ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/334844293

Deep Light-field-driven Saliency Detection from a Single View

Conference Paper · August 2019

DOI: 10.24963/ijcai.2019/127

CITATIONS
62

READS
712

5 authors, including:

Zhengkun Rong
Dalian University of Technology
6 PUBLICATIONS **708** CITATIONS

SEE PROFILE

All content following this page was uploaded by Zhengkun Rong on 24 January 2021.

The user has requested enhancement of the downloaded file.

Deep Light-field-driven Saliency Detection from a Single View

Yongri Piao[1*], **Zhengkun Rong**[1], **Miao Zhang**[2,3], **Xiao Li**[1] and **Huchuan Lu**[1]
[1]School of Information and Communication Engineering, Dalian University of Technology, Dalian, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, Dalian, China
[3]DUT-RU International School of Information and Software Engineering, Dalian University of Technology, Dalian, China
{yrpiao@, rzk911113@mail, miaozhang@, lixiaohao@mail, lhchuan@}dlut.edu.cn

Abstract

Previous 2D saliency detection methods extract salient cues from a single view and directly predict the expected results. Both traditional and deep-learning-based 2D methods do not consider geometric information of 3D scenes. Therefore the relationship between scene understanding and salient objects cannot be effectively established. This limits the performance of 2D saliency detection in challenging scenes. In this paper, we show for the first time that saliency detection problem can be reformulated as two sub-problems: light field synthesis from a single view and light-field-driven saliency detection. We propose a high-quality light field synthesis network to produce reliable 4D light field information. Then we propose a novel light-field-driven saliency detection network with two purposes, that is, i) richer saliency features can be produced for effective saliency detection; ii) geometric information can be considered for integration of multi-view saliency maps in a view-wise attention fashion. The whole pipeline can be trained in an end-to-end fashion. For training our network, we introduce the largest light field dataset for saliency detection, containing 1580 light fields that cover a wide variety of challenging scenes. With this new formulation, our method is able to achieve state-of-the-art performance.

1 Introduction

Salient object detection aims to extract the most relevant parts that grab human attention of what we see. As a fundamental task in computer vision, it has received increasing attention in recent years because of its great success in many fields, such as semantic segmentation, tracking and person re-identification.

The existing saliency detection algorithms can be roughly divided into three categories based on the 2D, 3D and 4D input images. Among 3D and 4D saliency detection methods [Li *et al.*, 2014; Li *et al.*, 2015; Qu *et al.*, 2017; Han *et al.*, 2017], there are different kinds of inputs, such as depth maps, focal stacks and multi-view images. They all provide the accurate geometric information, which plays a dispensable role in extraction of salient objects. Recent advances in 3D and 4D saliency detection show the most promise in challenging scenes.

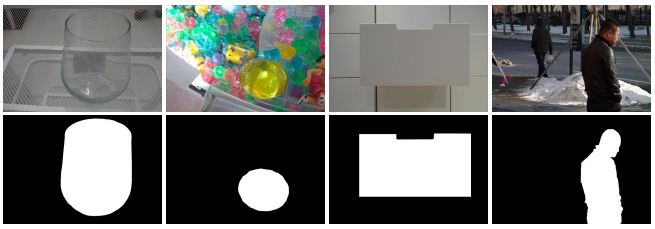

Figure 1: Challenging scenes in saliency detection (e.g. transparent objects, complex background, similar foreground and background and low intensity environment, etc.). First row is the center view image and second row is the corresponding ground truth.

On the other hand, among 2D saliency detection methods, the traditional approaches [Li *et al.*, 2013; Qin *et al.*, 2015; Tu *et al.*, 2016] mainly rely on various handcrafted features and prior knowledges. For example, the image boundary regions are mostly background (boundary prior), the color contrasts between foreground and background are high (contrast prior).The deep-learning-based methods benefit from the unique feature extraction capability of convolutional neural networks (CNNs). CNNs can extract both low-level and high-level features such as color, intensity, texture and semantic information. Many deep-learning-based methods [Liu and Han, 2016; Zhang *et al.*, 2017b; Hou *et al.*, 2017] contribute more meaningful feature representations to facilitate saliency detection. However, compared with the aforementioned 3D and 4D methods which are grounded by geometric information, 2D saliency detection methods may appear fragile when it comes to challenging scenes shown in Figure 1. The reasons for limited performance of 2D saliency detection methods are twofold. First, for traditional 2D methods, many prior knowledges are not fully effective in complex scenes. This significantly limits performance of traditional 2D methods. Second, current 2D deep-learning-based methods are empowered by the learning capability of CNNs. They directly relate multi-

