singh@jssaten.ac.in[4] }

IIT Ropar, Punjab[1], JSS Academy of Technical Education, Noida[2,3,4]

Abstract. With strong applications in object finding, video surveillance, image compression, and other fields, salient object detection, or SOD, is a solid and well-established field of contemporary computer vision. A number of SOD techniques sought to list items with more focused cues than the surroundings, including contrast before. However, deep learning frameworks began to dominate SOD due to the growth of media and sophisticated learning mechanisms. Stacking layers are used to process and then extract visual information in a deep learning-based SOD model. As a result, feature representation in SOD operations is more crucial and challenging. The primary goal of the paper is to increase the prominence of cutting-edge techniques, including both deep learning-based and classical-based approaches.

Keywords: salient object detection, saliency, deep learning.

1 Introduction

Salient object detection (SOD) is an image analysis technique that seeks to automatically identify the visually significant areas of an image. The goal of SOD is to determine the key areas of an image that evoke interest in viewers. The human visual system, which can swiftly and effectively identify important objects in a visual scene, serves as the model for the salient detection of objects. Therefore, by determining the areas that are most visually significant in a scene and improving their visibility for users of retinal prostheses, salient object detection plays a major role in assistive technologies for people with visual impairments, including those who use them. This helps these individuals overcome the limited resolution of current retinal prostheses.

Additionally, salient object identification has numerous applications in image processing and computer vision, like object recognition, compression, autonomous driving, and medical imaging. Saliency detection is based on two fundamental ideas. The first is the existence of objects. Next, indicating the location, i.e., localization. By dividing up complete images into superpixels or patches, searching has become more efficient. For example, segmentation-based saliency works quite well

IICTDS 2025, December 19-20, Chandigarh, India
Copyright © 2026 EAI
DOI 10.4108/eai.19-12-2025.2362583

to achieve this goal [1]. There is continuous data-driven research—more significantly, multimedia-driven research—along with learning frameworks to accomplish both goals. In essence, salient detection advances the processes of object classification and image segmentation.

As illustrated in Figure 1, two classes can be used to broadly classify contemporary SOD techniques: 1) traditional approaches and 2) deep-learning-based approaches. Traditional methods, such as local contrast-based, diffusion-based, and classical supervision, employ a few heuristics and low-level features to identify essential items. Furthermore, to enhance performance, deep learning-based methods can assist in extracting deep semantic characteristics. Inspired by the extensive research on these two categories, this paper presents a short review of the methods commonly used for SOD.

