Chapter 2

Literature Survey

Salient object detection (SOD) aims to find the most prominent and conspicuous object(s) in a given image, which is similarly identified by the human visual system. Saliency slightly differs from general image segmentation. The objective of salient object detection is to suppress the contents of background pixels while simultaneously highlighting the foreground image pixels to locate and detect the prominent objects. In contrast, Image segmentation generally divides images into various regions and picks up one region based on the thresholds. Salient object detection has attracted the interest of researchers in the field of computer vision saliency computation, which is an intermediate step in most complex vision-related applications like Object Classification [18], Object Recognition [19], Image/Video Summarization [20], semantic segmentation [21], Neurobiology of Attention [3], camouflaged object detection [22], and Dermoscopic Segmentation [23]. It has been used extensively

in various robotic applications like, object discovery [24], [2] and human-robot interaction [25]. It is also used in various Graphics applications like non-photorealistic rendering [26] image retargeting [27] and image cropping [28] etc.

This chapter presents the survey of state-of-the-art methods in salient object detection. Section 2.1 presents the literature survey in the field of 2D and 3D salient object detection. The Research Gaps of this domains are discussed in Section 2.3. Section 2.4 lists the databases used for experimental analysis. The Section 2.5 describes and define the various evaluating parameters, which are used in result analysis and experiments. The last section 2.6 of this chapter concludes the literature survey.

2.1 Literature Review

In this section, a comprehensive survey of the state-of-the-art methods has been done in the field of visual saliency computation and closely related other streams. This study profoundly investigates the existing domains, modalities, contributions, and associated research gaps.

2.1.1 Visual Saliency

The ***Visual Saliency*** has objective to design the computer algorithm which formulates the most essential and prominent parts of visual media. The visual saliency was later developed as salient object detection and aggressively used in Computer

Vision, Artificial Intelligence, Graphics, and Multimedia applications. The various computational domains of visual saliency are studied here. The computational domain is classified into statistical and deep-learning-based models. The statistical and probabilistic models have been studied in section 2.2.1. The summaries of the 3D conventional-based salient object detection models have been described and summarised in section 2.2.2. Finally, Section 2.2.3 has reviewed deep learning based models. The study of the visual saliency model can be broadly classified into two domains 1) the Visual saliency model and) 2 Saliency computation domains.

2.1.2 Visual saliency models

There are mainly two streams of visual saliency models: the first is salient object detection, and the second is eye-fixation prediction models. These models are heavily researched and widely used in all arena computer vision applications. The salient object detection focused on predicting and detecting a complete salient object, and the Eye Fixation prediction model focused on the localization of salient points.

2.1.2.1 Fixation Prediction models

Visual saliency computation in eye fixation prediction models focuses mainly on the localization of an object. These models identify some human fixation locations

on the images rather than precisely detecting the object in a well-defined boundary. Although, detecting well-defined boundaries and extracting uniform salient objects is the objective of *"salient object detection"*. In some parameters, the applications and computing models of both computation domains are different, and in some parameters, both are similar. Learning-based human fixation models have been proposed through various discriminative classifiers as training vectors. They produce a location or attention map, which is considered an eye-fixation location point. [3] [10] [29] [8] [30].The diversity of their application, methods, and evaluation criteria is so complex in salient object detection and human fixation prediction that it cannot be completely covered here. The detailed survey in the fixation domain was performed by Borji *et al.* [19] with 36 models and 3 datasets, and another survey was done by Judd *et al.* [31] with 9 models over only 300 images in MIT-Dataset. Our research mainly focuses on salient object detection, which has been heavily researched in recent years. Consequently, it has become the mainstream saliency computation domain.

2.1.2.2 Salient object detection models

The salient object detection model aims to detect, localize, and predict the whole object in the image. Initially, the computational models have two stages: saliency computation followed by segmenting through binarization. For example,the authors [4] [5] [11] [32] have used these two stages separately. They include saliency

computation followed by segmentation using mean [11], mean shift [33], fuzzy growing [32], graph-cut [34], saliency-cut [35], meaningful contrasted boundaries [36] and watershed segmentation [5] etc. In Contrast, recent researchers mainly focused on devolving algorithms and methods to produce salient objects directly, without using segmentation methods, which is an additional computational overhead.

The tremendous success of traditional and deep learning-based models have the objective now to predict the whole object without segmentation and similar to ground truth level. Initially, most highly reported work has been worked on one or two handcrafted low-level features. So all these methods gained popularity to produce the salient object in an image where the object is large-sized and color-based distinguishable. The challenge of the complexity of new datasets and the next benchmark level brings new computational domains like a probabilistic model and a deep learning-based approach. Most statistical and machine learning-based models either used one or combinations of various saliency priors to overcome the challenges of complex and cluttered backgrounds. The deep learning-based model has recently demonstrated enormous capabilities in improving the performance of various computer vision tasks, mainly pixel-level detection using Convolution Neural Network(CNN). The Deep CNN-based various models have been utilized to propose various saliency computation models. The performance of deep learning-based models dominate over statistical and machine learning models. Salient object detection model can be properly investigated in three domains, which is shown in Fig.2.1 as follows:

1. Models of saliency computation

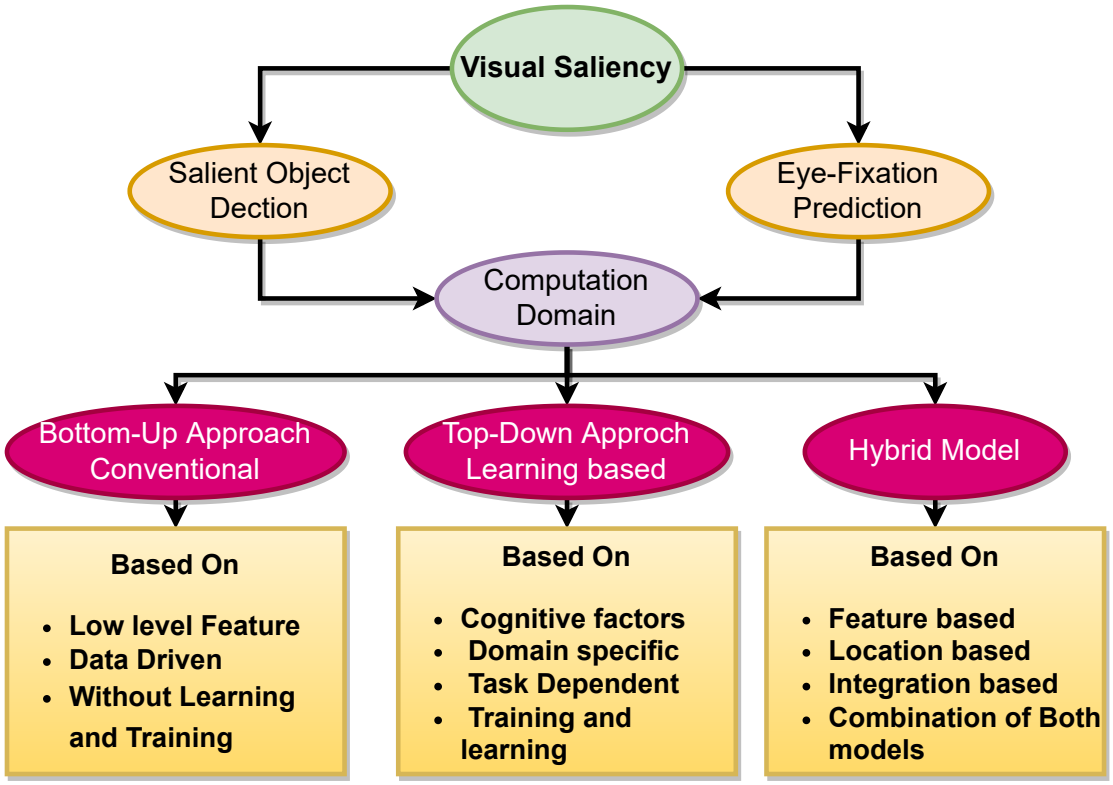

FIGURE 2.1: Visual saliency and Computational domains.

2. Methodological and

3. Dataset and Evaluation domain.

2.1.3 Saliency computation domains:

The saliency computation is broadly classified into three domains, which are described as follows:

2.1.3.1 Bottom-up-search dependent statistical models:

The early models [37] [38] [4] [39] [5] [6] of saliency computation is used for highlighting the object from their surroundings with data-driven, low-level image information, starting from an image without training and learning. The handcrafted features are low-level cues like visual information, such as color, intensity, texture, and orientation. Search in this manner uses some priors like center [40], and background connectivity [12] [41] [42], surroundedness [36], depth [43] and abjectness [44]. This model is computationally simple, fast and provide initial benchmark to the saliency computation. It does not depend upon prior specific or domain knowledge.

2.1.3.2 Top-down-task dependent deep learning based models:

This domain uses cognitive factors that are based on image relevance, that are attention-based and that help in training and learning with manually annotated ground truth data. The models like [45], [46], [47], [48], [49] are based on exploring high level, semantic, and contextual data depends on prior but domain-specific knowledge using supervised learning framework. This domain are widely used by recent

2.1.3.3 Hybrid and Probabilistic models :

It is a classic combination of top-down and bottom-up approaches. These models use, for example, probabilistic approaches like Bayesian principle [29], Matrix decomposition [50],and Fuzzy Theory, [51]; some other times it uses others integration strategy [52] [53]. In this model, bottom-up saliency is based on features, locations, and appearances. This is mainly appearance-driven, where final saliency is combined with specified integration methods. In this method, bottom-up integration explores the salient objects through multiple integration of saliency clues.

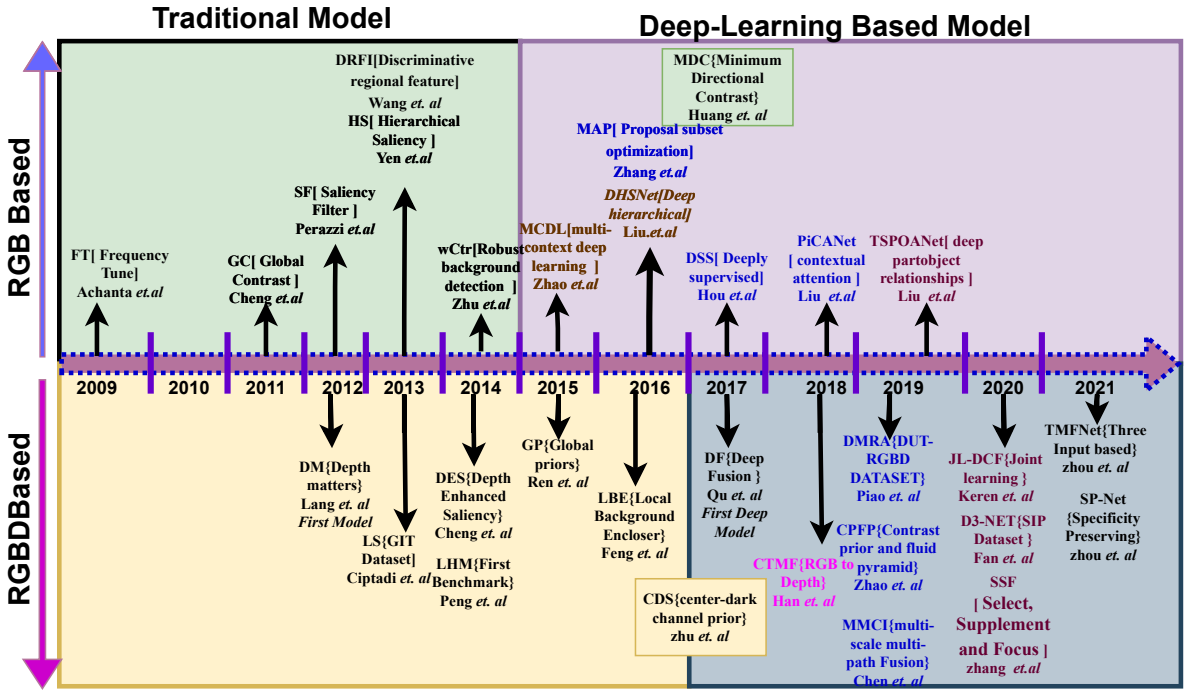

FIGURE 2.2: Computational domains and associated outstanding models in Salient Object Detection.