*Contact Author

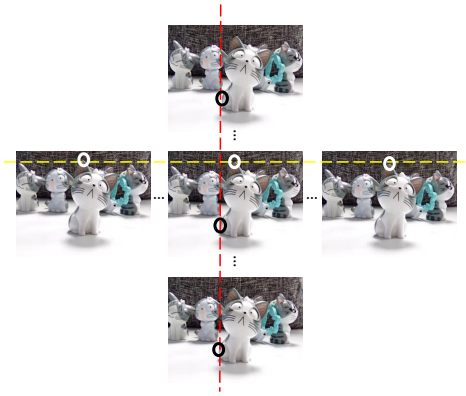

Figure 2: Multi-view images of one scene in the proposed dataset. There are seven images in horizontal and vertical directions respectively. White circles are in the same location of multi-view images in horizontal direction and black circles are in vertical direction.

level features to the ground truth but the relationship between scene understanding and salient objects can not be effectively established.

The light field provides images of the scene from an array of viewpoints which spread over the extent of the lens aperture. These different views provide abundant spatial parallax information as well as accurate depth information about the objects in the scene. Light field data has been demonstrated in favor of many applications in computer vision, including depth estimation [Zhou *et al.*, 2018; Guo *et al.*, 2017; Song and Lee, 2018], super resolution [Zhu *et al.*, 2019; Yeung *et al.*, 2019] and material recognition [Wang *et al.*, 2016b]. In this paper, inspired by light field, we create an end-to-end CNN framework, from a novel perspective, that decomposes the saliency detection problem into the subproblems: light field synthesis from a single view and light-field-driven saliency detection. The light field information (multi-view images and depth maps) can be automatically generated by the light field synthesis network. This allows 2D methods to have more reliable information for saliency detection in challenging scenes. The whole pipeline is shown in Figure 3. The key insights here are that i) both light field synthesis and light-field-driven saliency detection consider the geometric information. ii) the whole pipeline can be trained in an end-to-end fashion. Due to the limited number of light field datasets, we collected 1580 light fields that cover a wide variety of scenes, including multi-view images (shown in Figure 2) and a pixel-wise ground truth of central view. We discover that our saliency detection network is able to outperform all the previous methods on the proposed dataset.

In summary, we make following contributions:

- We collected the largest available light field dataset for saliency detection, containing 1580 light fields, divided into 1100 for training and 480 for testing, captured by the Lytro Illum camera. Each light field consists of multi-view images and a pixel-wise ground truth of the central view.
- We show for the first time that the saliency detection problem can be factorized into two subproblems: light field synthesis and light-field-driven saliency detection. This factorization can effectively improve the performance of saliency detection in challenging scenes.
- We propose a novel light-field-driven saliency detection network where the new saliency feature extraction technique facilitates saliency detection and the multi-view attention module helps integrate multi-view saliency maps in a view-wise way.
- The proposed method outperforms state-of-the-art 2D, 3D and 4D methods on the proposed light field dataset. The source code and light field saliency detection dataset can be found at https://github.com/OIPLab-DUT/.

2 Related Work

2.1 Saliency Detection Based on 3D and 4D Inputs

Recently, a small number of saliency detection works focus on using geometric information of the scene to help saliency estimation. The 3D approaches mainly exploit depth cues to help saliency detection. [Han *et al.*, 2017] utilize two CNNs to extract both RGB features and depth representations and make a fusion automatically to generate the final saliency prediction. [Qu *et al.*, 2017] motivated by the traditional saliency detection methods, first generate saliency feature vectors using RGB images and depth maps, and then use a CNN to learn from the existing features. [Zhu *et al.*, 2018] use a master network to process RGB values and a sub-network to exploit depth cues which can facilitate the prediction of master network. [Wang and Gong, 2019] propose a novel fusion module in which a switch map is learned to adaptively fuse the saliency maps predicted from RGB and depth map.

For the 4D approaches, there are only traditional methods at present. All of them concentrate on the low-level handcrafted features. [Li *et al.*, 2014] propose the first light field saliency detection approach in 2014, which compute the focusness and objectness of the focal stack firstly and then integrate focusness-based saliency candidates with other contrast cues using objectness as a weight. [Li *et al.*, 2015] develop a weighted sparse coding framework to handle the heterogenous types of input data. They first generate initial saliency candidate regions and then use an iterative method to refine the candidates. [Zhang *et al.*, 2015] utilize both background prior and location prior and introduce an additional depth cue into the contrast computation. [Zhang *et al.*, 2017a] integrate multiple saliency cues extracted from light field images by a random-serach-based weighting method. Those methods have better performance than 2D methods in some challenging scenes, such as similar foreground and background, transparent objects, complex background and low intensity environment, because light field information is involved. In our work, we adopt a high-quality light field rendering network in which the multi-view images can be automatically generated.

