Journal of Management & Entrepreneurship
ISSN 2229-5348

UGC Care Group I Journal
Vol-09 Issue-02 Dec 2020

Cross-Modal Feature Aggregation Network for Light Field Salient Object Detection

Dr. Subrat Kumar Mohanty,
College of Engineering Bhubaneswar

Abstract—Although existing CNN-based saliency identification techniques are especially made for RGB images, not light fields, light field saliency detection can take advantage of the rich visual properties of light field (LF) to highlight the salient regions. A three-stream cross-modal feature aggregation network is suggested for 4D light field saliency detection in order to address this issue. Three smaller networks are configured to analyze the depth map, all-focus image, and focal stack, respectively, in order to fully take advantage of the rich visual properties of light field. Next, cross-level features are aggregated top-down using feature aggregation modules. In the end, a cross-modal feature fusion module is made to combine the combined features of different modalities from the three sub-networks, enabling fast and accurate identification of salient objects. In comparison to state-of-the-art (SOTA) methods, extensive experiments on three benchmark datasets demonstrate the effectiveness and superiority of the proposed algorithm on five evaluation metrics, both qualitatively and numerically.

Index Terms: saliency, light field, feature aggregation, depth map, and salient object identification.

I. Introduction

SALIENT object detection(SOD) refers to detect the most informative object that grab human attention, which has attracted increasing attention due to its importance in different kinds of applications, such as visual recognition [5], object tracking [3], and camouflaged object detection [1]. As is known to all, light field includes the RGB color information, and the directions of all incoming light that contains abundant geometric information of scene objects. Recently, With the development of light field imaging technology, light field has been exploited to enhance the performance of various tasks, including material recognition [4], depth estimation [2], and etc.

Existing SOD methods can be roughly divided into 2D-RGB, 3D-RGBD and 4D-LF saliency detection according to their input data. Most of them fit into the first category, and the rest are part of the last two categories. Among them, 2D-RGB SOD algorithms [6]–[12] have achieved promising progress in recent years, which profits from the great development of convolutional neural networks(CNN). But these methods usually acquire mediocre performance when they encounter challenging scenes. The reasons are mainly about two aspects, one is the dependence of 2D-RGB methods on prior knowledge, the other is the deficiency of 3D visual information in limited RGB images. At the same time, 3D-RGBD SOD methods [13]–[21] have also attracted researcher's interest, because depth map equiped with rich geometric information is helpful for understanding of contextual information of salient objects, and can improve the accuracy of saliency detection to some extent. However, depth map with low quality severely degrade the performance of SOD algorithms.

Recently, with the popularity of light field camera, light field data is easy to obtain, which contains focal stack, all-focus image, and depth map. Focal stack displays various focus depth levels, and contains rich visual information. Consequently, light field has a good application prospect in 4D-LF SOD [23]–[30]. Still, most existing 4D-LF SOD methods have not taken full advantage of the rich visual information of light field image, which just extract some hand-crafted features from light field. On the whole, such approaches have less been explored mainly because 4D light field data is more difficult to tackle than 2D RGB image. Moreover, except for MoLF [30] and DLLF [29], CNN-based SOD methods have been ignoring from current researches in 4D LF-SOD. In view of this, it is of importance to open up research of the CNN-like framework for 4D-LF SOD, as do 2D-RGB and 3D-RGBD SOD approaches. Inspired by MoLF, to make use of CNN-like framework, a three-stream cross-modal feature aggregation network is proposed for 4D-LF SOD in this letter. In order to automatically extract discriminative features of various modalities, three sub-networks with similar network structure are designed in parallel to extract cross-level features from focal stack, depth map, and all-focus image separately. To make the best of the complementarity of cross-modal features, a specifically designed cross-modal feature fusion module is used to fuse these cross-modal features in a holistic perspective.