In this domain, the statistical model-based low-level features combined with high-level guidance are defined explicitly as low-rank matrix recovery theory (LRR) [54]. Shen and Wu [55] proposed a unified or hybrid approach, low-rank matrix recovery (ULR), which integrated the classic low-level feature with high-level guidance. Similarly, Peng *et al.* [56] proposed a matrix decomposition-based salient object detection with high-level or probabilistic priors, like tree-structured sparsity regularization and Laplacian term to capture structure features. Tang *et al.* [57] combined weighted low-rank recovery (WLRR) with a high-level background prior to proposed salient object detection.

The detailed survey of salient object detection presented here follows the chronology of various methods and models developments. Initailly, the focuses on traditional or conventional feature-based methods on 2D (RGB) and 3D (RGBD) saliency. The traditional-based approach initially used one or two handcrafted features; later, it used multiple handcrafted features followed by a complex fusion process to improve the performance in complex images. The fusion process is preferred by using machine learning or probabilistic models. The traditional-based methods in 2D domains are studied in section 2.2.1, while the same approaches in the 3D domain are studied in section 2.2.2. The current approaches of the deep learning-based model are studied in section 2.2.3. The traditional models are based on hand-crated features without using training and learning. At the same time, the deep learning-based models used exhaustive training and testing process. Due to training and learning with a large-sized data-set deep learning-based model improves tremendous performance even in

complex images. Nevertheless, the traditional model recently improved by proposing various fast computational algorithms. Therefore, both domains have their pros and cons. The computational domains, according to the modalities and saliency cues, are classified into four classes, which are shown in Fig. 2.2. The details comparisons and analysis of the domain mentioned above for finding the issue and challenge are summarized below.

2.2 Salient Object Detection

This subsection discusses the various techniques used for salient object detection. The techniques are classified into statistical, machine learning, and deep learning methods. Section 2.2.1 provides the review for statistical and machine learning-based methods and Section 2.2.3 reviews deep learning methods.

2.2.1 Statistical and Probabilistic Models(2D Saliency)

An extensive set of saliency methods have been proposed, and incredible success paved the way for more inventions in this domain to produce a better salient object in complex and cluttered background images. The traditional and probabilistic model has been studied first, and it is broadly classified based on the methodology proposed into three classes, described in the following sections.

2.2.1.1 Biologically Inspired Models

Initially, Itti *et al.* [10] used the center-surround contrast to differentiate the salient(foreground) and non-salient regions (background). Center-surround contrast is a computational framework inspired by Biologically and Neuroscience principles. It is the most widely preferred cue to distinguish between foreground and background. It is measured the contrast using multi-scale and multi-channel features. After that, various algorithms to measure contrast were proposed, such as color contrast, regional contrast, spatial contrast, depth-based contrast, discriminant center-surround contrast [58], center-surround divergence contrast [59], histogram-based contrast [4] minimum directional contrast [5]. Based on the computation, these methods divided global and local contrast. The global contrast-based [11] methods formulated the contrast over the whole image. In contrast, local contrast [60] is based on region or neighborhoods pixels relationship. The global contrast-based method generally performed better than the local contrast-based salient object detection.The topographical Primal Sketch [61] model classifies the grayscale image using first and second derivatives to find different surfaces. At the same time, the proposed probabilistic model is based on color images to produce a topographical surface that includes the salient object.

Highly referenced and first formally recognized method for silent object detection is proposed by Achanta *et al.* [11]. It defined the global color contrast as pixel saliency. The color contrast of each pixel is defined as the difference from the average image color. It is a straightforward and fast algorithm to generate the full-length

salient object. Similarly, histogram-based pixel-wise global contrast is proposed by Cheng *et al.* [4], which measures the saliency by computing color-wise histograms. The same author proposed a region-based saliency algorithm RC by segmenting the image into regions. In this model, saliency is defined as spatial and regional contrast. Another model of saliency using global contrast and spatial distribution is proposed by Perazzi *et al.* [62] and named uniqueness. The global contrast is defined as Euclidean color distance w.r.t entire image with Gaussian spatial weight. The minimum directional contrast (MDC) based salient object detection method is proposed by Huang *et al.* [5] with the assumption that the foreground pixel has high contrast from all directions compared to the background. A marker-based watershed algorithm follows it to estimate each pixel as foreground or background to predict the salient object. Additionally, diverse deep learning-based models have been proposed to learn the contrast [63], [28], [64] to predict the salient object. Although, these methods have great success over the statistical model in complex and cluttered background. The most dominating cue is global contrast. The detailed investigation of the global contrast-based method is separately discussed in the next sections:

2.2.1.2 Global Contrast-Based Models

The methodological variations in saliency computation have their own merits and demerits. They apply in general or specific cases. Initially, global contrast models [4] [11] [65] [66] are used. Then they are followed and further enhanced by regional

contrast [4] [43] [67] models for further enhancing the saliency. Many global contrast models compute the color contrast as principal cues for saliency. First revolutionary model is based on the global contrast of frequency, and is Frequency Tuned(FT) [11]. This first performed smoothing operation by the Difference of Gaussian (DoG) scale followed by center surround contrast. Next level of improvement was modeled by Maximum Symmetric Surround(MSS) [66]. In this model central surround contrast is estimated by proposing pixel level surround symmetry. These global contrast-based methods were further enhanced by proposing saliency filtering [62] based on spatial distribution for each color. These algorithms is failed in complex, small sized and multi-object images. The soft image abstraction approach, (GC) [68] focused to enhance global saliency by highlighting large-scale perceptually homogeneous region by using color histogram based Gaussian Mixture Model for regional decomposition. This is further enhanced by similarity-based Global Uniqueness (GU) [68] and color spatial distribution. So the global contrast based model cannot distinguish between the salient and nonsalient regions in complex image set. So to improve their performance, global contrast based model is supplemented by regional cues.

2.2.1.3 Regional Contrast-Based Models

The global contrast-based models failed in complex and cluttered images and produced internal saliency discrepancies. The input image is divided first into regions and superpixels to overcome this limitation. And these regions are used to compute regional saliencies such as spatial, depth, and color-based regional contrast. These

regional saliencies integrate to produce the final salient object, which is shown in Fig. 2.3. Cheng *et al.* [4] proposed fist regional models based on histogram. Histogram-based global contrast (HC) defined on the histogram of quantized color channels. In this method, global color contrast is computed for the most frequently used color in the form of the histogram to represent entire image. The author of this paper improved the computing by five levels of quantization of colors. This algorithm fails in most cases where the salient object is not prominent in color. This situation often occurs in complex images.

clues.

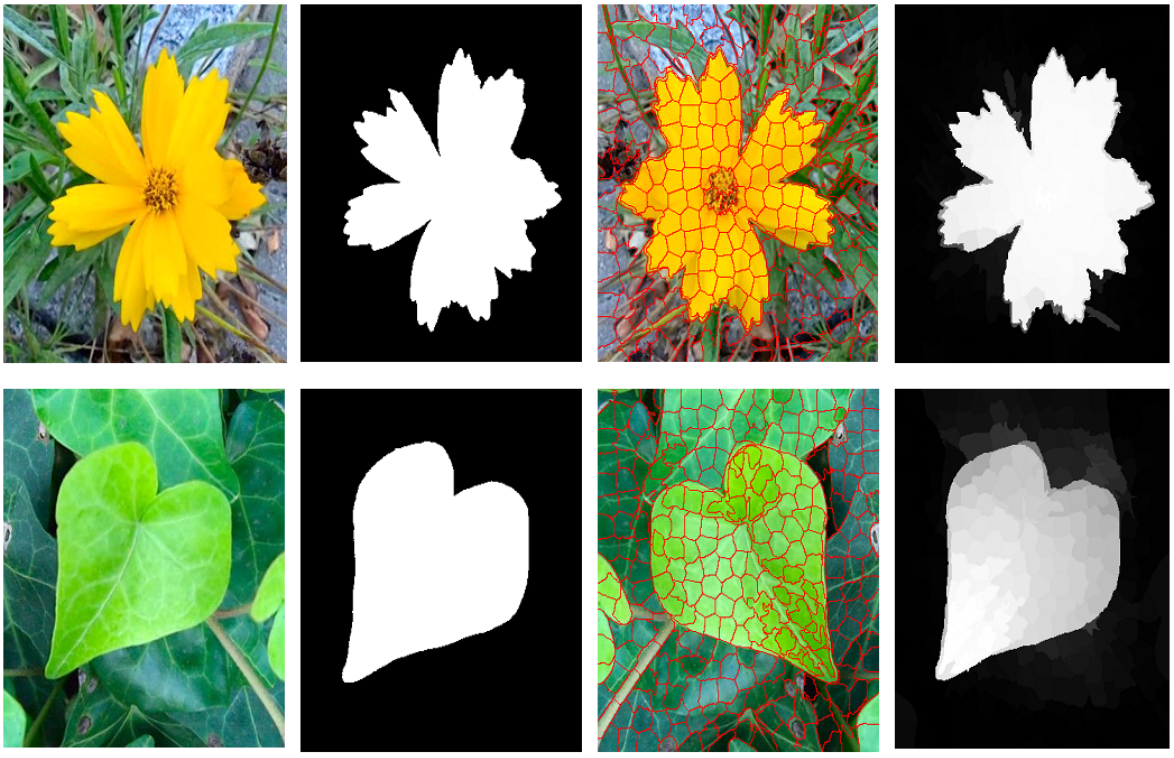

FIGURE 2.3: Regional saliency and Superpixel based salient object detection.

The same author proposed the regional saliency (RC) [4] based on the color and space contrast. Input image I divided into n region, where $R_i \quad (i = 1 \quad to \quad n)$. The regional saliency in Eq. 2.1 computed as the spatially weighted sum of contrasts between all regions, as follows:

$$S_i^{RC} = \sum_{j=1}^{n} e^{\frac{-D(r_i,r_j)}{\delta^2}} w(r_j) d_r(r_i, r_j)) \tag{2.1}$$

where $D(r_i, r_j)$ and $d(r_i, r_j)$ are regional contrast respectively by space and color while $w(r_j)$ looks at number of pixels in that region. In many complex image-sets, it suppressed the salient points to increase interior salient discrepancy. Therefore, this method required computationally expensive Graph- cut based saliency-cut algorithm to separate the objects.

Huang et al. [5] proposed spatial distribution based minimum directional global contrast for saliency computation. In this work, the author computed saliency in three steps. The first step was MDC, which was calculated based on two initial seed pixels, one from the foreground and the other from the background into four directional coordinates. The Direction Contrast (DC) of region ω for target pixel i was calculated through following equation 2.2:

$$DC_{i,\omega} = \sqrt{\sum_{j \in \omega} \sum_{ch=1}^{K} (I_{i,ch} - I_{j,ch})^2} \tag{2.2}$$

where I is original image and K is color channels in CIE-LAB space. Minimum directional contrast (MDC) is the minimum value of DC. The second step [4]is named as saliency smoothing while this step is not considered for computational efficiency as this is based on regional contrast (RC) based methods. In this method, object boundary is completely destroyed through marker-based watershed morphological segmentation because of multiple generations of the marker in marker-based watershed segmentation algorithm.

TABLE 2.1: Comprehensive details of exemplary models for salient object detection using statistics and probabilistic approaches.

