2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW) | 978-1-6654-4989-2/20/$31.00 ©2021 IEEE | DOI: 10.1109/ICMEW53276.2021.9456017

MULTI-GENERATOR ADVERSARIAL NETWORKS FOR LIGHT FIELD SALIENCY DETECTION

Hongyan Cai[*], *Xudong Zhang*[†], *Rui Sun*[†], *Ronald poppe*[†], *Jun Zhang*[†]

[*]hongyann.cai@gmail.com; [†]xudong@hfut.edu.cn; [†]sunrui@hfut.edu.cn;
[†]r.w.poppe@uu.nl; [†]zhangjun@hfut.edu.cn

ABSTRACT

Compared to traditional cameras, light field imaging can additionally record the direction of incoming light. Based on this, light fields can be represented in focus stacks that focus at different depths and all-in-focus images. The refocusing information of focal stacks can provide supplementary information for saliency detection. In this paper, we propose a novel multi-generator adversarial network for saliency detection that consists of a cascaded multi-generator and a discriminator. The multi-generator extract saliency features from all-in-focus images and focal stacks. Besides, we set the predicted map multiplied by the all-in-focus image as the input of the discriminator. With this multiplication, we can preserve color and texture information of the salient objects, and reduce the computational cost. We conduct experiments on three public datasets. Compared with existing methods, our method achieves competitive results.

***Index Terms*—** light field, saliency detection, multi-generator adversarial network, focal stack

1. INTRODUCTION

Saliency detection aims at separating the most relevant objects from backgrounds. The topic has raised much attention in the computer vision community [1, 2] and is widely applied in visual tracking [3] and semantic segmentation [4].

With the success of deep learning techniques, 2D/3D saliency detection has made great progress based on RGB [5, 6] or RGB-D images [1, 2]. Compared to 2D/3D cameras, 4D light field imaging [7] not only captures the 2D color and texture information but also records the 2D direction information of the incoming light. Based on these characteristics, the 4D light field can be represented in several 2D images, such as all-in-focus images [8], multi-view sub-aperture images, and focal stacks [9]. A focal stack is a collection of 2D images focused at different depth levels. Some focal slices focus on salient regions while others pay attention to non-salient regions, which suggests that refocusing information may be effective to separate salient objects from backgrounds. How to exploit these images needs to be considered.

A few recent works [10, 11, 12] are dedicated to improving 2D hand-crafted features of the light field. These works focus on exploiting low-level cues, which is difficult to detect salient objects in complex scenes. There have been few attempts [13, 14, 15, 16] address saliency detection benefiting from the hierarchical feature representation and the weight sharing mechanism of deep convolutional networks (DCNs). However,learning from small-scale light field saliency datasets [12, 13, 17] will lead to high-order inconsistency between ground truths and predicted maps. It is therefore important to develop algorithms that can deal with saliency detection on small-scale datasets.

Recently, Generative Adversarial Networks (GANs) [18] provide a unique and promising method to obtain more training data. In GANs, training is driven by two competing parts: the generator for synthesizing fake images and the discriminator for distinguishing real and fake images. Recent works [19, 20] show that saliency detection via GANs can also achieve state-of-the-art results. However, these methods only focus on 2D or 3D images. Until now, GANs have not been used to detect saliency from light fields.

In this paper, we propose a novel light field saliency detection model based on GANs. As shown in Fig. 1, our method consists of a cascaded multi-generator for integrating refocus information and a discriminator. We propose to combine two content losses with an adversarial loss as our objective function, and we show this is beneficial to learning the structure of salient objects. Besides, the all-in-focus image is important to the differentiation of the discriminator [6]. Inspired by cGAN [21], we multiply the saliency map by the all-in-focus image as the input of the discriminator so that the saliency region in the all-in-focus image is retained, and the pixel values of other areas are set to 0. In this way, the texture of salient objects can be well preserved and the computational cost can be largely reduced. In terms of metrics, our method is also superior to state-of-the-art methods on the newly proposed Lytro-Illum dataset [13] and competitive on the other two Lytro datasets [12, 17].

2. RELATED WORKS

We discuss related work on light field saliency detection from three aspects, including multi-cue-based methods, DCNs-based methods, and GANs-based methods.

