



Deep Learning Survey on Saliency Detection: Encoders to Decoder Refinements

Gaurav Aggarwal^{1*}, Ankush Joshi²

¹Department of Computer Science and Engineering COER
University, Roorkee, India
*gaurav.aggarwalcoer@gmail.com

²Department of Computer Science and Application
COER University, Roorkee, India ankushjoshi1987@sgmail.com

Abstract. This paper presents a thorough examination of the encoder-decoder architecture for detection of salient image regions, a key objective in the area of computer vision (CV) where it is necessary to locate and evaluate the primary visible features within an image. Beginning with a description of how design has progressed over the years from about 2010, the paper looks back to when CNN-type neural networks were initially developed and have been subsequently transformed into more sophisticated networks through the use of transformer technologies. Additionally, some important benchmarks and measures are included to allow the reader to consider the effectiveness of a particular network at accurately classifying images. Finally, discussion is provided regarding the current problems that exist within the field, such as the need for fast and effective algorithms and accurate boundary detection, as well as potential future research directions, including the development of mixed and composite architectures and new and improved methods of implementing foundation models.

Keywords: Image Saliency Detection, Encoder-Decoder Networks, CNN Architectures, Transformer Integration, Visual Attention, Evaluation Metrics, Hybrid Models.

1 Introduction

Recent advances in the development of attention mechanisms on the CNN-based encoder-decoder architectures have significantly improved the performance of models designed for image saliency detection. Attention mechanisms enhance the ability of models to identify salient objects in complex backgrounds by overcoming some of the drawbacks associated with traditional methods of saliency detection. The self-attention mechanism enables an effective manner to refine feature representations, enabling the effective use of the multi-scale representations that are critical for accurate prediction of visual saliency in different scenarios [3].

Further, the combination of attention and skip connections in the architecture results in retaining important spatial information and provides an effective way to combine high-level semantic information with low-level physical features, resulting in much better saliency maps [22]. Such an evolution in architecture shows that innovative approaches will be essential as researchers seek to develop a computational system capable of simulating human-like visual attention. As advancements in the field continue to be made, so too does the need to create models that identify object features that are important and also mimic the subtlety of attitudinal processing performed by human beings. Current evidence suggests that incorporating biologically-based mechanisms such as 'center-surround interaction' into models may allow for more contextually relevant methods of prioritizing attention over purely visual features[19]. This evidence correlates to findings from perceptual research, which highlights the significance of contrast in visual attention, as well as the role of contrast in the perception of objects that are salient among the clutter in a visual scene[7]. The investigation of neural architecture that supports top-down attention signals (as done by NeVA[16]) will provide insight into ways that task-specific guidance may reflect human cognition.

. Increased saliency detection through new ways of doing things allows for the use of saliency detection in places other than the lab where visual perception can be hindered by poor viewing conditions. As researchers continue to look into the intricacies of visual saliency detection, an increasingly common method for increasing model robustness is using multi-task learning frameworks. These frameworks can allow for the simultaneous training on multiple related tasks (such as object detection and semantic segmentation), which will make use of shared representations to produce better saliency predictions across multiple contexts. Research shows that multi-task learning methods improve performance of models developed using saliency detection within complex visual scenes (including competing salient objects), thereby reducing the number of mistakes associated with object misidentifications caused by nearby distracting objects [18]. Incorporating the temporal dynamics of visual scenes into a saliency model (i.e., how visual scenes change over time) will enable researchers to better understand how attention is allocated to visual stimuli in real-world environments.

These technological breakthroughs do more than just bring computational models closer to human visual processing. They also significantly enhance the models' performance in real-life situations, which demand quick reactions and adjustments to changes in stimuli.

