scientific reports

Check for updates

OPEN

LRNet: lightweight attention-oriented residual fusion network for light field salient object detection

Shuai Ma, Xusheng Zhu✉, Long Xu, Li Zhou & Daixin Chen

Light field imaging contains abundant scene structure information, which can improve the accuracy of salient object detection in challenging tasks and has received widespread attention. However, how to apply the abundant information of light field imaging to salient object detection still faces enormous challenges. In this paper, the lightweight attention and residual convLSTM network is proposed to address this issue, which is mainly composed of the lightweight attention-based feature enhancement module (LFM) and residual convLSTM-based feature integration module (RFM). The LFM can provide an attention map for each focal slice through the attention mechanism to focus on the features related to the object, thereby enhancing saliency features. The RFM leverages the residual mechanism and convLSTM to fully utilize the spatial structural information of focal slices, thereby achieving high-precision feature fusion. Experimental results on three publicly available light field datasets show that the proposed method surpasses the existing 17 state-of-the-art methods and achieves the highest score among five quantitative indicators.

Keywords Light field, Saliency detection, Lightweight attention, Residual ConvLSTM

Light field[1] is a novel imaging form that can simultaneously record the direction and intensity information of all rays, thereby collecting three-dimensional information of the scene. Meanwhile, some specific optical structures are developed to capture the light field, such as gantry system, programmable aperture light field camera, camera array system, and lenslet-based light field camera. The lenslet-based camera is formed by placing a microlens array in front of the detector of the traditional camera, which is widely applied in particle image velocimetry (PIV)[2], temperature field measurement[3], turbomachinery blades measurement[4]. The excellent application of light field imaging in these fields is mainly attributed to the flourishing development of light field technologies such as digital refocusing[5], super-resolution[6,7], depth estimation[8,9], image rendering[10,11], and saliency detection[12,13]. Among these light field technologies mentioned above, light field saliency detection attempts to detect the most attractive objects in the scene, which can greatly promote the application of light field imaging.

To achieve saliency detection, many methods based on RGB imaging[14–18] and RGB-D imaging[19–21] were proposed, as shown in Fig. 1(a,b). The methods based on RGB imaging mainly utilize the color and texture information of the image as prior knowledge to obtain saliency results. However, Due to the lack of geometric structure information in RGB images, the color and texture information often make it difficult to solve the problem of object detection caused by weak textures, resulting in poor experimental results. The methods based on RGB-D imaging obtain the geometric structure of the scene by introducing depth information, further improving the accuracy of saliency detection in weak texture and occlusion areas. However, unreliable depth maps will not provide valuable geometric spatial information and have a negative impact on the performance of saliency detection.

Light field imaging can capture the three-dimensional geometric information of the scene through a single exposure of the light field acquisition device. The geometric information of the scene can be encoded into the focal slices by refocusing technology. Therefore, these focal slices focused at different depths are used for the saliency detection to compensate for the faultiness of RGB-D imaging, as shown in Fig. 1(c). Li et al.[22] first proposed the saliency detection method based on light field imaging and demonstrated that the refocusing ability of light field imaging can provide various valuable clues. This method integrates position cues, contrast cues, and foreground cues to achieve saliency detection. Zhang et al.[23] introduced the depth cue into the light

ChengDu Aircraft Industrial (Group) Co., Ltd., Qingyang, Chengdu 610092, Sichuan, China. ✉email: msromam@163.com

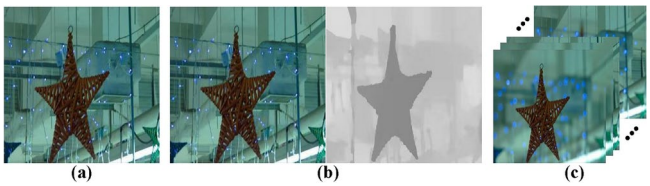

Fig.1. Salient object detection methods involve different types of input data. (**a**) The input data of the methods based on RGB imaging; (**b**) The input data of the methods based on RGB-D imaging; (**c**) The input data of some methods based on light field imaging.

field saliency detection to improve the accuracy. This method computed the saliency contrast by the depth cue and computed the background prior by the selected focal slice, further integrating the above information to obtain the saliency object. However, these methods leveraged hand-crafted features instead of semantic features to detect salient objects, resulting in low-precision salient object detection. Zhang et al.[24] first utilized the deep learning method to explore the salient object in light field focal slices and proposed a memory-oriented spatial method to prevent the disruption of the spatial relationship among focal slices during the fusion process, further achieving competitive saliency detection results. However, This method used channel-wise attention instead of pixel-wise attention to emphasize valuable information among focal slices, which may make it difficult to focus on the features related to salient objects and cut down the accuracy of saliency detection. Zhang et al.[25] introduced the idea of the residual connection to eliminate homogeneous features among focal slices and preserve more pure saliency features. However, during removing homogeneous features, salient features may be also eliminated due to not being treated differently from other useless features, leading to poor performance in salient object detection.

To solve the above problem, this paper proposes a lightweight attention and residual ConvLSTM network (LRNet) to achieve light field salient object detection, which can focus more on feature information related to salient detection through lightweight attention-based feature enhancement module and eliminate features unrelated to salient object in focal slices through residual ConvLSTM-based feature integration module, further improving the performance of salient object detection. In summary, the contributions of this paper are summarized as follows:

- We design a lightweight attention-based feature enhancement module (LFM) to enhance saliency features. The module leverages the lightweight attention mechanism to pay more attention to the features related to the saliency object in each focal slice, thereby assigning more weight information to saliency features.
- We propose a residual ConvLSTM-based feature integration module (RFM) to remove useless features. The module first utilizes the residual connections to eliminate redundant features between weighted focal slices, and further extracts more pure saliency features. The salient features obtained from the focal slice are integrated by ConvLSTM to compute the saliency object results.
- We provide a comparative analysis between the proposed LRNet and 18 state-of-the-art methods on the three benchmark datasets. The analysis results show that the proposed LRNet outperforms these state-of-the-art methods in the five quantitative metrics.

Related work

In this section, the related works of the methods based on RGB imaging, RGB-D imaging and light field imaging will be reviewed, respectively.