In short, our main contributions are summarized as follows:

1) We propose a three-stream cross-modal feature aggregation network for 4D-LF saliency detection, each

TABLE I
Quantitative Comparison by Removing Some Modules

Modules	DUTLF-FS				HFUT-Lytro				LFSD			
	$E_\phi\uparrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	MAE↓	$E_\phi\uparrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	MAE↓	$E_\phi\uparrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	MAE↓
$S_{NO\text{-}Both}$	**0.832**	**0.794**	**0.785**	**0.130**	**0.752**	**0.685**	**0.623**	**0.151**	**0.762**	**0.685**	**0.661**	**0.193**
$S_{NO\text{-}CFR}$	**0.856**	**0.824**	**0.807**	**0.112**	**0.761**	**0.723**	**0.648**	**0.139**	**0.796**	**0.737**	**0.687**	**0.168**
$S_{NO\text{-}CFFM}$	0.903	0.872	0.864	0.071	0.787	0.796	0.665	0.117	0.818	0.793	0.789	0.145
S_{Both}	0.951	0.914	0.912	0.038	0.832	0.816	0.724	0.068	0.834	0.821	0.810	0.087

sub-network aggregates multi-scale multi-level features from single modality, and the cross-modal feature fusion module makes full use of features from different modalities.

2) Compared with 20 SOTA 2D-RGB, 3D-RGBD and 4D-LF SOD approaches, extensive experiments on three datasets show that the proposed approach achieves superior performance on five evaluation metrics.

II. Related Works

As mentioned before, the existing SOD algorithms can be roughly summarized into 2D-RGB [31], 3D-RGBD, and 4D-LF saliency detection. From another point of view, these algorithms can be simply divided into traditional and CNN-based methods. The former is mainly focus on the hand-crafted features which cannot deal with the challenging scenarios where the tacit assumption is not satisfied with, here we mainly discuss the CNN-based methods, which has achieved excellent performance recently.

A. 2D-RGB CNN-Based Saliency Detection

With the development and wide application of CNNs, a variety of CNN-based saliency detection approaches have been proposed in recent years. These approaches have been combined with contextual features [11], post-treatment steps [32], attention modules [11], refinement model [12], and etc. Li *et al.* [33] propose an end-to-end deep contrast network, which produces pixel-level saliency maps, and then improve the fused saliency map by a fully connected CRF model. Hou *et al.* [10] introduce short connections to the skip-layer structures within the HED architecture, which combines multi-scale feature maps, and fuses these feature maps to segment salient objects. Zhang *et al.* [34] present a generic aggregating multi-level convolutional feature framework, which integrate multi-level features in multiple resolutions and combine them to predict saliency maps. Soon afterwards, Deng *et al.* [12] propose a recurrent residual refinement network equipped with residual refinement blocks to learn the complementary saliency information of the intermediate prediction. Li *et al.* [21] create a contour-to-saliency network with two branches to predict contours and estimate pixel-level saliency, then automatically transfer contour knowledge to saliency detection without using any manual saliency masks. Liu *et al.* [11] propose a pixel-wise contextual attention network to learn contextual features to generate saliency map in the global and local form.

By and large, CNNs can automatically extract low-level and high-level visual features, and conduct a mapping between images and prediction maps. But it is never wise to directly apply the existing 2D-RGB CNN-based models to light field, because these models are not well qualified for light field. In addition, 2D RGB image is deficient in 3D visual information. It is necessary to construct a novel CNN-based network for light field data. Detailed summaries about 2D-RGB CNN-based saliency detection can be found in [35].