2.2 Saliency Detection Based on 2D Input

Over the past decades, lots of 2D saliency detection methods have been developed. The 2D saliency models can be roughly divided into two categories: traditional methods and deep-learning-based methods. The traditional methods [Qin *et al.*, 2015; Li *et al.*, 2013; Tu *et al.*, 2016] create the groundwork for saliency detection. With the development of deep

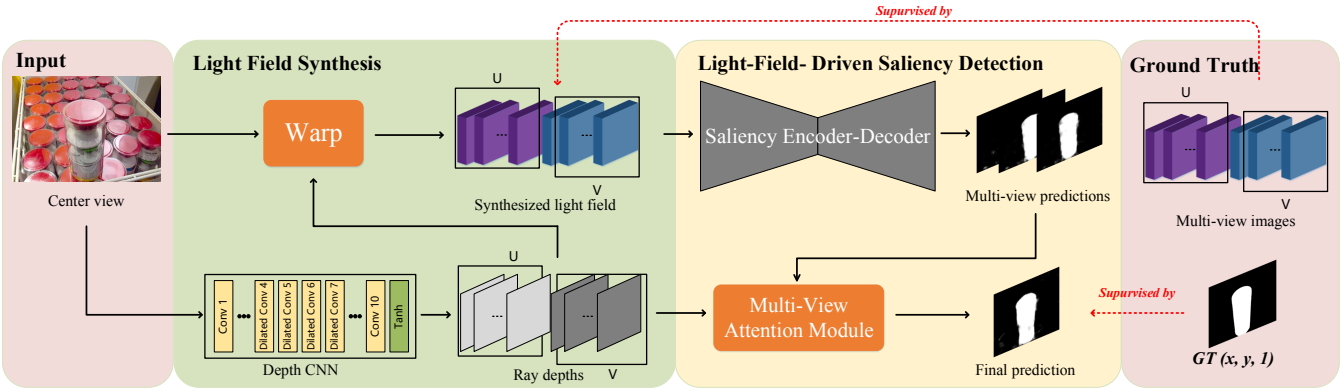

Figure 3: The whole pipeline.

learning, many new approaches of 2D saliency detection are proposed. In 2015, [Li and Yu, 2015] use three different CNNs with three kinds of inputs: the considered superpixel, the immediate neighboring segments and the whole image, and then aggregate all three information with fully connected layers to label the considered superpixel. [Li and Yu, 2016] propose a two-stream framework to extract both pixel-wise and segment-wise saliency maps and use a CRF to fuse them efficiently. [Liu and Han, 2016] develop a novel deep hierarchical saliency network to first generate a coarse saliency map and then using a RCL module to refine the details of saliency prediction step by step. [Wang *et al.*, 2016a] combine the FCN with a recurrent architecture, which can refine the saliency map progressively. [Zhang *et al.*, 2017b] integrate multi-level features into multiple resolutions and predict saliency maps hierarchically. A boundary preserved refinement is also applied to achieve accurate salient objects boundary inference. [Deng *et al.*, 2018] propose a series of residual refinement modules to alternately extract low-level and high-level features and use them to gradually refine the saliency map.

There is a strong relationship between scene understanding and salient objects. 2D saliency detection methods without consideration of that relationship are more likely to be compromised in those challenging scenes.

2.3 View Synthesis from Light Fields

Over the past decades, there are fewer works on light field rendering. [Levoy and Hanrahan, 1996] capture a densely-sampled 4D light field images of a scene and interpret the input images as 2D slices of a 4D light field. [Gortler *et al.*, 1996] use the silhouette information to compute the approximate geometry and utilize it to improve the quality of the rendered images. The unstructured lumigraph rendering framework [Buehler *et al.*, 2001] is designed to meet many specific goals, using a set of unstructured 2D slices of light field. The recent light field synthesis method [Srinivasan *et al.*, 2017] estimates depth maps, then uses the geometry to render the 4D light fields. But the huge amount of high dimensional data makes the processing of light field information time-consuming. For efficient learning, in this paper, we only synthesize the views along the horizontal and vertical directions. This is good enough to represent the original light field.