978-1-6654-4989-2/21/$31.00 ©2021 IEEE

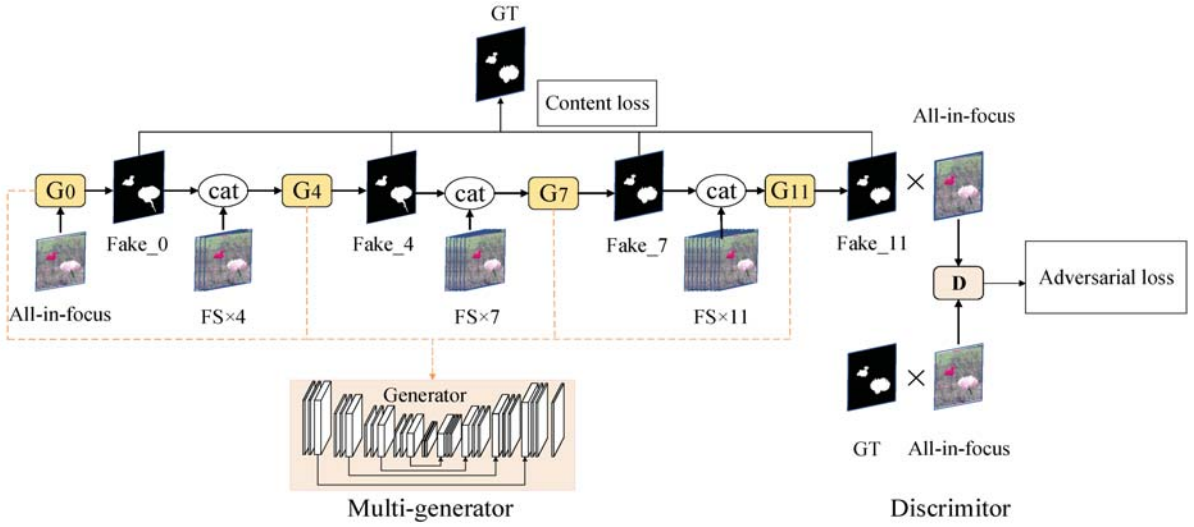

Fig. 1. Overview of our model: $\{G_0, G_4, G_7, G_{11}\}$ is the multi-generator, and every generator follows the structure of $Generator$. The outputs {Fake_0, Fake_4, Fake_7} of $\{G_0, G_4, G_7\}$ are concatenated with FS × 4, FS × 7 and FS × 11 respectively. And then the predicted saliency (Fake_11) and ground truth (GT) are multiplied by the all-in-focus image as the input of the discriminator.

Multi-cue-based methods. In recent years, multi-cue-based methods have been proposed for light fields saliency detection [10, 11, 12]. For example, Li *et al.* [10] exploited focusness, depth and objectness cues of the light field to detect salient objects. Zhang *et al.* [11] proposed to use focusness and depth information to extract the saliency cue. Zhang *et al.* [12] explored the role of light field flow information and location priors in the light field saliency detection. These methods are dedicated to extracting low-level information.

DCNs-based methods. Compared to traditional saliency detection methods [11, 12, 17], DCNs-based methods [13, 14, 15, 16] learn high-level semantic features automatically from light fields. According to the images used, the DCNs-based methods can be categorized into two categories: multi-view-based methods [13, 16] and focal-stacks-based methods [14, 15]. For example, Piao *et al.* [16] generated multi-view images from a single view and proposed a multi-view attention model to predict saliency. Zhang *et al.* [13] designed a network to model angular changes (MAC) from multi-view images. The advantage of the multi-view-based approach is that it combines perspective information from different angles. However, the performance of these method leaves room for improvement in complex scenarios, such as scenarios with high color contrast. The focal-stack-based approaches explore another research direction. Wang *et al.* [15] proposed a two-stream network, one stream for extracting features from the focal stacks, and another for learning saliency from all-in-focus images. Zhang *et al.* [14] proposed an attention mechanism to fuse features of the focal slices. This method utilizes high-level semantic features to guide the choice of low-level features.

GANs-based methods. Owing to the strong learning ability of GANs, many researchers attempt to predict saliency with GANs [6, 22]. Pan *et al.* [6] first proposed a saliency detection model with adversarial training, named SalGAN. Their network utilizes adversarial loss and a content loss for back-propagation. Zhang *et al.* [22] presented gaze prediction on egocentric videos using GANs. The generator network consists of two streams: one stream for extracting features of the foreground and another for the background. Inspired by these works, we propose a new GANs-based saliency detection model for light fields. Differing from [6, 22], our network is designed as a cascaded multi-generator with an encoder-decoder structure.

