Salient Object Detection

Subjects: Others

Contributor: Pritee Khanna

Detection and localization of regions of images that attract immediate human visual attention is currently an intensive area of research in computer vision. The capability of automatic identification and segmentation of such salient image regions has immediate consequences for applications in the field of computer vision, computer graphics, and multimedia. A large number of salient object detection (SOD) methods have been devised to effectively mimic the capability of the human visual system to detect the salient regions in images. These methods can be broadly categorized into two categories based on their feature engineering mechanism: conventional or deep learning-based. In this survey, most of the influential advances in image-based SOD from both conventional as well as deep learning-based categories have been reviewed in detail. Relevant saliency modeling trends with key issues, core techniques, and the scope for future research work have been discussed in the context of difficulties often faced in salient object detection. Results are presented for various challenging cases for some large-scale public datasets. Different metrics considered for assessment of the performance of state-of-the-art salient object detection models are also covered. Some future directions for SOD are presented towards end.

Keywords: salient object detection ; conventional salient object detection ; Deep learning-based salient object detection

1. Introduction

The behavior of SOD models is expected to mimic the pre-attentive stage of HVS which guides human attention to the highly interesting regions in the scene. The identified salient regions in images can facilitate subsequent high-level vision tasks for improved efficiency and optimal resource usage. As a preprocessing step, SOD has served many computer vision tasks such as, visual tracking[1][2], image captioning[3], image/video segmentation [4][5][6], and so forth.

The challenges and difficulties in SOD come from the very nature of the scenes captured in free viewing conditions. Several sample images from different SOD datasets can be seen in Figure 1. The accompanying pixel-wise annotations shown here are used for evaluation but clearly delineate the basic requirements for a salient object detector. A SOD method should keep the error metric values to their least by strictly attaining to the salient regions and missing the non-salient ones. It is further expected that the SOD method should be computationally inexpensive in producing a high resolution saliency map for accurate salient object localization [7]. Being an active research field over the past two decades, a large number of models have been attempted to satisfy the minimum requirements for image based SOD. Early efforts for saliency detection were focused at fixation prediction[8][9]. Fixation prediction aims to attend the spatial locations where an observer may fixate within few seconds of free-viewing. SOD is different from fixation prediction as models for the former should detect and segment the entire extent of salient regions/objects in the scene. A general approach adopted by conventional SOD models to accomplish this goal is to assign high probability values to salient elements in a scene while producing a saliency map. Once detected, techniques such as thresholding can be used to segment out the whole salient object. Conventional SOD models following Itti et al.[8] attempt to capture the notion of scene rarity or uniqueness mainly by devising center-surround contrast features. Regional contrast in terms of global and local schemes have been frequently used in conventional SOD. Various complementary heuristic saliency priors have also been deployed to effectively capture the most conspicuous object regions in images. These conventional models have been proven to be efficient and effective in relatively simple scenes with a single object and/or clean background.

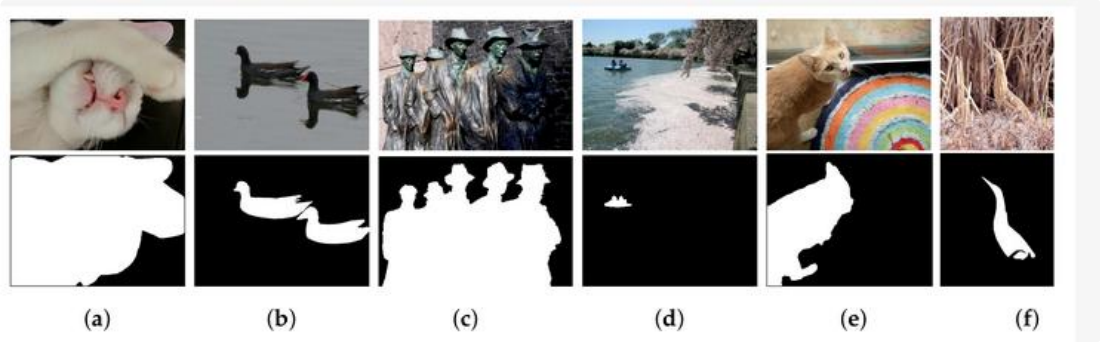

Figure 1. Sample challenging images for salient object detection with corresponding pixel-wise annotations shown below. (**a**) Large object, (**b**) Reflection, (**c**) Multiple objects, (**d**) Small object, (**e**) Complex scene, and (**f**) Low contrast.