3 Light Field Dataset

There is no light field dataset, in saliency detection, containing a large number of images for training. To solve the lack of light field dataset, we collected 1580 light fields to build the largest available light field dataset with the Lytro Illum camera. We decoded the light field format file using the Lytro Desktop, which makes better 2D renderings than other light field toolboxes. The decoded light field images have no distortion in shape, intensity and color. Each light field consists of multi-view images and a pixel-wise ground truth of the central view. Figure 2 shows a sample of the multi-view images in the proposed dataset. The spatial resolution of each view is 400×590 and the angular resolution in both horizontal and vertical directions is 7. We use a custom segmentation tool to manually label the central view in light field. Our dataset is randomly split into 1100 for training and 480 for testing.

This dataset includes light fields with a wide range of indoor and outdoor scenes. Each scene in our dataset was captured related to the surrounding environment of our daily life, such as supermarkets, offices, classrooms, streets and so on. Many challenging scenes are included in the proposed dataset (e.g. similar foreground and background, transparent objects, complex background and low intensity environment, etc.).

4 The Proposed Method

In this section, we formulate the saliency detection problem as two sub-problems, namely light field rendering and light-field-driven saliency detection. The whole pipeline is shown in Figure 3. We will show the network details of two parts in the following part, respectively.

4.1 Analysis of the Whole Pipeline

Light field contains both spatial and angular information of the light rays which benefits many tasks in computer vision such as scene flow estimation, lens aberrations corrections and refocusing. Inspired by light field, we design a light field rendering network to facilitate saliency detection with the light field information. Then we propose a light-field-driven saliency detection network to build the relationship between salient objects and scene understanding.

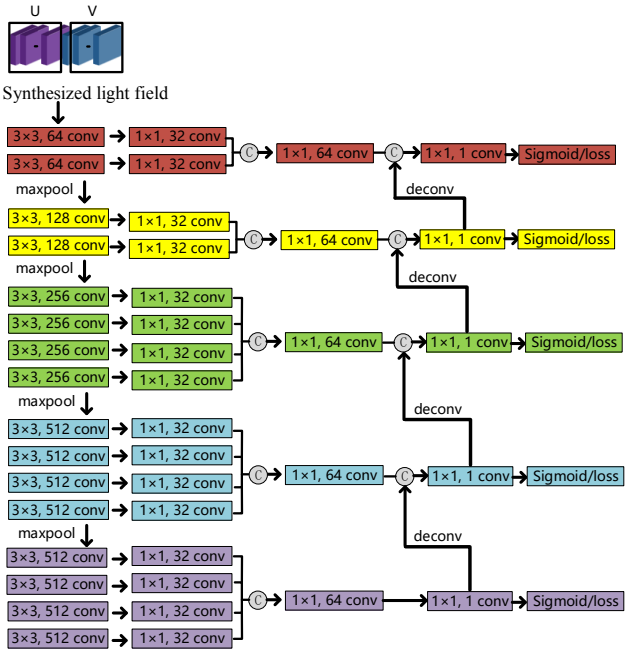

Figure 4: Details of multi-view saliency detection subnetwork. The inputs are the synthesized light field images with size of $14 \times 256 \times 256 \times 3$, and our network outputs multi-level saliency predictions which have 1, 1/2, 1/4, 1/8, 1/16 spatial resolution of the input images, respectively.

As shown in Figure 3, we expand a single view to an array of views firstly. However, the huge amount of high dimensional data makes existing light field processing ineffective and inefficient. Considering the redundancy of light field information, we only synthesize the views along the horizontal and vertical directions through the central view which are good enough to represent the original light field. Recently, [Srinivasan *et al.*, 2017] propose a new light field synthesis method, which is a learning-based method by warping a single image using the corresponding depth information to generate the light field. We train depth CNNs to estimate scene depths and render a Lambertian approximation of light field based on a physically-based warping layer. Inspired by [Srinivasan *et al.*, 2017], we adopt its design and develop our light field rendering network based on it.

After generating the light field images, we design a light-field-driven saliency detection network to estimate saliency maps. The light-field-driven saliency detection network consists of two parts: multi-view saliency detection subnetwork and multi-view attention module. It is natural to think that the richer convolutional features can be generated in a deeper network. And rich convolutional features are highly effective for saliency detection. However, the network is difficult to converge when going deeper because of vanishing/exploding gradients. To solve this problem, our first subnetwork adopts a novel rich feature extraction technique to facilitate saliency detection in each view. In the second module, to integrate multi-view saliency maps, we first warp each saliency map into the central view using the corresponding depth map. This warping operation is equivalent to the inverse procedure in light field synthesis. Then we integrate the warped multi-view saliency maps using the proposed attention mechanism. The view-wise attention vector is generated in our multi-view attention module (MVAM). We generate the final prediction by integrating the attentive multi-view saliency maps.

