Weakly-Supervised Salient Object Detection on Light Fields

Zijian Liang, Pengjie Wang, Ke Xu, Pingping Zhang, and Rynson W.H. Lau

***Abstract*—Most existing salient object detection (SOD) methods are designed for RGB images and do not take advantage of the abundant information provided by light fields. Hence, they may fail to detect salient objects of complex structures and delineate their boundaries. Although some methods have explored multi-view information of light field images for saliency detection, they require tedious pixel-level manual annotations of ground truths. In this paper, we propose a novel weakly-supervised learning framework for salient object detection on light field images based on bounding box annotations. Our method has two major novelties. First, given an input light field image and a bounding-box annotation indicating the salient object, we propose a ground truth label hallucination method to generate a pixel-level pseudo saliency map, to avoid heavy cost of pixel-level annotations. This method generates high quality pseudo ground truth saliency maps to help supervise the training, by exploiting information obtained from the light field (including depths and RGB images). Second, to exploit the multi-view nature of the light field data in learning, we propose a fusion attention module to calibrate the spatial and channel-wise light field representations. It learns to focus on informative features and suppress redundant information from the multi-view inputs. Based on these two novelties, we are able to train a new salient object detector with two branches in a weakly-supervised manner. While the RGB branch focuses on modeling the color contrast in the all-in-focus image for locating the salient objects, the Focal branch exploits the depth and the background spatial redundancy of focal slices for eliminating background distractions. Extensive experiments show that our method outperforms existing weakly-supervised methods and most fully supervised methods.**

***Index Terms*—Light field, salient object detection, weak supervision.**

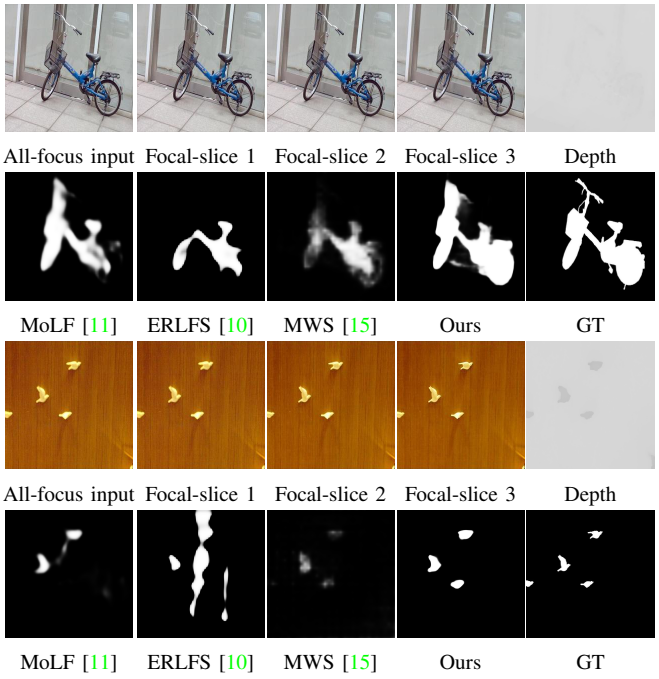

Fig. 1. State-of-the-art SOD (both 4D [11] [10] and 2D [15]) methods fail to accurately delineate the boundary of salient object with complex structure and detect small salient objects. Our method exploits the complementary information from light field, *i.e.*, the all-focus input and focal-slices, and the depth images in a weakly supervised learning manner, and can successfully address these two problems. We only show three out of twelve focal-slices here for illustration.

I. Introduction

Salient object detection (SOD) aims to detect objects (regions) of human interest in the input image. Existing saliency methods can be divided into three categories: *i.e.*, 2D (RGB) saliency detection, 3D (RGB-D) saliency detection, and 4D (light field) saliency detection, according to the types of input data. Among them, 2D saliency detection methods [1] [2] [3] [4] [5] [6] typically suffer from limited visual clues from RGB images for distinguishing salient objects out of cluttered background objects. This often results in unsatisfactory SOD performance, especially in processing complex scenes. 3D saliency detection methods [7] [8] [9] introduce depth information to improve the performance of saliency detection. However, depth information is often not reliable and can significantly degrade the SOD performance. 4D saliency detection methods [10] [11] [12] [13] [14] typically utilize the light field, *i.e.*, a series of images that record both light intensity and traveling directions with different focuses (referred to as focal slices), to achieve good SOD performance for complex scenes. The key advantage of light field is that light field contains abundant information, *i.e.*, focus, depth and color, which play the complementary role to the SOD task. For example, depth information may help determine the foreground object far away from background, while focal slices may provide rich color contrast information to help determine salient objects when they are not far away from background.

Despite that state-of-the-art performance has been achieved by previous 4D saliency methods [10] [11], they suffer from two problems. First, these two methods are fully supervised

Z. Liang is with the Department of Computer Science, Dalian Minzu University, China (e-mail: liang_zijian@foxmail.com)

P. Wang is with the Department of Computer Science, Dalian Minzu University, China (e-mail: pengjiewang@gmail.com). Corresponding author: Pengjie Wang.

K. XU is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: kkangwing@gmail.com).

P. Zhang is with the School of Artificial Intelligence, Dalian University of Technology (e-mail: zhpp@dlut.edu.cn)

R. W. H. Lau is with the Department of Computer Science, City University of Hong Kong, Hong Kong (e-mail: Rynson.Lau@cityu.edu.hk).

learning based so that they require time-consuming pixel-level manual annotation for training. Second, both methods rely on 4D representations learned by modelling the long range correlation of focal slices via ConvLSTM, to locate salient objects. However, such learned representations often contain distracting background information from light field images with different focal points. This fails these methods in accurately detect small salient objects from cluttered background and delineating their boundaries. As shown in the upper two rows of Figure 1, existing 4D [11] [10] methods cannot delineate the boundary of the salient object with complex geometric structure, when the depth information is not reliable. They also cannot detect small salient objects due to the foreground/background ambiguities introduced by different focal slices (see the last two rows of Figure 1). Nonetheless, considering the fact that the focus, depth and color information is complementary with respect to the localization of salient objects, it is possible that we can learn discriminative features from light field images including depths and RGB images with weak annotations.

In light of this, in this paper, we propose a novel weakly supervised saliency detection method on light field. To our knowledge, we are the first to explore this specific topic. Despite various visual features can be derived from the light field images, segmentation of the salient objects is still a challenge. Using the image-level category labels as in the previous weakly supervised 2D methods [15] [16] can only provide coarse salient object localisation (Figure 1). In [17], the object-level bounding box annotation is explored for 2D salient object detection. They show that using bounding boxes not only reduces the cost of labeling, but also provides accurate location clues of objects compared with using category labels. Hence, in our method, we use the bounding box as the weak supervision signals for learning 4D SOD representations. To fully exploit the rich 4D input information, we propose the ground truth label hallucination method to generate pseudo ground truth saliency maps by fusing complementary 4D information, *i.e.*, depth maps and light field slices. Compared to [17], our ground truth label hallucination method takes advantages of abundant information of depth maps and light field slices, and can produce better pseudo ground truth maps than those generated by only RGB images in [17].

In addition, we propose a fusion attention module to selectively focus on informative features and suppress the redundant ones across light field images of different focal points. Based on the proposed pseudo ground truth hallucination method and the fusion attention module, we design a new saliency detector with two parallel branches. The first branch takes the all focus image as input to exploit the RGB contrast information, while the second branch takes the focal slices as input to learn saliency features on light field with the help of the fusion attention module. The whole detector is trained using the pseudo ground truth saliency maps generated by our label hallucination method. We conduct extensive experiments to demonstrate the effectiveness of proposed method against existing weakly- and fully-supervised SOD methods.