2D imaging saliency detection

The input data of the methods based on RGB imaging is a single image, which mainly relies on the contrast, color, and edge information of the image to predict salient objects. Early methods[26,27] mainly adopt hand-craft features and strategies to predict the saliency object in images. The current methods mainly utilize deep learning to obtain more robust semantic features, thereby ensuring detection accuracy. To fully utilize the contextual information in 2D images, Chen et al.[16] proposed the context-aware feature interweaved aggregation module and global context flow module to integrate different level features and obtain the global context information, respectively. Siris et al.[28] presented a context-aware network composed of the context refinement module and contextual instance transformer to achieve saliency detection. Zhu et al.[29] developed a novel module to selectively utilize the local saliency features and global saliency features, and constructed two successive chain modules to fusion the saliency features from different dilated convolutional layers. Wei et al.[30] designed a comprehensive network that combines the fusion, feedback and focus functions to address the issue of low-quality saliency detection caused by differences in the receptive domains of different convolutional layers. Recently, some methods explored the uncertain distribution in salient object detection. Tian et al.[14] constructed a distributional uncertainty modeling by long-tail learning, test-time strategies and single-model uncertainty modeling to detect salient objects.

RGB-D imaging saliency detection

Although significant progress has been made in saliency detection methods based on 2D imaging, the lack of spatial information in 2D images limits the performance of such methods in some challenging scenes.. The input data of the method based on RGB-D imaging includes an image and a depth map which can provide the

geometric information to improve performance. Li et al.[21] designed a hierarchical alternate interaction network to detect salient objects in RGB-D imaging. The method mainly utilized the features from depth information to enhance the features from RGB image in turn by the proposed hierarchical alternate interaction module. To fully integrate feature information from RGB images and depth maps, Fu et al.[31] proposed a joint learning and densely-cooperative fusion network to execute RGB-D saliency detection, which can extract saliency features by joint learning module and discover complementary features by densely-cooperative fusion. Liu et al.[32] proposed a mutual attention framework to address the cross-modal feature fusion problem, which leveraged the non-local attention to propagate long-range contextual features and adopted selective attention to control the effect of depth cues. In recent years, some weakly supervised methods have also been proposed to perform saliency detection. Li et al.[33] presented a scribble-based weakly supervised network for RGB-D saliency detection. The method combined dynamic searching processing and dual-branch consistency learning to perceptively explore saliency features in RGB-D imaging.

Light field saliency detection

Light field imaging can also provide abundant spatial information to improve the performance of network models. The existing methods mainly use micro-lens images, sub-aperture array images, and focal slices as inputs to obtain spatial information. Zhang et al.[34] leveraged light field micro-lens image as the input data and presented an end-to-end deep learning network. The methods designed a model angular changes block to encode the geometric feature information of the micro-lens image as the input of the network, and provided a publicly available light field salient object detection dataset. Zhang et al.[35] leveraged sub-aperture array images as the input data and presented a multi-task network to explore the coherence among spatial, edge, and depth cues, further achieving saliency detection. In recent years, focal slices have gradually become the mainstream input type for network models to explore spatial information in light field imaging. Piao et al.[36] developed a teacher network and a student network to explore the saliency feature in the focal slices and all-focus RGB images, respectively, and transferred the focusness information obtained by the teacher network to the student network. Gao et al.[37] proposed a novel light field object detection benchmark, which included multiple formats of light field input data and annotation to meet the requirements of different light field tasks.

Lightweight attention and residual ConvLSTM network

The overall framework diagram of the proposed network (LRNet) is shown in Fig. 2. The proposed LRNet utilizes the VGG-19[38] as the backbone network to extract salient features of focal slices and the all-focus image. The last three blocks of the VGG-19 network are connected as initial input features, which are fed into LFM and RFM to fully capture and integrate salient features, thereby achieving high-precision salient object detection. The initial input features from focal slices and the all-focus image can be denoted as $\{f_i\}_{i=1}^{13}$, where $f_1, \cdots, f_{12} \in R^{C\times W\times H}$ represents the initial input features of focal slices, respectively. $f_{13} \in R^{C\times W\times H}$ represents the initial input features of the all-focus image.

Each focal slice and the all-focus image can provide abundant object clues, texture clues, and depth cues to increase the accuracy of salient object detection in weak texture and occlusion areas. However, focal slices are obtained by refocusing techniques along the depth direction and contain a large amount of redundant and useless feature information, making it different for the network model to distinguish useful salient features from these features. Therefore, The key task of light field saliency detection is how to accurately distinguish saliency features from massive focal slice features and effectively enhance them. In the next section, Section *A* will provide a detailed introduction that the LFM utilizes the lightweight attention mechanism to enhance the saliency features in each focal slice. Section *B* introduces that RFM will use the residual mechanism and ConvLSTM to integrate the salient features enhanced by LFM, and then calculate the saliency object results. The implementation details, parameters setting and loss function of the proposed network will be introduced in Section *C*.

Lightweight attention-based feature enhancement module

Focal slices and the all-focus image contain a total of thirteen feature maps, which provide a lot of useful feature information for salient object detection. However, there is also a large amount of similar, redundant and useless feature information in the feature maps among the focal slices. These features will hinder the network from selecting feature information related to saliency objects, thereby reducing the accuracy of the saliency detection. The common way to solve the above problem is to distinguish between features related to saliency objects and features unrelated to saliency objects in focal slices and the all-focus image. Attention mechanisms, such as the spatial attention mechanism[39], non-local network[40] and broad attentive graph fusion network[41], are widely regarded as prominent means of focusing more attention on the features related to the salient object. In order to better focus on the saliency-related regions and measure the correlation for each pixel-pair of these regions, these attention mechanism methods need to produce enormous attention feature maps, resulting in high computational complexity and memory capacity. However, focal slices and the all-focus image have a large number of feature maps, making it different for the above methods to provide an attention map for each feature map under limited computing power and memory space. One approach to solving the above challenge is to offer the same attention map for all feature maps. The feature maps from focal slices and the all-focus image have significant differences and a single attention map is different for the network to focus on the saliency-related features in each feature map. Therefore, it is necessary to offer an attention map for each feature map to enhance salient features and remove irrelevant features.

The criss-cross attention mechanism[42] is introduced to provide a lightweight attention map for each feature map to enhance salient features, where sparse connections are only made to each position of the feature map in the vertical and horizontal directions, with lower computational complexity and memory capacity. Hence, a lightweight attention-based feature enhancement module (LFM) is presented, as shown in Fig. 3. There are

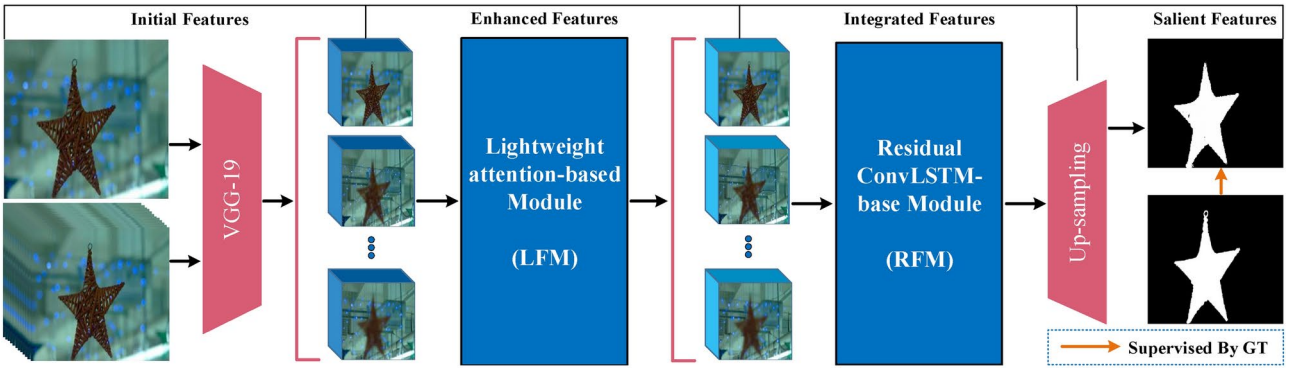

Fig. 2. The overview of the proposed light field saliency detection network.