4.2 Light Field Synthesis Network

Our light field synthesis network is illustrated in Figure 3. First, we apply depth CNNs, represented by $d_u(\bullet)$ and $d_v(\bullet)$, to estimate depth maps $D_u(x,y,u)$and $D_v(x,y,v)$ along horizontal and vertical directions, respectively:

$$\begin{aligned} D_u(x,y,u) &= d_u(I(x,y);\theta_u) \\ D_v(x,y,v) &= d_v(I(x,y);\theta_v) \end{aligned} \tag{1}$$

where (x,y) is the spatial coordinates and (u,v) are angular coordinates. $I(x,y)$ is the central view of light field. The two depth CNNs have the same structure but different parameters θ_u and θ_v. We develop our depth CNNs based on the view synthesis network [Srinivasan *et al.*, 2017], in which there are ten convolutional layers including four dilated convolutional layers. The dilated convolution is used to obtain a large receptive field. The output channels of the last two convolutional layers are modified to be the number of the views in horizontal and vertical directions. The detailed network architecture can be found in [Srinivasan *et al.*, 2017].

Then the central image and the predicted depth maps are fed in the warping layer to render other viewpoints of light field. The physically-based warping layer is the core part of the light field synthesis network. The process can be expressed as follows:

$$\begin{aligned} L_u(x,y,u) &= I(x+uD_u(x,y,u),y) \\ L_v(x,y,v) &= I(x,y+vD_v(x,y,v)) \end{aligned} \tag{2}$$

where $L_u(x,y,u)$ and $L_v(x,y,v)$ are the predicted multi-view images. After rendering the new viewpoints, we calculate the reconstruction error between the predicted images and the ground truth. Here, we use a simple L_1 loss function to supervise the reconstruction quality:

$$\ell_{re} = \left\| L_u(x,y,u) - \hat{L}_u(x,y,u) \right\|_1 + \left\| L_v(x,y,v) - \hat{L}_v(x,y,v) \right\|_1 \tag{3}$$

where $\hat{L}_u(x,y,u)$ and $\hat{L}_v(x,y,v)$ are the ground truth.

To further improve the quality of depth maps, the consistency regularization loss and total variation regularization loss proposed by [Srinivasan *et al.*, 2017] are applied in our light field rendering network. We update the parameters of the depth CNNs by minimizing the final loss:

$$\min_{\theta_u,\theta_v} \sum_T (\ell_{re} + \lambda_c \ell_c + \lambda_{tv} \ell_{tv}) \tag{4}$$

where T is the training set. λ_c and λ_{tv} are the weight of consistency regularization loss and the weight of total variation regularization loss, respectively.

4.3 Light-field-driven Saliency Detection Network

Our proposed light-field-driven saliency detection network can be divided into two parts: multi-view saliency detection subnetwork and multi-view attention module.

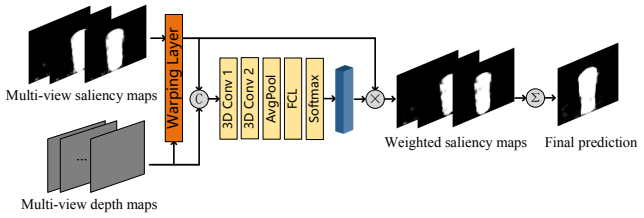

Figure 5: Details of multi-view attention module.

Multi-View Saliency Detection Subnetwork

Our multi-view saliency detection subnetwork is based on the VGG-19 model. The detailed architecture of our network is shown in Figure 4. Previous feature extraction techniques in saliency detection produce feature maps only from the last convolutional layer in each block. To exploit rich saliency cues, we make a combination of the features in the same block. We concatenate all the features in the same convolutional block, and use a convolutional layer with 1×1 kernel size to weight the importance of each feature map. The generated hybrid features contain rich saliency cues at each level.

Then we use the extracted rich convolutional features to predict the saliency map hierarchically. The recursive mechanism is applied to refine the saliency predictions from top convolutional layer to shallower layers. We adopt the deep supervision [Lee *et al.*, 2015] scheme to facilitate saliency map learning. In this way, the pixel-wise supervised information can guide the recursive saliency prediction at each level.

