Unsupervised Salient Object Detection on Light Field with High-Quality Synthetic Labels

Yanfeng Zheng, Zhong Luo, Ying Cao, Xiaosong Yang, Weiwei Xu, Zheng Lin, Nan Yin, Pengjie Wang†

***Abstract*—Most current Light Field Salient Object Detection (LFSOD) methods require full supervision with labor-intensive pixel-level annotations. Unsupervised Light Field Salient Object Detection (ULFSOD) has gained attention due to this limitation. However, existing methods use traditional handcrafted techniques to generate noisy pseudo-labels, which degrades the performance of models trained on them. To mitigate this issue, we present a novel learning-based approach to synthesize labels for ULFSOD. We introduce a prominent focal stack identification module that utilizes light field information (focal stack, depth map, and RGB color image) to generate high-quality pixel-level pseudo-labels, aiding network training. Additionally, we propose a novel model architecture for LFSOD, combining a multi-scale spatial attention module for focal stack information with a cross fusion module for RGB and focal stack integration. Through extensive experiments, we demonstrate that our pseudo-label generation method significantly outperforms existing methods in label quality. Our proposed model, trained with our labels, shows significant improvement on ULFSOD, achieving new state-of-the-art scores across public benchmarks.**

***Index Terms*—Light Field, Salient Object Detection, Unsupervised Model.**

I. Introduction

Salient Object Detection (SOD) has always been an essential task in the field of computer vision. Its goal is to identify pixels or regions in an image that capture the human attention the most. Detecting salient objects can enhance various computer vision and image processing applications, including visual tracking and image segmentation. In the past few years, SOD based on RGB images [1], [2] has made significant progress. To further enhance detection accuracy in challenging scenarios, recent efforts have applied emerging light field data to this task, known as LFSOD. The key distinction between LFSOD and traditional salient object detection is the wealth of information present in light field data, providing an abundance of visual cues that significantly enhance performance.

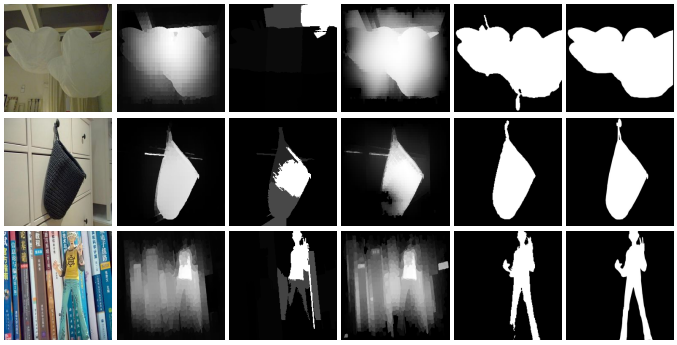


Fig. 1. Comparison of different pseudo-label generation methods. Notably, compared to hand-crafted methods, our method can generate higher-quality labels that are closer to the ground truth.

Based on these abundant cues, researchers have proposed competitive salient object detection methods on light field [6]–[8]. However, most methods necessitate supervision in the form of pixel-level labels, and acquiring these labels for a large quality of images is highly time-consuming. Feng et al. [9] made the first attempt at ULFSOD. However, their method employs traditional hand-crafted SOD methods [5] to generate labels, which are often inherently noisy. Generated pseudo-labels exhibit inaccurate localization, unclear boundaries, and lower confidence, as shown in Figure 1. This is because the traditional hand-crafted methods largely rely on low-level local features to detect salient objects. Directly training deep networks with such noisy labels is not appropriate, as the models can easily adapt to the corrupted labels [10], yielding unsatisfying results at test time. Hence, it is imperative to investigate more effective pseudo-label generation methods to offer unsupervised light field SOD models with pseudo-labels of higher quality, which helps further improve performance.

In light of this, we focus on ULFSOD, and propose a learning-based framework that generates pseudo-labels for SOD in an unsupervised manner. The key component to the success of our approach is a prominent focal slice identification module which is able to precisely locate the optimal focal slice on which true salient regions can be estimated reliably. Our approach is capable of hallucinating more accurate, complete and noiseless labels compared to existing methods as shown in Figure 1. While our label generation method is generic and can work with any SOD models trained with pixel-level annotations, we also propose a SOD model, to further

Yanfeng Zheng and Zhong Luo contributed equally to this work. Pengjie Wang is the corresponding author.

Yanfeng Zheng is with the Department of Computer Science, Dalian Minzu University, China (e-mail: zhengyanfeng1998@163.com).

Zhong Luo is with the College of Civil Engineering, Dalian Minzu University, China (e-mail: zhongluo98@163.com).

Ying Cao is with the School of Information Science and Technology, ShanghaiTech University China (e-mail: caoying59@gmail.com).

Xiaosong Yang is with the National Center for Computer Animation, Bournemouth University, UK (e-mail: xyang@bournemouth.ac.uk).

W. Xu is with the State Key Lab of CAD& CG, Department of Computer Science, Zhejiang University, Hangzhou, Zhejiang 310058, China (e-mail: xww@cad.zju.edu.cn).

Zheng Lin is with the Department of Computer Science and Technology, Tsinghua University, Beijing, China (e-mail: frazer.linzheng@gmail.com).

Nan Yin is with the Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi 108100, UAE (e-mail: yinnan8911@gmail.com).

Pengjie Wang is with the Department of Computer Science, Dalian Minzu University, China. He is also with the National Center for Computer Animation, Bournemouth University, UK (e-mail: pengjiewang@gmail.com).

boost performance, with two novel components: 1) a Multi-Scale Spatial Attention (MSSA) module to enhance focal stack features and reduce redundant information; 2) a Cross Fusion Module (CFM) based on coordinate attention [11] for the fusion of RGB features and focal stack features. The entire model is trained using pixel-level pseudo-labels generated by our unsupervised pseudo-labeling method. In contrast to [12], our method is unsupervised, avoiding manual labeling costs. We conducted extensive experiments to demonstrate the effectiveness of the proposed method compared to existing unsupervised, weakly-supervised, and fully-supervised methods.

Our main contributions can be summarized as follows:

- We propose a novel unsupervised deep learning framework for LFSOD that generates high-quality salient object maps by leveraging complementary visual cues from depth maps, focal stacks, and all-in-focus images. Our method, for the first time, can achieve pseudo-labels with reasonable quality and thereby greatly helps improve the performance of ULFSOD methods.
- We propose an improved model that is able to learn rich feature representations for SOD through the novel attention modules. Training the model on our generated labels yields further performance improvement.
- Experimental results demonstrate the superiority of our approach over the state-of-the-art unsupervised methods for LFSOD in both 2D and 3D domains.