3. METHOD

In this paper, we design a novel light field saliency detection model based on GANs, which learns a mapping from focal stacks to the saliency map. The proposed network consists of two parts, as shown in Fig. 1.

3.1. Basic model

Our basic model is built on the structure of GANs [18], but some changes have been made. Inspired by U-Net [23], we design the generator as an encoder-decoder structure, as shown in Fig. 1. In the decoder structure, transposed convolution is used to increase the resolution of the feature maps, which has been shown to improve the stability of GANs [18]. In DCNs, the input of each layer is susceptible to parameters of the previous convolutional layer. This dependency between front and back layers requires the network to use a small learning rate and suitable initialization parameters, which slows down the speed of network convergence. Our experiments show that adding batch normalization (BN) [24] after each convolution helps to improve the convergence speed of the network. Besides, differing from the U-Net [23], we utilize the Leaky ReLU to increase the nonlinearity for our network. Finally, we apply the sigmoid activation function in the last layer of the generator to output predictions.

The discriminator adopts the model of PatchGAN [25], which penalizes structure at the scale of patches. The input of the discriminator is the result of multiplying the all-in-focus image and the saliency map. Through a series convolution operation, the discriminator tries to classify if each patch of an input image is real or fake.

3.2. Multi-generator adversarial networks

As mentioned before, different focal slices focus at different depths. Our experiments also show that the more focal slices input, the richer information of saliency the network can learn. How to use these information needs us to consider.

In this paper, we propose a multi-generator network that contains four cascaded generators shown in Fig. 1, indicated as $\{G_0, G_4, G_7, G_{11}\}$. These generators follow the structure described in Sect. 3.1. The subscript indicates the quantity of input focal slices of each generator. G_0 is responsible for extracting features from the all-in-focus image. $\{G_4, G_7, G_{11}\}$ are designed for fusing features from different number of focal slices and the output of the previous generator. Specifically, the outputs named {Fake_0, Fake_4, Fake_7} of $\{G_0, G_4, G_7\}$ are concatenated with FS × 4, FS × 7, FS × 11 respectively. To avoid information redundancy and make full use of refocusing information, we randomly selected the focal slices based on a certain proportion a. And then the predicted saliency (Fake_11) and ground truth (GT) are multiplied by the all-in focus image as the input of the discriminator.

3.3. Objective

We train our multi-generator model over an adversarial loss and two content losses. Content loss is often used to calculate the quality of the reconstructed image obtained by the generator. Our experiments show that the combination of the adversarial loss and the content loss can obtain better results than using adversarial loss alone.

The objective function is described in Eq. 1,

$$\sum_{i=\{0,4,7,11\}} [\alpha L_{L_1}(G_i) + \beta L_{BCE}(G_i)] + L_{GAN}(G_{11}, D) \quad (1)$$

where $i = \{0, 4, 7, 11\}$, $L_{L_1}(G_i)$ and $L_{BCE}(G_i)$ represent the L1 loss and binary cross entropy (BCE) loss for each generator, α and β are the scale factors for $L_{L_1}(G_i)$ and $L_{BCE}(G_i)$, $L_{GAN}(G_{11}, D)$ represents the adversarial loss.

Specifically, adversarial loss is defined in Eq. 2,

$$L_{GAN}(G, D) = \min_G \max_D E_{x,y}[logD(x, y)] \quad (2)$$

where x and y represent the input of the generator and the ground truth. It aims to minimize the loss of the generator (G) and maximize the discriminator (D) loss so that the predicted result of the generator closer to the ground truth.

4. EXPERIMENT RESULTS

4.1. Datasets

We evaluate our method both qualitatively and quantitatively on three light field datasets: LFSD [17], HFUT-Lytro [12], and Lytro-Illum [13]. The LFSD dataset [17] includes 100 light fields and there is only one salient object in the foreground center. Zhang *et al.* [12] constructed the HFUT-Lytro dataset with 255 different light fields. Some of these include multiple salient objects. The Lytro-Illum dataset [13] provides 640 light fields. Many of them contain multiple salient objects either with a similar background or overlapping with the boundary of salient objects. Each focal stack of Lytro-Illum [13] contains 11 focal slices and the spatial resolution of each slice is 540×375. But the number of focal slices decoded by LFSD [17] and HFUT-Lytro [12] datasets is much less than 11. In this paper, we randomly copy the focal slices of the LFSD [17] and HFUT-Lytro [12] from a small number to 11 to meet the input requirements of our network.