Inspired by the way our brains alter our perception when exposed to the same stimulus for a long time, the authors highlight ways of improving the adaptability of saliency detection systems to various environments. Research also shows that the brain's ability to adapt over time to the statistics of the visual input helps it focus more on the most salient features by constantly tuning the sensitivity to the visual input. In other words, our brain's adaptation mechanisms led to improved performance in our brain's recognition of complex scenes in line with the

computational models’ human visual processing alignment in addressing the challenge of maintaining attention in the presence of a highly cluttered visual environment. Additionally, these adaptive strategies could be of great use in self-driving and surveillance systems, where it is extremely important to have a fast, situation-aware saliency detection system due to the ever-changing nature of visual stimuli.

2 Evolution of Architectures

The model architecture of salient object detection (SOD) has shifted dramatically over time, with the introduction of increasingly complex models that include attention mechanisms, multi-scale feature fusion, and, most recently, transformer-based components, starting from the very basic CNN encoder-decoder architectures. Fig. 1 presents a timeline of encoder–decoder saliency object detection (SOD) architectures, highlighting key developments over time. This progression demonstrates the constant search for better.



Fig.1. Timeline of Encoder-Decoder SOD Architectures

2.1 Early CNN Encoder-Decoder Designs

At the very beginning, deep learning was adopted for SOD mainly via simple CNN encoder-decoder architectures. Usually, these models consisted of a contracting

path (encoder) that would gradually downsample the input image, thus allowing the extraction of high-level semantic features, and an expansive path (decoder) that would upsample the features, thus reconstructing a pixel-level saliency map. The architectures, e.g., fully convolutional networks (FCNs) and U-Net, initially created for semantic segmentation, were the main influences. In fact, the U-Net was responsible for proposing the direct skip connections transmitting feature maps from the encoder to corresponding layers in the decoder. This was highly beneficial for maintaining the detail of spatial structures that are normally sacrificed during the downsampling process, which in turn helps the decoder to generate more accurate boundaries of salient objects[15]. Some early models utilized pre-trained classification networks (e.g., VGG) as their encoders and then converted them for the dense prediction task of saliency detection.

2.2 Attention-Enhanced and Multi-Scale Decoders

Development in the field made leading scientists understand that the fusion of multilevel features should be done effectively and the focus should be put on relevant features for maximum results. This is how the idea of attention mechanisms and multi-scale decoders came about.

As depicted in Fig. 2, different saliency detection models—from early CNN architectures to transformer-based variants—are compared against the ground truth, highlighting progressive improvements in feature representation and boundary refinement. On one hand, attention mechanisms were incorporated into decoder architectures in such a way that the network could assign degrees of importance to various features. Channel attention modules, which were inspired by such works as Squeeze-and-Excitation Networks (SENet), allow the model to discover inter-channel relationships, while spatial attention mechanisms direct the focus to salient regions and effectively disregard the background noise. Hence, the application of attention is an effective strategy for the decoder to generate neat and accurate saliency maps [5].

Multiscale decoders were created to solve the problem of recognizing salient objects of different sizes. Methods like Atrous Spatial Pyramid Pooling (ASPP) and Feature Pyramid Networks (FPN) were widely used. ASPP obtains multiscale contextual information by atrous convolutions with different dilation rates, whereas FPNs create a feature pyramid with rich semantic information at all scales. These multi-scale approaches give the decoder the capability to combine features from various levels of the encoder, thus enhancing the model's proficiency in managing both small and large salient objects[9].

Besides that, iterative refinement methods became popular. Models such as BASNet [13] and CPD-Net [25] utilize cascaded or recurrent structures in which an initial coarse saliency map is gradually refined through subsequent modules. For example, BASNet uses a residual refinement module to specifically improve the boundary

quality, whereas CPD-Net employs a cascaded partial decoder to efficiently refine features, thus showing that multi-stage processing is very effective in achieving higher accuracy and efficiency.