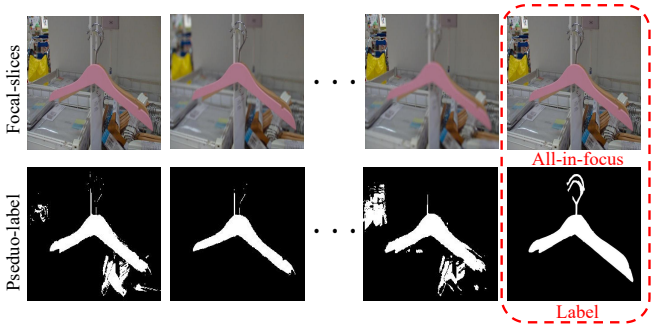


Fig. 2. Label predictions with different focal slices.

II. Related Work

Early SOD methods (e.g., [13], [14]) were based on traditional methods to manually craft low-level features, such as color and texture contrast. In recent years, with the rise of deep learning methods, the performance of SOD has significantly improved. Wang et al. [15] provide a comprehensive survey of deep learning-based salient object detection, covering network architectures, datasets, evaluation metrics and future challenges. This paper mainly discusses the SOD methods based on deep learning that have different inputs.

A. 2D Salient Object Detection

2D (based on RGB images) SOD is the largest family in the SOD field. These methods are usually fully/weakly supervised, requiring tedious manual annotations to obtain pixel-level labels or category labels [16], [17], bounding boxes [18], category labels and captions [19], and scribble [20]. Zhang et al. [20] proposed weakly supervised saliency method based on scribble annotation. They introduced an auxiliary edge detection task to locate target edges, and a gate structure perception loss to constrain the range of structures that need to be recovered. Xu et al. [21] utilized a mix of pseudo labels generated by unsupervised methods and real labels for weakly supervised salient object detection. [22] and [23] are salient object methods based on multimodality. While the former focuses on integrating thermal infrared features as complementary cues to RGB features, the latter focuses on constructing a framework to fuse two modalities to enhance the performance of salient object detection.

To mitigate the label annotation problem, various unsupervised methods have been proposed. Wang et al. [24] proposed a biologically inspired model that infers salient objects from human fixations, leveraging a fixation map to progressively optimize object saliency. Wang et al. [25] introduced an iterative and cooperative top-down and bottom-up network that enhances saliency detection by integrating high-level and low-level features. Furthermore, Wang et al. [26] developed PAGE-Net, which combines pyramid attention and salient edge detection for precise boundary delineation in salient object segmentation. Zhou et al. [27] introduced a pipeline that transforms activation maps from a pre-trained network into high-quality pseudo labels. Zhou et al. [28] developed a confidence-aware salient distilling method to extract rich and precise saliency information from noisy labels, effectively overcoming the limitation of existing methods in handling hard samples.

B. 3D Salient Object Detection

3D (based on RGB-D images) SOD combines depth information to help distinguish between foreground and background objects. Zhang et al. [29] proposed a framework (UC-Net) for RGB-D salient detection inspired by uncertainty. Ji et al. [30] introduced a deep RGB-D unsupervised salient object detection method (USOD) that decomposes depth features to reconstruct both saliency-guided and non-saliency-guided depths, which are used to update pseudo-labels. Yang et al. [31] proposed using depth maps to estimate the pseudo labels, and then they introduced a uncertainty-aware label optimization method to enhance these labels. Li et al. [32] presented a weakly-supervised RGB-D SOD model with scribble guidance. Nevertheless, when the depth information is unreliable, these methods may still fail in many scenes. Xu et al. [33] proposed a weakly-supervised RGB-D salient object detection method that employs prediction consistency training and active scribble boosting to improve detection accuracy while significantly reducing the annotation burden. Zhou et al. [34] provided a comprehensive survey on RGB-D salient object detection models, benchmark datasets, and challenges, offering insights into future directions for enhancing RGB-D SOD performance.

C. 4D Salient Object Detection

Traditional light field SOD techniques have demonstrated the effectiveness of using light field data. Li et al. [35]

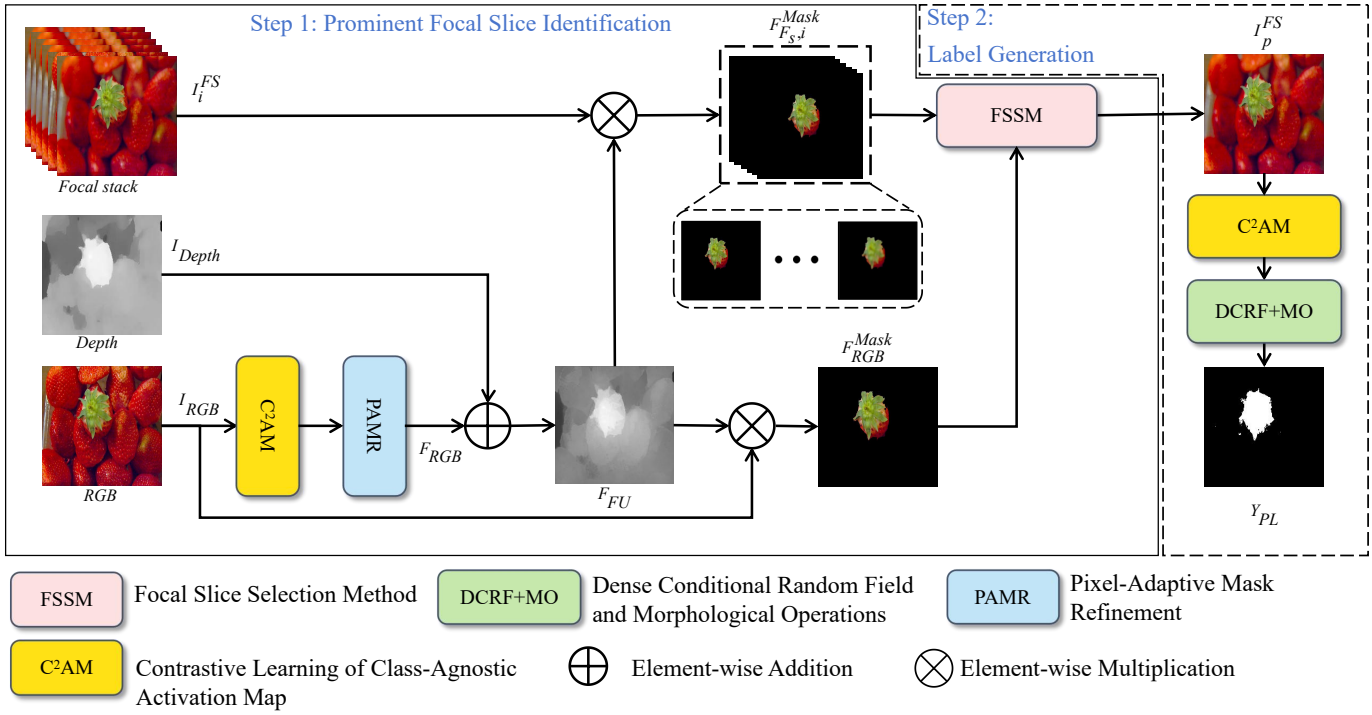


Fig. 3. Pseudo-label generation. The pseudo-label generation process consists of two stages: the prominent focal slice identification (PFSI) and the label generation. We first use the DUT-LFSD dataset to independently train the contrastive learning of class-agnostic activation map (C^2AM) module, and freeze its weights after training is completed. In the first stage (PFSI), we use the frozen C^2AM to select the optimal focal slice. In the second stage (label generation), we generate pseudo-labels from the selected focal slice to serve as supervisory signals for the subsequent network.

introduced the first light field SOD method and established the initial benchmark dataset. Later, Li et al. [36] developed a salient dictionary using weighted sparse coding to generate salient maps.