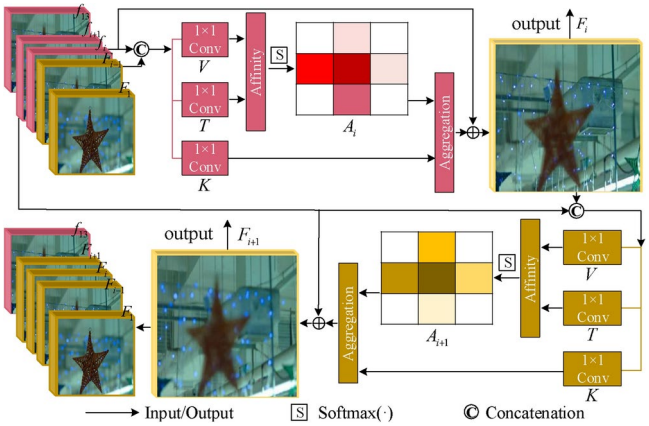

Fig. 3. Schematic illustration of the proposed LFM. The LFM can enhance saliency features in each focal slice, and distinguish saliency features from other massive features in focal slices.

similar saliency object features between adjacent focal feature maps, which can effectively assist the network in focusing on the saliency-related features. Then, each feature map from focal slices is fused with the adjacent feature map to obtain the attention map, further enhancing saliency features.

$$\widetilde{f}_i = w_1 * D(f_i, F_{i-1}) + b_1, \tag{1}$$

where D denotes the concatenation operation, $*$ represents the convolution operation, F_{i-1} denotes the focal feature enhanced by LFM, w_1 and b_1 are the weights and bias of three 3×3 convolutional layers, respectively. Inspired by[42], $\widetilde{f}_i$ is utilized to generate the attention map. Firstly, two 1×1 convolutional layers are applied on $\widetilde{f}_i$ to obtain two feature maps $V \in R^{C' \times H \times W}$ and $T \in R^{C' \times W \times H}$, where C' is less than C. Then, the affinity operation[42] is applied on V and T to obtain an attention map $A_i \in R^{(W+H-1) \times W \times H}$. The enhanced feature can be collected by the attention map A_i as follows.

$$F_i^p = \sum_{j=0}^{H+W-1} A_i^{j,p} \Psi^{j,p} + \widetilde{f}_i^p, \tag{2}$$

where $A_i^{j,p}$ denotes the value at position p and channel j in the attention map A_i, f_i^p represents a feature vector in the feature map f_i at position p, $\Psi^{j,p}$ is obtained by the $K \in R^{C \times W \times H}$. Two 1×1 convolutional layers are applied on $\widetilde{f}_i$ to generate K. Hence, each feature map can generate the enhanced feature map $(F_1, \cdots, F_{13})$ through the corresponding attention map $(A_1, \cdots, A_{13})$.

Residual convLSTM-based feature integration module

In LFM, the criss-cross attention mechanism is utilized to enhance the saliency features in focal slices and the all-focus image, further improving the accuracy the salient object detection. These enhanced features in focal slices and the all-focus image can be a more convenient way for the network to extract the features related to saliency objects. Meanwhile, these enhanced features are also distributed in various focal slices and the all-focus

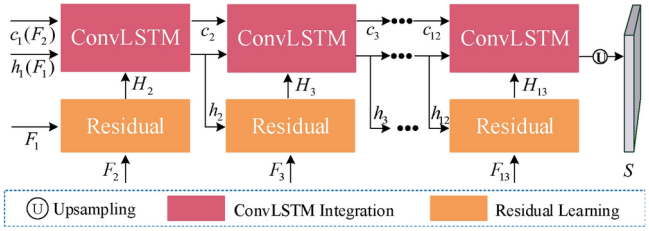

Fig. 4. Schematic illustration of the proposed RFM. The RFM can fully integrate focal slice features, eliminate redundant information.

	Metric	*w/o* LFM	*w/o* RFM	LRNet
HFUT-LFSD	$E_S\uparrow$	0.881	0.874	**0.887**
	$S_\alpha\uparrow$	0.810	0.800	**0.811**
	$\omega F_\beta\uparrow$	**0.735**	0.720	0.732
	$F_\beta\uparrow$	**0.769**	0.744	0.759
	$MAE\downarrow$	0.054	0.059	**0.053**
LFSD	$E_S\uparrow$	0.890	0.891	**0.897**
	$S_\alpha\uparrow$	0.849	0.847	**0.852**
	$\omega F_\beta\uparrow$	0.828	0.820	**0.829**
	$F_\beta\uparrow$	**0.858**	0.846	0.854
	$MAE\downarrow$	0.071	0.073	**0.070**

Table 1. Quantitative comparison of the proposed LRNet in ablation studies. Significant values are in bold.

image, and effective fusion of the features is crucial for obtaining the high-precision saliency object results. In the process of features fusion, although the features related to saliency objects are further enhanced to suppress the negative impact of other features on the salient object detection, those features unrelated to saliency objects still provide a large amount of worthless information to the network, leading to under-performing saliency detection results. In addition, these enhanced features originate from a series of refocused images, and contain plenty of spatial structural information. The spatial structural information can offer depth clues and object clues to improve the saliency object detection performance in weak texture and occlusion regions. However, inappropriate feature fusion can damage the spatial structure of refocusing sequence features, resulting in the inability to offer accurate depth clues.