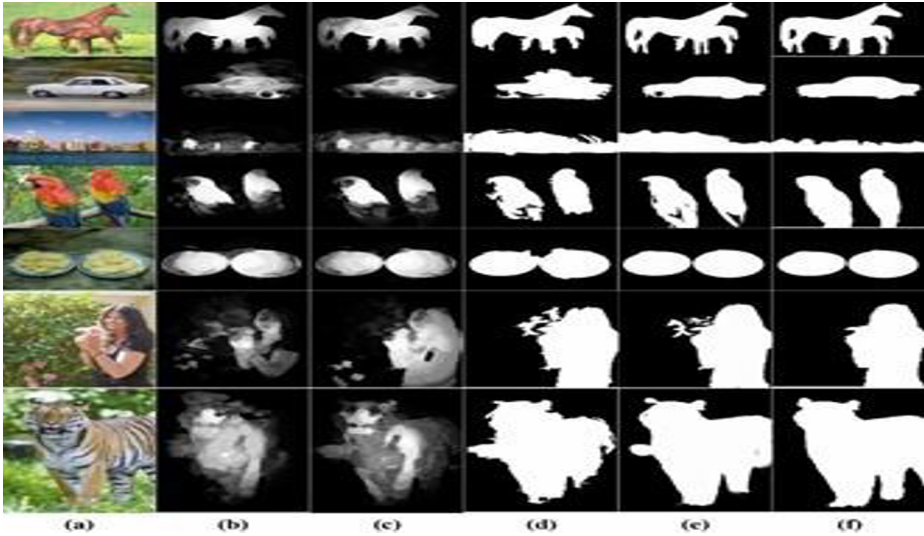


Fig.2. a) Input Image b) Early CNN Model c) Context-Aware Model d) Boundary-Aware Model e) Transformer Variant f) Ground Truth

2.3 The Rise of Transformer-Based Architectures

The newest paradigm of deep learning, Transformer, has brought a radical change to the saliency detection domain through a wave of innovation. The Transformer gained persistent acclaim in the literature, where it had previously dominated Natural Language Processing (NLP), particularly utilizing its selfattention mechanism to effectively capture long-range dependencies and global context, which are both critical to determining saliency.

Totally based on transformer architecture, e.g., Visual Saliency Transformer (VST), saliency can now be viewed as a sequence-to-sequence prediction task. VST can now treat spatially adjacent images as individual sequences, making use of the full extent of global relationships contained in the full extent of the image, which can be exploited more effectively by VST than traditional CNNs could through the locally constrained spatial extent of their respective receptive fields[11].

Still, the sole Transformers can be too heavy for the required computing power and may even struggle in capturing fine-grained local details (which CNNs are superb at); thus, researchers have designed hybrid architectures that fuse the strengths of the two methodologies. Usually, the CNNs are used first (due to their speed at representing local features), and then the new Transformer blocks are added for modeling global relationships and further feature enhancement. Further bridging this

with Hierarchical Transformers such as the Swin Transformer achieved even more, using windowed and shifted self-attention to allow efficient computation while still creating a hierarchical feature representation akin to CNNs, which makes them a robust backbone for saliency detection[24].

This combination is an extremely appealing field but also challenging because it involves finding a trade-off between local details and global context, which leads to more robust and reliable results.

3 KEY MODELS AND TECHNIQUES

3.1 Early Encoder-Decoder Adaptations

The U-Net family is integral to the early SOD models. The U-Net network was created with the purpose of biomedical image segmentation but adopts a symmetric encoder-decoder structure with skip connections because they enable the passing of feature maps directly from the encoder to the layer of the decoder to which they are linked [15].

Within the proposed survey framework, as shown in Table 1, scheduling logic is the manner in which feature extraction/refinement operations at different scales through several layers of encoder and decoder are ordered.

In the first place, the encoder extracts features from raw data in multiple scales. Then these features are scheduled to be decoded in a hierarchical order, i.e., from the coarse to the fine spatial detail. On the one hand, the decoder uti-

Table 1. Scheduling Logic Across Encoder–Decoder Stages

Stage	Operation	Purpose
Encoder Stage	Feature Extraction	Capture semantic information
Early Decoder	Upsampling + Fusion	Coarse map reconstruction
Mid Decoder	Multi-scale Aggregation	Context enhancement
Late Decoder	Refinement Layers	Boundary sharpening

lizes staged refinement techniques, that is, high-level features are upsampled and combined with corresponding low-level ones through skip connections. The scheduling makes sure that while global contextual information is maintained, boundary details are gradually restored. Also, multi-scale aggregation is carried out at certain stages to improve the saliency consistency at different resolutions. This well-organized scheduling of operations leads to better boundary localization as well as overall detection accuracy.