The advancement of deep learning technology has improved the performance of light-field salient object detection through the use of deep neural networks. A number of LFSOD techniques that leverage deep learning have succeeded in enhancing the quality of predicted salient object maps. Zhang et al. [37] proposed a memory-oriented decoder to fuse the complementary information between 4D light fields and RGB images. Piao et al. [38] explored focal slices in a regional manner and integrated focused salient regions. Liang et al. [12] proposed a weakly supervised learning framework for LFSOD based on bounding boxes. Feng et al. [9] proposed to create pixel-level labels through unsupervised salient object detection methods, and use the labels to train a SOD model, resulting in an USOD method. Unfortunately, their generated labels are often noisy and incomplete, and thus training on such low-quality labels leads to inferior performance.

In this paper, we also aim for ULFSOD with pseudo-label synthesis similar to [9]. However, different from [9], we do not employ traditional hand-engineered methods for label generation, and instead propose a deep learning-based approach to significantly improve the quality of generated labels, resulting in improved SOD performance.

III. The Proposed Method

Given a focal stack, a depth map, and an RGB image from light field data, the goal of LFSOD is to predicts masks for salient objects in the RGB image. Our unsupervised method for LFSOD consists of two stages: first, pseudo-labels (i.e., salient object masks) on light field images are generated; second, the generated labels are used to train a SOD network. In this section, we first present the proposed pseudo-label generation method in detail, which is based on deep learning instead of traditional hand-crafted features. Then, we elaborate on our proposed SOD network, which consists of two essential modules, the multi-scale spatial attention (MSSA) and cross fusion module (CFM).

A. Pseudo-label Generation

Our method is focused on improving the quality of generated pseudo-labels, whose design is essentially motivated by two key observations as follows. First, we observe that information from focal stack, depth maps and RGB images complements each other in locating salient objects, but state-of-the-art methods have not fully harnessed such fertile information. Secondly, different slices aid in effectively separating the foreground from the background. This is because the refocused region in one image includes salient objects, while some defocused regions solely represent the background [37], [39]. We define one slice among the focal stack which is the best for detecting the salient objects as prominent focal slice (PFS). In Figure 2, we present the pseudo-label results generated with different focal slice. If the PFS is used, higher-quality pixel-level pseudo-labels will be obtained compared to using non-prominent focal slice. Hence, we carefully design a prominent focal slice identification module to choose the prominent focal slice by employing rich attributes (focal stack, depth and RGB image) in the light field data. Specifically,

as shown in Figure 3, our pseudo-label generation method includes two components: a prominent focal slice identification module for fusing the complementary features in order to select the prominent focal slice, and a label generation module for generating the final pseudo label from the selected prominent focal slice.

As shown in Figure 3 (left), our Prominent Focal Slice Identification (PFSI) module first obtains a contrastive learning of class-agnostic activation map (C^2AM) for the input RGB image that highlights the foreground region using the approach of [40], which is learned with contrastive learning without image-level supervision. This approach acquires discriminative information between foreground and background by learning the distribution of semantic information in feature space varies across images. It can capture object boundaries, textures, and crucial features, resulting in a more precise delineation of the foreground and background on the activation map.

For our method, we train the C^2AM approach on the DUT-LFSD dataset to enhance its adaptation to light field images. Additionally, to further enhance the quality of the foreground region, we apply the Pixel-Adaptive Mask Refinement (PAMR) [41] to generate adaptive masks pixel by pixel, allowing for a more precise activation of the foreground region. Both the depth map and F_{RGB} have values between 0 and 1, with higher values indicating foreground pixels. Thus, adding them up will combine complementary information from the two to more accurately localize the foreground region. The fusion can be represented as follows:

$$F_{FU} = \left(F_{RGB} + I_{Depth}\right)/2, \tag{1}$$

where F_{RGB} represents the refined contrastive learning of class-agnostic activation map for the RGB image, I_{Depth} represents the depth image, and F_{FU} represents the fused representation of F_{RGB} and I_{Depth}.

Furthermore, we take full advantage of the unique properties of the focal stack in light field data. The focal stack consists of a series of images which focus on different depths, generated by processing the raw light field data. Among them are images with a clear foreground and blurred background, which are exactly what we need. We perform pixel-wise multiplication of the fused representation F_{FU} with the all-in-focus image and the focal stack, resulting in masked images with a black background that emphasize the foreground.

$$F_{RGB}^{Mask} = F_{FU} \otimes I_{RGB}, \tag{2}$$

$$F_{FS,i}^{Mask} = F_{FU} \otimes I_i^{FS}, \tag{3}$$

where $\otimes$ represents element-wise multiplication, F_{RGB}^{Mask} represents the masked RGB image, andI_i^{FS} represents the version of i-th focal stack, and $F_{FS,i}^{Mask}$ represents the masked version of i-th focal stack with only the foreground.

Based on the computed F_{RGB}^{Mask} and $F_{FS,i}^{Mask}$, we perform the best focal slice selection by measuring the consistency of $F_{FS,i}^{Mask}$ with respect to F_{RGB}^{Mask}. In particular, we compute the pixel-level distance between the masked RGB image and the masked focal stack and choose the focal slice with the minimum distance. Then the prominent focal slice is denoted as I_p^{FS} that has a clear foreground region and a blurred background area. Formally, the distance between F_{RGB}^{Mask} and $F_{FS,i}^{Mask}$ is written as:

$$D_{RGB,i} = \sum_{\mathrm{x,y}} \left| F_{RGB}^{Mask}(\mathrm{x,y}) - F_{FS,i}^{Mask}(\mathrm{x,y}) \right|, \tag{4}$$

We empirically found that the quality of the depth maps in the current dataset varies, which affects the effectiveness of the model. By contrast, the quality of focus stack is relatively stable, which contains not only focus information but also certain depth information. Therefore, we choose the prominent focal slice for further processing, which is sent to the label generation module.

In the label generation module, I_p^{FS} is first passed into the C^2AM method, resulting in an activation map of soft values. The activation map is then converted into a binary mask with the dense conditional random field [42] and morphological operation, giving the final pseudo-label Y_{PL}.