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
1	Manifold Ranking Based on Robust Foreground (Ma *et al.*)IJAC-2021 [69]	Probabilistic Modelling, Preliminary saliency map based on Graph based boundary background queries followed by Manifold Ranking with background queries.	*Internal and external saliency discrepancy and failed in complex image*	**MAE/F-Measure**: ASD: 0.059/0.904 : 0.179/0.805 ECSSD: 0.159/0.721
2	Minimization of Bilinear Factor Matrix Norm (Li *et al.*) IEEE-Access-2020 [45]	Probabilistic Modelling, matrix decomposition Based on non-convex weighted Schatten-p quasi-norm using matrix factorization formulation	*Based on multiple regularization and norm, Enhanced Internal saliency discrepancy, But failed in complex image*	**MAE/AUC**: SED2: 0.169/0.804 iCoSeg: 0.179/0.805 PASCAL-S: 0.275/0.711
3	Fuzzy Theory and Object-Level Enhancement (Zhou *et al.*) IEEE-T-2019 [51]	The integration of regional saliency measure and object-level information using fuzzy theory, object proposals.	*Based on prior saliency cues And do not produce complete object Unstructured salient object with background*	**MAE**: ASD: 0.0439 ECSSD: 0.1316 PASCAL-S: 0.1710
4	Reliable boundary seeds and saliency refinement (Wu *et al*) IET-Computer Vision- 2019 [46].	Background and foreground-based map are generated based on Reliable boundary seeds and saliency refinement.	*Failed on object border and complex images, Super-pixel based graphical model make it computationally slow*	**Precision/F-Measure/AUC**: ASD: 0.934/0.912/0.863 ECSSD: 0.818/0.734/0.792 PASCAL-S: 0.682/0.585/0.769 THUR-15K: 0.600/0.570/0.822 DUT-OMRON: 0.593/0.566/0.827
5	Fusing Foreground and Background Priors (Huang *et al*) IEEE-ICIP-2018 [70]	The integration of foreground and background prior and surroundedness cue with geodesic refinement	*Failed in estimating initial foreground and background seeds in complex images Failed to produce complete salient in complex images*	**MAE/AUC**: ASD: 0.0596/0.9446 PASCAL-S: 0.1868/ 0.6917
6	300-FPS Salient Object Detection Minimum Directional Contrast (Huang*et al.*) IEEE-T-2017 [5]	Spatial distribution based minimum directional contrast,300FPS, Marker based Watershed Segmentation,	*Structure and shape loss, Failed on corner based object*	**F-Measure/MAE**: MSRA-10K: 0.790/0.207 PASCAL-S: 0.618/0.228 ECSSD: 0.612/0.262 DUTOMORON: 0.22/0.532
7	Cognitive Neuroscience (Zhu *et al.*) IEEE-MBD-2017) [71]	Color and depth features is used to obtain preliminary detection map, which further enhance with center-bias	*Exterior regional saliency discrepancy*, and *Computationally slow*	**MAE**: RGBD1: 0.1065 RGBD2: 0.1007
8	Hierarchical Contour Closure (Liu *et al*) IEEE-T-2017 [72]	A hierarchical contour closure-based holistic model. The saliency based on closure completeness,and closure reliability fused to produce final salient object	*Exterior and Interior regional saliency discrepancy, Failed in complex image*	**F-Measure/MAE/AUC**: MSRA10K: 0.846/0.097/0.964 PASCAL-S: 0.631/0.185/0.846 DUT-OMRON: 0.571/0.115/0.889

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
9	Non-convex Structured Matrix Decomposition (Zhang *et al.*) IEEE-CIS-2017 [73]	L1 norm of logistic function on the singular values of a matrix to approximate rank function, relationship between each superpixel, Laplacian regularization	*Interior and exterior regional saliency discrepancy, Produced saliency including background Unstructured salient objects*	**MAE** ECSSD: 0.178
10	Focusness guided (Xiao *et al.*) IEEE-SMC-2017 [74]	Combining fine-grained contrast prior with rough-grained object consistency to produce Focusness Guided (FGS) algorithm	*False detection, Failed in complex image, Produced exterior regional discrepancy.*	**MAE/F-Measure**: MSRA10K: 0.1107/0.761 SED2: 0.1118/0.7595 PASCAL-S: 0.1935/ 0.5597 ECSSD: 0.2206/0.5928
11	Hierarchical Image Saliency (Shi *et al.*) IEEE-T-PAMI-2016 [13]	Based on fusion of different scales using hierarchical saliency *Correctly detect large size object and*	*Failed in small and complex image Produced internal discrepancy*	**MAE**: MSRA-1000: 0.0961 ECSSD: 0.2265
12	Cellular Automata (Qin *et al.*) IEEE-CVPR-2015 [14]	A background-based map using colorand space contrast with the clustered boundar seeds is constructed and integrate multiple saliency maps with Bayesian framework *Produced center bias seliency*	*Failed in low depth images Removed the border related salient regions*	**MAE**: MSRA-5000: 0.1499 MSRA: 0.078 ECSSD:0.183 DUTOMORON:0.196
13	Global Contrast (Cheng *et al.*) IEEE-CVPR-2015 [4]	A regional contrast based algorithm, which used simultaneously global contrast differences and spatial weighted coherence. scores.	*Produce full lenght saliency Computationally fast but loss object border.*	**F-measure**: MSRA-1000: 0.9287 MSRA-10k: 0.8878
14	Depth Enhanced Saliency (Cheng *et al.*) ICIMCS-2014) [43]	Proposed color and depth based regional saliency and enhanced with center and spatial contrast to represent the 3D saliency *Widely used Depth based RGB-D, Dataset, DES has been proposed.*	*Failed in low depth images Unstructured and internal saliency loss*	**MAE/F-measure**: RGBD-135: 0.299/0.765 STEREO: 0.295/0.700
15	Graph-Based Manifold Ranking Modeling (Yang *et al.* 2013) [75]	Image represent as close-loop graph with superpixels as node and it is ranked based on the similarity to background and foreground to compute saliency.	*Produced full length saliency with background Computationally slow but loss object border.*	**MAE/F-measure**: MSRA: 0.825 DUTOMRON: 0.623
16	Contrast Based Saliency Filtering (Perazzi*et al.*) IEEE-CVPR-2012 [62]	Decompose an image into perceptually four homogeneous elements and compute the global contrast using uniqueness and spatial distribution	*Produced full length saliency including background. Produced unstructured and inconsistent salient objects*	**MAE/F-measure**: MSRA: 0.380/0.715
17	Frequency-tuned (Achanta *et al.*) IEEE-CVPR-2009 [11]	Compute Global contrast after using DoG band pass filters to fine tune in frequency domain *First model to formally recognized salient object detection.*	*Produced full length saliency including background and exterior and internal discrepancies*	**F-measure**: MSRA: 0.7209

2.2.1.4 Background Approximation Based Models

The background measure-based methods have been proposed and achieved remarkable success. The backgroundness measure is based on photographic psychology. As per this principle, the regions related to the image border are considered as background. Inspired by this hypothesis, Wei *et al.* [76] proposed a geodesic distance-based identification of background and foreground regions. Similarly, Strand *et al.* proposed a minimum barrier distance MBD [77], which is used to measure the backgroundness. It is used to compute the distance of regions with the image border. Another graphical model is based on a minimum spanning tree (MST) [42]to produce salient objects based on backgroundness.

After that, various saliency detection models, based on backgroundness have been proposed. For example, Yang *et al.* [75] proposed the backgroundness of border regions by using the superpixels-based regions. The image divided into superpixels to produce a coarse salient object. This coarse saliency is further processed with various post-processing methods to produce the final salient object. The next representative model is based on graphical representation of image super-pixels. Graphical Ranking saliency (GMR) [75] is semi-supervised and based on the learning and ranking of regions. The SLIC based region is used in two-stage scheme for bottom-up saliency detection using ranking with background and foreground queries to distinguish between salient and non-salient super-pixels.

A recent hierarchy-based saliency estimation model (CS) addresses the issues of

complex background images and proposed a challenging dataset, named as Extended CSSD or (ECS) [13]. This hierarchical model focused on the analysis of region scale computation followed by region-merge approach. It is a graphical model for integrating the three levels of hierarchical regions by using local Euclidean distance based consistency. So, saliency improvements are continued by using region-based heuristics. This regional approach has increased the computational cost. These methods improved the interior saliency discrepancy, but failed on the boundary of an object in the complex scene and increased the exterior saliency discrepancy.

All the global contrast-based methods are computationally efficient, while this algorithm mainly suffers from the exterior and interior regional discrepancy. This is the main objective of multiple integration. The proposed methods used the region based saliency enrichment. The main focus is to increase the salient points in the interior, while suppressing the exterior saliency into a well defined concave topographical surface, which has failed in above mentioned models. Ranking saliency (GMR) [75], and BSA [14] used color and regional saliency by using SLIC based super-pixel, rather than color-based region. ECS [13] proposed a region generation algorithm, and proved that color based region generates better color saliency than super-pixel based on the object boundary. So, we used K-mean based color region for maintaining the object boundary characteristics. The super pixel-based approach improved the interior saliency, so we used super-pixel based interior saliency in saliency enhancement.

There are some other methods [78], [12], [79] which use the same principle to differentiate salient objects by using background measures. Although with great success in accurately detecting a salient object in simple and large-sized object, these models failed in complex and cluttered backgrounds where salient and non-salient regions are similar. Furthermore, these models are computationally slow due to extra backgroundness measures.

2.2.1.5 Psychology Inspired Models

The recommendations and suggestions of psychologists emphasized that the figures are more attended than the background. Based on this principle, salient object detection is formulated as a figure-ground separation problem. The various saliency cues like surroundedness [80], focusness [81], objectness [82] and convexity [83] were proposed to formulate the principle of separation of figure-ground. For example, Zhang et al. [80] proposed a Boolean map-based saliency computation model, which utilized the surroundedness cue. A set of binary images (Boolean Maps) was constructed using arbitrary thresholding on color channels. The topological structure of Boolean maps is used to compute salient object detection. Lu et al. [83] proposed a convexity context-based saliency model. This model produced the concave arcs, which were related to each contour with corresponding superpixels that define the concavity context window. This window encoded the figural(salient) and background objects. The focusness prior proposed by Jiang *et al.* [81]. The author defined focusness as a degree of focal blur. The de-focus blur was estimating the standard

deviation of the Gaussian kernel by scale-space analysis. Finally, the saliency values comprised a non-linear combination of global contrast, objectness, and focusness scores. Chang et al. [84] proposed a graphical model to integrate the objectness and regional saliency to produce salient object detection. The integration strategy with the above two parameters is jointly optimized by iteratively minimizing the energy function. Jang et al. [81] defined *regional objectness* as the average saliency of each region, which guides the regional saliency computation. These models improved the saliency of an image having simple, single, and large-sized objects. At the same time, these models are failed complex and cluttered backgrounds. Based on this dominant statistical clue, there are various efficient model have been proposed which are summarised in the Table 2.1.

Various other models which have been developed by researchers, to use the statistical and probabilistic properties of an image. Most of the researchers use ensemble feature because any particular low level features are not enough to predict the salient object. The problem of salient object detection has been approached in different ways. Some of the models view the problem as a segmentation framework. A few models try to locate objects in the image and then classify them for saliency. Prior based models are mostly found because they simplify the location of the salient object to some extent. They also correspond more to the human visual attention system. Various graph-based techniques have also been proposed for finding the salient object. Sparse representations are also used in detecting salient objects. Researchers have also used the matrix decomposition model where the feature matrix

is decomposed to reflect salient and non-salient regions.

Superpixel algorithms allow shallow segmentation of images. Borji *et al.* [85] used superpixels to get an idea about the complexity of the image. Simple images have less number of superpixels than a complex image. Some of the researchers emphasize using *focus* as an important prior for finding the salient object. The assumption is that salient objects in an image are always in focus. Thus, non-focused objects can be marked as background. Sun *et al.* [86] develop a sparse dictionary based on focus prior map. *background and foreground.* When the image is segmented as per these criteria, foreground regions are then processed to extract the salient object. Huang *et al.* [41] select foreground seeds based on surroundedness cue. Further, geodesic refinement is applied to the foreground region for salient object detection. Wang and Liu [87] separate background using geometrical interpretation.