B. 3D-RGBD Saliency Detection

In the last three years, 3D-RGBD saliency detection attracts more and more attention of many researchers. Qu *et al.* [13] design a CNN-based RGBD SOD to automatically learn the interaction mechanism, and exploit hand-crafted features to train the CNN-based SOD model. Chen *et al.* [15], [16], [21] exploit cross-level complementarity and cross-modal complementarity, and design multi-scale multi-path fusion network to fuse multi-level features from RGB or depth modality to predict saliency maps. Chen *et al.* [17] propose a three-stream attention-aware fusion network to extract RGB-D features and introduce channel-wise attention mechanism to adaptively select complementary feature maps. Zhu *et al.* [18] present an independent encoder network to process depth cue, and utilize RGB-based prior-model to guide the main learning stage. Wang *et al.* [19] propose an adaptive fusion scheme with two-streamed CNN to fuse saliency predictions generated from the RGB and depth modalities. Piao *et al.* [20] propose a depth-induced multi-scale recurrent attention network which includes a depth refinement block to extract and fuse complementary RGB and depth features, a depth-induced multi-scale weighting module, and a recurrent attention module to generate more accurate saliency results in a coarse-to-fine manner.

C. 4D-LF Saliency Detection

4D-LF SOD is still in the early stage, only a few available models fall into this category, but such models have exhibited good prospects in some complex scenes. The pioneering work by Li *et al.* [23] builds the first light field saliency dataset, and demonstrates the practicability of detecting salient objects utilizing all-focus images and focal stacks of light field. Li *et al.* [24] also propose a weighted sparse coding framework to handle the heterogenous types of input (RGB, RGB-D and light field image). Zhang *et al.* [25] extend 2D contrast-based SOD method by introducing depth cue in connection with location and background prior into 4D-LF SOD, which reveal the advantage and effectiveness of light field. Zhang *et al.* [27] also introduce a computational SOD scheme by integrating various visual cues from light-field image, and present a benchmark dataset, named HFUT-Lytro. Very recently, Wang *et al.* [29] propose a fusion framework with two CNN streams where the focal stacks and all-focus images serve as the input, adversarial examples is used to help train the deep network and improve the robustness of its approach. Zhang *et al.* [30] introduce a light filed saliency dataset with 1462 samples, named DUTLF-FS, and propose a novel memory-oriented decoder tailored for 4D-LF SOD, which can precisely predict salient objects by means of Mo-SFM and Mo-FIM modules.

Some visual examples are shown in Fig. 1. Obviously, benefit from the rich visual information of light field image, these 4D methods perform better than 2D and 3D SOD methods in some challenging scenes, for instance, complex and cluttered background, similar foreground and background, and low intensity environment. However, relatively few efforts have been spent in modeling 4D SOD by taking all input information into consideration, this leads to insufficient multi-modal fusion.

III. Proposed Method

A. Top-Down Feature Aggregation Stream

The overall architecture of the proposed approach is shown in Fig. 2. The three stream networks take all-focus image, depth

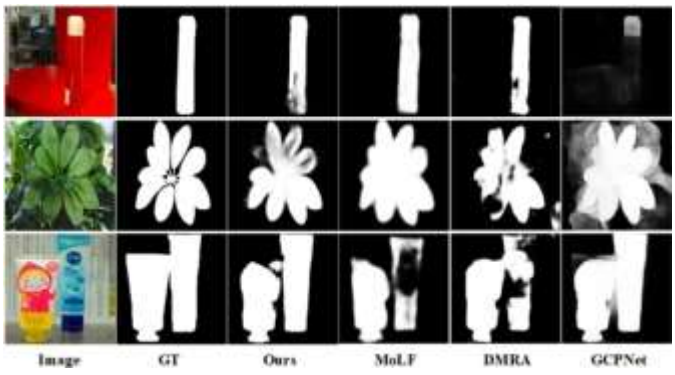

Fig. 1. Image and ground truth(GT) of three samples with the corresponding saliency map of our algorithm and other SOTA approaches including MoLF [30], DMRA [20], and GCPNet [6].