Hence, a residual convLSTM-based feature integration module (RFM) is presented to fuse these enhanced features, as shown in Fig. 4. Firstly, Residual learning[43] is introduced to extract valuable saliency features, further removing those useless features among focal slices.

$$H_{i+1} = \sigma(w_2 * D(h_i, F_{i+1}) + b_2) + F_{i+1}, \tag{3}$$

where σ is the PReLU function[44], h_i denotes the hidden state of ConvLSTM[45], w_2 and b_2 are the weight and bias of two 1×1 convolutional layers and one 3×3 convolutional layer, respectively. Meanwhile, the refocusing features on the depth axis can be regarded as the moving features on the timeline. ConvLSTM is considered an effective way to fuse these continuous variational features without compromising the spatial relationships among these features. Then, the refined feature H_i is input into ConvLSTM to compute the saliency map.

$$\begin{aligned} v_i &= \delta(W_{xv} * H_i + W_{hv} * h_{i-1} + W_{cv} \circ c_{i-1} + b_v) \\ f_i &= \delta(W_{xf} * H_i + W_{hf} * h_{i-1} + W_{cf} \circ c_{i-1} + b_f) \\ c_i &= B + v_i \circ tanh(W_{xc} * H_i + W_{hc} * h_{i-1} + b_c) \\ o_i &= \delta(W_{xo} * H_i + W_{ho} * h_{i-1} + W_{co} \circ c_i + b_o) \\ h_i &= o_i \circ tanh(c_i), \end{aligned} \tag{4}$$

where $B = f_i \circ c_{i-1}$, $\delta(\cdot)$ is the sigmoid function, $\circ$ is the Hadamard product. v_i, f_i and o_i denote the input gate, forget gate and output gate, respectively. c_i represents the memory cell, which can save past information, h_i is the hidden state, which contains saliency object information and serves as the output of the network, W_* and b_* are the parameters that need to be learned. The final salient object detection result S can be obtained by applying a 1×1 convolutional layer and upsampling operation on the output of network h_{13}. The hidden state h_i is initialized to F_1, i.e., $h_1 = F_1$. The memory cell c_i is initialized to F_2, i.e., $c_1 = F_2$.

Loss function

The proposed network is capable of predicting the saliency map, which is a binary image. Applying supervision signals to the predicted saliency map can promote the rapid convergence of the network, thereby obtaining high-precision salient object detection results. The loss between the predicted saliency map and the ground truth is utilized as the supervision signal. The loss function adopts the cross-entropy, as follows:

$$L = -\sum_{x=1}^{n}[S(x) \cdot logG(x) + (1 - S(x)) \cdot (1 - logG(x))] \tag{5}$$

where $S(x)$ denotes the saliency map, $G(x)$ denotes the ground truth, x is the pixel index of the predicted saliency map and the ground truth, n denotes the number of pixels.

Experiment

Three representative benchmark datasets for light field salient object detection are leveraged to train and test the proposed LRNet, i.e., DUT-LFSD[24], HFUT-LFSD[46], and LFSD[22]. DUT-LFSD contains 1462 light field data samples, of which 1000 samples are utilized for training, and 462 samples are used for testing. This dataset involves a large number of outdoor and indoor scenes, which places high demands on the detection ability of the network. HFUT-LFSD contains 255 light field data samples. This dataset involves plenty of weak texture and small object scenes, making it the most challenging dataset. LFSD contains 100 light field data samples and is also the first benchmark dataset for light field salient object detection, greatly promoting the development of this field.

The proposed LRNet is implemented using the Pytorch deep learning framework and Python language, and trained on a 1080Ti GPU, with a training time of approximately 2 days. The VGG-19 network is utilized to obtain the initial features of focal slices and the all-focus image, thereby serving as the input of LRNet to compute the saliency object. The stochastic gradient descent (SGD) optimizer is adopted to train the proposed LRNet. The weight decay, momentum, and learning rate of the optimizer are 0.0005, 0.9, and 1e-10, respectively.

The proposed LRNet is trained through the training set of DUT-LFSD and 181 light field data samples of HFUT-LFSD, and is tested through the testing set of DUT-LFSD, the remaining 74 light field data samples of HFUT-LFSD, and 100 light field data samples of LFSD. The size of the input image in all three datasets is 256×256. Five commonly used quantitative indicators are leveraged to evaluate the salient object detection results of the proposed LRNet and other methods, which are the enhanced-alignment measure (denoted as E_S)[47], the structure-measure (denoted as S_α)[48], the F-measure (denoted as F_β)[49], the weighted F-measure (denoted as wF_β)[50], and the mean absolute error (denoted as MAE).

Ablation studies

To verify the effectiveness of the two modules of the proposed LRNet, the ablation studies are conducted in this section. Two methods (denoted as *w/o* LFM and *w/o* RFM) are structured by removing the two modules of the proposed LRNet, respectively. The quantitative comparison results of these three methods are shown in Table 1. Compared with the proposed network, the method *w/o* LFM which removes the LRM achieves lower scores on most metrics. This directly demonstrates that removing the LFM can cut down the accuracy of salient object detection, and the criss-cross attention mechanism can focus on the features related to saliency objects, thereby improving the performance of salient object detection. Compared to the proposed LRNet, the method *w/o* RFM which removes the RFM obtains the lower scores on all metrics, as shown in Table 1. This indicates that the direct fusion of focal slice features will lose its spatial structure information and retain redundant feature information, resulting in the network not being able to fully utilize the depth clues and object clues to improve detection accuracy.

In addition, the visual comparison results of these three methods on three datasets are shown in Fig. 5. Among them, the HFUT-LFSD and LFSD datasets contain a large number of challenging scenarios such as small objects and complex backgrounds, as shown in rows 2, 3, and 4 of Fig. 5. The method *w/o* LFM was difficult to detect small objects in the distance of the image, as shown in the second row of Fig. 5. This indicates that removing LFM makes it difficult for the network model to focus on the subtle features in the focal slices, resulting in the failure of small object detection. Meanwhile, the method *w/o* RFM was also difficult to correctly detect objects in the complex background in the fourth row of Fig. 5, indicating that removing LFM makes it difficult for the network model to select the correct object information from the complex background. The method *w/o* RFM detected redundant object features as shown in the first and fourth rows of Fig. 5, and also lost some object information as shown in the third row of Fig. 5, indicating that removing RFM makes it difficult for the network model to correctly fuse focal features. On the contrary, the proposed LRNet can obtain the complete salient object and rarely contain irrelevant object regions, as shown in the third column of Fig. 5.