Deeply Supervised Saliency Detection (DSSD) is another very early work that also has VGG-16 as its backbone, and it added short connections between several side output layers and the final fusion layer. Thanks to it, deep supervision has become possible at different scales, and this has made it easier for the model to learn multilevel features, which, in turn, have resulted in better saliency prediction [4].

Likewise, Multi-Level Feature Deep Neural Networks (MLF-DNNs) are aimed at combining pixel-level and superpixel-level features in a CNN setting so as to achieve high precision in saliency detection [7].

3.2 Attention and Contextual Techniques: Enhancing Feature Selectivity

The Contextual Encoder-Decoder Network (CEDNet) was the first to apply dilated convolutions in its architecture to significantly extend the receptive field while maintaining the resolution. This enabled the model to accumulate more comprehensive global contextual information that plays a major role in correctly identifying salient objects even in very complicated backgrounds [6]. Effectively embedding the context, CEDNet enhanced the model's grasp of the whole scene. Saliency Fusion Attention U-Net (SalFAU-Net) is the name of a recent method that utilizes an explicit attention mechanism in a U-Net-like architecture. The fusion attention module integrates the network's ability to selectively focus on the most informative regions of an image while simultaneously suppressing the non-salient parts. This selective attention, together with the U-Net's feature to combine multi-scale features, results in better saliency map creation [12]. The Visual Saliency Transformer (VST) was a big step forward that introduced a completely transformer-based encoder-decoder architecture for SOD. The VST views saliency detection as a sequence-to-sequence prediction task and, using the transformer's self-attention mechanism, is able to capture long-range dependencies as well as global contextual information throughout the whole image, which is a major limitation of purely CNN-based models [11].

3.3 Attention Advanced Fusion Models: Precision and Efficiency

BASNet (Boundary, Aware Salient Object Detection) is a method that addresses the problem of blurry boundaries. To this end, a predict-refine architecture was used with a densely supervised encoder-decoder network for initial saliency prediction and a special residual refinement module (RRM) for boundary quality improvement only. BASNet also developed a new hybrid loss function (BCE + SSIM + IoU) that allows the network to learn pixel, patch, and map-level details; thus, sharper object boundaries result [13].

CPD, Net (Cascaded Partial Decoder), is a method that improves speed and accuracy simultaneously. It has a cascaded partial decoder that selectively throws away larger resolution features from shallow encoder layers, thus lowering the computational cost without degrading the performance substantially. Moreover, this technique

achieves efficient distractor suppression and feature enhancement by repeatedly using the deep layer features for optimization with the saliency map generated[25]. EGNNet (Edge Guidance Network) is a method that merges edge information with saliency detection for better output. It proposes a multitask learning model that simultaneously outputs saliency maps and edge maps. A special edge guidance module is also a part of the network to integrate the edge features into the saliency map prediction such that the resulting salient objects are bounded with correct and precise edges[27].

PoolNet investigates how far conventional pooling operations can lead to robust saliency detection. It proposes a Global Guidance Module (GGM) that delivers the global context and a Feature Aggregation Module (FAM) that elegantly integrates multi-scale features. Pooling roles are extended by PoolNet to acquire richer contextual information, which then allows the detection of salient features robustly at different scales [10]. Table 2 provides a comparative analysis of prominent encoder–decoder-based saliency detection models, highlighting their architectural characteristics and performance.