The main contributions in this paper can be summarized as:

- We propose a ground truth label hallucination method that can produce high quality saliency maps by fusing various visual clues from depths and RGB images from light field images. To our knowledge, we are the first to explore the weakly-supervised method for light field saliency detection.
- We propose a fusion attention method which studies both spatial and channel information to effectively fuse the various visual clues and eliminates unnecessary redundant information from the light field images.
- Experimental results demonstrate that our method outperforms state-of-the-art weakly-supervised methods and most fully-supervised methods.

II. Related Works

Early salient object detection methods (*e.g.*, [1] [2]) are based on hand crafted low-level features, such as color and texture contrasts. In recent years, the development of deep learning has been boosting the performance of salient object detection. In this paper, we mainly review the deep leaning based saliency detections methods and the weakly-supervised learning in saliency detection.

A. 2D Saliency Detection

2D (RGB image based) saliency detection is the largest family in the SOD field. We discuss the representative ones. Zhang *et al.* [18] proposed a finer-resolution saliency prediction method using an encoder-decoder network. It attached dropout layers in the encoder to learn uncertain features, and hybrid upsampling operations in the decoder to avoid checkerboard artifacts. Hou *et al.* [19] proposed a side-output deep architecture, which uses the high-level semantic information in the deep network for locating the salient regions, and the low-level features for refining the salient region boundaries via enriched details. Liu *et al.* [20] proposed to embed both global and local pixel-level context attention modules into the U-Net [21], for learning saliency-related spatial contextual information better. Li *et al.* [22] proposed a deep network with three branches for dealing with three different sizes of input images, and these three outputs were fused through a learnable attention module.

These methods are typically fully supervised that requires tedious human annotations of ground truth pixel-level labels. To address this problem, several works were proposed to learn weakly-supervised saliency detectors, via assigned category labels [16], bounding boxes [17], category labels and captions [15], and scribbles [23]. Despite their efforts, a fundamental limitation is that RGB images contain limited visual cues, which can easily fail these 2D methods in complex scenes, *e.g.*, when salient objects are small or have complex structures.

B. 3D Saliency Detection

3D (RGBD images based) saliency detection incorporates depth information to help differentiate the foreground and background objects. Their methods mainly differ in the ways of fusing RGB and depth information. Song *et al.* [24]

proposed a salient fusion strategy of multi-scale judgments with the bootstrap learning for RGBD salient object detection. Chen *et al.* [25] proposed a multi-modal fusion framework to perform top-down and bottom-up multi-level feature forecasting across models. Zhao *et al.* [26] proposed to first use a contrast-enhanced network to obtain a single-channel enhanced depth map, then use a pyramidal module to fuse cross-modal and cross-layer features. Cong *et al.* [27] proposed to incorporate depth information into the co-saliency detection, which produces RGBD co-saliency map via a refinement-cycle model. Zhou *et al.* [28] proposed to use an attention-guided bottom-up module to fuse RGB and depth information, and use a top-down module to refine the saliency prediction. Nonetheless, these methods may still fail in many scenes, when the depth information is not reliable.

C. 4D Saliency Detection

In recent years, some works proposed to incorporate light field information by exploiting its multi-view nature for detecting salient objects in complex scenes. Li *et al.* [29] proposed the first dataset, *i.e.*, the LFSD dataset, for salient object detection on light field. They also proposed to estimate the foreground/background information based on the focal and depth information, and combine it with color contrast information, for salient object detection. Zhang *et al.* [13] proposed a multi-cue approach to combine the saliency prior and location prior for salient object detection. Zhang *et al.* [14] proposed to exploit the depth and contrast cues derived from light field information to eliminate the background distracting objects and detect the salient foreground objects. Later works leveraged the deep learning techniques to learning better 4D SOD representations. Piao *et al.* [30] proposed to factorize the saliency detection problem into light field synthesis from a single view and light-field-driven saliency detection. Based on such formulation, a light field synthesis network is proposed to generate rich 4D light field information, and a light-field driven saliency detection network is used for effective saliency detection. Zhang *et al.* [11] proposed a memory-oriented decoder to fuse the complementary information between the 4D light field and the RGB image. Piao *et al.* [10] proposed a knowledge distillation based method to reduce the computation and memory cost of light field SOD. A teacher network was designed to fully exploit the focusness knowledge, and a student network was designed to absorb the knowledge and learn how to replace the knowledge from light field with the knowledge from a RGB input. Piao *et al.* [31] studied the importance of different regions in light field slicing, and proposed a multivariate study module and a patch-aware network to explore light field data in a regional manner. They also designed the sharpness recognition module to boost the performance via feature integration. Zhang *et al.* [32] proposed a light field refinement module and a light field integration module in order to fully utilize the light field information. The light field refinement module learns to refine the features according to the differences between the light field and RGB images, and the light field integration module learns the correlation between the refined light field features to boost the performance. Zhang *et al.* [33] proposed a new dataset using Lytro Illum camera, and obtained Micro lens image array for each light field. They proposed the Modal Angular Changes blocks to extract features from the Micro lens image and the light field sequence for SOD.

Our method differs from previous fully supervised methods in that we propose a weakly supervised learning framework for SOD on light field. Our ground truth label hallucination method leverages various visual clues from light field, depth, and RGB images, to generate high quality pseudo labels from bounding box annotations, for training our 4D saliency detector.

(a) (b) (c)

Fig. 2. Three cases of labeled saliency bounding-boxes (in blue) : (a) single salient object, labeled with a single bounding box; (b) multiple salient objects with overlaps, labeled with a single bounding box; (c) multiple salient objects without overlaps, labeled with multiple boxes.

D. Weakly-Supervised Learning in Saliency Detection

In order to reduce the high manual labeling cost, many endeavors have been made to develop weakly-supervised learning based methods for saliency detection. The key challenge is to design high-quality pseudo maps for guiding the networks to produce pixel-wise saliency maps. Some methods [17], [34], [35] proposed to use traditional SOD methods to generate pseudo labels for training deep saliency models. The image-level object class labels were used in [15], [36], [37] to produce pseudo maps guided by the class activation maps (CAMs) [38]. There are also some methods [23], [39] that proposed to combine pre-trained contour networks with segment proposals [39] or leveraging scribbles [23], [40] to generate pseudo labels for training saliency detection networks. In [17], bounding boxes were used in order to leverage their localization capability for salient object detection. In addition, subitizing was also explored as a weak supervision label for salient object detection [41] and salient instance detection in [5], [42].

All these methods were proposed for 2D saliency detection. In this work, we show that incorporating light field information can significantly boost the salient object detection performance while introducing minimal efforts.