Objectness proposals are also an interesting method to find salient objects. In these methods, the models generally first try to find as many complete objects they can find in an image and, after that, finalize a salient object among them. Zhang *et al.* [88] use objectness map and fuse it with the saliency map generated by the Markov chain using background seeds. Zhou *et al.* [51] uses fuzzy theory to integrate various regional saliency maps and objectness proposals.

Boundary is an important prior used by various models. The assumption is that generally, salient object lies away from the boundary. The boundary marks the background region of an image. Using this assumption, the background can be separated from the image, and this salient object can be extracted. Abkenar *et al.* [89] use

distribution-based boundary contrast map. The graph representation of the image is used to compute the connectivity of the image regions to the image boundary as well as to their local neighbors and the image foreground. Connectivity maps obtained are then fused with the boundary contrast map. Huo *et al.*

Another approach of locating salient objects is to use *graphs*. In using graphs, initially, some seed values are assigned that become the nodes, and edges connect these nodes. The nodes and edges are updated to converge to find the salient object. Nouri *et al.* [90] use contrast as a feature to generate graph. A threshold for edge weight is used to eliminate non-salient edges. Li *et al.* [91] generates a hypergraph using information from a pixel's similarity to its neighborhood and its dissimilarity from the background. Wang and Lv [92] use eye fixation prediction as a prior to apply graph-cut on an input image. Filali *et al.* [93] proposed multi-scale graph ranking and iterative local–global object refinement based graphical model. multi-layer graph is constructed based on superpixels, and optimized to diffuse saliency from image borders to salient objects. The iterative random forests and local boundary refinement using color, texture and edge to improve and enhance the final saliency.

Contour-based models have been primarily used for segmentation. The advantage of using them are getting neat and clear object boundaries. Du and Chen [94] use patch rarities to form the object contours. The patch rarities are computed by using the random forest. Liu *et al.* [72] use completeness and closure measure as saliency cues to form the contour of the object. Zhang *et al.* [73] build a nonconvex structure

matrix decomposition model. They explore the relationship between each super-pixel to make the salient object highlighted consistently. Laplacian regularization is used to increase the distance between salient regions and non-salient regions in feature space. Peng *et al.* [56] use structured matrix decomposition model with two structural regularization.

Spatial and Center prior from an image also helps to locate the salient object accurately. Zhang *et al.* [95] use spatial priors in addition to color and central bias to construct the graph. The graph is built by connecting nodes that are spatially close in an image. The edge weights are provided based on color similarity and spatial proximity. Qin. *et al.* [14] proposed as Cellular Automata based salient object detection model. A background based coarse map using color, space and regional contrast computed on super pixel of image with cluster boundary seeds. A factor matrix and coherence matrix have been designed to refine the cell saliency based on regional similarity and dissimilarity. Finally, Bayesian framework based integration framework to fused multiple saliency maps. Wang *et al.* [96] use spatial relationship to develop a geodesic weighted Bayesian model. A comprehensive analysis of highly referenced, and widely used methods is given in Table 2.1.

2.2.1.6 Summary

The statistical or probabilistic models provided the first well-recognized computation domain since 2007 and produce enormous saliency methods. The most contributing models are based on global contrast, Regional contrast, and background

approximation-based models. The merits, demerits, and summaries of the models mentioned above have been described here.

1. These contrast-based methods enrich the regional saliency on the cost of exterior regional discrepancy means a region having similar outer border region with background region can be enhanced rather than suppress in contrast computation.

2. The major drawback of the above methods is failed to produce the saliency in an image with a salient object with multiple colors regions and low depth and structural similarity of salient and non-salient regions in complex and clutter images.

3. The global contrast-based methods produce full-length saliency with exact object boundary, including backgrounds and missing some internal-salient regions.

4. The regional saliency produced better internal saliency while it destroyed the boundary of salient object and produced unstructured and inconsistent salient object.

5. The background elimination methods improved the performance by using various background approximation principle. At the same time, they destroyed the salient regions which are similar to backgrounds.

6. These entire 2-D models are failed in low-depth images. Distance-based spatial contrast is not sufficient to recognize salient objects accurately. These are the regions for proposing 3D saliency.

These models improved and contributed to saliency prediction while insufficient to create a salient object in complex and cluttered backgrounds. The probabilistic models based on global reference surface have been proposed to integrate the regional, spatial, depth, and background elimination based saliency to address these limitations.

2.2.2 RGBD(3D) Statistical Models

The invention of the depth-based sensor camera provides depth and spatial information, which ia absent in the RGB-based 2D plane. This RGBD-based 3D information adds a different and recent stream in the saliency computation domain. This depth-based information is the primary objective of retrieving the distance-based depth and spatial information using various statistical and deep learning models, which is shown in Fig. 2.4. The deep learning-based model is discussed separately in the next section. The enormous attributes based on depth information such as regional distance map, color, and spatial contrast, boundary cues, shape attributes, and geometrical information are gradually explored to improve the exact salient object detection in complex and cluttered backgrounds.

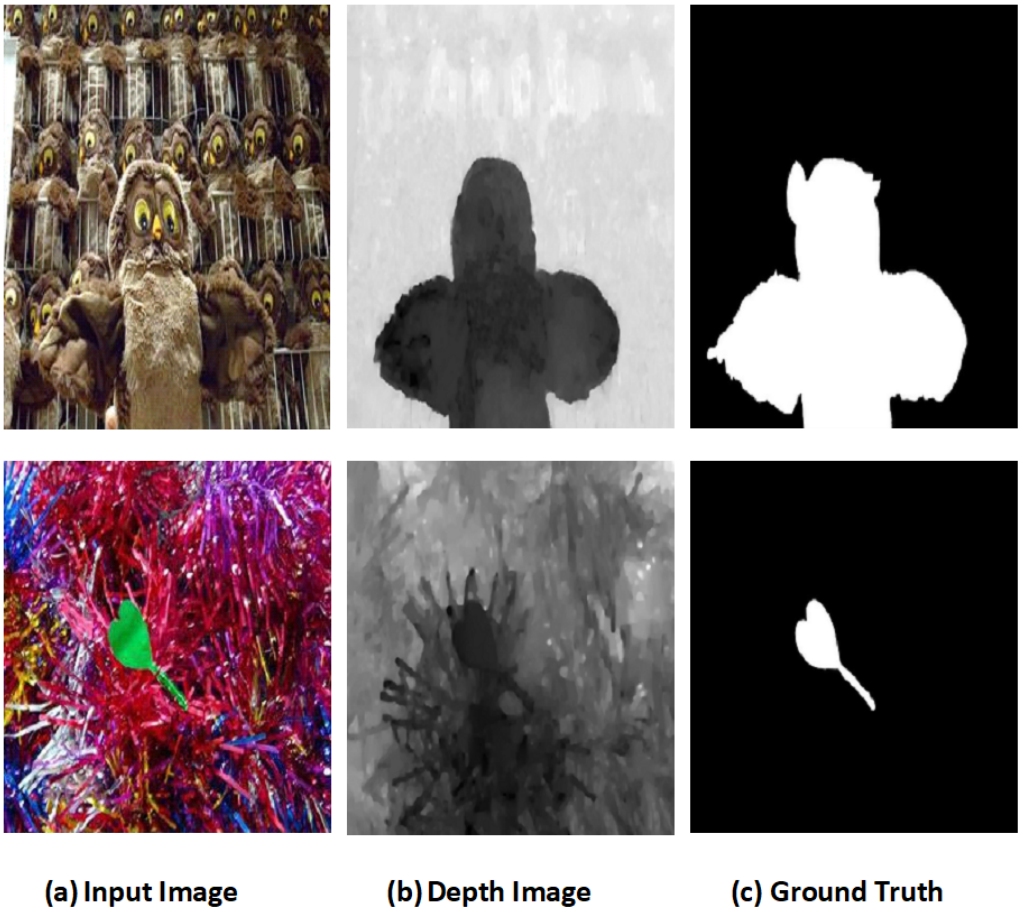

FIGURE 2.4: The 3D saliency and Depth Information.

These entire 2-D models are failed in low depth images. The depth clues in RGBD saliency provides an additional space to increase the saliency in low depth 3D images and it has discriminative power against the complex and clutter background. Therefore, to capture the low depth features RGBD saliency model is used. RGBD saliency model capture low level depth information while itself not sufficient for salient object detection. Therefore, Cheng *et al.* [43] used color and structural based regional saliencies for saliency. But this model produces saliency which minimized exterior saliency and improved the interior saliency but failed on border region discrepancy. Zhu *et al.* [38] used dark channel, central channel and other purifying

saliency feature along with depth features.

Niu *et al.* [97] first time used a stereoscopic device to design the STEREO dataset. Depth maps are constructed based on the disparity between the left and right views of a stereoscopic image in this work. Using this dataset, some researchers [98], [99] computed the depth-based spatial and color contrast to produce the saliency maps. A different approach from above has been proposed by Li *et al.* [100] to utilize the depth information in light field images. Afterward, Cheng *et al.* [43] designed the DES dataset collected from the Kinect device. In this model, various feature contrast like regional contrast , spatial contrast, color contrast, and depth contrast used to produce salient object detection in complex and cluttered background. Similarly, Peng *et al.* [98] proposed NLPR dataset and RGBD based salient object detection model using depth-based 3D information.

The advent of depth-based RGBD datasets stimulate the various RGBD based salient object detection(SOD) models. Ren *et al.* [101] proposed global prior, like anisotropic center surround difference based SOD computed on distance based saliency from surroundings. Subsequently, Feng *et al.* [102] proposed re-weighted the local background enclosure-LBE based saliency using depth and spatial contrast to separate the background and foreground. Guo *et al.* [103] proposed two saliency maps based on color and depth contrast. These two saliency maps were finally fused and refined using cellular automata-based central saliency. Similarly, Wang *et al.* [104] used minimum barrier distance to optimize the final saliency through depth-map-based depth saliency, depth bias, and 3D prior. The various pre-segmentation and

multi-level saliency maps from color image and depth map fused to produce RGBD saliency by Song *et al.* [105] Similarly, Cong *et al.* [106] used a transformation strategy to integrate depth-map-based depth cues into color-based regional saliency.

Zhu *et al.* [38] proposed an RGBD based center-dark channel prior based salient object detection. Color and depth based regional saliency fused with center and dark channel based saliency to produce the final salient object. Song *et al.* [105] proposed a depth enhanced model using multiscale discriminative saliency fusion and bootstrap learning for RGB images and depth images. In this model various low-level feature contrasts, mid-level feature weighted factors, and high-level location priors are combined using a random forest regressor to learn the discriminative saliency feature maps. Finally, these saliency maps are optimized by using support vector machine. Liang *et al.* [107] a 3D SOD based on contrast and depth-guided background prior. These prior assumed boundary super-pixels as background. Finally, color, spatial and regional contrast, with 2D spatial dissimilarity features, are further utilized to refine the final saliency map.

TABLE 2.2: Comprehensive details of Top Performing conventional 3D model of salient object detection.