Many diverse datasets have surfaced in the past ten years to challenge these SOD models. The presence of multiple salient objects, heterogeneous salient objects with variations in shape, size and position, low-contrast objects, and much cluttered background in datasets are challenging issues to address while adhering to high prediction requirements of SOD. However, the recent rapid development of deep learning-based techniques in the field has been highly successful in tackling most of the aforementioned issues. Fully convolution neural networks (FCN) lies at the core of deep learning-based SOD[10]. The powerful hierarchical multi-scale feature representation of FCN has been utilized in various ways for a coarse saliency prediction and its refinement for boundary accurate saliency map in a data-driven manner. However, the conventional models for SOD have the advantage of providing real-time performance and can be applied in the wild. Meanwhile, several deep models have leveraged saliency priors to improve the representational ability of multi-layer features and to speed-up the training process. Wang et al.[11] combined saliency estimate of multiple conventional methods as the prior knowledge informative of salient regions to guide saliency detection. Chen et al.[12] utilize saliency priors as an initial prediction for saliency refinement. Zhang et. al.[13] devised a deep unsupervised saliency detection with noisy supervision from multiple conventional SODs. Simple heuristic operator such as contrast in Reference[14] has been adopted for contrast modelling of multi-scale features in References[15][16]. These adaptations suggest that despite tremendous progress and superior performance demonstrated by deep learning based SOD, the tools of conventional saliency detection can be useful for further raising the performance bar of deep models.

Overview of Salient Object Detection

Saliency detection has been an interdisciplinary field. The fundamental investigations on cognitive and psychological theories of HVS attention [17][18][19] were contributed by cognitive psychologists and neuroscientists. Such theories preliminarily formed the base for development of the early saliency models. A major milestone in visual saliency was achieved when the complete implementation of the computational attention architecture[19] was realized by Itti et al.[8]. The feed-forward model proposed in Reference[8] computes and combines multi-scale color contrast, intensity contrast, and orientation contrast to direct computational mechanism to highlight the salient locations in a low-resolution saliency map. Further, a winner-take-all (WTA) neural network is invoked multiple times to shift the focus of attention to the next most conspicuous location by employing inhibition of return mechanism after the first WTA invocation. This ability to shift from location to location in a fixation map is vital for tasks such as image understanding. Nevertheless, the computation of center-surround contrast using low-level features and their integration for attention guidance provided great insight for further research in the conventional SOD paradigm.

It is widely accepted that the seminal work of Liu et al.[20] and frequency tuned approach proposed in Reference[14] brought novel contributions to boost up research in SOD. Liu et al.[20] introduced the computational methods for extracting local, regional, and global features that capture different aspects of saliency information. A binary segmentation is achieved using conditional random fields (CRFs) with all extracted features. In addition to that, the first large-scale dataset was also presented in Reference[20]] with bounding box annotations for training and evaluation of SOD models. Contributions by Reference[14] include in-depth frequency analysis of sub-sampled features used for contrast computation and generation of full-resolution saliency maps using a frequency-tuned approach.

Deep convolutional neural networks (CNNs) have demonstrated exceptional performance in many vision tasks such as image classification [21][22], semantic segmentation[23][24][25], object detection[26][27], and object tracking[28][29]. Deep CNNs have also benefited SOD and delivered a huge performance gain compared to the conventional SOD models. This data-driven approach generates a hierarchy of multi-scale feature representation automatically from the input image. The stacking of convolution and pooling operation in deep CNNs allows the receptive field of the network to grow gradually with depth. Due to the large receptive field, deep layers in the network could capture the global semantics and provide a

holistic estimation of the salient regions. The shallow layers retain more spatial details useful for the localization of fine structures and salient object boundaries. Different deep learning-based SOD models utilize these complementary multi-layer features in various ways to learn robust saliency representations with a powerful end-to-end learning[24]. Figure 2 shows a sudden rise in the number of papers published in SOD from images since 2015 when the first few deep learning-based SOD models were proposed.