Quantitative and visual comparison

To further demonstrate the effectiveness of the proposed LRNet, the quantitative and visual comparisons are made between the salient object detection results obtained by the proposed LRNet and the results of 18 state-of-the-art methods. Meanwhile, the input for the other 17 methods includes the light field data, RGB-D data, and RGB data, respectively. The methods for inputting light field data include MEANet[51], ERNet[36], DLFS[52], LFS*[22], and DILF*[23]. The methods for inputting RGB-D data include HAINet[21], JL-DCF[31], D^3Net[53], HDFNet[54], and S2MA[55]. The methods for inputting RGB include F^3Net[30], MINet[56], GCPANet[16], EGNet[57], BASNet[58],

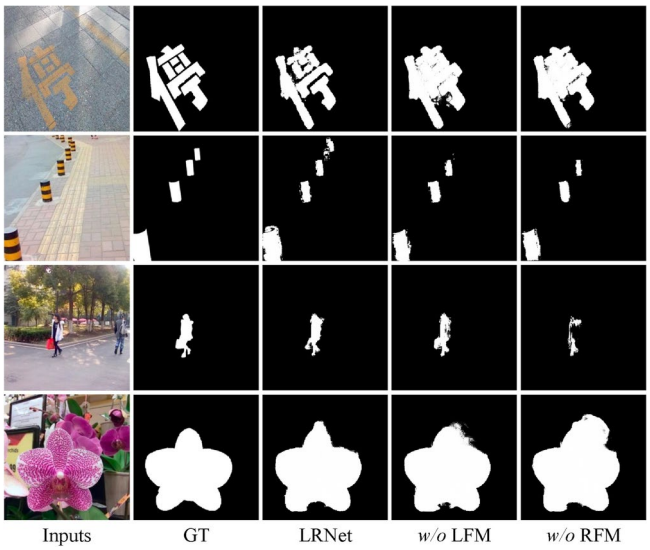

Fig. 5. Visual comparison of the proposed LRNet in ablation studies.

Methods	DUT-LFSD				HFUT-LFSD				LFSD			
	$E_S\uparrow$	$\omega F_\beta\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_S\uparrow$	$\omega F_\beta\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$	$E_S\uparrow$	$\omega F_\beta\uparrow$	$F_\beta\uparrow$	$MAE\downarrow$
LRNet	*0.951*	**0.907**	*0.924*	*0.034*	**0.887**	**0.732**	**0.759**	**0.053**	**0.897**	**0.829**	**0.854**	**0.070**
MEANet	**0.962**	*0.905*	**0.936**	**0.031**	–	–	–	–	*0.893*	0.803	*0.849*	0.077
ERNet	0.943	0.880	0.889	0.040	–	–	–	–	0.889	0.809	0.842	0.080
DLFS	0.891	0.763	0.801	0.076	0.802	0.585	0.617	0.080	0.806	0.657	0.715	0.147
LFS*	0.728	0.288	0.484	0.240	0.672	0.242	0.413	0.199	0.771	0.479	0.740	0.208
DILF*	0.805	0.494	0.641	0.168	0.726	0.411	0.511	0.137	0.810	0.604	0.728	0.168
HAINet	0.932	0.891	0.897	0.042	0.778	0.531	0.599	0.083	0.882	0.814	0.851	*0.072*
JL-DCF	0.926	0.858	0.881	0.047	0.810	0.603	0.646	0.077	0.882	*0.822*	**0.854**	0.078
D^3Net	0.844	0.670	0.757	0.098	0.754	0.569	0.595	0.086	0.862	0.760	0.810	0.095
HDFNet	0.910	0.808	0.871	0.061	0.750	0.598	0.610	0.094	0.891	0.806	0.843	0.076
S2MA	0.920	0.853	0.874	0.048	0.719	0.503	0.538	0.119	0.873	0.772	0.820	0.094
F^3Net	0.912	0.846	0.884	0.054	0.786	0.647	0.688	0.059	0.818	0.754	0.801	0.104
MINet	0.897	0.817	0.861	0.061	0.786	0.620	0.642	0.073	0.821	0.751	0.799	0.094
GCPANet	0.901	0.833	0.868	0.062	0.795	0.644	0.674	0.079	0.834	0.765	0.811	0.097
EGNet	0.914	0.829	0.870	0.053	0.801	0.602	0.651	0.067	0.850	0.775	0.828	0.085
BASNet	0.927	0.872	0.879	0.042	0.823	0.661	0.670	0.076	0.874	0.805	0.834	0.082
AADFNet	0.901	0.856	0.880	0.052	0.786	0.670	0.712	0.088	0.817	0.748	0.782	0.095
PoolNet	0.919	0.832	0.868	0.051	0.831	0.635	0.665	0.063	0.849	0.763	0.826	0.092
PiCANet	0.899	0.766	0.825	0.073	0.748	0.541	0.585	0.092	0.836	0.694	0.764	0.124

Table 2. Quantitative comparison of the proposed LRNet with other methods. The best result is highlighted in bold, and the second best result is highlighted in italic.

AADFNet[29], PoolNet[59], and PiCANet[60]. The experimental results of other 17 methods are collected through literature[36,61]. The methods LFS*[22] and DILF*[23] are hand-designed feature methods rather than deep learning methods.

The quantitative comparison of the proposed LRNet and other methods is shown in Table 2. The proposed LRNet achieves the highest scores in most of four quantitative evaluation indicators on three datasets. This strongly proves the effectiveness of the proposed LFM and RFM, which have strong saliency feature enhancement and fusion capabilities and can cope with salient object detection in different complex scenes. In addition, the experimental results of the methods based on light field imaging and RGB-D imaging are significantly better than those based on RGB imaging, which directly proves that the spatial structural information can provide additional clues for the network to improve the accuracy of the salient object detection results. Meanwhile, the experimental results of deep learning methods based on light field imaging are also superior to those based on RGB-D imaging, as shown in Table 2. The low-quality depth map of RGB-D imaging will directly reduce the

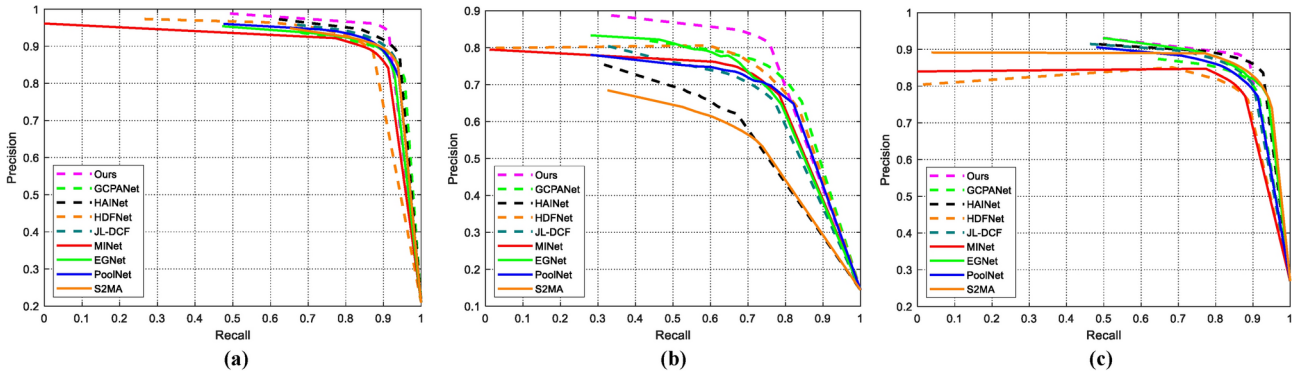

Fig. 6. The precision-recall (PR) curves of the proposed LRNet and other 8 methods on three light field datasets. (**a**) DUT-LFSD. (**b**) HFUT-LFSD. (**c**) LFSD.

accuracy of salient object detection results, while the methods based on light field imaging are not affected by the quality of the depth map.