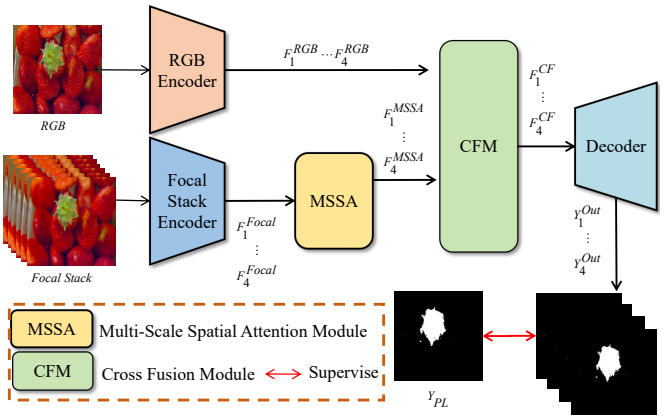


Fig. 4. **Our proposed LFSOD network with pseudo-labels.** The network uses the pseudo-labels generated in the first stage for supervised learning. We propose MSSA in order to enhances the focal stack features and eliminate redundant information, the proposed CFM fully fuses the RGB image and focal stack features.

B. Unsupervised Salient Object Detection with Pseudo-labels

As shown in Figure 4, our proposed SOD model consists of two parallel feature extraction networks, encoding the RGB image and focal stack respectively. These feature extraction networks are constructed based on the ResNet-50 [43] architecture, without the fully connected layer used for classification. Through four feature extraction layers, we extract features from both the RGB image and focal stack, generating four feature maps of four scales for each denoted as $\left(F_1^{Focal}, \ldots, F_4^{Focal}\right)$ and $\left(F_1^{RGB}, \ldots, F_4^{RGB}\right)$, which serve as inputs for subsequent modules. In order to enhance the focal stack features and eliminate redundant information, we introduce MSSA, which weights the focal stack features, to emphasize their importance at different spatial scales, thus capturing scene characteristics more effectively. We further introduce CFM to fuse the RGB image features with the focal stack features. This fusion combines features from different modalities to create a more informative representation.

1) Multi-Scale Spatial Attention: Interference within the focal stack may potentially degrade the performance of SOD, although they contain rich light field cues. This interference

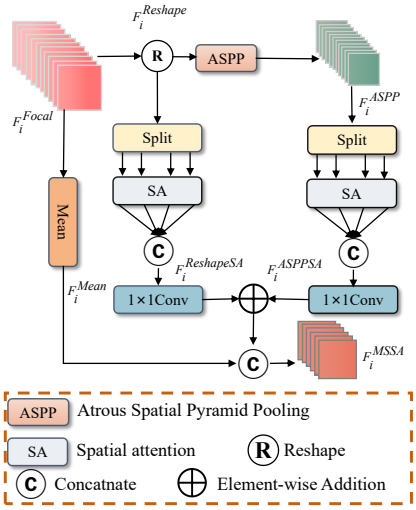


Fig. 5. Multi-Scale Spatial Attention (MSSA).

is usually a result of focal slices that are focused on unrelated depth levels. In these cases, the focused regions often correspond to the background, while the blurred region contains the desired objects. To address this issue, we enhance the focal stack features with the MSSA before feeding them into CFM.

The MSSA integrates the Atrous Spatial Pyramid Pooling (ASPP) [44] and spatial attention [45]. The ASPP structure uses dilated convolutions with different dilation rates to capture multi-scale features, enabling wide receptive fields across various spatial ranges. Such multi-scale features help to capture the depth and structure information of the scene, and reduce information loss to some extent. Moreover, spatial attention not only emphasizes the semantic information of individual features but also model the positional relationships of features over spatial dimension, thereby enhancing the richness of information available for subsequent tasks.

Figure 5 illustrates the structure of the proposed MSSA module. Given the focal stack feature $F_i^{Focal} \in R^{S\times C\times H\times W}$, where S is the number of focal stack and C and (H, W) represent the channel number and spatial resolution of the feature map. We first apply averaging across stack to it and then reshape it, resulting in two features $F_i^{Reshape} \in R^{CS\times H\times W}$ and $F_i^{Mean} \in R^{C\times H\times W}$, , which serve as inputs to the module. We feed $F_i^{Reshape}$ to the ASPP that contains dilated convolutions (with dilation rates of 3 and 6) followed by adaptive average pooling, and the results are concatenated along channel dimension, resulting in a feature F_i^{ASPP} with context information at different scales.

Due to the observed feature redundancy in the focal stack and significant differences in features at different focus positions, we split the features along stack channel and input them into the spatial attention module, resulting in $F_i^{ReshapeSA}$ and F_i^{ASPPSA}:

$$F_i^{ReshapeSA} = Conv\left(Cat\left(SA\left(\phi\left(F_i^{Reshape}\right)\right)\right)\right), \quad (5)$$

$$F_i^{ASPPSA} = Conv\left(Cat\left(SA\left(\phi\left(F_i^{ASPP}\right)\right)\right)\right), \quad (6)$$

where ϕ represents stack-wise split, SA represents spatial attention, Cat denotes concatenation along channel dimension,

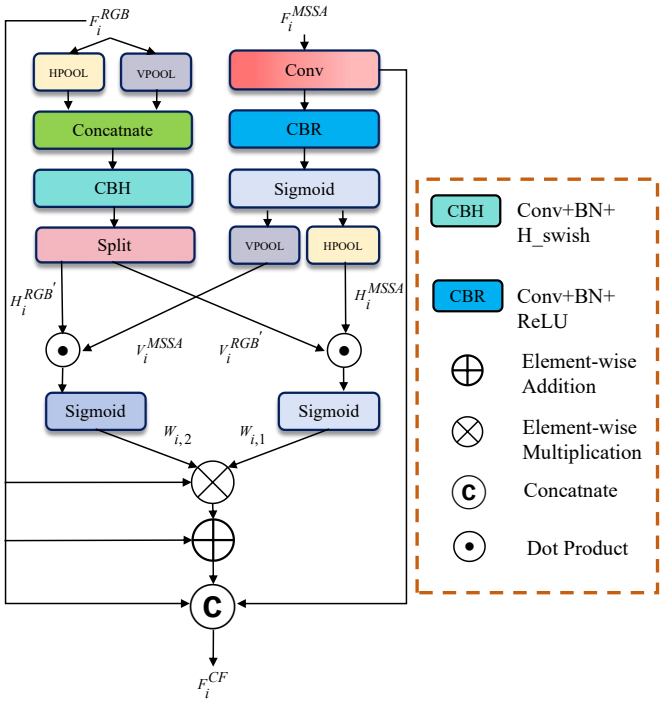


Fig. 6. Cross Fusion Module (CFM).

and C_1 represents 1×1 convolutions. Then, the two features $F_i^{ReshapeSA}$ and F_i^{ASPPSA} are transformed by two 1×1 convolutions, added element-wise, and concatenated with F_i^{Mean} that captures global information across stack to obtain the final output feature F_i^{MSSA} .

$$F_i^{MSSA} = Cat\left(F_i^{Mean}, \left(F_i^{ReshapeSA} + F_i^{ASPPSA}\right)\right), \quad (7)$$

The above operations fuse global information of the stack, multi-scale features and spatial relational information, which gives rise to an expressive feature representation for the stack and thus improves the model performance.