Table 2. Comparison of Key Encoder-Decoder Based Saliency Detection Models					
Model	Year	Backbone	Key Technique	DUTS F-measure (max)	DUTS MAE
U-Net var.	2015	VGG-16	Skip connections	~ 0.820	~ 0.085
DSS	2017	VGG-16	Short connections	0.825	0.056
MLF-DNN	2016	CNN	Pixel/ Superpixel fusion	~ 0.840	~ 0.065
CEDNet	2019	ResNet-50	Dilated convolutions for context	0.890	0.042
VST	2021	ViT-Base	Pure Transformer encoder-decoder	0.896	0.037
SalFAU-Net	2024	Custom	Saliency fusion attention	0.921	0.038
BASNet	2019	ResNet-34	Predict-Refine & Hybrid Loss	0.859	0.048
CPD-Net	2019	ResNet-50	Cascaded Partial Decoder	0.865	0.043
EGNet	2019	ResNet-50	Edge Guidance Module	0.887	0.039
PoolNet	2019	ResNet-50	Global Guidance & Feature Aggregation	0.887	0.039

4 DATASET

4.1 Benchmark Datasets

To promote the progress of SOD, some datasets have been established, and each has the unique abilities of the models being referred to. They generally contain natural images with pixel-wise ground truth annotations that indicate the visually attractive regions.

Whereas, as in Fig. 3, the superlative performance of several saliency detection approaches on the DUTS data set established with the majority of metrics. DUTS (DUT, Saliency) is a very large and popular dataset for salient object detection. It has a training set (DUTS, TR) of 10553 images and a testing set (DUTS, TE) of 5019 images. The DUTs are collected from the ImageNet DET training and validation sets. This provides a huge variety of objects and scenes. A large dataset like this provides an excellent training set for deep learning models and a large database for performance analysis [20].

ECSSD (Extended Complex Scene Saliency Dataset), which is focused on centerbased saliency detection in complex, cluttered scenes. It features 1000 images, which include numerous semantically meaningful yet structurally complex objects observed in cluttered backgrounds. Due to such complexity, ECSSD is an excellent standard for evaluating how well a model can distinguish salient objects in visually challenging settings [17].

HKU-IS is one of the well-known datasets in the field, comprising a total of 4,447 images. What makes HKU, IS stand out is that it often contains multiple disconnected salient objects throughout the images or salient objects that are located near the image edges. These images, in particular, put sharp focus on the smartness of the models in correctly segmenting and recognizing all the salient areas, even those that are divided or partly hidden [7].

PASCAL S is a relatively small but quite important dataset; it was created using 850 images from the PASCAL VOC 2010 validation set. The special feature of PASCAL-S is that it uses human eye-tracking data; thus, it gives a ground truth of saliency that is more perceptually oriented. Hence, this dataset can be used to check if a model's behavior is consistent with human visual attention patterns[2].