III. The Proposed Method

Previous weakly-supervised methods are mostly based on category labels [16] [36], which can not provide accurate location information of the salient objects. We note that bounding box is another type of cheap annotation labels but is able to provide accurate localization information. Hence,

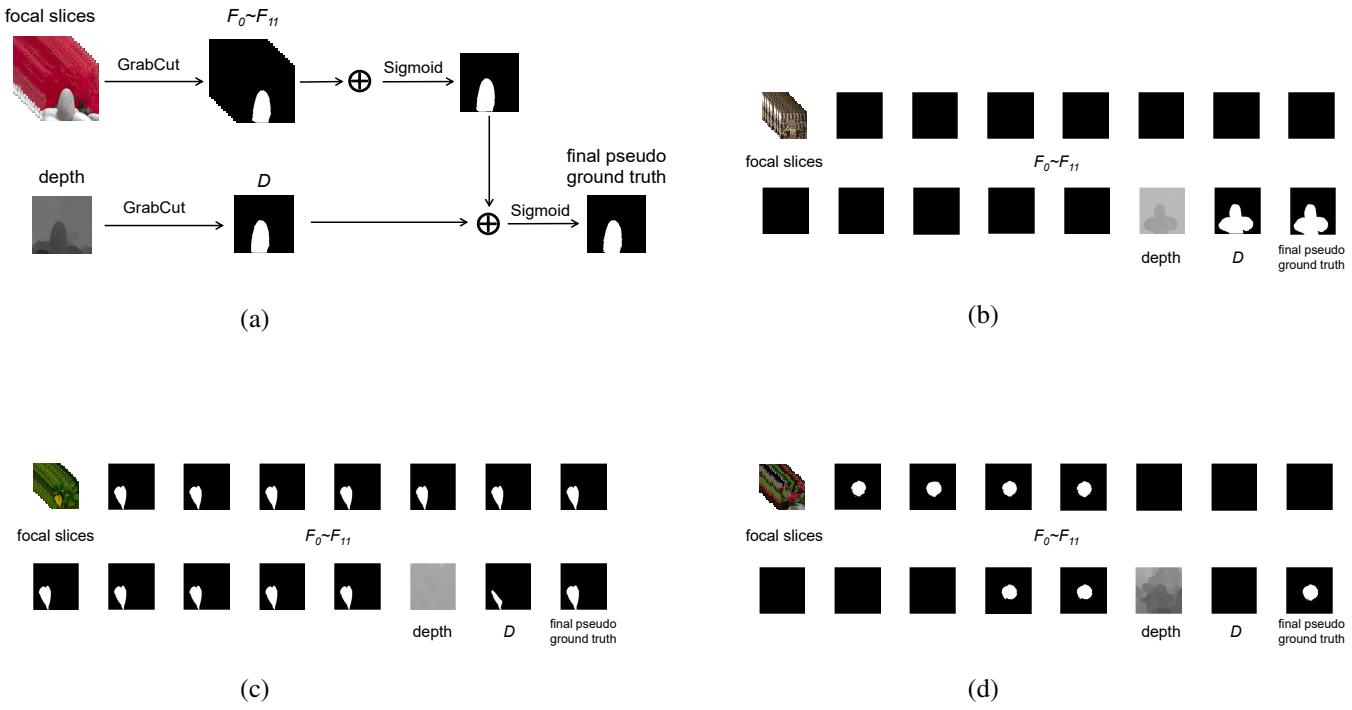

Fig. 3. The pipeline of our pseudo ground truth label hallucination method (a), and three different cases (b), (c) and (d). "Focal slices" represents 12 slices of light field images. "$F_0 \sim F_{11}$" represents the initial saliency maps generated by GrabCut from the 12 focal slices. "Depth" represents depth map, and "D" represents saliency map produced by GrabCut applied on the depth map.

we propose a new weakly-supervised method using saliency bounding-boxes for light field salient object detection.

Specifically, we label the salient objects based on the following rules:

- The saliency bounding-box is a rectangular box which tightly surrounds the salient area and ensures that all salient objects are within the box.
- For multiple salient objects, if the objects are close or overlap with each other, we use one salient bounding-box to label all of them. Otherwise, we label each object with individual bounding boxes.

In Figure 2, we show three cases of labeled salient bounding-box : (a) only one salient object in the image, labeled with a single bounding box; (b) multiple salient objects with overlaps in the image, labeled with a single bounding box; (c) multiple salient objects in the image, labeled with multiple individual bounding boxes. For each image, all the foreground pixels are inside the bounding box and the pixels outside the bounding box are background pixels.

A. Ground Truth Label Hallucination Method

Visual clues from color and depth information from light field images can complement with each other. In light of this, we propose a ground truth label hallucination method, which can fuse various clues and produce a pseudo ground truth label, as shown in Figure 3 (a). First, based on labeled saliency bounding boxes, we apply GrabCut [43] to process 12 slices of light field and depth maps, and obtain the initial salient maps $F_0 \sim F_{11}$ and D respectively. Second, $F_0 \sim F_{11}$ are fused with element-wise addition successively, which is then activated via the sigmoid function, and weighted with D, to produce the final pseudo ground truth salient map S_{final}, as:

$$S_{final} = Bin(Sigmoid(\alpha \times Sigmoid(\sum_{i=0}^{11} F_i) + (1-\alpha) \times D)), \quad (1)$$

where we set α to 0.7 in our implementation. Bin represents the binarization operation, and we set the threshold to 128, *i.e.*, pixel values larger than 128 are set to 255 and 0 otherwise. We use the S_{final} to train our salient object detector.

Figure 3 (b,c,d) show three examples to reveal the complementary nature of visual and depth cues in light field.

- In Figure 3 (b), when the foreground inside the box and the background outside the box have similar colors, the foreground can not be detected via the GrabCut using color contrast information. This is reflected by the poor performance as shown in $F_0 \sim F_{11}$. However, in the depth map, the contrast between foreground and background is obvious. Hence, we can still obtain good result S_{final} after fusion.
- Figure 3 (c) shows the case where the color contrast is more than the depth contrast, and our S_{final} can be benefited from the depth cue.

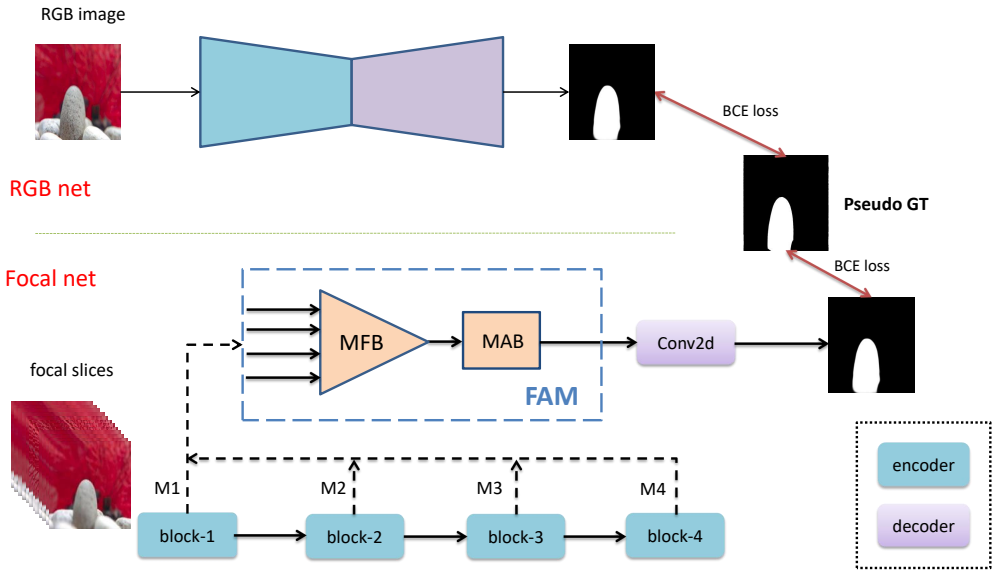

Fig. 4. The overall architecture of our proposed network, which contains two parallel sub-networks, the RGB net and the Focal net. The RGB network takes an RGB image as input, and the Focal network takes light field slices as inputs. The focal slices are processed by four convolution blocks to obtain four feature maps $M_1 \sim M_4$ at different scales. Those feature maps then go through the fusion attention module (FAM), which fuses various visual clues and eliminate unnecessary redundant information from the light field maps. In the test stage, we discard the RGB net, and only use the Focal net for inference.