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
1	Contrast and depth guided background prior (Liang *et al.* Neurocomputing-2018) [107]	3D stereoscopic saliency model based on both contrast and depth-guided-background prior. Fused these features by measuring contrast of disparity and color	*Efficient in removing background while produced internal and border region discrepancies.*	**MAE**: NLPR:0.114 NJUD:0.203 DES:0.100 STERE:0.166 LFSD:0.225
2	MDSF:Multiscale discriminative saliency fusion and bootstrap learning (Song *et al.* IEEE-TIP-2017) [105]	Utilizing low-level feature contrasts, mid-level feature weighted factors, and high-level location priors through random forest regressor based are learned discriminative saliency fusion (DSF) model to DSF saliency maps across multiple scales.	*Used muti-stage fusion model while missing global integrating reference surface. Produced internal discrepancies and shape loss in complex low depth 3D images*	**MAE**: NLPR:0.095 NJUD:0.157 DES:0.122 STERE:0.176 LFSD:0.197
3	Multilayer backpropagation based on depth mining (Zhu *et al.* ICCAIP-2017) [108]	A multilayer back-propagation saliency detection algorithm based on depth mining by which exploit depth cue from four different layers of images	*Back-propagation model is not fully utilised on RGB and depth features. Inconsistent in background estimation, Structural and shape loss in complex 3D images*	**MAE**: NLPR:0.089 NJUD:0.202 DES:0.102 :0.178 LFSD:0.218
4	CDCP:Center-dark Channel Prior. (Zhu *et al.* ICCAIP-2017) [38]	Proposed a Super-pixel based center prior, dark prior, and a innovative two stage fusion model to predict RGBD saliency	*Used multiple saliency cues and missing global integrating reference surface. Baised on the central regions. Inconsistent in complex 3D images*	**MAE**: NLPR:0.114 NJUD:0.181 DES:0.119 STERE:0.149 LFSD:0.199
5	Depth Confidence Analysis and Multiple Cues Fusion (Cong *et al.* IEEE-SPL-2016) [109]	Depth based graph is designed to measure the color and depth-based compactness. The detection errors in compactness are minimized in background and foreground saliency and finally fused to predict the saliency.	*Used multiple cue based fusion while missing global integrating reference surface. Produced internal discrepancies and shape loss in complex 3D images*	**MAE**: NLPR:0.196 NJUD:0.167 DES:0.194 STERE:0.150 LFSD:0.155
6	ACSD:Anisotropic center-surround difference (Ju *et al.* ICIP-2014) [110]	Depth images based on anisotropic center-surround difference (Difference of Gaussian (DoG)) and global contrast with the surrounding method for 3D saliency.	*Lacking in extracting Depth based saliency cuesand, Inconsistent in features integration of 3D with 2D*	**MAE**: NLPR:0.1386 NJUD:0.1072 DES:0.0701 STERE:0.1318

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
7	Saliency detection for stereoscopic images (Fang *et al.* IEEE-TIP-2014) [99]	Color, luminance, texture, and depth features are extracted, and the Gaussian model is used to compute local and contrast-based RGBD saliency.	*Lacking in extracting Depth, spatial and regional based saliency cues and, Inconsistent in background estimation, Produced unstructured saliency in complex 3D images*	**MAE**: NLPR:0.1319 NJUD:0.1124 DES:0.1334 STERE:0.1147
8	DES:Depth enhanced saliency detection method (Cheng *et al.* ICIMCS-2014) [43]	Proposed a Super-pixel based regional color, spatial and depth contrast based RGBD salient object detection. proposed DES Data-sets and highly referenced in literature	*Used multiple contrast and and missing global integrating reference surface. Produced unstructured saliency in complex 3D images*	**MAE**: NLPR:0.301 NJUD:0.448 DES:0.289 STERE:0.417 LFSD:0.415
9	NLPR:RGBD A Benchmark and Algorithms (Peng *et al.* ECCV-2014) [98]	Proposed RGBD depth enhanced low-level feature contrast, mid-level region grouping, and high-level priors enhancement features for 3D salient object detection. They also design a large-scale RGBD dataset.	*Used muti-stage fusion model while missing global integrating reference surface. Produced internal discrepancies and shape loss in complex 3D images*	**MAE**: NLPR:0.119 NJUD:0.201 DES:0.097 STERE:0.179 LFSD:0.211

2.2.2.1 Summary

The 3D-based RGBD models enhanced the performance of salient objects at another level. It used distance-based saliency using a depth map to complement the RGB features. The spatial, regional, depth, color, background, and structural-based features are integrated to overcome the challenges of 2D saliency. Although, these models failed in the complex and cluttered background because the integration was only based on the simple addition of various saliency cues. A global reference surface is missing, which incorporates all the saliency cues to produce the exact salient object. These models produced inaccurate saliency in complex scenarios like an image having a salient object with multiple colors, regions, low depth, and structural similarity of salient and non-salient regions in complex and cluttered images. But, the main advantage of these models are to produce full-length global saliency. Consequently, a global reference surface has been proposed to integrate all regional saliency cues and produce an exact salient object to overcome these challenges.

2.2.3 RGBD(3D) Deep learning Models

Convolution neural networks have achieved tremendous success when Krizhevsky *et al.* [111] proposed an 8-layer model. It is recognized globally after winning the ***ImageNet: Large Scale Visual Recognition Challenge (ILSVRC)*** in 2012 with a very high margin. Subsequently, many models [112], [113], [114] with 8 to 350 deep layers have been proposed and used in all areas of computer vision. Within

a short span, various deep SOD frameworks from different perspectives have been proposed, including network architecture, level of supervision, learning paradigm, and object-/instance-level detection. The initial deep SOD models mainly utilized the multi-layer perceptron (MLP) [115] to detect the salient object by the saliency score of deep features. In contrast, the recent approaches utilized the fully convolutional network (FCN) [116], [30] to design SOD architectures. Many variations in Simple CNNs have evolved into Fast R-CNN and Mask R-CNN. A variety of network models - AlexNet [117], ResNet [118], GoogleNet, inception architecture [119], and VGGnets [120] are available for implementing the algorithms. General Adversarial Networks (GANs) [121] are brought into play a newer variety of these networks. Deep learning-based models are classified according to the modalities. Some models preferred RGB-based deep features and used the week supervision learning paradigm. At the same time, the recent approach is based on RGBD(3D) saliency with end-to-end supervision for salient object detection in complex and cluttered backgrounds. The outstanding models have been summarized in Table 2.3 A variety of models have been developed, which are discussed below.

2.2.3.1 Multi-Layer Perceptron (MLP) Models

MLP based models extract deep features from superpixels [16] or Patches [122] as inputs units to produce the corresponding score. Afterward, MLP-classifier used these input units based on saliency scores on predicting salient object detection. He *et al.* [16] proposed superpixels based on two low-level statistic features and

corresponding CNN streams to extract deep features. These models initiate the Deep learning-based approach to salient object detection. MLP-based models improve the performance of SOD over non-deep learning-based models. At the same time, they cannot fully apply the deep CNN model to extract spatial, high-level semantic, and contextual information.

2.2.3.2 Fully Convolutional Network (FCN) Models

The limitations of MLP-based models became the motivation to design the FCN architecture [123]. It uses pixel-level information without decomposing the input image into regions or superpixels to overcome the limitation in MLP-based models. These models extract high-level semantic and spatial information. The architectures design of FCN-based models has been revolutionalized in recent times and categorized as: single-stream, multi-stream, side-output/skip connections, bottom-up/top-down, and branched networks models. These models are heavily researched and dominate over other models. Hou *et al.* [124] introduces short connections to the skip-layer structures within the Holistically-Nested Edge Detector (HED) architecture. In this model, multi-level and multi-scale features extracted from Fully Convolutional Neural Networks (FCNs) are utilized in the multistage decoder. These models add another dimension to deep learning-based models. Due to the emergence of various 3D-based RGBD datasets, the recent focus shifted to 3D salient object detection.

2.2.3.3 RGBD 3D SOD Models

The emergence of depth cameras such as RealSense and Kinect 3D saliency (RGBD, D for "depth") has become a more attractive new stream of 3D saliency computations. Numerous research papers have been published in RGB SOD models, while RGBD based model is a current stream of saliency computations that utilize depth and RGB-based deep saliency features. Some works have been done in RGBD models, while some challenging issues still make it appealing for saliency computations. Qu *et al.* [125] proposed the first model of deep fusion based CNN for RGB-D SOD in 2017, which utilized shallow level features from both modalities to predict salient objects. Han *et al.* [126] proposed two stream model(RGB, and Depth) to extract the corresponding features and simultaneously fully connected layer-based fusion model to obtain the final saliency. Chen *et al.* [116] proposed a progressive fusion model to fuse the RGB and depth-based multi-scaled CNN features. Fusion models used skip-connection to integrate low-level to high-level CNN features. Architectural point of view RGBD Deep CNN model is classified into a single stream and multi-stream models. Some recent models used the Attention Mechanism to enhance the CNN features. The details analysis of RGBD SODs models has appropriately been studied in next following sections.

2.2.3.4 Single-stream Models

The early models focused on designing [127], [105], [128], [63] a single-stream network to extract the saliency features to predict the salient object. In these models, input image is composed by concatenation of RGB and depth frames. Shigematsu *et al.* [128] developed a single stream CNN model to integrate bottom-up, and top-down CNN features to predict the salient object. This model refined the final saliency by fusing background enclosure distribution and histogram-based global and depth contrast features. Similarly, Wang *et al.* [63] presented the depth contrast-based subsidiary network to guide the mainstream network to predict saliency. Single-stream models failed to extract cross-complementary features, which is essential for predicting the salient object in complex and cluttered backgrounds. It also failed in producing salient object low depth images.

2.2.3.5 Multi-stream Models

Designing a multi-streams model [45], [129], [130] aims to effectively and efficiently extract and fuse the cross-complementary features to produce another high-level contextual and semantic features for middle stage or end to end fusion. Consequently, most recent models and current researchers are working on these multi-stage and multi-scale cross-complementary fusions to predict the exact salient object in complex and cluttered backgrounds. Zhao *et al.* [64] proposed a contrast-enhanced depth map, which is fed into subsequent fluid pyramid integration. Fan *et al.* [130] proposed

a depth-depurator unit to eliminate the low depth CNN features during fusion with RGB features. The same authors also designed SIP (Salient Identification Person) RGBD dataset. Fu *et al.* [131] proposed a JL-DCF model for joint learning through cross-complementary features. The RGB and Depth based complementary features are fused with a siamese network-based densely cooperative fusion. Similarly, Pang *et al.* [132] utilized densely connected structures to explore cross-complementary and refined the final saliency through a hierarchical dynamic filtering network. The multi-stream model extracts the multi-resolution and hierarchical features in RGB and Depth modality. The recent models proposed various fusion methods to fuse the cross-complementary features.

2.2.3.6 Fusion Model

The cross-complementary fusion model is another parameter to classify Deep CNN models into three categories: (1) Early Fusion Model,(2) Late Fusion model, and (3) Middle Fusion model.

Early Fusion Model The early fusion-based model [101], [98], [105] directly concatenates the RGB and depth into four input channels, which are collectively fed into a designed network to extract CNN features. Some other model [116] [64] independently extract RGB and Depth shallow CNN features on low-level, handcrafted features. These models of fusion do not fully utilize CNN. Consequently, the middle and late fusion-based models have been preferred in cross-complementary fusion.

Late Fusion Model The late fusion model is also classified into a) Later fusion strategy and b) Late fusion strategy. First CNN model [126] extracts RGB and Depth stream, multi-stage saliency features that concatenate simultaneously and use to predict salient objects further. In contrast, the Late fusion strategy [133] produces RGB and Depth saliency, which combine to produce final salient object detection.

Multi-scale Fusion Model This strategy is the most efficient and widely preferred recent fusion models [134]. It is used to explore the correlations among cross-complementary features. It is also divided into two categories. In the first strategy, the cross-complementary interactions between RGB and depth CNN features are further processed into another learning network. Chen *et al.* [129] designed a multi-scale multi-path fusion network to integrate both modalities. Then, the final fusion process explores cross-complementary between low and high-level semantic features. The second and current strategy is to extract RGB and Depth based CNN features which are further processed through skip connection into a specifically designed Decoder for the final fusion of cross-complementary features. For example,Fan *et al.* BBS-Net [135] proposed a bifurcated backbone strategy to segregate the multi-level feature into teacher and student features. These features are purified and enhanced by using spatial and channel-wise view through a depth-enhanced module.