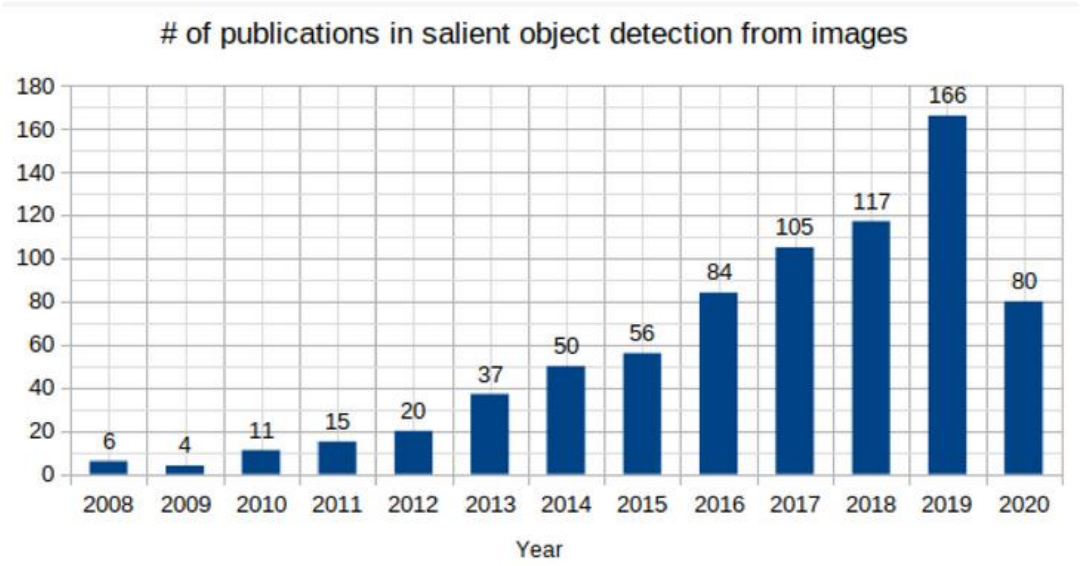

Figure 2. The trend of publications in salient object detection from still images from 2008–2020 (July).

Recently, the most advanced models in SOD have been devised from the field of computer vision. Table 1 compares SOD with some related computer vision tasks such as fixation prediction[30][31], image segmentation[32][33], semantic segmentation[23][24][25], object proposals generation[34], object detection[26][27], and salient object subitizing[35]. Table 2 highlights various research tasks in the similar fields compare with SOD in terms of objective and approach taken. Although this survey focuses mainly on single RGB image based SOD models, closely related fields such as co-saliency detection(CoSOD), RGB-Depth (RGB-D) SOD, video SOD, and SOD on light field have also experienced a great deal of interest in the recent past. The CoSOD task aims at the automatic detection of the salient object(s) that are common among multiple related images. Given an image group, a co-salient object should be salient in each image along with a high chance of repeatability and appearance similarity among the related images [36]]. Classical approaches to CoSOD resort to inter-image correspondence modelling strategies[37][38] to represent the common attributes among multiple images. Recent deep learning-based CoSOD models[39] learn co-salient object representations jointly, and have utilized deep-CNN models to achieve outstanding performance. Typical applications of CoSOD include collection-aware crops[40], co-segmentation[41] and video foreground detection[42]. The RGB-D based SOD models utilize important complementary information of depth along with color measurements for detecting salient objects on RGB-D images. Similar to SOD, traditional RGB-D models[43][44] rely heavily on hand-crafted features while combining RGB image with depth maps. Models[45][46] that exploit the implicit shape and contour information in depth maps to refine saliency results have shown promising performance. Deep learning-based, end-to-end RGB-D models[47][48] [80,81] are becoming more and more popular as they can effectively exploit multi-modal correlations, and multi-layer information hierarchy for robust RGB-D saliency detection[49]. Video SOD models leverage the sequential, motion, and color appearance information contained in a video sequence to detect targets that are repeated, dynamic, and salient[36] [48]. Video SOD has many applications viz., action recognition[50] and compression[51]. Very similar to other related fields, current state-of-the-art models in video SOD are deep learning-based which capture and focus on combining the spatial and temporal saliency information efficiently[41]. Efforts have also been made to deal with data insufficiency problem in the supervised video-SOD models through novel data augmentation techniques[41] or introducing new datasets[52]. The detection of saliency on 4-D light field (LF) is another interesting task related to the RGB-SOD task. A light field is an array of 2-D images which includes focal stacks, depth maps and all-focus images captured through handheld light field camera Lytro Illum[53]. In absence of a large-scale LF-SOD dataset, low-level cues have been utilized to tackle the task. Recently, Reference[54] proposed a new dataset and deep learning based model for the LF-SOD task. Interested readers may refer to References[36][49][52][54][55] for further information on these related tasks.

Table 1. Comparison of salient object detection with other computer vision tasks (GT - Ground truth).