- In Figure 3 (d), both color and depth cues provide insufficient information individually, our method exploit both of them and obtain satisfactory S_{final} after fusing them.

B. Proposed Network

1) The Pipeline of Proposed Network: As shown in Figure 4, our proposed network consists of two parallel sub-networks, the RGB network, which takes an RGB image as input, and the Focal network, which takes light field slices as inputs. The RGB network has the ResNet-101 [44] like network structure but removes the final fully connected layer. The Focal network also follows the Resnet-101 design, by removing the last pooling layer, convolution layer, and fully connected layer. The focal slices are processed by four convolution blocks to obtain four feature maps $M_1 \sim M_4$ at different scales. Those feature maps then go through the fusion attention module (FAM), which can effectively fuse various visual clues and eliminate unnecessary redundant information from the light field maps. The features produced by FAM are then fed into the decoder to produce the salient object maps. The decoder consists of four convolutional blocks, each of which contains the convolution, RELU activation, and the upsampling operations. We use skip connections to enrich the decoder features with their corresponding encoder features.

2) Fusion Attention Module(FAM): We observe that while light-field images with the focus on the foreground may benefit salient object detection, images focusing on the background may have a negative impact on the detection. Based on this observation, we propose the FAM module. Our FAM module (as shown in Figure 5) has a multi-features fusion block to aggregate visual cues of different scales. To eliminate unnecessary redundant information from the light filed images, we design the multi-branches attention block, which computes a non-local affinity map along the channel dimension and uses this channel-wise attention map to reweight the features.

Multi-features Fusion Block(MFB). The structure of MFB is shown in Figure 5. We first perform maxpool with stride 2 on the M_1, 1×1 convolution on the M_2, M_3 and M_4, respectively. M_3 and M_4 are then upsampled with a scaling factor 2 and 4, respectively. In this way, we obtain four feature maps with size of $12 \times 64 \times 64 \times 64$. We fuse these feature maps by element-wise addition and then reshape it from "$B \times C \times W \times H$" to "$1 \times BC \times W \times H$". Finally we obtain the final output S_{fusion} of the MFB with the size of $1 \times 768 \times 64 \times 64$.

Multi-branches Attention Block(MAB). We use the S_{fusion} obtained from the MFB in the previous step as the input of MAB. The MAB consists of three branches as shown in Figure 5. In the first branch, we obtain correlation matrix $S_{ChanAttn}$ by multiplicating the S_{FeatA} with the transpose of S_{FeatA}, while S_{FeatA} is computed according to Eq. 2. In the second branch, we obtain S_{FeatB} by a 1×1 convolution according to Eq. 3. $S_{ChanAttn}$ is activated via the softmax function, and then combined with S_{FeatB} via matrix multiplication under the form of self attention [48] in order to reweight the features S_{FeatB} along the channel dimension to produce $S_{AttnFeat}$. The obtained result is further fed to a 1×1 convolution to obtain the second branch result, $S'_{AttnFeat}$. Note that the dimensionality reduction in the first two branches follows the bottleneck design of [44], which effectively reduce the computation of this module. In the third branch, S_{fusion} is fed to the global max-pooling to obtain $S_{maxpool}$ as described in Eq. 4. We compute the output of the

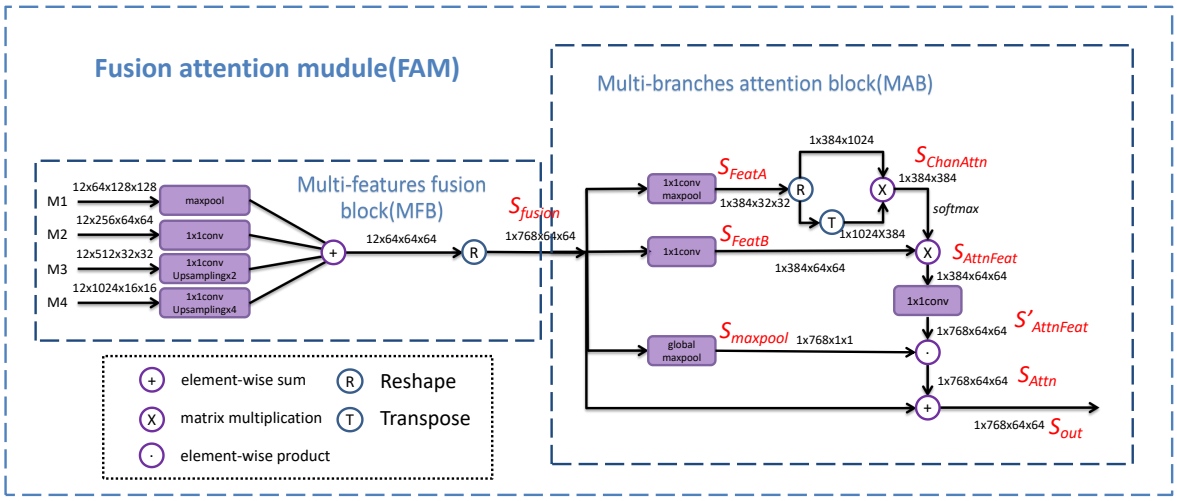

Fig. 5. Fusion attention module. "⊗" represents matrix multiplication, "⊙" represents element-wise product, and "⊕" represents element-wise sum of the matrix.

TABLE I
THE PROPOSED METHOD IS COMPARED QUANTITATIVELY WITH OTHER 2D, 3D AND 4D METHODS. THE BEST RESULTS AMONG THE WEAKLY-SUPERVISED METHODS ARE SHOWN IN BOLD.