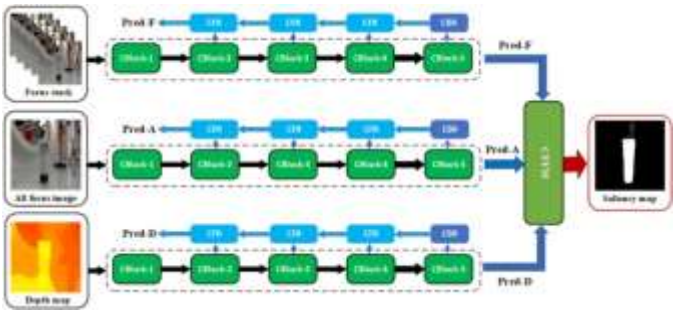

Fig. 2. Overall pipeline of the proposed approach for 4D-LF SOD.

map, and focal stack as input respectively. The all-focus image across RGB channels is fed into the first stream, all slices of focal stack are concatenated and then fed into the second stream. Follow the practice in [15], [17], the depth map is encoded into three-channel HHA representation and then fed into the third stream. Following EGNet [8], we insert three convolutional layers on each side path to get more discriminative context features, each convolutional layer is followed a ReLU layer to ensure the nonlinearity. As everyone knows, high-level feature is helpful to locate salient objects and remove noises. By contrast, low-level feature can provide more spatial structure information. Obviously, these two level features are complementary with each other [6], [9]. Furthermore, the global contextual information is also conducive to detecting more complete and accurate salient objects. Therefore, a cross-level feature refinement module, denoted as CFR, is introduced to refine features in top-down manner. Multiple CFR modules are connected in series from top to bottom to obtain more discriminative features.

Specifically, the CFR module receives the feature f_l^i passed the three convolutional layers on each side path and the feature f_h^{i+1} from the contiguous upper-level CFR module, f_l^i is the feature generated by the i-th side path. Through this top-down supervision, these features are gradually aggregated and refined. Unlike previous works (PoolNet [22]) that often integrate these features by concatenation or addition operation, we directly multiply them together to restrain the background noises as follows.

$$f_h^i = f_h^{i+1} \otimes Conv_{1\times1}(f_l^i) \quad (1)$$

Specially, the last convolution block, denoted as CBlock-5, followed by a CBR module that is composed of a convolutional layer, a Batch Normalization layer, and a ReLU operation, which can generate more accurate high-level semantic feature.

B. Cross-Modal Refined Feature Fusion

As shown in Fig. 2, in order to make full use of the complementarity of cross-modal features, a cross-modal feature fusion module, denoted as CFFM, is designed to effectively fuse cross-modal features from the three sub-networks to generate the final saliency prediction map.

In particular, to sufficiently capture the cross-modal complementary information, the CFFM is designed as a two-stage fusion process. In the first stage, we learn a prediction map for feature fusion of the all-focus stream and focal stack stream. In the second stage, the fused prediction map in first stage, and the feature map of depth stream, are further fused to generate the final saliency map in the same way. In each fusion stage, inspired by AFNet [19], we first concatenate the features of two streams, and then the fused prediction map is fed into two 1×1 convolutional layers to suit the intermediate supervision and reduce interference during training. It is described by the following formulations.

$$\cancel{F^{s1}_{sw}} = Conv_{1\times1}(Concat(F^{FS}_{ref}, F^{AF}_{ref})) \quad (2)$$

$$F^{s1}_{fuse} = F^{AF}_{ref} \otimes \cancel{F^{s1}_{sw}} + F^{FS}_{ref} \otimes (1 - \cancel{F^{s1}_{sw}}) \quad (3)$$

$$\cancel{F^{s2}_{sw}} = Conv_{1\times1}(Concat(F^{s_1}_{fuse}, F^{DM}_{ref})) \quad (4)$$

$$F^{s2}_{fuse} = F^{s1}_{fuse} \otimes \cancel{F^{s2}_{sw}} + F^{DM}_{ref} \otimes (1 - \cancel{F^{s2}_{sw}}) \quad (5)$$

Here F^{FS}_{ref}, F^{AF}_{ref} and F^{DM}_{ref} are features from the three all-focus, focal stack and depth stream sub-networks, respectively. $F^{s_1}_{fuse}$ is fused prediction map in first stage, and $F^{s_2}_{fuse}$ is the final saliency map. $\otimes$ is pixel-wise multiplication operator.