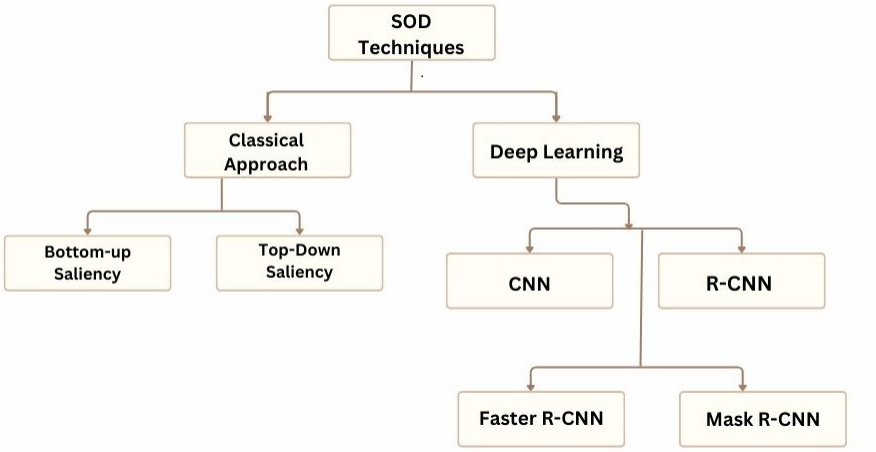


Fig. 1. Taxonomy of SOD Techniques

2 Classical Approaches

According to Li et al. [2], three levels—low-, mid-, and high-level—can be used to visualize the features of an object. Primitive object features like color, edge, and texture are considered low-level features. In contrast, mid-level features depend on spatial information and object contours, whereas high-level features are linked to the semantic analysis of object interaction. Accurately designed to capture image data and its semantics, these features or cues are then used to model a multitude of SOD approaches. The groundbreaking research of Itti et al. [3] applied its unique framework for salient object detection to both natural settings and artificial images. Artificial visual saliency was discovered to function in two ways: 1) top-down saliency and 2) bottom-up saliency. In the subsequent sections, the following techniques are covered in detail.

2.1 Bottom-up Saliency Method

A group of researchers [4] has attempted to use the Density Maximization (DM) Saliency method to address the difficulties associated with SOD. It uses the bounding-box search that matches the distinctive object's dimensions and form. The best bounding box can also be found by applying the branch-and-bound algorithm. The hypergraph representation produced by dynamic feature selection was proposed in the study [5].

The bottom-up saliency with mid-level vision, including spatial context, was summarized in another study [6]. Existing network optimization techniques were used to model cognitive characteristics explicitly. According to the study, a single graph cannot adequately capture the whole image, particularly when complicated. Thus, the design needs several graphs that illustrate a single feature. The edges will come from the spatial proximity property, which specifies which pixels belong to which classes.

2.2 Top-down Saliency Method

The top-down approach is exposed to high-level information to motivate saliency computation, whereas the bottom-up method is partially motivated by eye fixation prediction or visual information. Such approaches necessitate prior domain-specific expertise. The training data must be manually labelled as part of the algorithms' supervised learning strategy. The most basic techniques to facilitate the mapping of saliency masks and salient detection are background prior, boundary prior [7], and contrast prior [8]. Additionally, the contrast prior functions in two ways: locally, or edge/corner-based contrast, and globally, or region-based contrast. According to studies [9], [10], and [1], the most important objects are typically located in the middle. [3] suggests centre-surround operations to create a saliency network. When using a bottom-up approach, having too many low-level elements might negatively impact cost performance, while having too few can leave critical factors unused. Johnson and Subha [11] expand on the study of optimal feature selection by incorporating supervised learning.

The ideal feature set for saliency demonstrates resilience. Supervised learning approaches can produce good results but are unsuitable for real-time applications. Labelling becomes time-consuming as the dataset grows in size. It can be challenging and time-consuming to learn about visual saliency, which makes it ineffective for some applications.

3 Deep Learning

Because of their exceptional performance, convolutional neural networks (CNNs) are frequently utilised for image processing. CNN extracts features using hidden layers and automatically adjusts weights to determine saliency. Hu et al. [12] used VGG16net and ReLU transfer layers to create a saliency map with unambiguous borders and reduced blurring. It was followed by end-to-end guided filtering to improve operations. Deeper neural networks are prone to training errors. Wang et al. [13] developed a residual network to deepen the network and lower learning errors. Zhang et al. [14] suggested a practical method for optimising global contextual features in multi-path recurrent networks. The study demonstrated an effective method for suppressing background

Table 1: Advantages and Disadvantages of Deep Learning

S.No.	Model	Advantages	Disadvantages
1.	CNN	As handcrafted features are not worthy, CNN can represent high-level features	Difficult to trace across all spatial locations within an image
2.	R-CNN	Fragmented into regions that contain objects we are looking for instead of feature engineering	Time-space trade-off is unbalanced
3.	Faster R-CNN	Does not consume memory space	Fails to determine the orientation of objects
4.	Mask R-CNN	A saliency mask is obtained that consists of the exact shape, size, and location of distinct objects	Applying a huge number of photos at once will result in a slow function.

noise and achieving acceptable results.