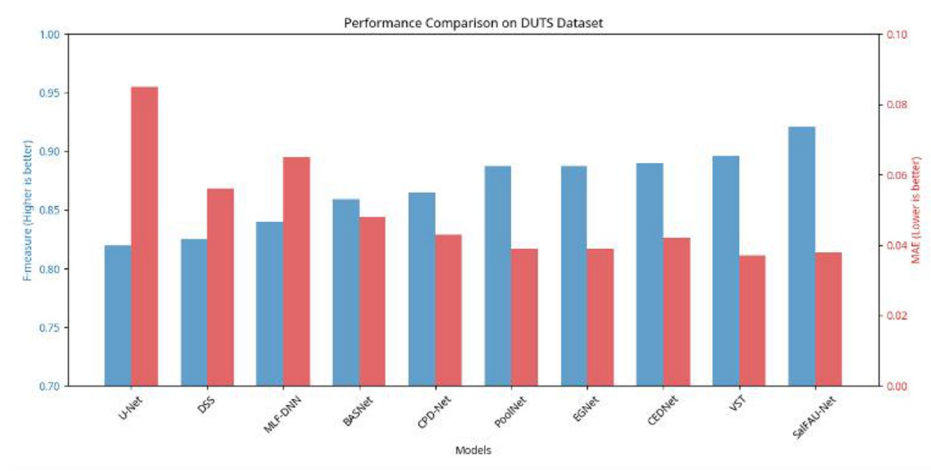


Fig.3. Performance Metrics Bar Chart on DUTS Dataset

5 CURRENT CHALLENGES

Real-time efficiency: A significant number of the leading SOD models, especially those which are based on complicated CNN architectures or transformers, are characterized by high computational costs. This makes it very difficult to use such models on limited hardware devices like mobile phones or embedded systems. Maintaining the level of accuracy while significantly increasing the speed to real-time is a critical challenge that is continuously faced and may require a compromise between the complexity of the model and the speed of its inference. The demand for lightweight backbones (e.g., MobileNet, EfficientNet) and, well, tuned decoder designs is very high in real-world applications.

Boundary Preservation: Considerable efforts have been made to bring the generation of saliency maps to a level of high accuracy; however, there are still issues outlining object boundaries with exact precision, particularly in the case of complicated features such as hair, very thin objects, or objects with complex textures. It is known that the extent of the reliability of SOD gets limited due to the use of coarse or inaccurate boundaries when it is necessary to accomplish the segmentation task at a very detailed level[23]. Some work has been done on the introduction of boundary-aware loss functions and refinement modules, but still going for pixel-level perfect boundary prediction is a necessary task that requires more research.

Complex Scenarios: Typically the SOD models of the present time have a hard time coping with very challenging visual environments. The situations that include images with hardly any contrast between the salient objects and the background, or the cases of transparent or even translucent objects in the scenes, or simply the scenarios having a lot of background clutter, to the extent that it becomes fairly ambiguous to distinguish the salient foreground from the distractors. Differences in

object appearance, lighting conditions, and scene composition remain the biggest challenge for the generalization of the methods used[8].

6 EMERGING TRENDS AND OPPORTUNITIES

Integration of Transformers with CNNs is expected to be more common in the future. Hybrid architectures that utilize CNNs for the efficient extraction of local features and Transformers for the capture of global context can provide a better balance between preserving local detail and obtaining a global understanding. The combination of these two types of architectures will address the issues associated with solely using either a CNN-based model or a Transformer-based model [26].

The primary constraint in the development of deep SOD models is the extensive labor and time required to create a large, pixel-level annotated dataset to train deep learning models. As a result, future research will likely focus on the development of weakly or unsupervised learning techniques that can effectively train deep learning-based SOD systems using less precise annotation types (e.g., image-level tags, bounding boxes) and potentially no annotations at all. thereby reducing the cost and effort associated with data preparation [1]. Using information from different representations, e.g., depth (RGB, D), thermal (RGB, T), or event data, can greatly improve saliency detection in situations where RGB data alone are not enough. For example, depth features may allow the system to separate the main object from the background more efficiently. Moreover, extending the task of SOD to videos (Video Object Saliency, VOS) means also using temporal coherence to create more accurate and stable predictions [21].

The advent of huge general-purpose models, such as the Segment Anything Model (SAM), opens up new avenues in SOD. By repurposing these broadly trained models, which were pre-trained on very diverse and extensive datasets, for specialized tasks in saliency, we could achieve a remarkable leap in the generalization of models and hence lessen model development for each individual task. Further work will be done to find out how these strong models can be used for SOD through methods like fine-tuning and prompt engineering[14].

7 CONCLUSION

In conclusion, building upon SOD's state-of-the-art achievements, research into real-time, highly accurate, and robust to various and complicated visual situations is still our goal. It is foreseeable that future SOD algorithms will be a fusion of architectural innovations, better learning schemes, and multimodal data combinations. Although deep learning has represented one big leap for saliency detection, most of the existing work largely focuses on the nature of the encoders and scarcely considers the contribution of each decoder at the decoder side for spatial details recovery and boundary accuracy.

What's more, no objective comparison is ever concluded between different encoderdecoder configurations in terms of evaluation measures, such as IoU and Fmeasure. Our research will make a contribution to this area by thoroughly reviewing the evolution of decoder frameworks with the encoders and presenting a comparative performance evaluation.