Multi-View Attention Module (MVAM)

The purpose of our multi-view attention module is to integrate multi-view saliency maps in a view-wise attention fashion. The attention mechanism can learn the importance of saliency maps from different views by considering the geometric information (multi-view depth maps). The detailed structure is shown in Figure 5.

The salient objects shift slightly in different views. For effective integration, we first warp the multi-view saliency maps into the central view using the corresponding depth maps. Then we concatenate the warped multi-view saliency maps and depth maps in the color channel and generate a 4D vector (height $\times$ width $\times$ angular channels $\times$ color channels). We connect two 3D convolutional layers in which each filter has access to every 2D view. The kernel size of the 3D convolutional layer is $3\times3\times3$. Next, we use an average pooling layer and a fully connected layer to predict the weight w_n of the saliency map S_n from the n-th view. A softmax operation is applied to normalize w_n spatially:

$$\tilde{w}_n = \exp(w_n) / \sum_{n=1}^{N} \exp(w_n) \tag{5}$$

where N is the number of view images. Finally, the integrated saliency map $\tilde{S}$ is calculated by:

$$\tilde{S} = \sum_{n=1}^{N} \tilde{w}_n S_n \tag{6}$$

Model	Input	S-measure	F-measure	E-measure	MAE
DSR	2D	0.640	0.574	0.767	0.173
BSCA	2D	0.673	0.605	0.777	0.198
MST	2D	0.637	0.548	0.738	0.179
ACSD	3D	0.675	0.637	0.792	0.188
RGBD	3D	0.535	0.567	0.732	0.179
LFS	4D	0.538	0.423	0.717	0.242
UCF	2D	0.788	0.709	0.814	0.136
UCF+	2D	0.792	0.715	0.821	0.131
Amulet	2D	<u>0.801</u>	0.734	0.839	0.104
DSS	2D	0.740	0.709	0.228	0.112
DSS+	2D	0.731	0.674	0.795	0.141
R^3Net	2D	0.793	**0.749**	<u>0.851</u>	<u>0.089</u>
PDNet	3D	0.761	0.692	0.827	0.126
AFNet	3D	0.731	0.687	0.822	0.109
Ours	2D	**0.806**	**0.749**	**0.861**	**0.088**

Table 1: Quantitative comparison of S-measure, F-measure, E-measure, and MAE scores. The retrained models are denoted as "XX+". (bold: best; underline: second best).

5 Experiments

5.1 Experimental Setup

Implementation Details

We implement our method based on the Tensorflow toolbox with one NVIDIA 1080Ti GPU. We train the whole network end-to-end using the Adam optimization algorithm with default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 1\mathrm{e}-08$, respectively. For light field synthesis, the learning rate is 0.0001, weight of consistency regularization loss λ_c and total variation regularization loss λ_{tv} are 0.01 and 0.001 respectively. For saliency detection, the learning rate is 0.001. The minibatch size is set to 1 and the maximum iteration is set to 200000.

Datasets

For the training, the general data augmentation schemes are employed on our light field dataset to improve the varieties, including the flipping, cropping and rotating operations. In this way, we produce 12100 training images totally including the original image. We conducted performance evaluations on the proposed dataset.

Evaluation Metrics

We adopt four of main metrics to evaluate the performance of our method, including the S-measure, F-measure, E-measure and mean absolute error (MAE) scores. We apply the implementations of [Hou *et al.*, 2017] to compute the F-measure and MAE and the definitions can be found in their paper. The S-measure evaluates structure similarities of saliency maps and the E-measure evaluates the pixel-level matching and image-level statistics. The definitions of S-measure and E-measure can be found in [Fan *et al.*, 2017] and [Fan *et al.*, 2018], respectively.

5.2 Comparisons with the 2D Methods

We compare our method with 7 state-of-the-art 2D saliency detection methods, including the deep-learning-based meth-

Figure 6: Visual comparsion of saliency maps on the proposed dataset.

Model	S-measure	F-measure	E-measure	MAE
RSF$^-$	0.801	0.731	0.849	0.097
MVAM$^-$	0.803	0.730	0.845	0.094
Ours	**0.806**	**0.749**	**0.861**	**0.088**

Table 2: Ablation analysis on the proposed dataset.

ods (R^{3}Net [Deng *et al.*, 2018], Amulet [Zhang *et al.*, 2017b], DSS [Hou *et al.*, 2017], UCF [Zhang *et al.*, 2017c]) and the traditional methods (MST [Tu *et al.*, 2016], BSCA [Qin *et al.*, 2015], DSR [Li *et al.*, 2013]). For a fair comparison, we use the recommended parameter settings provided by the authors. Table 1 shows the quantitative results in terms of S-measure, F-measure, E-measure and MAE. Our proposed method achieves the best results compared with other methods. Besides, we retrain two representative networks using the released codes on our proposed dataset.