C. Loss Function

In the training stage, training set can be represented as $T = (FS_i, AF_i, DM_i, GT_i)^N$, N is the total number of samples with M piexls in the training set, FS_i, AF_i, DM_i, GT_i are focal stack, all-focus image, depth map, and ground truth map respectively. There are three saliency maps generated by the proposed model, denoted as S_m^{AF}, S_m^{FS}, and S_m^{DM} respectively. The proposed three-stream feature aggregation network is trained to extract and fuse cross-level and cross-modal features in each stream sub-network, which hearten Complementary integration and refine the multiple-level feature map gradually. The cross-entropy loss of the i^{th} level is defined by

$$L_s^i = -\sum_{i=1}^{N}\sum_{j=1}^{M} gt_{ij} \log F_s^{ij} + (1 - gt_{ij}) \log\ 1 - F_s^{ij} \quad (6)$$

where $F_s^{ij} = \omega_i F_{ref}^i + \omega_{i+1} F_{ref}^{i+1}, i = 2, 3, 4.$

Consequently, the cross-modal multiple-level feature loss between predicted feature map and ground truth map can be represented as

$$L_s = L_s^{AF} + L_s^{FS} + L_s^{DM} \quad (7)$$

In the two-stage fusion process, same to AFNet [19], the switch loss between single-modal prediction map and pseudo ground truth, is calculated by cross-entropy loss, which can be represented as

$$L_f = L_f^{s1} + L_f^{s2} \tag{8}$$

Therefore, the total loss fuction is defined by

$$L = L_f + L_s \tag{9}$$

IV. Experiments and Analysis

A. Experiments Setup

We conduct experiments on three benchmark datasets designed for 4D-LF SOD: LFSD [23], HFUT-Lytro [27], and DUTLF-FS dataset [29], [30]. The train set of DUTLF-FS is used for training, and all images are uniformly resized to 256×256. Six widely-accepted metrics are adopted to verify the effectiveness of the proposed method, including precision-recall(PR) curves, F-measure [36], Weighted F-measure [37], mean absolute error(MAE) [38], Structure-measure(Sm) [39] and Enhanced-alignment measure(Em) [40].

B. Ablation Analysis

To validate the proposed CFR and CFFM modules, we further analyze the results of the following cases: the model without CFR and CFFM module, the model without CFR but with CFFM module, the model with CFR but without CFFM module, the model with CFR and CFFM module, denoted by $S_{NO-Both}$, S_{NO-CFR}, $S_{NO-CFFM}$, S_{Both}. From the Table I, it can be seen that the CFR and CFFM modules are indispensable for saliency detection. The former is used to integrate multiple level visual features, and the latter is used to fuse cross-modal features or prediction maps.

C. Comparison With the State-of-The-Art

The proposed method is compared with 11 state-of-the-art 2D, 3D and 4D SOD approaches, including four 2D CNN-based methods:EGNet [8], CPD [7], GCPNet [6], and F3Net [9]; two 3D methods: AFNet [19], and DMRA [20]; and five 4D methods for light field: LFS [23], WSC [24], DILF [25], MCA [27], and MoLF [30]. For a fair comparisons, the saliency maps of other SOTA methods are generated by official codes by using the recommended parameter settings provided by the authors or directly provided by authors.