4.2. Implementation and experimental setup

We implement our method based on the Pytorch framework [26] with Python 3.6 and train it on one TITAN X (Pascal) GPU. The iteration number is set to 200 epochs and a batch size of 1. The weights in the network are initialized with a gaussian distribution, whose setting follows zero mean and standard deviation of 0.02. We set the proportion of focus slices selection a as 3/11 or 4/11, so that we can select the focal slices with a roughly equal proportion. As suggested in [18], we apply the iterative training on GANs, which is helpful to keep stability for training. After training, we use the multi-generator to evaluate the network performance on the testing light fields.

We apply geometric transformations (i.e., mirror and rotation) to augment training data of the Lytro-Illum dataset. After data augmentation, the number of training data increases from 640 to 1920. Every image is resized to 256 × 256 to leverage the utilization of global information by U-Net [23].

To avoid overfitting, we train our network on the Lytro-Illum dataset with 5-fold cross-validation as[13]. During training, we randomly divide data into five equal parts, one of which is used as the test, and the rest is used as the training data. We take the average of the five evaluation results as our evaluation result.

4.3. Evaluation Metrics

In this paper, we evaluate our method with four measures: F_β-measure (F-measure), Mean Absolute Error (MAE), weighted F_β^ω-measure (WF-measure) [27] and Structure-measure (S-measure) [28]. F-measure is calculated from precision and

Table 1. Quantitative evaluation results of the influence of focal slices on the Lytro-Illum dataset

Method	F-measure	WF-measure	MAE	S-measure
2D_IN	0.782	0.740	0.072	0.804
FS_6	0.824	0.791	0.055	0.842
FS_11	**0.835**	**0.807**	**0.049**	**0.857**

Table 2. Quantitative evaluation results of the loss combination on the Lytro-Illum dataset

Method	F-measure	WF-measure	MAE	S-measure
CON_LOSS	0.833	0.807	0.051	0.853
AD_LOSS	0.530	0.478	0.194	0.603
Ours	**0.845**	**0.817**	**0.049**	**0.857**

recall, as shown in in Eq. 3 where β^2 is set to 0.3 as suggested in [29].

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{3}$$

WF-measure [27] takes the inter-pixel dependence into account and introduces a weight function into the F-measure. MAE shows the per-pixel difference between a predicted saliency map and the corresponding ground truth. The above three metrics compute errors at the pixel level and usually ignore the structural similarity between the predicted map and the ground truth. For this purpose, S-measure [28] is proposed. It is defined in Eq. 4

$$S = \alpha \times S_o + (1-\alpha) \times S_r \tag{4}$$

where α is set to 0.5, S_o and S_r represent the global similarity for salient objects and similarity of block regions in the saliency maps, respectively.

4.4. Ablation Studies

To verify the design choices, we conduct some experiments on ablated versions of our model. Results are reported on the Lytro-Illum dataset [13]. For a fair comparison, 5-fold cross-validation is applied as [13].

Influence of focal slices. To explore the role of focal slices in our approach, we compare the 2D all-in-focus image (2D_IN) with different focal slices (FS_6, FS_11) using the basic model. Compared to the 2D input, the depth information of focal slices is useful to separate salient objects from complex backgrounds as shown in Table 1 and Fig. 2. Additionally, comparing FS_6 and FS_11, input more slices, the network can learn more saliency information. We further visualize the gradients and features of 2D_IN and FS_11 from the pre-trained model by extracting back-propagation parameters of the first layer [30], as shown in Fig. 3. From the vanilla back-propagation saliency (BP) [31], we find that FS_11 can guide the network to focus more on salient objects. Besides, focal stacks can separate the salient objects from the background more effectively as presented in the visual features, which is due to the contribution of depth information of focal stacks.