The precision-recall (PR) curve is also regarded as a metric to further validate the superiority of the proposed LRNet. The comparison results between the proposed LRNet and other methods are shown in Fig. 6, which shows that the proposed LRNet outperforms other methods on the PR curves of three light field datasets.

The visual comparison of the proposed LRNet and other methods is shown in Fig. 7. The proposed LRNet obtains complete and detail-rich saliency object maps in scenes with weak textures and complex backgrounds. Meanwhile, these saliency object maps are highly similar to the ground truth, which once again verifies the competitiveness of the proposed method. On the contrary, other methods achieve sub-optimal results in challenging scenes. In rows 1, 3, 8, and 11 of Fig. 7, other methods detect a large number of areas that do not belong to the saliency object, mainly due to the complex background of these scenes or the high similarity between the object and the background. Furthermore, in rows 4 and 7 of Fig. 7, some methods also find it different to detect the salient object due to the presence of weak texture areas. In rows 1 and 6 of Fig. 7, the performance of methods based on light field imaging and RGB-D imaging significantly surpasses those based on RGB imaging, thereby proving that the spatial structural information can assist the network in selected feature information related to the salient object. In row 9 of Fig. 7, the proposed LRNet can detect the details of the salient object, but other methods (ERNet[36]) based on light field imaging cannot detect the corresponding detail information.

Computational cost and model size

The proposed LRNet is compared with other methods based on light field data in terms of computational cost and model size, as shown in Table 3. The computational cost of all methods was tested on a GTX 1080Ti GPU and the running efficiency was represented using FPS(frame per second). The proposed LRNet outperforms MoLF[24] and ERNet[36], but is inferior to PANet[62]. The LRNet, MoLF[24] and ERNet[36] use the ConvLSTM[45] to fuse focal features, resulting in higher computational cost. Therefore, efficient and high-precision focal feature fusion methods are worth studying to improve the time performance of methods based on light field data.

Limitation of the method

Although the proposed LRNet achieved high scores in all three publicly available datasets, there are still some failed cases, especially in weak texture and small target scenes, as shown in Fig. 8. In the case where the object has weak texture and is similar in color to the background, the proposed LRNet is difficult to correctly select the object from the background, resulting in the detection of redundant areas or the loss of object areas, as shown by the red boxes in columns 1 and 2 of Fig. 8. Although focal slices can provide geometric information to constrain some special scenes, both the target and background have weak textures and the same color, resulting in the same texture in the focus slice at any depth. Similarly, the performance of the proposed LRNet in obtaining object details is also unsatisfactory, as shown by the red boxes in columns 3 and 4 of Fig. 8. The main reason for this is that the input data is resized as 256×256, resulting in the loss of image details.

Conclusions

In this paper, the lightweight attention and residual convLSTM network is presented to perform the light field salient object detection, which mainly includes the lightweight attention-based feature enhancement module (LFM) and residual convLSTM-based feature integration module (RFM). Specifically, the LFM provides a reliable attention map for each focal slice through the criss-cross attention mechanism to focus on the feature information related to the salient object, thereby enhancing salient features. the RFM utilizes the residual mechanism to remove redundant features from focal slices and extract valuable object feature information. ConvLSTM is leveraged to fuse the refined focal slice features, further preventing the destruction of the spatial structural information in focal slices to obtain high-quality salient object detection results. The experimental results show that the proposed LRNet detects the salient object in different scenes and can also handle the difficulty of salient object detection in weak texture and occlusion regions. The proposed LRNet is compared to