Regions with CNN (R-CNN)-based 3D object identification system that could predict vehicle orientation was proposed by Braun et al. [15]. The bounding boxes of items, including cars or pedestrians, as well as the direction of objects being identified, were provided by the exceptional work. It enhanced in-car systems' real-time detection. When working on bounding-boxes annotations, Gao et al. [16] suggested R-CNN in contrast to the studies on semi-supervised object detection. The bounding-boxes mining framework decreased the dependence on manual labeling. Rather, it used numerous picture labels that were less expensive for this purpose. To get around the issues with face detection, R-CNN additionally used pattern recognition [17].

Ren et al. [18] highlighted the limitations of multi-layer perceptron classifiers for features in the context of contemporary object detectors. By optimizing the usage of deeper convolutional architectures for feature classification, it expanded the Faster R-CNN model. It substantially improved the detection performance. Additionally, Wang et al. proposed a unique object detector that operates under invariances using Fast R-CNN. Having a dataset with occlusions and deformations was the main idea behind the project. The majority of them do not occur often. However, it forced the model to learn these challenges and evaluate its performance for a few images that were challenging for earlier object detectors to classify.

The Mask R-CNN, which sits at the top of the Faster R-CNN, generates an object segmentation mask for a given image using the object's box coordinates and class label. Because of its encouraging results, Guo et al. [19] were interested in using Mask R-CNN to help with hepatic steatosis segmentation inspection. It is characterized by an abnormal accumulation of lipids inside the tissues. Liver disease detection and transplantation are made possible by precise liver quantification analysis. Mask R-CNN is suggested to address steatosis detection and segmentation together. The alternative study by Kopelowitz and Engelhard [20] used 3D CT images to detect lung nodules.

Because organs vary in shape and have unclear borders, Shu et al. [21] found that single-organ segmentation is insufficient for an appropriate radiation strategy. In order to concentrate on multi-scale semantic characteristics that were effective for multi-organ segmentation, an enhanced Mask R-CNN framework was proposed. Table 1 shows the advantages and disadvantages of DL-based frameworks.

3.1 Challenges in Deep Learning

Although the positive aspects of deep learning approaches for SOD, segmentation, and classification have so far been demonstrated, significant work is still needed to address the apparent problems. Annotated samples have a significant influence on the accuracy of deep learning models. The network, which usually determines the model performance, is trained using a large dataset. However, there is a dearth of easily accessible annotated data. Model deterioration and a low prediction rate are caused by inadequate information. Time is the main obstacle when training the deep learning model. Faster convergence results from longer training times, while slower convergence results from shorter training times. A small area of an image is frequently given to anatomy-of-interest in medical monitoring. The retrieved patterns, however, are part of the non-interested region. High importance is given to these small areas, which may also be taken into account for saliency detection. Nevertheless, a training network biased toward the least interesting region is produced when training with unbalanced data.

4 Conclusion

In natural images, our visual system favors items that add more sharpness and distinction than the background. We frequently overlook a number of additional elements that are probably included in the picture but may not catch our notice because they are obscure or unclear. Salient item detection is intended for both artificial and natural situations. An image is created by weaving together many features. To investigate objects across spatial locations of an image, isolate them from their surroundings, and ultimately classify them into one of the labels, features like color contrast, centre prior, object contour, or texture are sufficient.

Following the global craze for social media and multimedia-driven platforms, DL began to soar in the market. By taking advantage of many hidden layers and automatically modifying the network weights to calculate the saliency, DL was included to improve the dataset's training. Different DL-based salient object proposal techniques are rich in extracting high content, fortifying the model network, and significantly improving performance, whereas traditional SOD methods were built upon handcrafted features. In object identification, deep learning frameworks like CNN, R-CNN, and Faster R-CNN are well known. However, Mask R-CNN generates its saliency mask (pixel-level segmentation) and displays a distinct rectangle around the object, which other DL techniques cannot do.