To further verify the effectiveness of the proposed method, we provide a visual comparison of our method and the state-of-art methods in Figure 6. It can be seen that the proposed method can obtain more complete and accurate salient objects, when salient objects are transparent as shown in the 2nd and 3rd rows, when foreground and background are similar or background is cluttered as shown in rows 4-8. Further, our method can block background effectively in different complex scenes.

5.3 Comparisons with the 3D and 4D Methods

In this section, we compare the proposed method with 5 state-of-the-art 3D and 4D methods, including four 3D methods (AFNet [Wang and Gong, 2019], PDNet [Zhu *et al.*, 2018], RGBD [Qu *et al.*, 2017], ACSD [Ju *et al.*, 2015]) and one 4D method (LFS [Li *et al.*, 2014]). For a fair comparison, we provide all the needed inputs for those methods. The results are shown in Table 1. As we can see, our method outperforms the 3D and 4D methods. For qualitative evaluation, the visual results are shown in Figure 6. In the complex scenarios, our method based on a single input can achieve better performance than 3D and 4D methods.

5.4 Ablation Study

The Effectiveness of Rich Saliency Features Extraction

To verify the importance of our feature extraction technique, we evaluate the proposed network without the rich saliency feature extraction technique, named RSF$^-$. For a fair comparison, we add one additional convolutional layer in which we change the output channels to keep the number of parameters approximately unchanged. We perform a detailed com-

parison of their performance using S-measure, F-measure, E-measure and MAE in Table 2. It can be observed that the proposed feature extraction technique can effectively facilitate saliency detection. This is mainly because the extracted rich convolutional features contain more salient cues.

The Effectiveness of Multi-View Attention Module
To verify the contribution of our proposed MVAM, we simply average the warped multi-view saliency maps to generate the final prediction, named MAVM$^-$. We show the quantitative comparison of our method with MAVM$^-$ in Table 2. We can see that the good results benefit from the attentive integration of multi-view saliency maps in MVAM. Our MVAM can learn the weight distribution of different views, which leads to different views corresponding to different degrees of contribution.

6 Conclusion

In this paper, we propose a novel end-to-end framework to detect saliency object in challenging scenes. We show for the first time that the saliency detection is decomposed into two sub-tasks: light field synthesis and light-field-driven saliency detection. The light field synthesis network generates high-quality 4D light fields from a single view. The light-field-driven network extracts rich saliency representations and builds the relationship between salient objects and scene understanding. Meanwhile, we collected the largest light-field saliency detection dataset, containing 1580 light fields that cover a wide variety of challenging scenes. Extensive quantitative and qualitative evaluations demonstrate that the proposed method outperforms the state-of-the-art 2D, 3D and 4D methods on the proposed dataset and is capable of capturing salient objects in challenging scenes.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61605022 and U1708263) and the Fundamental Research Funds for the Central Universities (DUT19JC58).

References

[Buehler *et al.*, 2001] Chris Buehler, Michael Bosse, Leonard McMillan, Steven J. Gortler, and Michael F. Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001.

[Deng *et al.*, 2018] Zijun Deng, Xiaowei Hu, Lei Zhu, Xuemiao Xu, Jing Qin, Guoqiang Han, and Pheng-Ann Heng. R^{3}net: Recurrent residual refinement network for saliency detection. In *IJCAI*, pages 684–690, 2018.

[Fan *et al.*, 2017] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4558–4567, 2017.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018.

[Gortler *et al.*, 1996] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996.

[Guo *et al.*, 2017] Xinqing Guo, Zhang Chen, Siyuan Li, Yang Yang, and Jingyi Yu. Deep depth inference using binocular and monocular cues. *arXiv: Computer Vision and Pattern Recognition*, 2017.

[Han *et al.*, 2017] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion. *IEEE transactions on cybernetics*, (99):1–13, 2017.

[Hou *et al.*, 2017] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H.S. Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 5300–5309, 2017.

[Ju *et al.*, 2015] Ran Ju, Yang Liu, Tongwei Ren, Ling Ge, and Gangshan Wu. Depth-aware salient object detection using anisotropic center-surround difference. *Signal Processing-image Communication*, 38:115–126, 2015.