2) Cross Fusion Module: To obtain more informative and discriminative representations for detection, it is crucial to fuse the rich feature information from the focal stack and the RGB image. The focal stack contains comprehensive information about the depth, shape, and position of objects in light field data. On the other hand, RGB images provide color and texture information. In the two-stream network structure, the RGB and focus stack feature extraction branches produce different and rich features, which we opt to merge together to produce a more powerful representation. To this end, we propose a cross fusion model which adapts the coordinate attention [11] to fuse features of different modalities: the focal stack feature and the RGB image feature, for each scale.

As shown in Figure 6, the inputs to the CFM include the focal stack feature F_i^{MSSA} output from the MSSA module and the RGB image feature F_i^{RGB} for scale i. We apply a 1×1 convolution to F_i^{MSSA}, obtaining $\hat{F}_i^{MSSA}$. Then, $\hat{F}_i^{MSSA}$ is fed through a convolution followed by a sigmoid operation, and the outputs are pooled along the horizontal and vertical

TABLE I
COMPARISON WITH PREVIOUS TRADITIONAL METHODS (T), FULLY SUPERVISED METHODS (F), WEAKLY SUPERVISED METHODS (W), AND UNSUPERVISED METHODS (U). - REPRESENTS UNAVAILABLE RESULTS, AND THE UNSUPERVISED BEST RESULTS ARE BOLDED IN BLACK.

Method	Type/Sup	DUT-LFSD				LFSD				DUTLF-V2				PKU-LF			
		$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$	$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$
WSC [36]	/T	0.158	0.591	0.776	0.652	0.150	0.722	0.787	0.702	0.156	0.484	0.738	0.609	0.117	0.598	0.743	0.698
PiCA [46]	2D/F	0.083	0.763	0.898	0.830	0.133	0.689	0.824	0.764	0.083	0.664	0.869	0.776	0.067	0.703	0.871	0.814
UCNet [29]	3D/F	0.087	0.758	0.858	0.792	0.149	0.689	0.800	0.722	0.057	0.799	0.899	0.844	0.054	0.790	0.891	0.842
MoLF [37]	4D/F	0.052	0.854	0.922	0.886	0.115	0.749	0.846	0.777	0.065	0.745	0.867	0.826	0.084	0.676	0.825	0.769
WSS [20]	2D/W	0.069	0.822	0.900	0.838	0.098	0.789	0.857	0.800	0.082	0.721	0.855	0.786	0.061	0.757	0.878	0.816
MFNet [47]	2D/W	0.099	0.752	0.854	0.783	0.131	0.736	0.804	0.749	0.095	0.679	0.833	0.758	0.070	0.715	0.864	0.784
WSLF [12]	4D/W	0.043	0.881	0.937	0.889	0.080	0.861	0.880	0.831	0.065	0.743	0.857	0.803	-	-	-	-
A2SV1 [27]	2D/U	0.077	0.820	0.889	0.827	0.103	0.808	0.855	0.797	0.079	0.732	0.857	0.790	0.056	0.787	0.882	0.835
DSU [30]	3D/U	0.109	0.729	0.853	0.776	0.128	0.760	0.834	0.781	0.109	0.645	0.824	0.743	0.075	0.734	0.852	0.803
DLM [31]	3D/U	0.075	0.808	0.897	0.846	0.098	0.806	0.875	0.824	0.088	0.694	0.847	0.784	0.065	0.749	0.858	0.826
A2SV2 [28]	2D/U	0.056	0.859	0.920	0.869	0.084	0.819	0.877	0.828	0.072	0.764	0.873	0.815	0.057	0.793	0.887	**0.839**
A2SV2 [28]	3D/U	0.058	0.865	0.920	0.865	**0.081**	0.834	**0.885**	**0.838**	0.066	0.775	0.884	0.820	0.055	0.793	**0.891**	0.835
Ours	4D/U	**0.052**	**0.889**	**0.924**	**0.877**	0.093	**0.842**	0.866	0.819	**0.061**	**0.796**	**0.888**	**0.826**	**0.054**	**0.804**	0.887	**0.839**

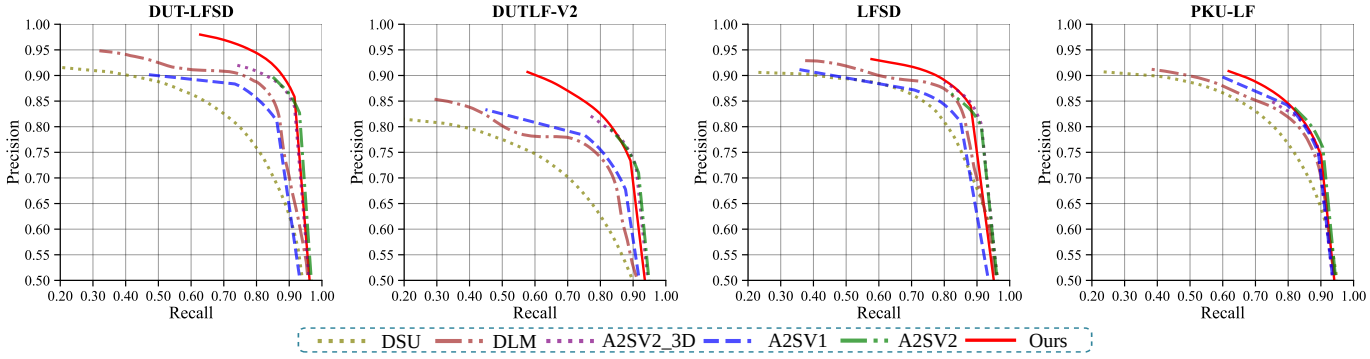


Fig. 7. Precision-recall curves of different unsupervised methods on four datasets.

directions:

$$H_i^{MSSA} = HPOOL\left(\sigma\left(CBR\left(\hat{F}_i^{MSSA}\right)\right)\right), \quad (8)$$

$$V_i^{MSSA} = VPOOL\left(\sigma\left(CBR\left(\hat{F}_i^{MSSA}\right)\right)\right), \quad (9)$$

where CBR represents the Convolution + BatchNorm + ReLU operation, σ represents the sigmoid function, $HPOOL$ and $VPOOL$ represent Horizontal Pooling and Vertical Pooling, respectively, and are specialized pooling operations designed to capture vertical and horizontal contextual information, respectively. These operations improve the network's ability to extract directional features, leading to more accurate object detection. Furthermore, RGB image feature F_i^{RGB} is pooled along the horizontal and vertical directions, resulting in H_i^{RGB} and V_i^{RGB}. These two pooled features are then concatenated and fed through a convolution with Swish nonlinearity, and the result is further split into two separate features $H_i^{RGB'}$ and $V_i^{RGB'}$ along the spatial dimension. Then, dot product is performed between V_i^{MSSA} and $H_i^{RGB'}$, as well as between V_i^{MSSA} and $V_i^{RGB'}$, resulting in two different attention weight matrix $W_{i,1}$ and $W_{i,2}$.