#	Task	Aim	GT Map	Vs. SOD
1	Fixation prediction	Finds where human look in a scene.	Several fixation dots in human fixation map.	Pixel-wise GT maps with clear boundaries are seldom used.
2	Image/Semantic segmentation	Assigns a label to each pixel in the image.	Each pixel has an associated category label.	Scope is the entire image, not just the salient objects.
3	Object proposals	Generates overlapping candidate region proposals.	Rectangular bounding-box annotation.	Objectness prior have been utilized in heuristic SOD models.
4	Object detection	To locate object(s) from fixed category list.	Rectangular bounding-box annotation.	Locates all instances of desired type, not just salient.
5	Salient object subitizing	Find existence and the number of salient objects.	Pixel accurate annotation with a count.	Indexing of individual objects as salient.

References

1. Ma, C.; Miao, Z.; Zhang, X.P.; Li, M. A saliency prior context model for real-time object tracking. IEEE Trans. Multimed. 2017, 19, 2415–2424.
2. Lee, H.; Kim, D. Salient region-based online object tracking. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, NV, USA, 12–15 March 2018; pp. 1170–1177.
3. Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 2048–2057.
4. Qin, C.; Zhang, G.; Zhou, Y.; Tao, W.; Cao, Z. Integration of the saliency-based seed extraction and random walks for image segmentation. Neurocomputing 2014, 129, 378–391.
5. Fu, H.; Xu, D.; Lin, S. Object-based multiple foreground segmentation in RGBD video. IEEE Trans. Image Process. 2017, 26, 1418–1427.
6. Donoser, M.; Urschler, M.; Hirzer, M.; Bischof, H. Saliency driven total variation segmentation. In Proceedings of the 2009 IEEE 12th International Conference on Computer Vision, Kyoto, Japan, 29 September–2 October 2009; pp. 817–824.
7. Borji, A.; Cheng, M.M.; Hou, Q.; Jiang, H.; Li, J. Salient object detection: A survey. Comput. Vis. Media 2019, 5, 117–150.
8. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. 1998, 20, 1254–1259.
9. Hou, X.; Zhang, L. Saliency detection: A spectral residual approach. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
10. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
11. Wang, L.; Wang, L.; Lu, H.; Zhang, P.; Ruan, X. Salient object detection with recurrent fully convolutional networks. IEEE Trans. Pattern Anal. Mach. Intell. 2018, 41, 1734–1746.
12. Chen, S.; Tan, X.; Wang, B.; Lu, H.; Hu, X.; Fu, Y. Reverse Attention-Based Residual Network for Salient Object Detection. IEEE Trans. Image Process. 2020, 29, 3763–3776.
13. Zhang, J.; Zhang, T.; Dai, Y.; Harandi, M.; Hartley, R. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9029–9038.
14. Achanta, R.; Hemami, S.; Estrada, F.; Susstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1597–1604.
15. Luo, Z.; Mishra, A.; Achkar, A.; Eichel, J.; Li, S.; Jodoin, P.M. Non-local deep features for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6609–6617.
16. Feng, G.; Bo, H.; Sun, J.; Zhang, L.; Lu, H. CACNet: Salient Object Detection via Context Aggregation and Contrast Embedding. Neurocomputing 2020, 403, 33–44.
17. Treisman, A.M.; Gelade, G. A feature-integration theory of attention. Cogn. Psychol. 1980, 12, 97–136.

18. Wolfe, J.M.; Cave, K.R.; Franzel, S.L. Guided search: An alternative to the feature integration model for visual search. J. Exp. Psychol. Hum. Percept. Perform. 1989, 15, 419.

19. Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. In Matters of Intelligence; Springer: Dordrecht, The Netherlands, 1987; pp. 115–141.

20. Liu, T.; Sun, J.; Zheng, N.; Tang, X.; Shum, H. Learning to Detect A Salient Object. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 7–22 June 2007; pp. 1–8.

21. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012; pp. 1097–1105.

22. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.

23. Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 39, 640–651.

24. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. 2017, 40, 834–848.

25. Noh, H.; Hong, S.; Han, B. Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1520–1528.

26. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 91–99.

27. Shen, Y.; Ji, R.; Zhang, S.; Zuo, W.; Wang, Y. Generative adversarial learning towards fast weakly supervised detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 5764–5773.

28. Chen, B.; Li, P.; Sun, C.; Wang, D.; Yang, G.; Lu, H. Multi attention module for visual tracking. Pattern Recognit. 2019, 87, 80–93.

29. Wang, L.; Ouyang, W.; Wang, X.; Lu, H. Visual tracking with fully convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3119–3127.