TABLE 2.3: Comprehensive Survey of some Deep learning based RGBD Salient object detection model

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
1	UC-Net: Uncertainty Inspired Conditional Variational Autoencoders (Zhang *et al.* CVPR-2020) [136]	Conditional variational autoencoders are designed for generating multiple probabilistic saliencies for each input image using sampling in the latent space. Used a depth correction network to decrease noise in final saliency. Improved the performance due to modeling of uncertainty in the saliency domain	*It produced multiple annotations, which improved the learning process, while all other methods are based on a single ground truth map.*	**MAE**: NLPR: 0.025 NJUD: 0.043 DES: 0.019 LFSD: 0.066 SIP:0.051
2	S2MA:Learning Selective Self-Mutual Attention (Liu *et al.* CVPR-2020) [137]	Non-Local Network-based self-attention model for long-range contextual dependencies to enhance deep localized features and fused using a proposed residual fusion-based decoder.	*Enhanced the performance by improving deep localized features. Failed in complex and challenging images. Because cross-complementary and modality specific features are not exploited*	**MAE**: NLPR: 0.030 NJUD: 0.053 DES: 0.021 LFSD: 0.094
3	JL-DCF:Joint Learning and Densely Cooperative Fusion Framework (Liu *et al.* CVPR-2020) [137]	The joint learning (JL) module provides robust saliency feature learning, at the same time, the densely cooperative fusion(DCF) introduces for complementary feature discovery. Improved the performance by exploiting cross-complementary features.	*Failed in some images that are distinguishable using modality-specific features and improved deep features, which are absent here.*	**MAE**: NLPR: 0.022 NJUD: 0.043 DES: 0.022 LFSD: 0.078 stere:0.042
4	cmSalGAN:Cross-View Generative Adversarial Networks (Jiang *et al.* IEEE-T-M-2020) [138]	An optimal view invariant and consistent pixel-level representation for RGB and depth images via a novel adversarial learning framework	*The inconsistency of GAN encoded features and fusion model sometimes leads to inaccurate prediction in low depth and occlusion-based images.*	**MAE**: NLPR: 0.0267 NJUD: 0.0462 STERE:0.0496
5	D3NET: SIP Data-set (Fan *et al.* IEEE-T-M-2020) [130]	The Depth-Depurator Network (DDU) and three-stream feature learning module (FLM) are designed to filter out low depth issues and cross-modality fusion.	*The inconsistency of DDU unite, because it is based on only fixed threshold. Cross-complementary and improved deep features is absent.*	**MAE**: NLPR: 0.0267 NJUD: 0.041 STERE:0.046 DES:0.031 SIP:0.063
6	SSF:Select, supplement and focus (Zhang *et al.* CVPR-2020) [139]	Global location, fine edge, and local detail in complementarities view from two modalities are fused by designing a complementary interaction module (CIM).	*Region-wise features selection and channel-wise attention mechanism improved the performance while failed in low-depth and disrupted depth map.*	**MAE**: NLPR: 0.026 NJUD: 0.04 STERE:0.044 DES:0.025 LFSD:0.066

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
7	PGANet:Progressively Guided Alternate Refinement (Chen *et al.* ECCV-2020) [129]	A lightweight depth stream by learning from scratch FOR extracting complementary features more efficiently with less redundancy and fed into guided residual (GR) blocks to reduce their mutual degradation.	*It is failed in the complex image due to the simple Fusion model, which is based on concatenation and simple element-wise addition.*	**MAE**: NLPR: 0.024 NJUD: 0.042 STERE:0.041 DES:0.028 LFSD:0.074
8	DANet:Single Stream Network (Chen *et al.* ECCV-2020) [127]	A single stream model to directly use the depth map to guide early fusion and middle fusion between RGB and depth for lightweight and real-time model.	*Failed in complex images and multiple object images that are distinguishable using modality-specific features a improved deep features, which are absent here.*	**MAE**: NLPR: 0.028 NJUD: 0.045 STERE:0.041 DES:0.023 SIO:0.054
9	cmMS:Cross-Modality modulation and selection (Li *et al.* ECCV-2020) [134]	A cross-modality feature modulation (cmFM) is designed to enhance features by taking the depth features as prior. An adaptive feature selection (AFS) module and saliency-guided position-edge attention (sg-PEA) are used to select saliency-related features and suppress the inferior ones.	*Failed in complex images and multiple object, that are distinguishable using modality-specific features. AFS module used features concatenation which leads inaccurate prediction in complex image*	**MAE**: NLPR: 0.028 NJUD: 0.040 STERE:0.039 SSD:0.050 LFSD:0.064
10	Asymmetric Two-Stream Architecture (Zhang *et al.* ECCV-2020) [140]	A flow ladder module (FLM) for the RGB stream and a depth attention module (DAM) is designed for the depth stream to fully extract global and local information in both streams for accurate prediction of the salient object.	*Failed in complex images and low depth images used two separate backbones, which caused synchronization issues and produced an unstructured salient object.*	**MAE**: NLPR: 0.0273 NJUD: 0.0442 STERE:0.0422 SSD:0.0524 LFSD:0.072
11	A2dele: Adaptive and attentive depth distiller (Piao *et al.* ECCV-2020) [141]	A depth distiller (A2dele) is designed for prediction and attention to transfer the depth-based features to the RGB stream only in the training phase to speed up the testing process.	*Failed in complex images and low depth images. Produce unstructured and border region discrepancies*	**MAE**: NLPR: 0.028 NJUD: 0.051 STERE:0.043 DES:0.043
20	CAS-GNN:Cascade Graph Neural Networks (Luo *et al.* ECCV-2020) [142]	A unified framework of a set of cascade graphs is proposed to learn powerful representations of RGB and Depth features and a novel Cascade Graph Reasoning (GCR) module to learn powerful, dense features to predict saliency maps.	*Using a high-level graphical correlation model to improve the performance while producing unreliable and unstructured saliency in challenging situations such as occlusions and ambiguities.*	**MAE**: NLPR: 0.025 NJUD: 0.051 STERE:0.039 DES:0.028
12	CPFP:Contrast Prior and Fluid Pyramid Integration (Zhao *et al.* ICCV-2019) [64]	They utilize contrast prior ,which is heuristics-based method, and use it into CNNs to integrate it with RGB features through fluid pyramid integration.	*This model is based on an early fusion-based method that does not fully utilize CNN features.*	**MAE**: NLPR: 0.037 NJUD: 0.053 DES:0.036 LFSD:0.088

S.No.	Model Pub-Year	Novelty/Techniques/clues	Disadvantage/Limitations	Summary of Result
23	AF:Adaptive Fusion (Wang *et al.* IEEE-ACCESS -2019) [133]	An adaptive fusion scheme using switch map to fuse RGB and Depth saliency. It is also proposed a three-loss function to refine the final fused saliency.	*It is based on the late fusion model, and cross-complementary fusion is absent.*	**MAE**: NLPR: 0.0327 NJUD: 0.0534 STERE:0.0462
13	CTMF:Cross-View Transfer and Multiview Fusion (Han *et al.* IEEE-ACCESS -2019) [126]	RGB and depth-based deep neural network is designed for the deep representations of both views. A multiview CNN fusion is introduced to fuse RGB view and depth view to improve saliency prediction.	*It is failed in complex images due to inconsistency in cross-modality fusion. Structural and regional discrepancies loss.*	**MAE**: NLPR: 0.0756 NJUD: 0.0453 STERE:0.0742 DES:0.0390
14	Progressively Complementary-aware Fusion Network (Chen *et al.* CVPR-2018) [116]	Cross-modal residual functions and complementary-aware supervisions are introduced in each CA-Fuse module to learn complementary information.	*It is failed in the complex image due to the simple CA-Fuse model, which is based on concatenation and simple element-wise addition.*	**MAE**: NLPR: 0.044 NJUD: 0.059 STERE:0.064 DES:0.0490 LFSD:0.119
15	Deep Fusion (Qu *et al.* IEEE-TIP-2017) [125]	The first convolutional neural network (CNN) fuses different low-level saliency cues into hierarchical features for automatically detecting salient objects in RGBD images.	*The CNN features are missing based on low-level features and cross-complementary features.*	**MAE**: NLPR: 0.100 NJUD: 0.151 STERE:0.141 DES:0.120 LFSD:0.142

2.2.3.7 Skip-Connection based models

The encoder and decoder's various stages of interaction have been achieved using Long-range skip connections or side outputs to recover image details in pixel-level prediction tasks. It has been preferred in the most recent RGBD salient object detection models. The stage-wise CNN features from the encoder to the corresponding decoder have been established for cross correlation-based fusion. Cross-correlation fusion performs through the features produced by skip connection or side outputs. The fusion performs by using simple feature-wise multiplication addition or concatenation. Chen *et al.* [143] first time formulate an efficient deep network for salient object detection using deepest layer residual learning to learn side-output for saliency refinement with less number of CNN parameters. This model also proposes a reverse attention module to erase the current predicted salient regions from side-output features for residual learning. Consequently, it achieved noticeable improvements in saliency prediction. Then after numerous models have been proposed to exploit the skip connection or side outputs. For example, Piao et al. [144] use skip connections and corresponding features to design depth refinement block. This block is used to integrate cross-complementary multi-resolution RGB and depth features. Finally, a dense decoder is designed for this work. The performance improves by densely connecting the higher-level side outputs from the encoder. Li et al. [134] develop a cmMS (cross-modality modulation and selection) to integrate the side outputs to extract cross-correlation features in a coarse-to-fine way. In JLDCF, the authors propose a cross-modality fusion block to incorporate the side outputs. This model first

compressed the concatenated side outputs to a fixed number of channels, followed by cross modality fusion to generate the final salient object.

2.2.3.8 Attention mechanism

The attention mechanism is the internal intelligence of the human vision system, which describes focusing or emphasizing the most attractive components in the visual scene. The computer vision system developed the algorithm to formulate visual media's most essential and prominent part and is recognized as ***Visual Saliency***. The prominent and conspicuous part of visual media is emphasized through the tendency or intelligence is invented as the ***Attention mechanism***. The visual saliency was later developed as a salient object detection computational domain, and the attention mechanism became integral in computation domains. This mechanism is primarily studied by neuroscientists and cognitive scientists and has recently attracted the interest of researchers from computer vision, Artificial Intelligence, graphics, and multimedia applications. Attention mechanism has achieved noteworthy success in various vision-related applications such as image classification [145] and visual question answering [146], neural machine translation [147], and image captioning [148]. Various models focused on prominent visual or textual media features through these mechanisms. Visual attention is selectively enhanced, and pay attention to extract relevant features in specified regions. At the same time, it reduces irrelevant processing.

Recently visual attention mechanism is adopted to improve the performance in visual question answering [146], image classification [145] and image captioning [148] *etc.* In contrast, textual attention mechanisms are preferred to handle long-term dependency in language processing. The semantic or syntactic input-output alignments under an encoder-decoder framework are improved using textual attention. This attention mechanism solved the various prevailing problems in language processing, such as machine translation [147] text generation [149], sentence summarization [150] and question answering [146]. Lu et al. [146]proposed a co-attention based deep model to learn the image attention, which helps in the question answering. Paulus et al. [151] proposed a fusion framework of reinforcement Learning to fuse inter-and intra-attention mechanisms in improving abstractive text summarization. The spatial attention mechanism [152] pays attention to specified regions to enhance and extract core regional features. The channel attention mechanism [153] pays attention to learning each channel's correlative features. The self-attention mechanism is based on a nonlocal Network to capture long-distance contextual dependency. Liu *et al.* [154] proposed self mutual attention based on a nonlocal network among deep RGB and Depth features to enhance the deep feature and fusion model. Consequently, the recent RGBD model improves performance by using an attention mechanism, while it has not been fully utilized to improve deep localized features and enhance the encoded features. It has ample scope for improvements in proposing enriched encoded and deep localized features.