References

[1] Liu T, Sun J, Zheng N, Tang X, Shum HY. Learning to detect a salient object. In: Proceedings of the IEEE International Conference on Computer Vision Pattern Recognition; 2007. p. 1-8.

[2] Li X, Zhao L, Wei L, Yang MH, Wu F, Zhuang Y, et al. DeepSaliency: Multi-Task Deep Neural Network Model for Salient Object Detection. IEEE Transactions on Image Processing. 2016;25(8):3919-30.

[3] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell. 1998;20(11):1254–1259.

[4] Luo Y, Yuan J, Xue P, Tian Q. Saliency Density Maximization for Object Detection and Localization. In: ACCV 2010, Part III, LNCS 6494; 2011. p. 396–408.

[5] Liang Z, Chi Z, Fu H, Feng D. Salient object detection using content-sensitive hypergraph representation and partitioning. Pattern Recognition. 2012;45:3886–3901.

[6] Zhang J, Ehinger KA, Wei H, Zhang K, Yang J. A novel graph-based optimization framework for salient object detection. Pattern Recognition. 2017;64:39-50.

[7] Zhang F, Du B, Zhang L. Saliency guided unsupervised feature learning for scene classification. IEEE Transactions on Geoscience and Remote Sensing. 2015;53(4):2175-84.

[8] Miyazato K, Kimura A, Takagi S, Yamato J. Real-time estimation of human visual attention with dynamic bayesian network and mcmc-based particle filter. In: IEEE Intl. Conf. on Multimedia and Expo.; 2009. p. 250–257.

[9] Cheng M, Zhang GX, Mitra NJ, Huang X, Hu SM. Global contrast based salient region detection. In: CVPR; 2011. p. 409-16.

[10] Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S. Salient object detection: a discriminative regional feature integration approach. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, Portland, Oregon, USA; 2013. p. 2083—2090.

[11] Johnson S, Subha TD. A Study on Eye Fixation Prediction and Salient Object Detection in Supervised Saliency. In: Materials Today: Proceedings. vol. 4; 2017. p. 4169-81.

[12] Hu P, Shuai B, Liu J, Wang G. Deep Level Sets for Salient Object Detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Honolulu, HI, USA; 2017. p. 540-9.

[13] Wang H, Dai L, Cai Y, Sun X, Chen L. Salient object detection based on multi-scale contrast. Neural Networks. 2018;101:47–56.

[14] Zhang X, Wang T, Qi J, Lu H, Wang G. Progressive Attention Guided Recurrent Network for Salient Object Detection. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, UT, USA; 2018. p. 714-22.

[15] Braun M, Rao Q, Wang Y, Flohr F. Pose-RCNN: Joint object detection and pose estimation using 3D object proposals. In: 2016 IEEE International Conference on Intelligent Transportation Systems (ITSC). Rio de Janeiro, Brazil; 2016. p. 1546-51.

[16] Gao J, Wang J, Dai S, Li LJ, Nevatia R. NOTE-RCNN: NOise Tolerant Ensemble RCNN for Semi-Supervised Object Detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Seoul, Korea (South); 2019. p. 9507-16.

[17] Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. p. 580–587.

[18] Ren Y, Zhu C, Xiao S. Object Detection Based on Fast/Faster RCNN Employing Fully Convolutional Architectures. Mathematical Problems in Engineering. 2018;2018(1):3598316.

[19] Guo X, Wang F, Teodoro G, Farris AB, Kong J. Liver Steatosis Segmentation With Deep Learning Methods. In: 2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019). IEEE, Venice, Italy; 2019. p. 24-7.

[20] Kopelowitz E, Engelhard G. Lung Nodules Detection and Segmentation Using 3D Mask-RCNN. In: MIDL 2019 Conference; 2019. .

[21] Shu H, Nian D, Yu H, Li X. An Improved Mask R-CNN Model for Multiorgan Segmentation. Mathematical Problems in Engineering. 2019;2020(1):8351725.