Method	Supervision	Type	DUT-LFSD					HFUT-LFSD					LFSD					DUTLF-V2				
			$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$
PiCANet [20]	fully-supervised	2D	0.821	0.083	0.829	0.892	0.736	0.618	0.115	0.781	0.726	0.556	0.671	0.158	0.729	0.780	0.621	0.761	0.114	0.787	0.776	0.753
EGNet [45]	fully-supervised	2D	0.870	0.053	0.886	0.914	0.829	0.672	0.094	0.772	0.794	0.672	0.762	0.118	0.784	0.776	0.717	0.772	0.085	0.783	0.796	0.765
CSF [46]	fully-supervised	2D	0.893	0.045	0.906	0.933	0.868	0.699	0.082	0.798	0.806	0.677	0.814	0.102	0.819	0.842	0.761	0.791	0.054	0.814	0.764	0.782
CPFP [26]	fully-supervised	3D	0.730	0.101	0.741	0.808	0.634	0.594	0.096	0.701	0.768	0.594	0.524	0.186	0.599	0.669	0.465	0.692	0.167	0.703	0.740	0.686
MMCI [47]	fully-supervised	3D	0.750	0.116	0.785	0.853	0.629	0.645	0.104	0.741	0.787	0.540	0.796	0.128	0.799	0.848	0.685	0.709	0.173	0.735	0.752	0.697
TANet [25]	fully-supervised	3D	0.771	0.096	0.803	0.861	0.702	0.638	0.096	0.744	0.789	0.587	0.804	0.112	0.803	0.849	0.727	0.726	0.154	0.743	0.758	0.717
DLFS [30]	fully-supervised	4D	0.801	0.076	0.841	0.891	0.763	0.615	0.098	0.741	0.783	0.590	0.715	0.147	0.737	0.806	0.657	0.742	0.115	0.764	0.773	0.738
MoLF [11]	fully-supervised	4D	0.843	0.052	0.887	0.923	-	0.627	0.095	0.742	0.785	-	0.819	0.089	0.830	0.886	-	-	-	-	-	-
ERLFS [10]	fully-supervised	4D	0.889	0.040	0.899	0.943	0.880	0.705	0.082	0.777	0.831	0.682	0.842	0.080	0.838	0.889	0.842	0.791	0.053	0.808	0.832	0.761
LFS [29]	traditional method	4D	0.484	0.052	0.563	0.728	0.288	0.416	0.222	0.559	0.650	0.264	0.740	0.208	0.680	0.771	0.479	0.457	0.086	0.542	0.698	0.221
WSS [36]	weakly-supervised	2D	0.743	0.126	0.771	0.840	0.637	0.602	0.122	0.713	0.744	0.533	0.771	0.140	0.779	0.837	0.664	0.720	0.158	0.747	0.791	0.708
MWS [15]	weakly-supervised	2D	0.793	0.104	0.829	0.875	0.687	0.604	0.127	0.723	0.732	0.529	0.785	0.130	0.809	0.834	0.687	0.736	0.109	0.757	0.829	0.725
SCA [23]	weakly-supervised	2D	0.814	0.075	0.825	0.880	0.771	0.633	0.098	0.726	0.779	0.596	0.782	0.107	0.786	0.823	0.725	0.751	0.097	0.763	0.825	0.732
SBB [17]	weakly-supervised	2D	0.819	0.076	0.832	0.895	0.769	0.628	0.099	0.723	0.785	0.597	0.816	0.097	0.816	0.858	0.764	0.768	0.081	0.782	0.837	0.752
SCWS [40]	weakly-supervised	2D	0.844	0.063	0.853	0.888	0.819	**0.668**	0.098	0.727	0.788	**0.621**	0.794	0.099	0.802	0.822	0.761	0.778	0.071	0.784	0.832	0.761
Ours	weakly-supervised	4D	**0.884**	**0.043**	**0.889**	**0.937**	**0.870**	0.652	**0.097**	**0.727**	**0.794**	0.609	**0.835**	**0.080**	**0.831**	**0.880**	**0.802**	**0.783**	**0.065**	**0.803**	**0.857**	**0.770**

third branch, S_{Attn}, via element-wise production of $S'_{AttnFeat}$ to spatially reweight the features. Finally, a residual connection (element-wise summation) is used between S_{fusion} and S_{Attn} to get the final output S_{out} as described in Eq.5.

$$S_{FeatA} = Maxpool(Conv_{1\times1}(S_{fusion})), \tag{2}$$

$$S_{FeatB} = Conv_{1\times1}(S_{fusion}), \tag{3}$$

$$S_{maxpool} = GlobalMaxpooling(S_{fusion}), \tag{4}$$

$$S_{out} = Conv_{1\times1}(softmax(S_{FeatA} \times S_{FeatA}^T) \times S_{FeatB}) \cdot S_{maxpool} + S_{fusion}. \tag{5}$$

The process of obtaining $S'_{AttnFeat}$ takes the form of the Non-local design in [49], but it differs from it in the followings aspects. The non-local module models the long range spatial correlations of features of a single scale, while our multi-features fusion block (MFB) fuses features of multiple scales. In addition, in order to model the correlations of feature maps of light field images with different focal points, we compute the long-range affinity map along the channel dimension instead of the spatial dimensions. To locate the salient objects, we also use $S_{maxpool}$ to spatially reweight the features. Table II shows that our method performs better than the original Non-local method in our weakly-supervised SOD task.

IV. EXPERIMENTS

A. Experimental Setup

1) Training Details: Our training dataset is the DUT-LFSD [11], which contains 1000 training samples and 462 test samples. Each sample contains an all-in-focus image, 12 focal slices focusing at different focal points, a depth map, and a ground truth salience map. In our experiment, we do not use the pixel-level ground truth labels of this dataset, but use our saliency bounding-box labels instead. Note that the 12 focal slices share the same saliency bounding-box, as they should contain the same salient objects, although they have different

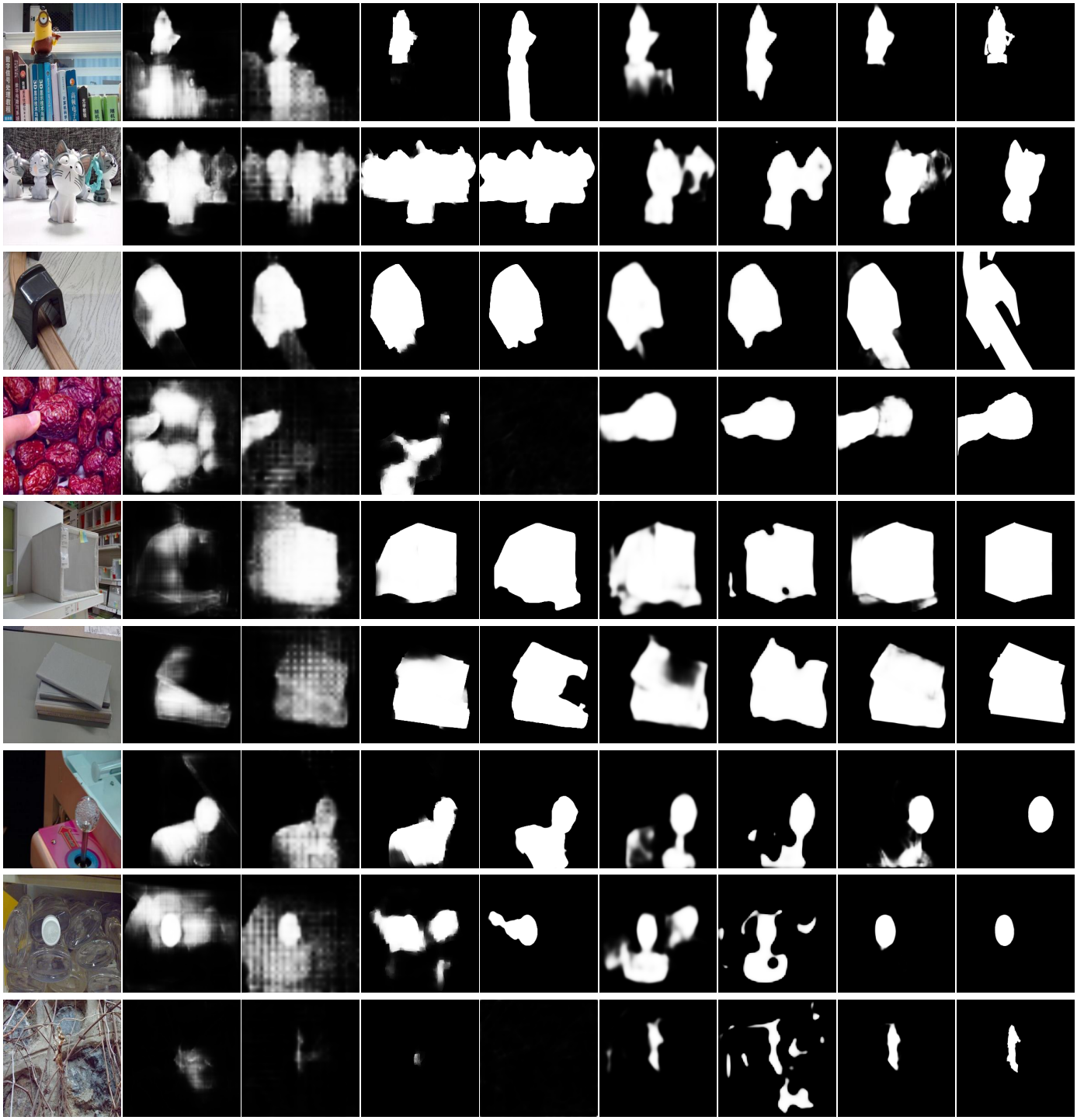

Fig. 6. Qualitative comparison of our method with other methods in some challenging scenes.