References

1. Du, L., Ma, C., Chen, J., Zheng, H., Nie, X., Gao, Z.: Survey on deep learning-based weakly supervised salient object detection. *Expert Systems with Applications* **281**, 127497 (2025)
2. Everingham, M., et al.: The pascal visual object classes (voc) challenge. *International Journal of Computer Vision* **88**, 303–338 (2010)
3. Fei, F., Liu, W., Shu, L.: A lightweight convolutional neural network for salient object detection. *Tehnički vjesnik* **31**(4), 1402–1410 (2024)
4. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. *IEEE TPAMI* **41**(4), 815–828 (2019). <https://doi.org/10.1109/TPAMI.2018.2815688>
5. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 7132–7141 (2018)
6. Kroner, A., Senden, M., Driessens, K., Goebel, R.: Contextual encoder–decoder network for visual saliency prediction. *Neural Networks* **129**, 261–270 (2020)
7. Li, G., Yu, Y.: Visual saliency detection based on multiscale deep cnn features. *IEEE transactions on image processing* **25**(11), 5012–5024 (2016)
8. Liang, B., Luo, H., Wang, J., Shark, L.K.: Multi-scale attention-edge interactive refinement network for salient object detection. *Expert Systems with Applications* **275**, 127056 (2025)
9. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2117–2125 (2017)
10. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3917–3926 (2019)
11. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 4722–4732 (2021)
12. Mulat, K.A., Feng, Z., Eshetie, T.S., Hasen, A.E.: Salfau-net: Saliency fusion attention u-net for salient object detection. *arXiv preprint arXiv:2405.02906* (2024)
13. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 7479–7489 (2019)
14. Ren, S., Zhao, N., Wen, Q., Han, G., He, S.: Unifying global-local representations in salient object detection with transformers. *IEEE Transactions on Emerging Topics in Computational Intelligence* **8**(4), 2870–2879 (2024)
15. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241. Springer (2015)

16. Schwinn, L., Precup, D., Eskofier, B., Zanca, D.: Behind the machine's gaze: Neuralnetworks with biologically-inspired constraints exhibit human-like visual attention. arXiv preprint arXiv:2204.09093 (2022)
17. Shi, J., Yan, Q., Xu, L., Jia, J.: Hierarchical image saliency detection on extendedccsdd. IEEE transactions on pattern analysis and machine intelligence **38**(4), 717–729 (2015)
18. Song, X., Tan, Y., Li, X., Hei, X.: Gdvifnet: A generated depth and visible image fusion network with edge feature guidance for salient object detection. Neural Networks **188**, 107445 (2025)
19. Su, Z., Liu, L., Müller, M., Zhang, J., Wofk, D., Cheng, M.M., Pietikäinen, M.: Rapid salient object detection with difference convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
20. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR (2017)
21. Wang, S., Jiang, F., Xu, B.: Swin transformer-based edge guidance network for rgb-d salient object detection. Sensors **23**(21) (2023)
22. Wang, Y., Wang, R., Liu, J., Xu, R., Wang, T., Hou, F., Liu, B., Lei, N.: Tfgnet: Frequency-guided saliency detection for complex scenes. Applied Soft Computing **170**, 112685 (2025)
23. Wang, Z., Zhang, Y., Liu, Y., Wang, Z., Coleman, S., Kerr, D.: Tf-sod : A novel transformer framework for salient object detection. Neural Computing and Applications (Feb 2022)
24. Wu, S., Zhang, G., Liu, X.: Swinsod: Salient object detection using swin transformer. Image and Vision Computing **146**, 105039 (2024)
25. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3907–3916 (2019)
26. Yao, C., Feng, L., Kong, Y., Xiao, L., Chen, T.: Transformers and cnns fusion network for salient object detection. Neurocomputing **520**, 342–355 (2023)
27. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 8779–8788 (2019)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