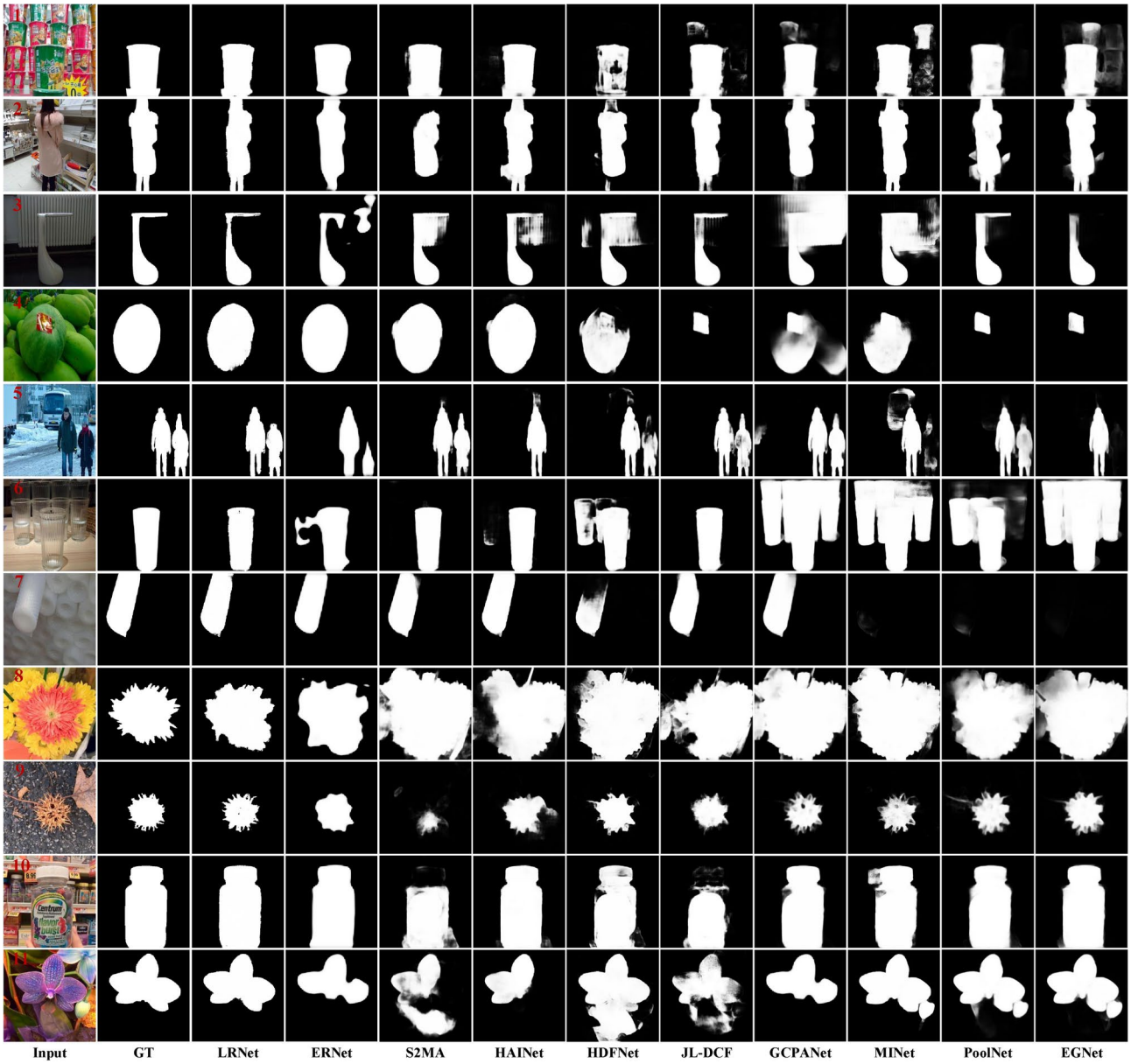

Fig. 7. Visual comparison of the proposed LRNet with other methods.

Method	Resolution	FPS	model size
PANet [62]	256× 256	9	60
ERNet [36]	256× 256	5	88
MoLF [24]	256× 256	4	177
LRNet	256× 256	7	71

Table 3. Computational cost and model size comparison of the proposed LRNet and other methods.

other 17 state-of-the-art methods on 5 quantitative indicators on 3 publicly available datasets, and the optimal scores are obtained, further proving its superiority.

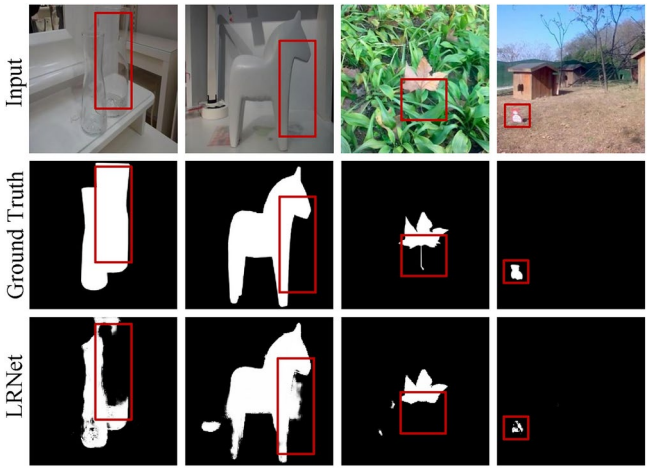

Fig. 8. Failure cases of the proposed LRNet.

Data availibility

All data generated or analysed during this study are included in this published article.

Received: 5 August 2024; Accepted: 17 October 2024
Published online: 29 October 2024

References

1. Levoy, M. & Hanrahan, P. Light field rendering. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 31–42 (1996).
2. Lynch, K., Fahringer, T. & Thurow, B. Three-dimensional particle image velocimetry using a plenoptic camera. In: 50th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition, p. 1056 (2012).
3. Sun, J. et al. Three-dimensional temperature field measurement of flame using a single light field camera. *Opt. Exp.* **24**(2), 1118–1132 (2016).
4. Ding, J., Li, H., Ma, H., Shi, S. & New, T. H. A novel light field imaging based 3d geometry measurement technique for turbomachinery blades. *Meas. Sci. Technol.* **30**(11), 115901 (2019).
5. Ng, R. Fourier slice photography. ACM Trans.Graph **24**, 735–744
6. Zhang, S., Lin, Y. & Sheng, H. Residual networks for light field image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11046–11055 (2019).
7. Liang, Z., Wang, Y., Wang, L., Yang, J. & Zhou, S. Light field image super-resolution with transformers. *IEEE Signal Process. Lett.* **29**, 563–567 (2022).
8. Wang, T.-C., Efros, A. A. & Ramamoorthi, R. Depth estimation with occlusion modeling using light-field cameras. *IEEE Trans. Pattern Anal. Mach. Intell.* **38**(11), 2170–2181 (2016).
9. Wang, Y., Wang, L., Liang, Z., Yang, J., An, W. & Guo, Y. Occlusion-aware cost constructor for light field depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19809–19818 (2022).
10. Yu, X., Wang, R. & Yu, J. Real-time depth of field rendering via dynamic light field generation and filtering. In: Computer Graphics Forum, vol. 29, pp. 2099–2107 Wiley Online Library (2010).
11. Zhang, B., Sheng, B., Li, P. & Lee, T.-Y. Depth of field rendering using multilayer-neighborhood optimization. *IEEE Trans. Visual Comput. Graph.* **26**(8), 2546–2559 (2019).
12. Zhang, M., Xu, S., Piao, Y. & Lu, H. Exploring spatial correlation for light field saliency detection: Expansion from a single view. *IEEE Trans. Image Process.* **31**, 6152–6163 (2022).
13. Zheng, X., Li, Z., Liu, D., Zhou, X. & Shan, C. Spatial attention-guided light field salient object detection network with implicit neural representation. IEEE Transactions on Circuits and Systems for Video Technology (2024).
14. Tian, X., Zhang, J., Xiang, M. & Dai, Y. Modeling the distributional uncertainty for salient object detection models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 19660–19670 (2023).
15. Wang, Y., Wang, R., Fan, X., Wang, T. & He, X. Pixels, regions, and objects: Multiple enhancement for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10031–10040 (2023).
16. Chen, Z., Xu, Q., Cong, R. & Huang, Q. Global context-aware progressive aggregation network for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 10599–10606 (2020).
17. Ge, Y., Zhang, Q., Xiang, T.-Z., Zhang, C. & Bi, H. Tcnet: Co-salient object detection via parallel interaction of transformers and cnns. *IEEE Trans. Circuits Syst. Video Technol.* **33**(6), 2600–2615 (2022).
18. Zheng, P. et al. Gconet+: A stronger group collaborative co-salient object detector. *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(9), 10929–10946 (2023).
19. Wu, Z., Allibert, G., Meriaudeau, F., Ma, C. & Demonceaux, C. Hidanet: Rgb-d salient object detection via hierarchical depth awareness. *IEEE Trans. Image Process.* **32**, 2160–2173 (2023).
20. Wei, W., Xu, M., Wang, J. & Luo, X. Bidirectional attentional interaction networks for rgb-d salient object detection. *Image Vis. Comput.Bold">138*, 104792 (2023).
21. Li, G. et al. Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Trans. Image Process.* **30**, 3528–3542 (2021).
22. Li, N., Ye, J., Ji, Y., Ling, H. & Yu, J. Saliency detection on light field. *IEEE Trans. Pattern Anal. Mach. Intell.* **39**(8), 1605–1616 (2017).
23. Zhang, J., Wang, M., Gao, J., Wang, Y., Zhang, X. & Wu, X. Saliency detection with a deeper investigation of light field. In: IJCAI, pp. 2212–2218 (2015).