30. Ramanathan, S.; Katti, H.; Sebe, N.; Kankanhalli, M.; Chua, T.S. An eye fixation database for saliency detection in images. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, Greece, 5–11 September 2010; Springer: Berlin/Heidelberg, Germany, 2010; pp. 30–43.

31. Huang, X.; Shen, C.; Boix, X.; Zhao, Q. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 262–270.

32. Shi, J.; Malik, J. Normalized cuts and image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 2000, 22, 888–905.

33. Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. 2002, 24, 603–619.

34. Alexe, B.; Deselaers, T.; Ferrari, V. Measuring the objectness of image windows. IEEE Trans. Pattern Anal. Mach. Intell. 2012, 34, 2189–2202.

35. Zhang, J.; Ma, S.; Sameki, M.; Sclaroff, S.; Betke, M.; Lin, Z.; Shen, X.; Price, B.; Mech, R. Salient object subitizing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4045–4054.

36. Cong, R.; Lei, J.; Fu, H.; Cheng, M.M.; Lin, W.; Huang, Q. Review of visual saliency detection with comprehensive information. IEEE Trans. Circuits Syst. Video Technol. 2018, 29, 2941–2959.

37. Cao, X.; Tao, Z.; Zhang, B.; Fu, H.; Feng, W. Self-adaptively weighted co-saliency detection via rank constraint. IEEE Trans. Image Process. 2014, 23, 4175–4186.

38. Huang, R.; Feng, W.; Sun, J. Color feature reinforcement for cosaliency detection without single saliency residuals. IEEE Signal Process. Lett. 2017, 24, 569–573.

39. Wei, L.; Zhao, S.; Bourahla, O.E.F.; Li, X.; Wu, F. Group-wise deep co-saliency detection. arXiv 2017, arXiv:1707.07381.

40. Jacobs, D.E.; Goldman, D.B.; Shechtman, E. Cosaliency: Where people look when comparing images. In Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology, New York, NY, USA, 3–6 October

2010; pp. 219–228.

41. Wang, W.; Shen, J.; Sun, H.; Shao, L. Video co-saliency guided co-segmentation. IEEE Trans. Circuits Syst. Video Technol. 2017, 28, 1727–1736.
42. Fu, H.; Cao, X.; Tu, Z. Cluster-based co-saliency detection. IEEE Trans. Image Process. 2013, 22, 3766–3778.
43. Chen, H.; Li, Y. Progressively complementarity-aware fusion network for RGB-D salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3051–3060.
44. Audet, F.; Allili, M.S.; Cretu, A.M. Salient Object Detection in Images by Combining Objectness Clues in the RGBD Space. In Proceedings of the International Conference Image Analysis and Recognition, Montreal, QC, Canada, 5–7 July 2017; Springer: Cham, Switzerland, 2017; pp. 247–255.
45. Wang, A.; Wang, M. RGB-D salient object detection via minimum barrier distance transform and saliency fusion. IEEE Signal Process. Lett. 2017, 24, 663–667.
46. Sheng, H.; Liu, X.; Zhang, S. Saliency analysis based on depth contrast increased. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 1347–1351.
47. Chen, H.; Li, Y.F.; Su, D. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 6821–6826.
48. Zhao, X.; Zhang, L.; Pang, Y.; Lu, H.; Zhang, L. A single stream network for robust and real-time rgb-d salient object detection. arXiv 2020, arXiv:2007.06811.
49. Fan, D.P.; Lin, Z.; Zhang, Z.; Zhu, M.; Cheng, M.M. Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. IEEE Trans. Neural Netw. Learn. Syst. 2020.
50. Feichtenhofer, C.; Pinz, A.; Wildes, R.P. Dynamically encoded actions based on spacetime saliency. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 2755–2764.
51. Li, Z.; Qin, S.; Itti, L. Visual attention guided bit allocation in video compression. Image Vis. Comput. 2011, 29, 1–14.
52. Fan, D.P.; Wang, W.; Cheng, M.M.; Shen, J. Shifting more attention to video salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 8554–8564.
53. Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; Hanrahan, P. Light Field Photography with a Hand-Held Plenoptic Camera. Ph.D. Thesis, Stanford University, Stanford, CA, USA, 2005.
54. Wang, T.; Piao, Y.; Li, X.; Zhang, L.; Lu, H. Deep learning for light field saliency detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 8838–8848.
55. Fan, D.P.; Lin, Z.; Ji, G.P.; Zhang, D.; Fu, H.; Cheng, M.M. Taking a Deeper Look at Co-Salient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 2919–2929.