[Lee *et al.*, 2015] Chen-Yu Lee, Saining Xie, Patrick W. Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. *international conference on artificial intelligence and statistics*, pages 562–570, 2015.

[Levoy and Hanrahan, 1996] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996.

[Li and Yu, 2015] Guanbin Li and Yizhou Yu. Visual saliency based on multiscale deep features. In *CVPR*, pages 5455–5463, 2015.

[Li and Yu, 2016] Guanbin Li and Yizhou Yu. Deep contrast learning for salient object detection. In *CVPR*, pages 478–487, 2016.

[Li *et al.*, 2013] Xiaohui Li, Huchuan Lu, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. Saliency detection via dense and sparse reconstruction. In *ICCV*, pages 2976–2983, 2013.

[Li *et al.*, 2014] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014.

[Li *et al.*, 2015] Nianyi Li, Bilin Sun, and Jingyi Yu. A weighted sparse coding framework for saliency detection. In *CVPR*, pages 5216–5223, 2015.

[Liu and Han, 2016] Nian Liu and Junwei Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *CVPR*, pages 678–686, 2016.

[Qin *et al.*, 2015] Yao Qin, Huchuan Lu, Yiqun Xu, and He Wang. Saliency detection via cellular automata. In *CVPR*, pages 110–119, 2015.

[Qu *et al.*, 2017] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. Rgbd salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017.

[Song and Lee, 2018] Gwangmo Song and Kyoung Mu Lee. Depth estimation network for dual defocused images with different depth-of-field. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1563–1567, 2018.

[Srinivasan *et al.*, 2017] Pratul P. Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *ICCV*, pages 2262–2270, 2017.

[Tu *et al.*, 2016] Wei-Chih Tu, Shengfeng He, Qingxiong Yang, and Shao-Yi Chien. Real-time salient object detection with a minimum spanning tree. In *CVPR*, pages 2334–2342, 2016.

[Wang and Gong, 2019] Ningning Wang and Xiaojin Gong. Adaptive fusion for rgb-d salient object detection. *arXiv preprint arXiv:1901.01369*, 2019.

[Wang *et al.*, 2016a] Linzhao Wang, Lijun Wang, Huchuan Lu, Pingping Zhang, and Xiang Ruan. Saliency detection with recurrent fully convolutional networks. In *ECCV*, pages 825–841, 2016.

[Wang *et al.*, 2016b] Ting-Chun Wang, Jun-Yan Zhu, Ebi Hiroaki, Manmohan Chandraker, Alexei A. Efros, and Ravi Ramamoorthi. A 4d light-field dataset and cnn architectures for material recognition. *european conference on computer vision*, pages 121–138, 2016.

[Yeung *et al.*, 2019] Henry Wing Fung Yeung, Junhui Hou, Xiaoming Chen, Jie Chen, Zhibo Chen, and Yuk Ying Chung. Light field spatial super-resolution using deep efficient spatial-angular separable convolution. *IEEE Transactions on Image Processing*, 28(5):2319–2330, 2019.

[Zhang *et al.*, 2015] Jun Zhang, Meng Wang, Jun Gao, Yi Wang, Xudong Zhang, and Xindong Wu. Saliency detection with a deeper investigation of light field. In *IJCAI*, pages 2212–2218, 2015.

[Zhang *et al.*, 2017a] Jun Zhang, Meng Wang, Liang Lin, Xun Yang, Jun Gao, and Yong Rui. Saliency detection on light field: A multi-cue approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 13(3):32, 2017.

[Zhang *et al.*, 2017b] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, pages 202–211, 2017.

[Zhang *et al.*, 2017c] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Baocai Yin. Learning uncertain convolutional features for accurate saliency detection. In *ICCV*, pages 212–221, 2017.

[Zhou *et al.*, 2018] Wenhui Zhou, Linkai Liang, Hua Zhang, Andrew Lumsdaine, and Lili Lin. Scale and orientation aware epi-patch learning for light field depth estimation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2362–2367, 2018.

[Zhu *et al.*, 2018] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. Pdnet: Prior-model guided depth-enhanced network for salient object detection. *arXiv preprint arXiv:1803.08636*, 2018.

[Zhu *et al.*, 2019] Hao Zhu, Mantang Guo, Hongdong Li, Qing Wang, and Antonio Robles-Kelly. Breaking the spatio-angular trade-off for light field super-resolution via lstm modelling on epipolar plane images. *arXiv preprint arXiv:1902.05672*, 2019.

View publication stats