24. Zhang, M., Li, J., Wei, J., Piao, Y. & Lu, H. Memory-oriented decoder for light field salient object detection. Adv. Neural Inf. Process. syst.**32** (2019).
25. Zhang, M. et al. Lfnet: Light field fusion network for salient object detection. *IEEE Trans. Image Process.* **29**, 6276–6287 (2020).
26. Yang, C., Zhang, L., Lu, H., Ruan, X. & Yang, M.-H. Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013).
27. Mahadevan, V. & Vasconcelos, N. Biologically inspired object tracking using center-surround saliency mechanisms. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**(3), 541–554 (2012).
28. Siris, A., Jiao, J., Tam, G. K., Xie, X. & Lau, R. W. Scene context-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4156–4166 (2021).
29. Zhu, L. et al. Aggregating attentional dilated features for salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **30**(10), 3358–3371 (2019).
30. Wei, J., Wang, S. & Huang, Q. F^3net: fusion, feedback and focus for salient object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 12321–12328 (2020).
31. Fu, K., Fan, D.-P., Ji, G.-P. & Zhao, Q. Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3052–3062 (2020).
32. Liu, N., Zhang, N., Shao, L. & Han, J. Learning selective mutual attention and contrast for RBD-d saliency detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(12), 9026–9042 (2021).
33. Li, L., Han, J., Liu, N., Khan, S., Cholakkal, H., Anwer, R. M. & Khan, F. S. Robust perception and precise segmentation for scribble-supervised rgb-d saliency detection. IEEE Trans. Pattern Anal. Mach. Intell.**46**(1) (2023).
34. Zhang, J., Liu, Y., Zhang, S., Poppe, R. & Wang, M. Light field saliency detection with deep convolutional networks. *IEEE Trans. Image Process.* **29**, 4421–4434 (2020).
35. Zhang, Q. et al. A multi-task collaborative network for light field salient object detection. *IEEE Trans. Circuits Syst. Video Technol.* **31**(5), 1849–1861 (2020).
36. Piao, Y., Rong, Z., Zhang, M. & Lu, H. Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11865–11873 (2020).
37. Gao, W., Fan, S., Li, G. & Lin, W. A thorough benchmark and a new model for light field saliency detection. IEEE Trans. Pattern Anal. Mach. Intell.**45**(7) (2023).
38. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014).
39. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W. & Chua, T.-S. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5659–5667 (2017).
40. Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7794–7803 (2018).
41. Xie, Z., Zhang, W., Sheng, B., Li, P. & Chen, C. P. Bagfn: broad attentive graph fusion network for high-order feature interactions. *IEEE Trans. Neural Netw. Learn. Syst.* **34**(8), 4499–4513 (2021).
42. Huang, Z. et al. Ccnet: Criss-cross attention for semantic segmentation. *IEEE Trans. Pattern Anal. & Mach. Intell.* **45**(06), 6896–6908 (2023).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016).
44. He, K., Zhang, Y., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1026–1034 (2015).
45. Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K. & Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. Adv. Neural Inf. Process. Syst.**28** (2015).
46. Zhang, J., Wang, M., Lin, L., Yang, X., Gao, J. & Rui, Y. Saliency detection on light field: A multi-cue approach. ACM Trans. Multimed. Comput. Commun. Appl. (TOMM) 13(3) 1–22 (2017).
47. Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M. & Borji, A. Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018).
48. Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T. & Borji, A. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4548–4557 (2017).
49. Achanta, R., Hemami, S., Estrada, F. & Susstrunk, S. Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1597–1604 (2009). IEEE
50. Margolin, R., Zelnik-Manor, L. & Tal, A. How to evaluate foreground maps? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2014).
51. Jiang, Y., Zhang, W., Fu, K. & Zhao, Q. Meanet: Multi-modal edge-aware network for light field salient object detection. *Neurocomputing* **491**, 78–90 (2022).
52. Piao, Y., Rong, Z., Zhang, M., Li, X. & Lu, H. Deep light-field-driven saliency detection from a single view. In: IJCAI, pp. 904–911 (2019).
53. Fan, D.-P., Lin, Z., Zhang, Z., Zhu, M. & Cheng, M.-M. Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**(5), 2075–2089 (2020).
54. Pang, Y., Zhang, L., Zhao, X. & Lu, H. Hierarchical dynamic filtering network for rgb-d salient object detection. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16, pp. 235–252 (2020). Springer
55. Liu, N., Zhang, N. & Han, J. Learning selective self-mutual attention for rgb-d saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13756–13765 (2020).
56. Pang, Y., Zhao, X., Zhang, L. & Lu, H. Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9413–9422 (2020).
57. Zhao, J.-X., Liu, J.-J., Fan, D.-P., Cao, Y., Yang, J. & Cheng, M.-M. Egnet: Edge guidance network for salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8779–8788 (2019).
58. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M. & Jagersand, M. Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7479–7489 (2019).
59. Liu, J.-J., Hou, Q., Cheng, M.-M., Feng, J. & Jiang, J. A simple pooling-based design for real-time salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3917–3926 (2019).
60. Liu, N., Han, J. & Yang, M.-H. Picanet: Learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3089–3098 (2018).
61. Ma, S. et al. Arfnet: Attention-oriented refinement and fusion network for light field salient object detection. *IEEE Syst. J.* **16**(4), 5950–5961 (2022).
62. Piao, Y., Jiang, Y., Zhang, M., Wang, J. & Lu, H. Panet: Patch-aware network for light field salient object detection. *IEEE Trans. Cybern.* **53**(1), 379–391 (2021).

Acknowledgements
This work was supported by National Key Research and Development Program of China (2020YFB2010704).

Author contributions
Shuai Ma proposed the idea of the article and implemented them through code. Xusheng Zhu wrote the manuscript text and provided feedback on the experimental design. Long Xu conducted the experiment of the article and organized the experimental data. Li Zhou drew the diagram of the article and organized the table. Daixin Chen provided valuable suggestions on the structure of the article and carefully reviewed it.

Declarations

Competing interests
The authors listed in this article declare that they have no conflict of interest.

Additional information
Correspondence and requests for materials should be addressed to X.Z.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.