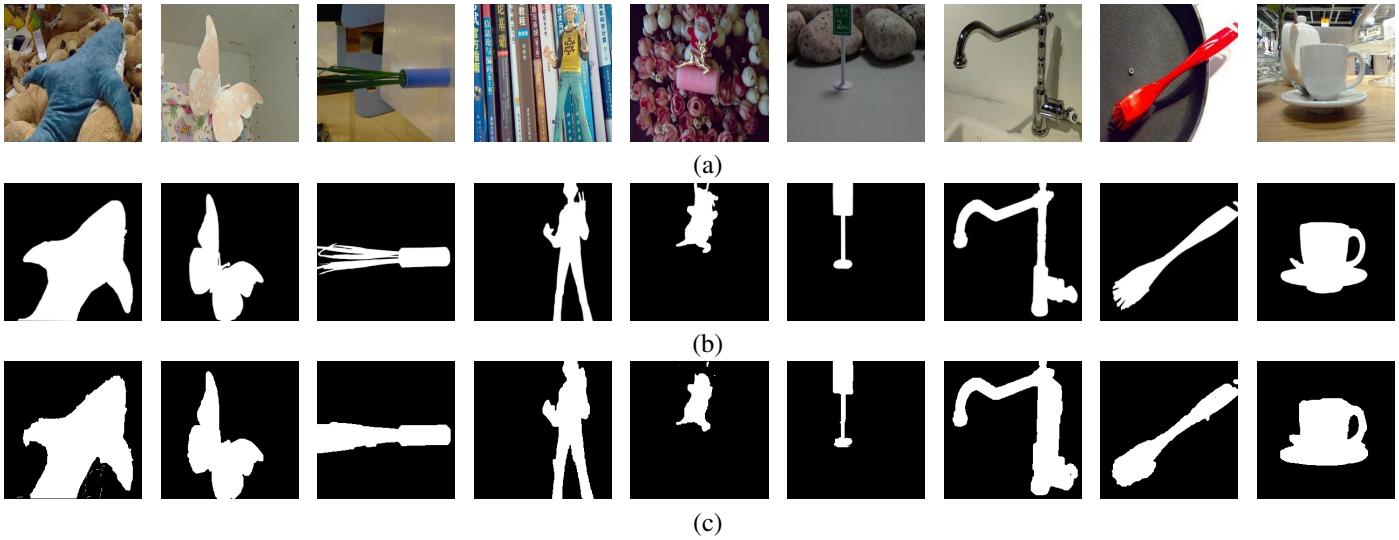

Fig. 7. Qualitative comparison of our pseudo-labels (c) with Ground Truth labels (b). Our hallucination method can produce high-quality pseudo-labels.

TABLE II
COMPARISON WITH THE WIDELY-USED NON-LOCAL METHOD [49].

Module	DUT-LFSD					HFUT-LFSD					LFSD					DUTLF-V2				
	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE $\downarrow$	$S_m\uparrow$	$E_m\uparrow$	$F_\beta^w\uparrow$
Non-local	0.851	0.061	0.855	0.912	0.801	0.634	0.101	0.708	0.772	0.552	0.801	0.104	0.789	0.856	0.731	0.762	0.083	0.784	0.827	0.741
Ours	0.884	0.043	0.889	0.937	0.870	0.652	0.097	0.727	0.794	0.609	0.835	0.080	0.831	0.880	0.802	0.783	0.065	0.803	0.857	0.770

focal points. Based on the labeled saliency bounding box, we use the GrabCut [43] to process 12 focal slices and the depth map to obtain the initial salient maps. Then we apply our pseudo ground truth hallucination method to these salient maps to obtain the pseudo ground truth for training our salient object detector. Our experiments are conducted using the PyTorch toolbox and on one GTX 1080Ti GPU. The optimizer is set to the Adam [50] with the learning rate $1e^{-4}$. The batch size is set to 1. All training images are resized to 256×256. The encoders of the two networks are initialized with ResNet-101 [44] pretrained on ImageNet, and other parameters are initialized by Gaussian kernels.

We train our method using the standard binary cross-entropy (BCE) loss function, defined as:

$$L_s = BCE(x_1, y) + BCE(x_2, y), \tag{6}$$

where x_1 and x_1 are the outputs of Focal net and RGB net, respectively. y is the pseudo ground truth.

The pseudo ground truth hallucination process takes about 5 hours, and the training process takes about 40 hours. Our model consumes about 7GB of GPU memory to test a single image.

2) Benchmark Datasets: To evaluate the performance of our proposed network, we conduct experiments on four widely-used light field benchmark datasets. DUT-LFSD [11] contains 1462 samples, which are devided into 1000 for training and 462 for testing. Each sample contains an all-in-focus image, 12 focal slices, a depth map and a ground truth label. The DUTLF-V2 [51] dataset contains 4204 samples in which 2957 samples are for training and 1247 samples are for testing. The HFUT-LFSD [13] dataset contains 255 samples and the LFSD [29] contains 100 samples. We use the test sets of DUT-LFSD 4 [11] and DUTLF-V2 [51], and the whole datasets of HFUT-LFSD [13] and LFSD [29], for evaluation.

3) Evaluation Metrics: Five widely-used salient evaluation metrics are used to evaluate the performance: F-measure (F_β) [1], mean absolute error (MAE), E-measure (E_m) [52], S-measure (S_m) [53], and weight F-measure (F_β^w) [54]. F-measure is defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{7}$$

where β^2 is set to 0.3.

MAE measures the average pixel-wise absolute difference between a predicted saliency map S and the ground truth G as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - G(x,y)|. \tag{8}$$

S-measure evaluates spatial structure similarity based on region-aware structural similarity S_r and object-aware structural similarity S_o as:

$$S_m = \alpha \times S_o + (1-\alpha) \times S_r, \tag{9}$$

where we set α to 0.5 in our experiment.

E-measure is an enhanced alignment measure to jointly capture image-level statistics and local pixel matching information with an alignment matrix ϕFM as:

$$E_m = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi FM(x,y), \tag{10}$$

TABLE III
ABLATION EXPERIMENTS ON DUT-LFSD DATASETS. IN THE COLUMN OF "LABEL TYPE", "GRABCUT" REPRESENTS THE PSEUDO GROUND TRUTH LABELS GENERATED BY THE GRABCUT, AND "OUR HALLUCINATION METHOD" REPRESENTS THE PSEUDO GROUND TRUTH LABEL HALLUCINATION METHOD IN SECTION III-A. "✓" REPRESENTS INCLUDE THIS NETWORK OR BLOCK.