2.2.3.9 Summary

Deep learning-based models have addressed the most prevailing issues in 2D and 3D conventional-based models. The most dominating contributions are cross correlative features exploration and fusion models, which improved internal and external saliency discrepancies. Depth-based features are utilized as complementary features to localize the salient object precisely. These deep learning-based models achieved noticeable success. Nevertheless, This domain still has some challenging issues which still not been thoroughly addressed.

1. The low-Depth issues are long prevailing issues because depth maps complement the RGB saliency by borders, edges, and localized features.

2. The cross-complementary features achieved great success, while there is scope for further improvements through the mutual attention mechanism.

3. Most exiting models used a standard backbone to extract features, which is insufficient in complex and challenging scenarios.

These existing research gaps address by proposing mutual attention-based cross-complementary fusion and composite backbone to enhance the encoded features.

2.3 Research Gaps

Salient object detection has wide applications in all domains of computer vision. The improvements have been noticed from cognitive [10], [3], [2] or statistical computational domains [11], [96] [79] to current deep learning-based approaches [45] [137] and 2D(RGB) [5] to 3D (RGBD) [38] based modality. However, the above success brings various research gaps and limitations in each domain, which motivate further improvements. Various research gaps have been identified through the above study in salient object detection. These can be summarized below:

- The global contrast-based methods generate saliency with the interior regional discrepancy, which is defined as" an object has clutter and complex interior region, similar to the background is suppressed not enhanced as salient points or regions."

- Regional contrast-based methods generate the exterior regional discrepancy, which is defined as "an object having outer region similar to the background will not preserve the outer boundary of objects till the realm of similarity and also increase the non-salient points."

- Background prior-based methods suffer from distinguishing the salient and non-salient regions. Their performance depends on identifying the region associated with the border.

- All these approaches and features fail individually in complex and cluttered images. Therefore various integration strategies are used, but these integration strategies are domain-specific rather than generic.

- The integration strategy simply fused the various low-level saliency features. These strategies have no initial reference surface, which is used to differentiate the salient and non-salient regions.

- Some salient object detection methods also highlight the non-salient regions or pixels and use a separate segmentation algorithm to find the salient object.

- Initially, most algorithms provide full-length saliency, including highlighted salient objects additionally produce minimized backgrounds.

The global contrast [4], [64] based methods compute pixel or region-wise contrast. The region-wise methods [43] increase the computation complexity because they divide the image into regions, and then after regional saliency, low-level features are integrated to produce a salient object. The regional definition increases the computational cost; at the same time, it achieves remarkable improvements. The regional-based models bring the various saliency features. These saliency features further integrate to produce the final salient object, while the global contrast-based models finally used some segmentation-based algorithm to produce a salient object. Regional definition brings the backgrounness [36] and surroundedness [35] based saliency features which enhance the saliency, but It does not highlight the salient object uniformly and fails in complex and cluttered background. These 2D(RGB)

models are stuck in complex and challenging images. The next improvements have been noticed by using depth information, formerly defined, ***3D or RGBD salient object detection***. The extracting accurate depth information is still a challenging task. It brings various limitations in 3D saliency computations in statistical models. These limitations are summarized here.

- It is challenging to design the discriminative features, which have the capabilities to distinguish regional disparity.
- It is challenging to formulate spatial and depth-based discriminative features to differentiate salient and non-salient regions.
- It is challenging to design an efficient integration strategy that incorporates all essential 3D features like Spatial contrast, depth contrast, regional contrast, and color contrast to produce salient objects correctly in complex and cluttered backgrounds.
- It is challenging to design a topographical reference surface used as an initial integrating surface plain for various regional and depth-based saliency features.

The various regional and super pixel-based saliency features are used to improve the saliency. The recent success of the CNN-based deep learning models in diverse applications has become a factor in adopting a deep learning-based model in saliency computation. The deep learning model improved the performance remarkably compared to statistical and probabilistic-based models, although these also suffer from various limitations. These limitations are summarized here:

- The depth modality is an essential input parameter in CNN-based RGBD salient object detection. The 3D sensor camera produces depth maps. These depth maps have boundaries, edges, and other structural pieces of information; they also have some low depth, unstructured and irrelevant depth maps. The destruction and occlusions destroy the object boundary. These issues bring the challenge of designing an efficient network to complement these research gaps.

- It is challenging to develop the model to preserve the Non-complementary features along with complementary fusion.

- It is challenging to enhance the salient regions and minimize the irrelevant features in the non-salient regions during the encoding stage.

- It is challenging to model depth modality's geometrical, structure, and boundary information to purify the RGB features during the encoding stage.

- It is challenging to develop the multi-model multi-stage fusion strategy to provide equal importance to low-level purified encoder features and high-level semantics.

- It is challenging to develop a strategy to explore complementary features between two modalities at multiple stages.

- It is challenging to enhance the deep localized features which guide the fusion process so that salient objects pop-out in complex and cluttered backgrounds.

- It is challenging to design the enhanced encoder feature because existing models use standard exiting backbone networks to extract the encoded features.
- It is challenging to design an improved dense decoder that combines all essential encode features and utilizes deep localized features to predict exact salient object detection.
- Current architecture requires large-scale pre-trained CNNs or needed large-scale training set to produce efficient results.

The research gaps mentioned above are a beacon for proposing various salient object detection models. The research gaps identified here are guided motivation for complex salient object detection.

2.4 Dataset

The challenge of the salient object is rapidly increasing by proposing more and more complex datasets. Initial datasets are mainly focused on single and large-sized objects, then after various parameters, including variable-sized, diverse domains, complex and cluttered background, and low depth, have been included. The study of existing datasets has been studied into two domains: the first is an RGB or 2D dataset, and the second is a 3D or RGBD dataset. Various publicly available datasets are used to evaluate the salient object detection algorithms.

TABLE 2.4: RGB Dataset used in various saliency computations

SN.	Dataset	No. of Images	Year	Resolution	Obj./Sub.	Domain
1.	ASD [154]	1000	2009	[400×300]	1/1.	SOD
2.	MSRA [85]	5K	2011	[400×300]	1/1.	SOD
3.	MSRA10K [154]	10K	2013	[400×300]	1/1.	SOD
3.	THUR15K [154]	15k	2013	[400×300]	1/1.	SOD
4.	DUTOMRON [75]	5K	2013	[400×400]	5/5.	SOD
5.	PASCAL-S [155]	850	2014	VAR	5/12.	SOD
7.	ImgSaL [156]	235	2013	[640×480]	2/19.	SOD
8.	ECSSD [157]	1000	2013	[400×300]	1/1.	SOD
9.	SOD1 [158]	100	2007	[300×225]	1/3.	SOD
10.	SOD2 [158]	100	2007	[300×225]	2/3.	SOD
11.	SOD [159]	300	2010	[481×321]	3/7.	SOD
12.	CSSD [75]	200	2013	[400×300]	1/1.	Eye-Fixation
13.	Infrared [160]	900	2011	[1024×768]	5/2.	Eye-Fixation
14.	Bruce-A [161]	120	2013	[681×511]	4/70.	Eye-Fixation
15.	Judd-A [161]	900	2014	[1024×768]	5/2.	Eye-Fixation

2.4.1 RGB(2D) Dataset

The 2D(RGB) dataset consists of RGB color images and annotated pixel-wise level ground-truth images. RGB datasets are of two types. The eye fixation-based datasets used for localization and salient object detection-based datasets contain complete object and ground-truth maps. The salient object detection and fixation prediction model used various datasets to validate their proposed algorithm. We summarised the characteristics, resolution, size, and other information in Table 2.4 and visually shown in Fig. 2.5. However, the salient object detection-based datasets used in this thesis are described here.

- ***Microsoft Research Asia (MSRA) Dataset*** [154]: It is the most widely used dataset. It has 10000 images with a single salient object, mostly located

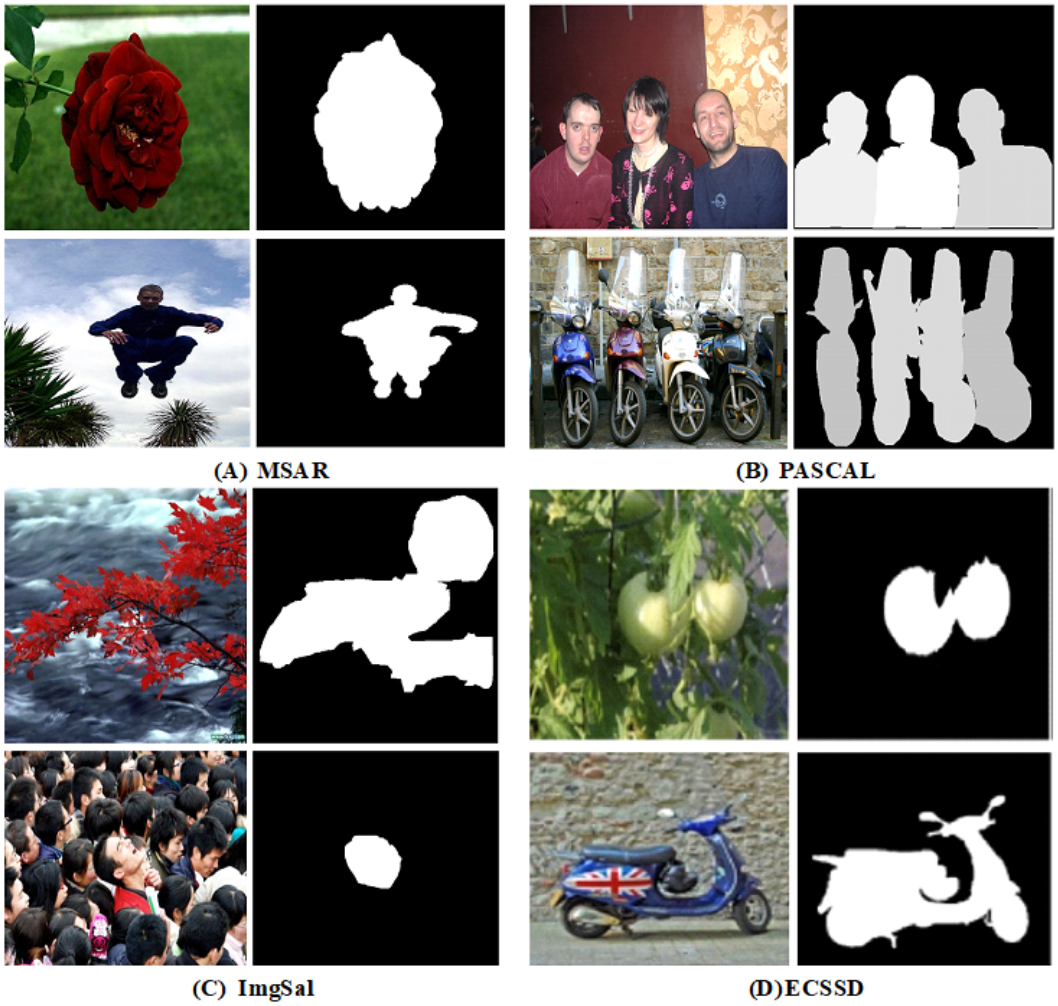

FIGURE 2.5: A set of RGB(2D) Datasets.

in the center of the image. The background of the images is simple and clutter-free. Researchers have also used subsets of this dataset with 5000 and 1000 images. They are called as MSRA-5k [154] and ASD [154], respectively.

- ***Extended Complex Scene Saliency Dataset (ECSSD)*** [157]: The dataset has 1000 images with complex background. It includes multiple salient object in complex background and some images come from the challenging Berkeley-300 dataset

- ***DUT-OMRON Image Dataset*** [75] : The dataset contains 5168 images with a highly complex background. It is presently most challenging available dataset with the objective of designing a dataset for evaluating the robustness of salient object detection models.