For qualitative evaluation, we draw the Precision, Recall, F-measure, Weighted F-measure, and MAE scores in Fig. 3. As we can see, the proposed algorithm achieves better results compared with other SOTA approaches, which achieves the lowest MAE score, and obtains the best scores on HFUT-Lytro datasets across all four evaluation metrics. Some selected representative samples of the visual comparison of our method and the current SOTA methods are further shown in Fig. 4. It can be observed that the proposed approach is able to accurately detect complete salient objects from various challenging scenes, including big salient object as shown in the 3 rd row, clutter background as shown in row 4 and 5, small salient objects as shown in rows 6,

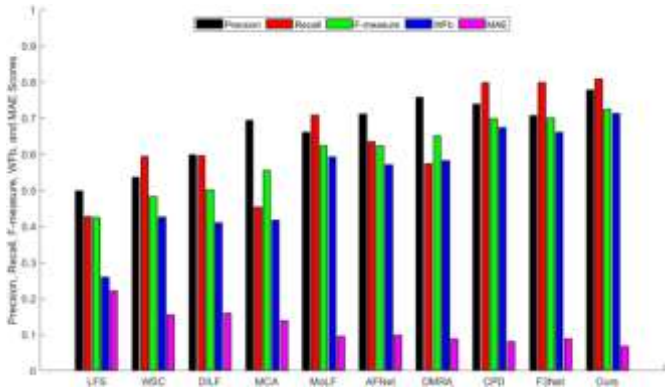

Fig. 3. Precision, Recall, Weighted F-measure, and MAE scores of our method and other SOTA approaches on HFUT-Lytro [27] datasets.

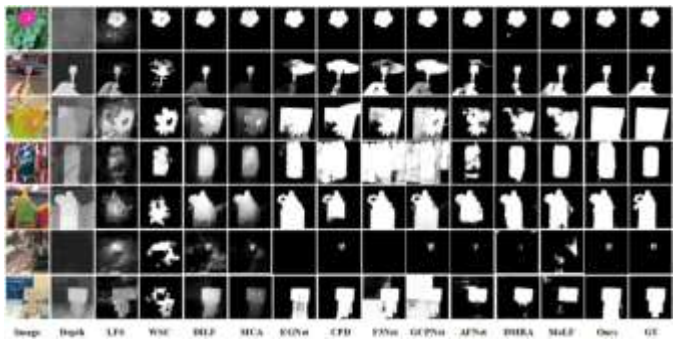

Fig. 4. Visual comparison of saliency maps from ours and other SOTA approaches for 4D-LF SOD.

and similar foreground and background color as shown in the last row.

Based on above-mentioned comparison, an interesting observation should be noted: the latest CNNs-based SOD methods, especially those methods that make use of the characteristics of light field, which obtains better results than others. This indicates that both cross-modal light field features and cross-level features extracted by CNN-based model are significant and promising for SOD. It is a wise choice to take advantage of both cross-modal features from light field data, and cross-level multi-scale features from CNN-based method.

V. Conclusion

Our goal in this letter is to maximize light field visual information for SOD. It is suggested to use a three-stream cross-modal feature aggregation network to identify objects of interest in a light field. In order to combine and enhance the multiple-level features within each sub-network, a few cross-level feature aggregation modules are introduced. A cross-modal feature fusion module is meant to fuse the prediction maps from various modalities in order to further capitalize on the complementarity among the three modalities. According to experimental data, the suggested strategy outperforms SOTA approaches across a range of assessment measures. It has also been demonstrated that SOD may be effectively and profitably achieved by exploiting feature extraction capabilities and mining rich visual information of light fields.

REFERENCES

[1] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2774–2784.

[2] W. Zhou, L. Liang, H. Zhang, A. Lumsdaine, and L. Lin, "Scale and orientation aware EPI-patch learning for light field depth estimation," in *Proc. 24th Int. Conf. Pattern Recognit.*, Aug. 2018, pp. 2362–2367.

[3] H. Lee and D. Kim, "Salient region-based online object tracking," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1170–1177.

[4] T. Wang, J. Zhu, H. Ebi, M. Chandraker, A. A. Efros, and R. Ramamoorthi, "A 4D light-field dataset and CNN architectures for material recognition," *Comput. Vis. Lecture Notes Comput. Sci.*, vol. 9907, pp. 121–138, 2016.