Table 3. Quantitative evaluation results of the influence of the discriminator on the Lytro-Illum dataset

Method	F-measure	WF-measure	MAE	S-measure
NO_AF	0.828	0.798	0.06	0.843
Ours	**0.845**	**0.817**	**0.049**	**0.857**

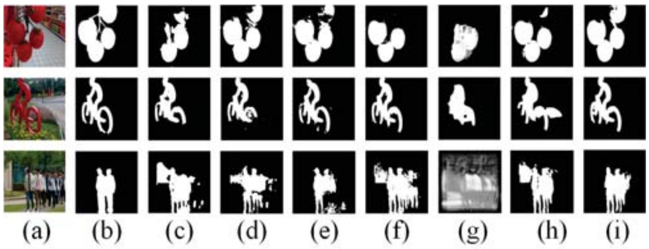

Fig. 2. Qualitative comparisons of predicted saliency maps. (a) All-in-focus, (b) GT, (c) 2D_IN, (d) FS_6, (e) FS_11, (f) CON_LOSS, (g) AD_LOSS, (h) NO_AF, (i) Ours.

Effect of loss function. To understand the role of each loss, we conduct ablation studies about only adversarial loss (AD_LOSS), only content losses (CON_LOSS), and both of them. From Table 2 and Fig. 2, we can see that edges of the salient objects can be better preserved by adding the content loss.

Input of the discriminator. In experiments, we investigate different inputs of the discriminator and their influence on final performance. From Table 3, we observe that all metrics are improved by more than 1% by introducing the all-in-focus image than saliency map input (NO_AF). Results suggest that color and texture information of salient objects have a positive effect on saliency detection.

4.5. Comparison with state-of-the-art

We compare our method to state-of-the-art light field saliency detection methods: LFS [17], WSC [10], DILF [11], Multicue [12], MAC [13], DeepLFSV [16], MOLF [14], DLLFSD [15]. Following WSC [10], we train our network on the Lytro-Illum dataset [13] and evaluate it on LFSD [17] and HFUT datasets [12] with 5-fold cross-validation.

The quantitative results on three light field datasets are shown in Table 4. We can find that our method outperforms most methods on the Lytro-Illum dataset. Comparing with the recent state-of-the-art method MAC [13], our method improves F-Measure by 3.4% and WF-Measure by 6.2% on the Lytro-Illum dataset. However, the performance is not so well on LFSD and HFUT datasets. The main reason is that the number of focal slices is limited in these two datasets. When testing on these two datasets, the multi-generator network may ignore the depth information because of the usage of the same copies.

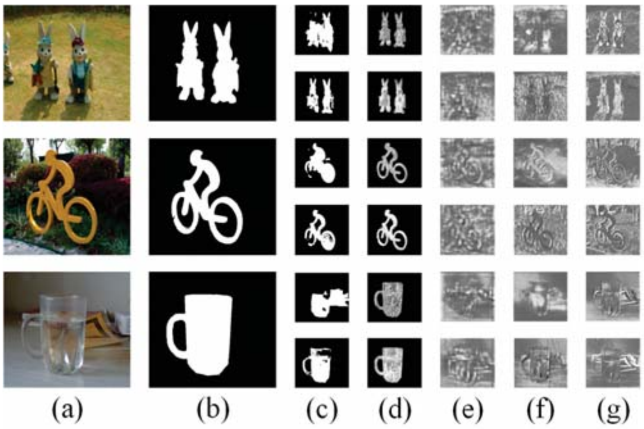

Fig. 3. Visualization of 2D_IN and FS_11. (a) All-in-focus, (b) GT, (c) Prediction, (d) BP, (e) 13th layer, (f) 16th layer, (g) 19th layer. In last five columns, the odd rows represent 2D_IN and even rows represent FS_11. The features of the 13th layer, 16th layer, and 19th layer are extracted from the first convolutional filter.

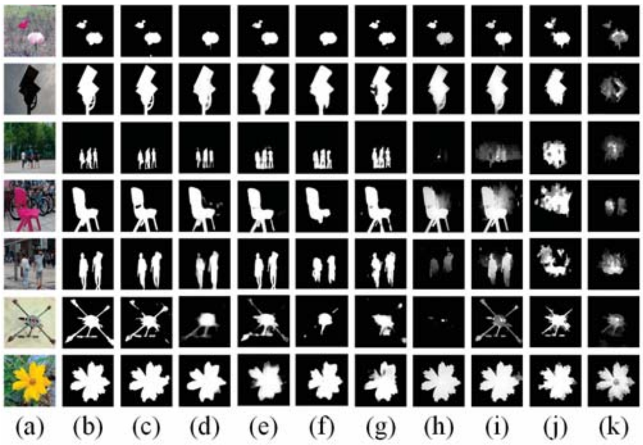

Fig. 4. Comparison with other advanced methods on three datasets. (a) All-in-focus, (b) GT, (c) Ours, (d) MAC, (e) DL-LFSD, (f) MOLF, (g) DeepLFSV, (h) Multi-cue, (i) DILF, (j) WSC, (k) LFS.