- ***PASCAL-S Dataset*** [155]: The dataset has 850 images on eight subjects. It contains confusing background and varying size salient object.

- ***Salient Objects Dataset (SOD) Dataset*** [159]: This dataset extracts salient object boundaries from segmentation boundaries present in Berkeley Segmentation Dataset (BSD). It has 300 images. And the two variations of this dataset, SOD1,and SOD2 [158] of resolution 300×225 is used by some researchers in their SOD tasks.

- ***THUR 15K Dataset*** [35]: The database contains five categories of images. For each group, there are 3000 images. The salient regions are marked at the pixel level.

- ***Judd Annotation (Judd-A) Dataset*** [161]: The dataset has 300 images of outdoor and indoor scenes and is broadly used for eye-fixation evaluation.

- ***Bruce-Annotation (Bruce-A) Dataset*** [161]: This dataset contains 120 indoor and outdoor scenes for eye-tracking prediction.

- ***ImgSal Dataset*** [156]:This is class specific image dataset and has 6 classes: (1) 50 images with large size salient object, (2) 80 images with intermediate

size salient object,(3) 60 images with small size salient object,(4) 15 images with complex backgrounds,(5) 15 images with repeating tract, (6)15 images with multiple salient objects.

These datasets are used in different contexts; some datasets have similar characteristics, while others are used in eye-fixation prediction applications. The most preferred datasets in salient object detection are MSRA, PASCAL, ECSSD, DUT-MORON, and ImgSal. In this thesis, these datasets have been used, and one set of examples is shown in Fig.2.5.

2.4.2 RGBD(3D) Dataset

The 3D salient object detection is formulated, executed, and validated on the RGBD datasets and is currently in recent trends in visual saliency computations. There are nine datasets have been designed from 2014 to till now. The RGBD-based salient object detection has been summarized in Table 2.4 and visually shown in Fig.2.6. These datasets have three components:(1) RGB image, (2) Depth image, and (3) Ground truth image. The emergence of depth-sensing devices addresses various issues of 2D saliency. The detailed attributes of these datasets are described in Table 2.5 and visually shown in Fig. 2.6. At the same time, the individual characteristics of each dataset are described below.

- ***STERE Dataset*** [97]: This is the first stereoscopic RGBD dataset, designed and annotated by three users. The authors initially collected 1,250 stereoscopic

TABLE 2.5: RGBD Dataset for Salient Object Detection

S.No.	Dataset	Year	Pub.	Source/Type of Sensor	Size	Objects	Resolution
1.	STERE [97]	2012	CVPR	Internet /Stereo //camera+sift flow// Left and Right depth view	1000	One	[251∼1200] [222∼900]
2.	GIT [162]	2013	BMVC	Multiple Home environment/ Microsoft Kinect	80	Multiple	[640∼480]
3.	DES [43]	2014	ICIMCS	Indoor /Microsoft Kinect	135	One	[640∼480]
4.	NLPR [98]	2014	ECCV	Multiple Indoor/outdoor Microsoft Kinect	1000	Multiple	[640∼480] [480∼640]
5.	LFSD [100]	2014	CVPR	Indoor/outdoor Lytro Illum camera	100	One	[360∼360]
6.	NJUD [110]	2014	ICIP	Movie/internet/photo FujiW3 camera+optical flow	1985	One	[231∼1213] [274∼828]
7.	SSD [163]	2017	ICCVW	Movies/ Sun's optical flow	80	Multiple	[960∼1080]
8.	DUT-RGBD [144]	2019	ICCVW	Multiple Indoor/outdoor	1200	Multiple	[400∼600]
9.	SIP [130]	2020	TNNLS	Person in the wild /Huawei Mate10	929	One	[992 ∼744]
10.	ReDWeb-S [137]	2020	ArXiv	Web stereo images+flownet2.0+post processing	3179	Multiple	[640∼480]

from various sources like Flickr, NVIDIA 3D Vision Live, and Stereoscopic Image Gallery. The annotated dataset sorted and selected 1000 images on the overlapping salient regions with a corresponding depth map and ground truth pairs.

- ***GIT Dataset*** [162]: It is collected from a mobile-manipulator robot in a real-world home environment. It consists of 80 RGB images and corresponding depth, and ground truth images. However, this dataset is not utilized and is

FIGURE 2.6: A set of RGBD(3D) Datasets.

less reported in the literature.

- ***NJUD Dataset*** [110]: This dataset was designed by a Fuji W3 stereo camera and collected from the internet, 3D movies, and photographs. It was initially designed with 1,985 stereo images; later, it was available in 2000 stereo image pairs with corresponding depth and ground truth images.

- ***NLPR Dataset*** [98]: It consists of 1,000 RGB images and their corresponding ground truth and depth maps. It is also reported with the name RGBD-1000 dataset in various papers. It is designed by a standard Microsoft Kinect, which includes indoor and outdoor images.

- ***DES Dataset*** [43]: It consists of 135 indoor RGB-D stereo images. It is designed by a Microsoft Kinect camera and annotated by three users. The overlapping area based fine label ground truth map with the corresponding depth map is produced.

- ***LFSD Dataset*** [100] :It consists of 40 outdoor and 60 indoor images. The images were taken from a Lytro light field camera and annotated by three users by manual segmentations. The annotation became ground truth when the overlap of the three results was over 90

- ***SSD Dataset*** [163]:It consists of 80 images with a resolution of 960×1080 and collected from three stereo movies and includes indoor and outdoor scenes. This dataset is the most challenging dataset present now.

- ***DUT-RGBD*** [144]: It consists of 1200(800 indoor and 400 outdoor scenes) images with corresponding depth and ground-truth maps. It has 800 indoor and 400 outdoor scenes with corresponding depth images. This dataset is designed to overcome the challenges of complex and cluttered backgrounds by including the following factors: *i.e.,* multiple or transparent objects, complex

backgrounds, similar foregrounds and backgrounds, and low-intensity environments.

- ***SIP Dataset*** [130]: It consists of 929 annotated high-resolution images taken by a smartphone, Huawei Mate10, focusing on multiple poses of persons in each image. It is annotated with pixel-level ground truths and covers diverse scenes, including various challenging factors.

- ***ReDWeb-S Dataset*** [137]:It consists of 3600 stereo images with very challenging and complex images with high-resolution ground truth and depth maps. It is collected from many web stereo images. It is designed with the Flownet2.0 algorithm followed by deep CNN to remove various noises and produce improved depth maps.

2.5 Performance Evaluation Metrics

Salient object detection tasks are verified, validated, and evaluated using various performance metrics. The evaluation aims to compute the similarities and dissimilarities between computed saliency maps and corresponding ground truth. The following metrics are used for the evaluation of proposed salient object detection models:

- ***Receiver Operating Characteristic (ROC-Curve)***: Graphical analysis of false positive (FPR) versus true positive rates (TPR) can be computed using

multiple fixed thresholds which changes from 0 to 255 and defined in Eq. 2.3 as follows:

$$TPR = \frac{|S_m \cap B_m|}{|S_m|}, FPR = \frac{|S_m \cap B_m|}{|S_m \cap B_m| + |\overline{S_m} \cap \overline{B_m}|} \tag{2.3}$$

where S_m, B_m, $\overline{S_m}$ and $\overline{B_m}$ represent true salient points, true ground truth points, false salient points and false ground truth points, respectively.

- ***Area Under Curve (AUC)***: It is the area under the curve of Receiver Operating Characteristics (ROC) Curve.

- ***Precision-Recall Curve(PR-Curve)*** Precision and recall are widely used performance measures. The Precision denotes the correct assignment of a percentage of salient points, while recall denotes the detection of the percentage of the salient pixels. For computation and evaluation of saliency map S_m, and corresponding binary mask B_m, the Precision and Recall is defined in Eq. 2.4 as:

$$Precision = \frac{|S_m \cap B_m|}{|S_m|}, Recall = \frac{|S_m \cap B_m|}{|B_m|} \tag{2.4}$$

Where $|-|$ represents the intersection and positive entry between B_m and S_m. This is the most robust analysis to bipartite saliency S_m by using multiple fixed thresholds which changes from 0 to 255. Precision and Recall are computed on each threshold and are combined to form a precision-recall (PR) curve. It describes the various model performances.

- ***Mean Absolute Error(MAE)*** The mean absolute error(MAE) is calculated between normalized (in the range $[0, 1]$) saliency map S_m and the ground-truth binary mask B_m in Eq. 2.5 as follow:

$$MAE = \frac{1}{n} \sum_{j \in n} (S_m(j) - B_m(j)) \tag{2.5}$$

- ***F-Measure*** The representation of relevancy of parameters, precision and recall through weighted harmonic for overall performance measurement is computed through F-Measure as in Eq. 2.6 follows:

$$F - Measure = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{2.6}$$

  For the uniform comparison for various salient object detection method, we use β^2 is $=0.3$, because same value is used in almost all salient computation method. We use adaptive rather than fixed threshold for F-Measure, which is two times of mean of saliency.

- ***F-Score:***It is the harmonic mean of precision and recall in Eq. 2.7.

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.7}$$

- ***Area Under Curve (AUC)***: It is the area under the curve of Receiver Operating Characteristics (ROC) Curve.

- ***Intersection-over-union (IoU) score:*** It is defined as the ratio between area of overlap and area of union between actual and predicted salient region. It is defined in Eq. 2.8.

$$\text{IoU} = \frac{\text{Area of overlap}}{\text{Area of union}} \tag{2.8}$$

- ***S-Measure*** The recent evaluation metric, S-measure [164], computes the structural similarity and dissimilarity. It computes region-aware S_{reg} and object-aware S_{obj} structural similarity between computed saliency map and ground truth map. This metric is defined in Eq. 2.9 as follows:

$$S_{measure} = \alpha S_{obj} + (1 - \alpha) S_{reg} \tag{2.9}$$

where $\alpha \in [0, 1]$ is set to 0.5.

- ***E-measure(E_ψ)*** : E-measure is recently defined as an Enhanced alignment measure, and the detailed definition and formulation are available here [165]. This measure is based on cognitive vision studies. It uses image-level statistics(mean) and local level pixel matching information. To demonstrate a comprehensive evaluation, we use the mean value of E-measure. The bias matrix is defined in Eq. 2.10 between Image I and global mean μ to compute the difference between local and global statistics.

$$\wp_I = I - \mu_I : \mathbb{B} \tag{2.10}$$

The $\mathbb{B}$ matric is computed for saliency map($\wp_{Sm}$) and ground truth map($\wp_{Gt}$).The Hadamard product ($\odot$)between $\wp_{Sm}$ and $\wp_{Gt}$ is computed to eliminate the luminance effect and quantify the bias matrix similarity. It is defined an alignment matrix in Eq. 2.11 as follows:

$$\alpha_{Sm} = \frac{2\wp_{Sm} \odot \wp_{Gt}}{2\wp_{Sm} \odot \wp_{Sm} + 2\wp_{Gt} \odot \wp_{Gt}} \tag{2.11}$$

The final E-measure is defined to measure pixel-level matching and image-level statistics by using an enhanced alignment matrix ϕ_{Sm}. It is defined in Eq. 2.12 as follows:

$$E - measure_{Sm} = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} \phi_{Sm}(i,j) \tag{2.12}$$

Where ϕ_{Sm} is enhanced alignment matrix and defined as $\phi_{Sm} = f(\alpha_{Sm})$, and $f(x) = ((1+x)^2)/4$.

2.6 Conclusion

In this chapter, closely related, and outstanding state-of-the-art methods in 2D and 3D, conventional and deep learning based salient object detection have been studied and reviewed. It also explored the issues and challenges of these fields. The benchmark databases on which the proposed models are evaluated are also described. Finally, the metrics on which the performance of the proposed model is evaluated are explained.