[5] T. Huang *et al.*, "Salient region detection and segmentation for general object recognition and image understanding," *Sci. China Inf. Sci.* vol. 54, pp. 2461–2470, 2011.

[6] H. W. F. Yeung, J. Hou, X. Chen, J. Chen, Z. Chen, and Y. Y. Chung, "Light field spatial super-resolution using deep efficient spatial-angular separable convolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 231–2330, May 2019.

[7] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3907–3916.

[8] J. Zhao, J. Liu, D. Fan, Y. Cao, J. Yang, and M. Cheng, "EGNet: Edge guidance network for salient object detection," *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct./Nov. 2019, pp. 8778–8787.

[9] J. Wei, S. Wang, and Q. Huang, "F3Net: Fusion, feedback and focus for salient object detection," *Proc. AAAI Conf. Artificial Intell.*, vol. 34, no. 7, pp. 12321–12328.

[10] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5300–5309.

[11] N. Liu, J. Han, and M. Yang, "PICANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[12] Z. Deng *et al.*, "R3Net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

[13] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGB-D salient object detection via deep fusion," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2274–2285, May 2017.

[14] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Process. Lett.*, vol. 24, no. 5, pp. 663–667, May 2017.

[15] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3051–3060.

[16] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection," *Pattern Recognit.*, vol. 86, pp. 376–385, 2019.

[17] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[18] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, "PDNet: Prior-model guided depth-enhanced network for salient object detection," in *Proc. IEEE Int. Conf. Multimedia Expo.*, Jul. 2019, pp. 199–204.

[19] N. Wang and X. Gong, "Adaptive fusion for RGB-D salient object detection," *IEEE Access*, vol. 7, pp. 55277–55284, 2019.

[20] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7253–7262.

[21] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[22] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3912–3921.

[23] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2806–2813.

[24] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5216–5223.

[25] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field," in *Proc. 24th Int. Joint Conf. Artif. Intell.*, IJCAI 2015, Q. Yang and M. J. Wooldridge, Eds., Buenos Aires, Argentina: AAAI Press, Jul. 25–31, 2015, pp. 2212–2218.

[26] A. Wang, M. Wang, X. Li, Z. Mi, and H. Zhou, "A two-stage Bayesian integration framework for salient object detection on light field," *Neural Process. Lett.*, vol. 46, no. 3, pp. 1083–1094, 2017.

[27] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency detection on light field: A multi-cue approach," *ACM Trans. Multim. Comput. Commun. Appl.*, vol. 13, no. 3, pp. 32-1–32-22, Jul. 2017.

[28] A. Wang, M. Wang, G. Pan, and X. Yuan, "Salient object detection with high-level prior based on Bayesian fusion," *IET Comput. Vis.*, vol. 11, no. 3, pp. 199–206, 2017.

[29] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8837–8847.

[30] M. Zhang, J. Li, J. Wei, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Proc. Adv. Neural Inf. Process. Syst. 32: Annu. Conf. Neural Inf. Process. Syst.*, NeurIPS Dec. 2019, pp. 8–14, Vancouver, BC, Canada (H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alche-Buc, E. B. Fox, and R. Garnett, eds.), 2019, pp. 896–906.

[31] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 186–202.

[32] G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Trans. Image Process.*, vol. 25, no. 11, pp. 5012–5024, Nov. 2016.

[33] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Patt. Recognit.*, Jun. 2016, pp. 478–487.

[34] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "AMULET: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 202–211.

[35] Q. Wang, L. Zhang, Y. Li, K. Kpalma, "Overview of deep-learning based methods for salient object detection in videos," *Pattern Recognition*, vol. 104, 2020, doi: 10.1016/j.patcog.2020.107340.

[36] R. Achanta, S. S. Hemami, F. J. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[37] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

[38] F. Perazzi, P. Krahenb uhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

[39] D. Fan, M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4558–4567.

[40] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.