$$W_{i,1} = \sigma\left(H_i^{MSSA} \odot V_i^{RGB'}\right), \quad (10)$$

$$W_{i,2} = \sigma\left(V_i^{MSSA} \odot H_i^{RGB'}\right), \quad (11)$$

We element-wise multiply the RGB image feature with $W_{i,1}$ and $W_{i,2}$ sequentially, and add a skip connection from the RGB image feature to the multiplication result, producing an enhanced image feature. Finally, we concatenate the RGB image feature, focal stack feature and the enhanced image feature to obtain the output of the CFM, denoted as F_i^{CF}:

$$F_i^{CF} = Cat\left(\hat{F}_i^{MSSA}, F_i^{RGB}, \left(F_i^{RGB} \otimes W_{i,1} \otimes W_{i,2}\right) + F_i^{RGB}\right), \quad (12)$$

3) Decoder: The output features of the CFM for each scale is sent into a decoder D to produce four salient map predictions:

$$Y_i^{out} = D\left(F_i^{CF}\right), i = 1, 2, 3, 4, \quad (13)$$

The decoder consists of two convolutions, followed by the batch normalization and the ReLU activation function.

4) Loss function: We utilize a combination loss function consisting of a BCE loss, a SSIM loss and an IoU loss for training [48].

$$\mathcal{L} = \sum_{i=1}^{4}\Big(\mathcal{L}_{BCE}\left(Y_i^{\text{out}}, Y_{PL}\right) + \mathcal{L}_{SSIM}\left(Y_i^{\text{out}}, Y_{PL}\right) + \mathcal{L}_{IoU}\left(Y_i^{\text{out}}, Y_{PL}\right)\Big), \quad (14)$$

Furthermore, we implement a deep supervision strategy by introducing supervision signals at different network layers, further improving the model's performance.

IV. Experiments

A. Experimental Setup

Datasets and Evaluation Metrics. In our experiment, we use 1000 training images from the DUT-LFSD dataset [37] as our training dataset. In our experiment, we do not use the ground-truth labels of this dataset, but use pixel-level pseudo-labels generated from the proposed method in this paper as supervisory signals to training the network. To further evaluate the performance of our proposed method, we conduct experiments on the following datasets: DUT-LFSD, LFSD [35], DUTLF-V2 [49], and PKU-LF [8]. We employ five commonly used evaluation metrics in LFSOD to evaluate the performance of different models, including mean F-measure (F_β^m) [13], Mean Absolute Error (MAE) [4], S-measure [50], E-measure [51].

Implementation Details. Our experiments were conducted on a single GTX 3090 GPU with the model implemented with the PyTorch framework. To be more specific, the backbone networks of our second stages are initialized with a pre-trained ResNet-50 using DINO [52]. The hyperparameter settings for the network in the first stage are the same as in C^2AM [40]. All the training images are resized to 224 × 224. We use the ASGD [53] optimizer, with a batch size of 4, an initial learning rate of 0.01. We optimize for 100 epochs, with a 10% learning rate decay every 5 epochs.

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

We primarily compare our method with 2D, 3D and 4D fully, weakly and unsupervised methods, including a traditional method (WSC [36]), 2D methods (PiCANe [46], MFNet [47], WSS [20], A2S-V1 [27], A2S-V2 [28]), 3D methods (UCNet [29], A2S-V2(RGBD) [28], DLM [31], DSU [30]), and 4D methods (MOLF [37], WSLF [12], NoiseLF [9]). Following the evaluation protocol of [54], we run these methods directly on the test set without re-training them.

Quantitative Evaluation. Table I presents the results of comparison with other approaches, excluding NoiseLF. It can be observed that our method achieves superior results, setting state-of-the-art scores across various evaluation metrics on the dataset.

We additionally compare our approach with other unsupervised methods in terms of precision-recall curves, as depicted in Figure 7. It is evident that the precision-recall curve coordinates of our method more closely approach (1, 1) compared to other methods, which again confirms the outstanding performance of our method.

To the best of our knowledge, NoiseLF was the pioneer in using noisy label for ULFSOD, while we seek to improve pseudo-labels by making the first attempt to employ a deep learning-based approach to higher-quality pseudo-labels. We compare our method with NoiseLF in Table II, and the results show that our proposed method significantly outperforms NoiseLF across several evaluation metrics on the DUT-LFSD and LFSD datasets.

TABLE II
Comparison with the Unsupervised 4D Method NOISELF.

Method	Supervision	DUT-LFSD		LFSD	
		$MAE \downarrow$	$F_\beta \uparrow$	$MAE \downarrow$	$F_\beta \uparrow$
NoiseLF [9]	U	0.148	0.689	0.152	0.723
Ours	U	**0.052**	**0.891**	**0.093**	**0.846**

Qualitative Evaluation. As shown in Figure 8, our method performs significantly better than the previous methods across various types of images. For example, in the first two rows, our method succeeds in localizing salient objects as well as their boundaries accurately, whereas the other methods either produce many false positives or fail to identify the precise shapes of some objects such as the two candies in the second row. The input images for the third to fifth rows have low contrast between the salient objects and the background. For the natural scenes and buildings in the last lines, the proposed method also demonstrates its robustness in processing images with various textures and complexities. Our method consistently performs well in these challenging cases while most of the other methods struggle with discriminating the salient objects from the visually similar background. DSU [30] can also produce reasonable results but has difficulty in finding precise object boundaries.

C. Ablation Study

To validate the effectiveness of the different components of our method, we conducted an ablation study on the DUT-LFSD dataset. We first established a simple baseline model by (1) using only C^2AM and DCRF for pseudo-label generation, and (2) training a variant of our SOD model without the MSSA and CFM modules. We then progressively added our proposed modules to the baseline model to analyze their contributions. In addition, we introduced two new variables: the type of pseudo-label used and the pre-trained weights employed for the backbone network. As shown in Table III, each of these components plays a crucial role in improving overall performance. Notably, our full pseudo-label generation method (PFSI) combined with the SOD baseline (without MSSA and CFM) already delivers strong results, demonstrating the effectiveness of our pseudo-label generation. Furthermore, integrating MSSA and CFM provides additional performance gains. The table also highlights the importance of choosing the appropriate backbone weights, with DINO-pretrained weights yielding the best results. We additionally experimented with MoCo and SimCLR as alternative pre-training approaches, observing that while these methods also enhanced performance, they were slightly outperformed by DINO, suggesting that the richer feature representations learned by DINO are particularly beneficial for our SOD task.