Fig. 4 shows qualitative visual comparisons with other approaches. Our predicted results are very close to the ground truths in various challenging scenarios.

4.6. Failure cases

Although our method can achieve excellent results in many challenging scenes, there is still room for improvement in some complex scenes. For example, there are some difficulties in extracting saliency from similar backgrounds, and salient objects with complex shapes, as shown in Fig. 5. The reason for the better performance of Multi-cues [12] and DILF [11] in these cases is that these methods explicitly use the depth prior map. Due to the limited number of focal slices in HFUT and LFSD datasets, our network cannot exploit enough refocus information, as shown in the last two examples of Fig. 5. Additionally, methods based on DCNs focus on extracting global information and ignore the differences between local regions.

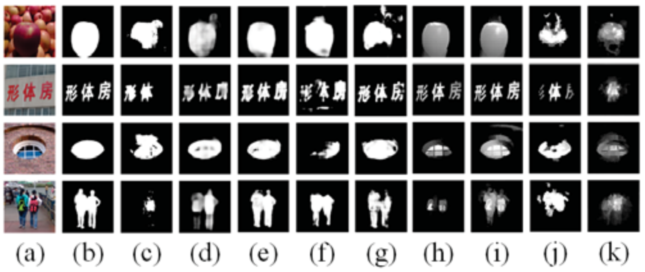

Fig. 5. Some failure cases are shown in this figure. These scenes are selected from three different datasets. (a) All-in-focus, (b) GT, (c) Ours, (d) MAC, (e) DL-LFSD, (f) MOLF, (g) DeepLFSV, (h) Multi-cue, (i) DILF, (j) WSC,(k) LFS.

In the future, it will be very promising to design a DCNs method that can combine background prior knowledge from the depth map or super-pixel algorithms. Super-pixel in image segmentation refers to irregular pixel blocks with similar texture, color, brightness, and other characteristics. Most of these small areas retain effective information for image segmentation and generally do not destroy the boundary information of objects. Also, if a large-scale light field saliency dataset can be established, it will be more helpful to improve the generalization ability of light field saliency model.

5. CONCLUSION

In this paper, we propose a novel multi-generator adversarial network to detect salient objects from focal stacks. Additionally, we adopt a strategy of combining the adversarial loss and two content losses for training. We verify that the joint loss is beneficial to learning the structure and edge information of salient objects. Comprehensive quantitative and qualitative show that our method is competitive to the existing methods on the Lytro-Illum dataset. Specifically, our approach can better preserve salient object boundaries and capture high contrast information. Although the performance of our method is lower on the other two tested datasets, it provides a new feasible method to segment the salient objects with refocusing information.

References

[1] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji, “Rgbd salient object detection: A benchmark and algorithms,” in *ECCV*, 2014.

[2] Karthik Desingh, Madhava K, Deepu Rajan, and CV Jawahar, “Depth really matters: Improving visual salient region detection with depth,” in *BMVC*, 2013.

[3] Rui Zhao, Wanli Ouyang, and Xiaogang Wang, “Unsupervised salience learning for person re-identification,” in *CVPR*, 2013.

[4] Yongquan Hu, Wei Zhou, Shuxin Zhao, Zhibo Chen, and Weiping Li, “Sdm: Semantic distortion measurement for video encryption,” 2018.

Table 4. Quantitative comparison of three light field datasets: LFSD, HFUT, and Lytro-Illum. Red: the best. Blue: second-best. Green, third-best.