Label Type	RGB net	Focal net	MFB	MAB	DUT-LFSD				
					$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
GrabCut	✓	✓	✓	✓	0.810	0.074	0.840	0.864	0.783
Our Hallucination Method	✓	✓	✓	✓	0.884	0.043	0.889	0.937	0.870
Our Hallucination Method	✓				0.812	0.078	0.829	0.880	0.774
Our Hallucination Method	✓	✓			0.858	0.056	0.878	0.912	0.821
Our Hallucination Method	✓	✓	✓		0.862	0.051	0.880	0.919	0.832
Our Hallucination Method	✓	✓	✓	✓	0.884	0.043	0.889	0.937	0.870

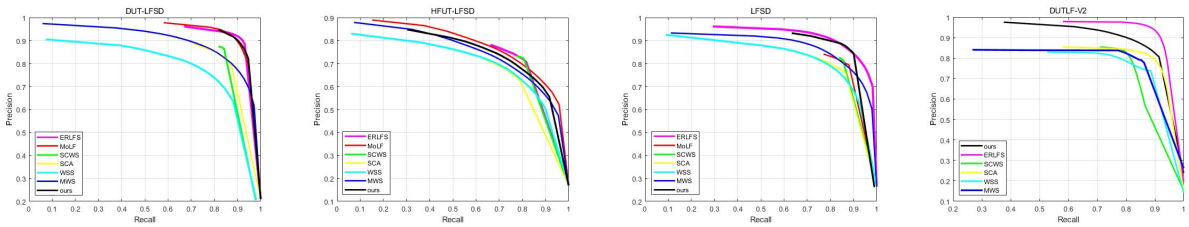

Fig. 8. Precision-recall curves of our method compared against four weakly-supervised methods and two fully-supervised methods.

where W and H are the width and height of the image respectively.

Considering the neighborhood information, F_β^w assigns different weights to errors in different locations as:

$$F_\beta^w = \frac{(1+\beta^2) \times Precision^w \times Recall^w}{\beta^2 \times Precision^w + Recall^w}. \quad (11)$$

B. Comparison with State-of-the-art Methods

We compared seven 2D methods (EGNet [45], PiCANet [20], CSF [46], WSS [36], MWS [15], SCA [23], SCWS [40]), three 3D methods (CPFP [26], TANet [25], MMCI [47]) and four 4D methods (LFS [29], DLFS [30], MoLF [11], ERLFS [10]) on three datasets with five metrics. For a fair comparison, we directly use the pre-trained models or the prediction results provided by their authors in our experiments.

1) Quantitative Evaluation: As shown in Table I, our 4D method outperforms all 2D weakly-supervised methods by a large margin in term of five evaluation metrics across all four datasets. This demonstrates the advantage of incorporating light field information for weakly-supervised salient object detection. Our method also performs favorably against fully-supervised 4D methods. Note that our training set contains only 1000 training samples, which is much smaller than DUTS-TR (of 10553 training samples). Our method achieves consistently superior performances on the currently largest DUTLF-V2 dataset. These results generally demonstrate the effectiveness of our pseudo ground truth hallucination method and our FAM module.

We further compare our method to the widely used Non-local Method [49], which models the long-range spatial correlations. We replace our FAM module with the Non-local block in [49] for comparison. Table II shows that our FAM performs consistently better than the original Non-local block across four datasets. This demonstrates the effectiveness of our FAM in modeling both spatial and channel-wise long-range correlations for the light field SOD task.

Figure 8 shows the precision-recall curves of our method compared with four weakly-supervised methods and two fully-supervised methods on the three datasets. The precision-recall curves of our method are closer to the coordinates (1,1), which means that our method can detect more ground truth foreground pixels with high accuracy.

2) Qualitative evaluation: Figure 6 shows the qualitative comparisons. We can see that our results are significantly better than those from the previous methods across various types of images. For example, in the first three rows, our method can precisely locate the salient regions and obtain high-quality saliency maps. The images in the 4th $\sim$ 6th rows exhibit low contrasts between foreground and background, and our method can delineate the salient objects' boundaries. In the 7th $\sim$ 9th rows, when the salient objects are small, our method can still detect them accurately.

The superiority of our method is due to two reasons. First, with the proposed ground truth label hallucination method, we can produce high quality pseudo saliency maps. Second, our FAM module can effectively suppress the redundant information of light field images and improve the performance of our results, especially for small objects and in complex backgrounds.

TABLE IV
EFFECTIVENESS OF THE DEPTH INFORMATION IN PRODUCING HIGH-QUALITY PSEUDO-LABELS.

	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
Without depth	0.902	0.045	0.922	0.898	0.887
With depth	0.935	0.039	0.943	0.925	0.913

C. Ablation Study

In Table III, we analyze the effects of our pseudo ground truth label hallucination method, the RGB net, the Focal net, and the fusion attention module (FAM). In the first column

TABLE V
COMPARISON BETWEEN FOCAL NET AND TWO FUSION STRATEGIES.

Output type	DUT-LFSD					HFUT-LFSD					LFSD					DUTLF-V2				
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
S_{focal}	0.884	0.043	0.889	0.937	0.870	0.652	0.097	0.727	0.794	0.609	0.835	0.080	0.831	0.880	0.802	0.783	0.065	0.803	0.857	0.770
$S_{fusion1}$	0.872	0.048	0.882	0.922	0.865	0.636	0.106	0.712	0.787	0.596	0.813	0.104	0.825	0.857	0.780	0.767	0.068	0.786	0.844	0.762
$S_{fusion2}$	0.876	0.050	0.891	0.935	0.869	0.647	0.103	0.721	0.784	0.607	0.811	0.102	0.829	0.868	0.794	0.780	0.063	0.812	0.855	0.768

of the table, "GrabCut" represents the pseudo ground truth labels generated by GrabCut, "Ours" represents the pseudo ground truth labels derived by our method introduced in Section III-A. The third row in Table III shows that by only using the RGB net, the 2D version of our method can produces on par performance compared to existing 2D weakly-supervised detection methods on the DUT-LFSD [11] dataset (refer to Table I for comparison). This demonstrates the effectiveness of incorporating 4D information for producing high-quality pseudo-labels. By incorporating more designs, the performance improves continuously. Note that MFB is designed to provide the input of MAB, using MFB alone may not help improve the performance significantly. MAB is the most important part of our FAM module. It fuses various visual clues and eliminates unnecessary redundant information from the light field maps. With the combination of MAB and MFB, FAM can greatly improve the performance. We further study the branches in MAB by removing the maxpool of S_{FeatA} and $S_{maxpool}$. In the first branch of MAB, we use maxpool to reduce the size of the feature map S_{FeatA}. If we remove maxpool of S_{FeatA}, the performance will decrease slightly and it will take a longer time for computation. In the third branch of MAB, We use $S_{maxpool}$ to spatially reweight the features $S'_{AttnFeat}$. The results (W $S_{maxpool}$) and (W/O $S_{maxpool}$) are shown in Table VI. The experiment results verify the design of MAB.

TABLE VI
THE RESULTS OF W AND W/O $S_{maxpool}$.

Type	DUT-LFSD				
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
W $S_{maxpool}$	0.884	0.043	0.889	0.937	0.870
W/O $S_{maxpool}$	0.879	0.047	0.882	0.926	0.861

We also conduct an experiment to evaluate the effectiveness of the depth information in producing high-quality pseudo-labels. Table IV shows that removing the depth information degrades the detection performance. The key reason is that the depth information can play a complementary role in our pseudo-label generation process, in addition to the light field slice. The depth information helps distinguish foreground objects when objects have low contrasts to the cluttered background.

Finally, we conduct experiments to study if it is necessary to use both RGB net and Focal net for inference. Specifically, we have explored a naive way to average the results of two branches to obtain the final detection result (indicated as $S_{fusion1}$). We have also adopted a shallow CNN (four convolutional layers) to fuse the results of two branches to produce the final one (indicated as $S_{fusion2}$). The comparison in Table V shows that our Focal net outperforms the other two strategies. Hence, we only use the RGB net to guide the Focal net with the color contrast information of the all-in-focus image during training. Table VII shows that if we remove RGB net, the fused features may lack the color contrast information, leading to decreased performances.