D. Ablation Study on the Input Modalities of PFSI

We investigate the importance of the RGB and depth inputs of PFSI by disabling each of them at a time and reporting

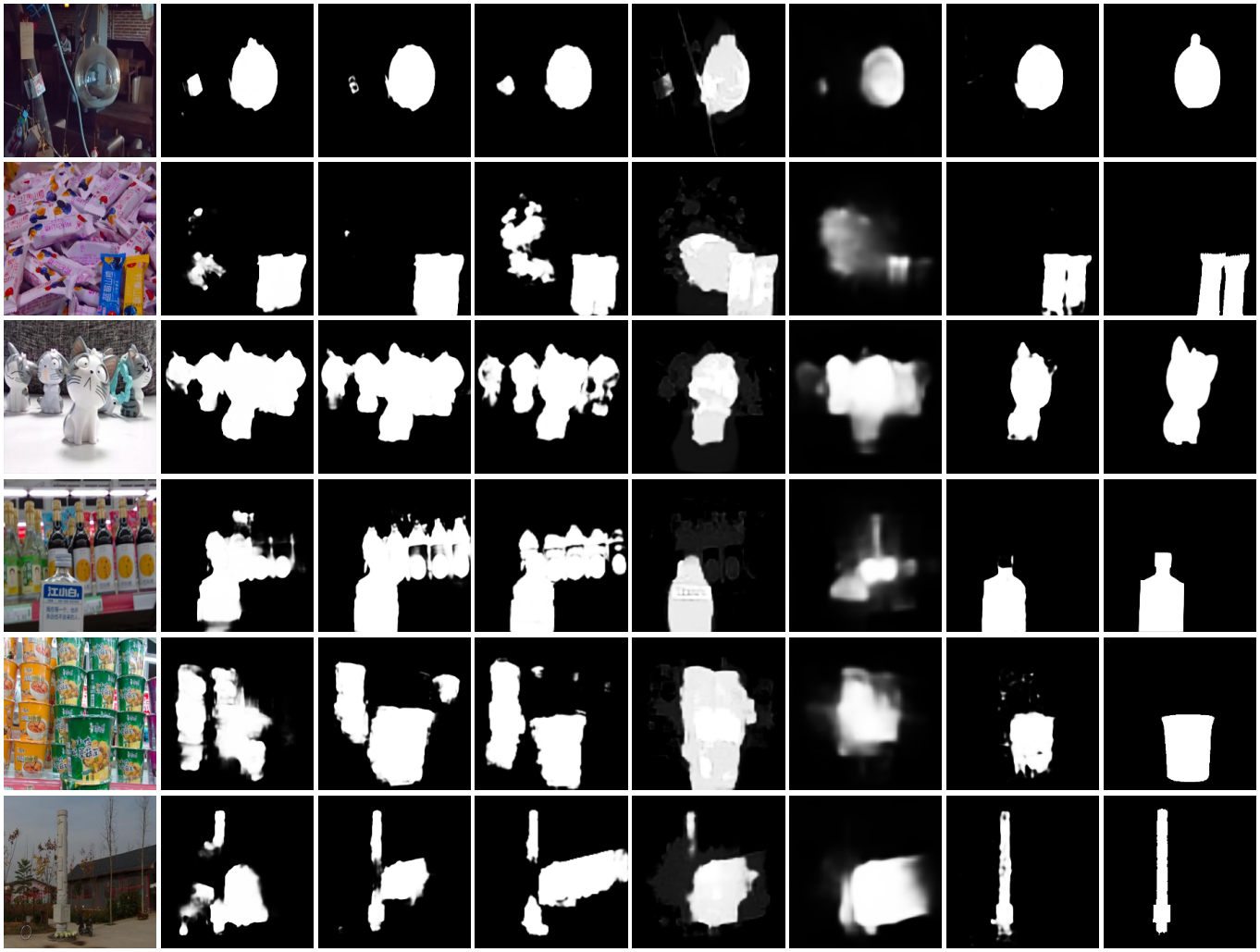


Fig. 8. Qualitative comparisons of our method with other methods.

TABLE III
ABLATION EXPERIMENTS ON DUT-LFSD.

PFSI	MSSA	CFM	Lable	Weight	$MAE \downarrow$	$F_\beta^m \uparrow$	$E_m \uparrow$	$S_m \uparrow$
×	×	×	Ours	DINO	0.080	0.831	0.900	0.831
✓	×	×	Ours	DINO	0.059	0.868	0.918	0.867
✓	✓	×	Ours	DINO	0.056	0.878	0.920	0.873
✓	✓	✓	MR	DINO	0.158	0.619	0.803	0.716
✓	✓	✓	SF	DINO	0.171	0.513	0.762	0.632
✓	✓	✓	wCtr	DINO	0.176	0.594	0.790	0.716
✓	✓	✓	Ours	MOCO	0.053	0.887	0.923	0.875
✓	✓	✓	Ours	SimCLR	0.064	0.864	0.908	0.858
✓	✓	✓	Ours	DINO	**0.052**	**0.889**	**0.924**	**0.877**

the results in Table IV. We can see that both input modalities are crucial to the performance of our method. And we have experimented with randomly selecting a focal slice from the input focal stack and using it as the input to the label generation stage. Figure 9 visually demonstrates that combining RGB and depth information improves focal slice selection, resulting in higher-quality pseudo-labels.

TABLE IV
ABLATION EXPERIMENT ON THE RGB AND DEPTH INPUTS OF PFSI ON DUT-LFSD.

	$MAE \downarrow$	$F_\beta^m \uparrow$	$E_m \uparrow$	$S_m \uparrow$
w/o Depth	0.120	0.762	0.818	0.762
w/o RGB	0.093	0.780	0.830	0.774
Random Selection	0.127	0.712	0.783	0.728
PFSI	**0.087**	**0.798**	**0.855**	**0.795**

E. Training Fully Supervised Methods Using Pseudo-labels.

To further demonstrate the validity of our model, we have trained LF-TransNet [54] and OBGNet [55] using the pseudo-labels generated by our method. As shown in Tables V and VI, although our model does not achieve the highest FPS, it consistently outperforms other methods in key metrics such as MAE and F_β^m. These results clearly demonstrate the superior accuracy and overall performance of our method on both the DUT-LFSD and DUTLF-V2 datasets.

TABLE V
COMPARISON OF DIFFERENT METHODS TRAINED ON OUR GENERATED PSEUDO-LABELS (DUT-LFSD).

Method	$FPS\uparrow$	$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$
LFTransNet [54]	18.12	0.053	0.879	**0.927**	0.876
OBGNet [55]	**34.16**	0.129	0.684	0.849	0.783
Ours	18.38	**0.052**	**0.889**	0.924	**0.877**

TABLE VI
COMPARISON OF DIFFERENT METHODS TRAINED ON OUR GENERATED PSEUDO-LABELS (DUTLF-V2).

Method	$FPS\uparrow$	$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$
LFTransNet [54]	18.96	0.074	0.753	0.866	0.804
OBGNet [55]	**36.76**	0.130	0.589	0.816	0.740
Ours	20.07	**0.061**	**0.796**	**0.888**	**0.826**

F. Label Quality Evaluation

We also evaluate the quality of our generated labels by treating the pseudo-labels output by the first stage as the final predictions, which are compared with the ground truth. The results on DUT-LFSD are reported in Table VII. It can be seen that our generated labels are of greater quality compared to those by existing hand-crafted label generation methods, having high degree of similarity to the ground truth. Furthermore, our proposed PFSI is crucial to label quality.