Dataset	Metrics	Traditional method				DL-based method				
		LFS	WSC	DILF	Multi-cue	DeepLFSV	MOLF	DL-LFSD	MAC	Ours
LFSD	F-Measure	0.7525	0.7729	0.8173	0.8249	0.7975	0.7707	0.7975	0.8105	0.7489
	WF-Measure	0.5319	0.7371	0.6695	0.7155	0.723	0.7354	0.7284	0.7378	0.7014
	MAE	0.2072	0.1453	0.1363	0.1503	0.1172	0.1347	0.1223	0.1164	0.1390
	S-Measure	0.6698	0.712	0.8037	0.7377	0.7782	0.7086	0.7757	0.7860	0.7071
HFUT	F-Measure	0.4868	0.5552	0.5543	0.6135	0.6286	0.6009	0.6362	0.6721	0.5716
	WF-Measure	0.3023	0.508	0.4468	0.5146	0.566	0.5586	0.5682	0.6087	0.5235
	MAE	0.2215	0.1454	0.1579	0.1388	0.1102	0.1203	0.1100	0.1029	0.1420
	S-Measure	0.56	0.6223	0.6522	0.6404	0.7069	0.6329	0.7223	0.6666	0.6163
Lytro-Illum	F-Measure	0.6107	0.6451	0.6395	0.6648	0.7839	0.7834	0.7860	0.8116	0.8452
	WF-Measure	0.3596	0.5945	0.4844	0.542	0.7403	0.7538	0.7329	0.7540	0.8165
	MAE	0.1697	0.1093	0.1389	0.1197	0.0606	0.0632	0.0553	0.0551	0.0489
	S-Measure	0.6409	0.7181	0.7395	0.7065	0.8364	0.8037	0.8677	0.8664	0.8570

[5] Ting Zhao and Xiangqian Wu, "Pyramid feature attention network for saliency detection," in *CVPR*, 2019.

[6] Junting Pan, Cristian Canton, Kevin McGuinness, Noel O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giró-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *ArXiv*, 2017.

[7] Edward Adelson and John Wang, "Single lens stereo with a plenoptic camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, pp. 99–106, 1992.

[8] Congyan Lang, Tam Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*, 2012.

[9] M. Levoy and P. Hanrahan, "Light field rendering," 1996.

[10] N. Li, Bilin Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *CVPR*, 2015.

[11] Jun Zhang, Meng Wang, Jun Gao, Yi Wang, Xudong Zhang, and Xindong Wu, "Saliency detection with a deeper investigation of light field," 2015.

[12] 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*, vol. 13, 2017.

[13] Jun Zhang, Yamei Liu, Shengping Zhang, Ronald Poppe, and Meng Wang, "Light field saliency detection with deep convolutional networks," *IEEE Transactions on Image Processing*, 2020.

[14] Zhang M, Li J, and Ji W, "Memory-oriented decoder for light field salient object detection," *NIPS*, 2019.

[15] Tiantian Wang, Yongri Piao, Huchuan Lu, Xiao Li, and Lihe Zhang, "Deep learning for light field saliency detection," in *ICCV*, 2019.

[16] Yongri Piao, Zhengkun Rong, Miao Zhang, Xiao Li, and Huchuan Lu, "Deep light-field-driven saliency detection from a single view," in *IJCAI*, 08 2019.

[17] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu, "Saliency detection on light field," *Transactions on Pattern Analysis and Machine Intelligence*, 2016.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Y. Bengio, "Generative adversarial nets," *ArXiv*, 2014.

[19] Prashant Patil, Omkar Thawakar, Akshay Dudhane, and Subrahmanyam Murala, "Motion saliency based generative adversarial network for underwater moving object segmentation," in *ICIP*, 2019.

[20] Yan Bing, Wang Haoqian, Wang Xingzheng, and Zhang Yongbing, "An accurate saliency prediction method based on generative adversarial networks," in *ICIP*, 2017.

[21] Mehdi Mirza and Simon Osindero, "Conditional generative adversarial nets," *ArXiv*, 2014.

[22] Mengmi Zhang, Keng Ma, Joo Lim, Qi Zhao, and Jiashi Feng, "Deep future gaze: Gaze anticipation on egocentric videos using adversarial networks," in *CVPR*, 2017.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *LNCS*, 2015, vol. 9351.

[24] Sergey Ioffe and Christian Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," 2015.

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[26] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, and Soumith Chintala, "Pytorch: An imperative style, high-performance deep learning library," 2019.

[27] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal, "How to evaluate foreground maps," in *CVPR*, 2014.

[28] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017.

[29] R. Achantay, S. Hemamiz, F. Estraday, and S. Süsstrunky, "Frequency-tuned salient region detection," *CVPR*, 2009.

[30] Utku Ozbulak, "Pytorch cnn visualizations," 2019.

[31] Yann Lecun, Bernhard Boser, John Denker, Don Henderson, R. Howard, W.E. Hubbard, and Larry Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, 1989.