TABLE VII
RESULTS WHEN RGB NET IS NOT INVOLVED IN TRAINING.

Type	DUT-LFSD				
	$F_\beta \uparrow$	MAE $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^w \uparrow$
W RGB net	0.884	0.043	0.889	0.937	0.870
W/O RGB net	0.873	0.050	0.880	0.915	0.859

D. Effects of Less Accurate Pseudo Ground Truth Labels

We study the effects on the final saliency maps qualitatively by using less accurate pseudo ground truth labels. The results are shown in Figure 9. Although the results training with less accurate pseudo labels can not perform as well as the results training with ground truths, our framework can still localize the foreground accurately. However, such cases only take a very small fraction of the whole training dataset. This shows that our method can achieve comparable performance compared to that trained with pixel-level annotations, but our method can significantly reduce the labeling efforts.

(a) Image (b) Training vis (c) Pseudo GT (d) GT

Fig. 9. (a) Images. (b) Training visualization on that how model performs with less accurate pseudo labels. (c) Low accurate pseudo ground truth labels. (d) Ground truth.

V. CONCLUSION

In this paper, we propose a weakly supervised light field salient object detection method based on saliency bounding-box annotations. First, We label the saliency bounding-boxes for the salient objects in the images. Second, in order to obtain high quality pseudo ground truth, we propose a ground truth

label hallucination method, which can produce high quality pseudo saliency maps by fusing visual cues from light field images. Third, we propose a fusion attention module to reduce the negative impact of redundant information from light field images. Based on these efforts, we propose a novel neural network for detecting salient objects on light fields. Experimental results show that the proposed method outperforms all the weakly-supervised methods and most fully-supervised methods on three datasets with five evaluation metrics. In the future, we are interested in exploring distillation methods to train a 2D weakly-supervised detector using light field information.

REFERENCES

[1] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in CVPR, 2009, pp. 1597–1604. 1, 2, 8

[2] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," TPAMI, vol. 37, no. 3, pp. 569–582, 2014. 1, 2

[3] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in ICCV, 2015, pp. 1404–1412. 1

[4] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in CVPR, 2016, pp. 2334–2342. 1

[5] X. Tian, K. Xu, X. Yang, B. Yin, and R. W. Lau, "Weakly-supervised salient instance detection," in BMVC, 2020. 1, 3

[6] X. Tian, K. Xu, X. Yang, L. Du, B. Yin, and R. W. Lau, "Bi-directional object-context prioritization learning for saliency ranking," in CVPR, 2022. 1

[7] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in CVPR, 2020. 1

[8] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient rgb-d salient object detection," in CVPR, 2020. 1

[9] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," TNNLS, pp. 1–16, 2020. 1

[10] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection," in AAAI, 2020. 1, 2, 3, 6, 7, 9

[11] M. Zhang, J. Li, J. Wei, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in NeurIPS, 2019, pp. 898–908. 1, 2, 3, 6, 7, 8, 9, 10

[12] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep learning for light field saliency detection," in ICCV, 2019, pp. 8838–8848. 1

[13] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency detection on light field: A multi-cue approach," TOMM, vol. 13, no. 3, pp. 1–22, 2017. 1, 3, 8

[14] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field." in IJCAI, 2015, pp. 2212–2218. 1, 3

[15] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in CVPR, 2019. 1, 2, 3, 6, 7, 9

[16] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in AAAI, 2018. 2, 3

[17] Y. Liu, P. Wang, Y. Cao, Z. Liang, and R. W. Lau, "Weakly-supervised salient object detection with saliency bounding boxes," TIP, 2021. 2, 3, 6

[18] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in ICCV, 2017, pp. 212–221. 2

[19] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in CVPR, 2017, pp. 3203–3212. 2

[20] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in CVPR, 2018, pp. 3089–3098. 2, 6, 9

[21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in MICCAI. Springer, 2015, pp. 234–241. 2

[22] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in CVPR, 2017, pp. 2386–2395. 2

[23] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in CVPR, 2020. 2, 3, 6, 7, 9

[24] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning," TIP, vol. 26, no. 9, pp. 4204–4216, 2017. 2

[25] H. Chen and Y. Li, "Three-stream attention-aware network for rgb-d salient object detection," TIP, vol. 28, no. 6, pp. 2825–2835, 2019. 3, 6, 9

[26] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for rgbd salient object detection," in CVPR, 2019, pp. 3927–3936. 3, 6, 9

[27] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, "An iterative co-saliency framework for rgbd images," TCYB, vol. 49, no. 1, pp. 233–246, 2017. 3

[28] X. Zhou, G. Li, C. Gong, Z. Liu, and J. Zhang, "Attention-guided rgbd saliency detection using appearance information," IVC, vol. 95, p. 103888, 2020. 3

[29] Nianyi, Li, Jinwei, Ye, Yu, Ji, Haibin, Ling, Jingyi, and Yu, "Saliency detection on light field," TPAMI, vol. 39, no. 8, pp. 1605–1616, 2017. 3, 6, 8, 9

[30] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep light-field-driven saliency detection from a single view." in IJCAI, 2019, pp. 904–911. 3, 6, 9

[31] Y. Piao, Y. Jiang, M. Zhang, J. Wang, and H. Lu, "Panet: Patch-aware network for light field salient object detection," IEEE TC, 2021. 3

[32] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu, "Lfnet: Light field fusion network for salient object detection," TIP, vol. 29, pp. 6276–6287, 2020. 3

[33] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," TIP, 2020. 3

[34] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in ICCV, 2017. 3

[35] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in CVPR, 2018. 3

[36] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in CVPR, 2017. 3, 6, 7, 9

[37] Y. Piao, J. Wang, M. Zhang, and H. Lu, "Mfnet: Multi-filter directive network for weakly supervised salient object detection," in CVPR, 2021. 3

[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in CVPR, 2016. 3

[39] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in ECCV, 2018. 3

[40] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, "Structure-consistent weakly supervised salient object detection with local saliency coherence," arXiv:2012.04404, 2020. 3, 6, 7, 9

[41] X. Zheng, X. Tan, J. Zhou, L. Ma, and R. Lau, "Weakly-supervised saliency detection via salient object subitizing," TCSVT, 2021. 3

[42] X. Tian, K. Xu, X. Yang, B. Yin, and R. W. Lau, "Learning to detect instance-level salient objects using complementary image labels," IJCV, 2022. 3

[43] C. Rother, V. Kolmogorov, and A. Blake, "" grabcut" interactive foreground extraction using iterated graph cuts," TOG, vol. 23, no. 3, pp. 309–314, 2004. 4, 8

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in CVPR, 2016, pp. 770–778. 5, 8

[45] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in ICCV, 2019, pp. 8779–8788. 6, 9

[46] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in ECCV, 2020. 6, 9

[47] H. Chen, Y. Li, and D. Su, "Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection," PR, vol. 86, pp. 376–385, 2019. 6, 9

[48] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," NeurIPS, 2017. 5

[49] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in CVPR, 2018, pp. 7794–7803. 6, 8, 9

[50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980, 2014. 8
[51] Y. Piao, Z. Rong, S. Xu, M. Zhang, and H. Lu, "Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection," arXiv:2012.15124, 2020. 8
[52] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," arXiv:1805.10421, 2018. 8
[53] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in ICCV, 2017, pp. 4548–4557. 8
[54] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in CVPR, 2014, pp. 248–255. 8