TABLE VII
QUANTITATIVE EVALUATION OF DIFFERENT LABEL GENERATION METHODS ON DUT-LFSD.

	$MAE\downarrow$	$F_\beta^m\uparrow$	$E_m\uparrow$	$S_m\uparrow$
MR [3]	0.213	0.466	0.750	0.646
SF [4]	0.226	0.362	0.671	0.502
wCtr [5]	0.229	0.472	0.743	0.630
Ours w/o PFSI	0.121	0.739	0.825	0.754
Ours	**0.087**	**0.798**	**0.855**	**0.795**

V. LIMITATION AND DISCUSSION

Despite these promising results, our approach has limitation as a light-field saliency method compared to RGB saliency methods. The computational complexity of processing large-scale light field data can be time-consuming. However, this challenge is common to all light-field saliency methods. To address this challenge, several directions could be explored. One approach is to develop a light-field pre-processing framework that efficiently handles light-field data prior to the main processing stages. Another potential solution involves designing lightweight feature fusion structures to enhance the overall efficiency of light-field salient object detection. Last but not the least, recent research in implicit representation [56], [57] encodes images into a latent space and then processes and fuses multi-modality features in the frequency domain [58]. Based on this research, a promising direction might be to fuse the abundant light-field features with RGB features in the latent space rather than in pixel space.

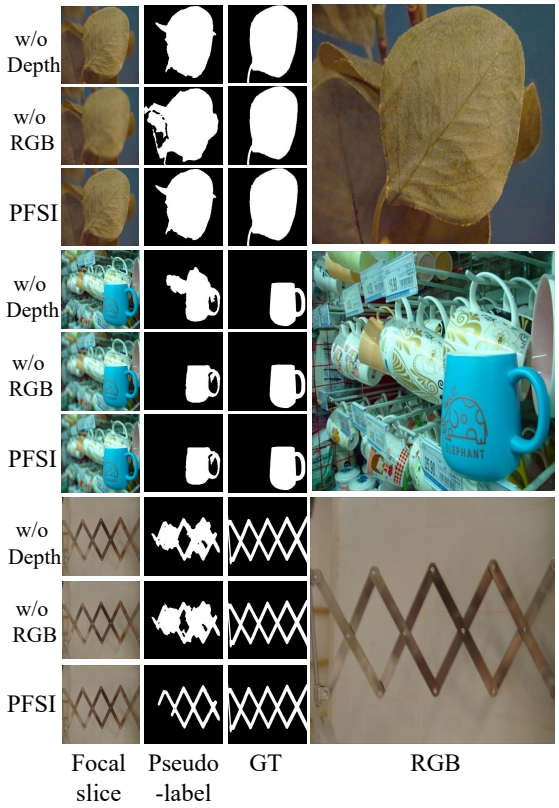


Fig. 9. The importance of the RGB and depth inputs of PFSI.

VI. CONCLUSION

In this paper, we propose a novel deep learning-based framework for unsupervised pseudo-label generation for LFSOD. Following a label synthesis scheme, we propose an unsupervised framework for pseudo-label generation that learns to select the optimal focal stack to maximize the quality of generated labels. To further boost performance, we propose an improved LFSOD model, with a multi-scale spatial attention module to learn multi-scale and contextual features for focal stack and a cross fusion module for deep fusion of multi-modal features. Experimental results demonstrate that our approach outperforms existing unsupervised methods.

REFERENCES

[1] J. Zhang, Q. Liang, and Y. Shi, "Kd-scfnet: Towards more accurate and efficient salient object detection via knowledge distillation," *arXiv preprint arXiv:2208.02178*, 2022.
[2] M. Ma, C. Xia, C. Xie, X. Chen, and J. Li, "Receptive field broadening and boosting for salient object detection," *arXiv preprint arXiv:2110.07859*, 2021.
[3] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2013, pp. 3166–3173.
[4] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 2012, pp. 733–740.
[5] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2014, pp. 2814–2821.

[6] S. Ma, L. Zhu, X. Chen, X. Yan, S. Wang, P. Yang, and B. Xu, "Arfnet: Attention-oriented refinement and fusion network for light field salient object detection," *IEEE Systems Journal*, vol. 16, no. 4, pp. 5950–5961, 2022.

[7] X. Wang, S. Chen, G. Wei, and J. Liu, "Tenet: Accurate light-field salient object detection with a transformer embedding network," *Image and Vision Computing*, vol. 129, p. 104595, 2023.

[8] W. Gao, S. Fan, G. Li, and W. Lin, "A thorough benchmark and a new model for light field saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[9] M. Feng, K. Liu, L. Zhang, H. Yu, Y. Wang, and A. Mian, "Learning from pixel-level noisy label: A new perspective for light field saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1756–1766.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning requires rethinking generalization." in *ICLR*, 2017.

[11] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 13 713–13 722.

[12] Z. Liang, P. Wang, K. Xu, P. Zhang, and R. W. Lau, "Weakly-supervised salient object detection on light fields," *IEEE Transactions on Image Processing*, vol. 31, pp. 6295–6305, 2022.

[13] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection, 2009," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1597–1604.

[14] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2014.

[15] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 3239–3259, 2022.

[16] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, no. 1, 2018.

[17] X. Zheng, X. Tan, J. Zhou, L. Ma, and R. W. Lau, "Weakly-supervised saliency detection via salient object subitizing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 11, pp. 4370–4380, 2021.

[18] Y. Liu, P. Wang, Y. Cao, Z. Liang, and R. W. Lau, "Weakly-supervised salient object detection with saliency bounding boxes," *IEEE Transactions on Image Processing*, vol. 30, pp. 4423–4435, 2021.

[19] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 6074–6083.

[20] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 12 546–12 555.

[21] B. Xu, H. Liang, W. Gong, R. Liang, and P. Chen, "A visual representation-guided framework with global affinity for weakly supervised salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 248–259, 2024.

[22] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "Rgbt salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 12, pp. 4421–4433, 2020.

[23] B. Tang, Z. Liu, Y. Tan, and Q. He, "Hrtransnet: Hrformer-driven two-modality salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 728–742, 2023.

[24] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1913–1927, 2020.

[25] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5961–5970.

[26] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 1448–1457.

[27] H. Zhou, P. Chen, L. Yang, X. Xie, and J. Lai, "Activation to saliency: Forming high-quality labels for unsupervised salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 2, pp. 743–755, 2022.

[28] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 7257–7267.

[29] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8582–8591.

[30] W. Ji, J. Li, Q. Bi, C. Guo, J. Liu, and L. Cheng, "Promoting saliency from depth: Deep unsupervised RGB-d saliency detection," in *International Conference on Learning Representations*, 2022.

[31] T. Yang, Y. Wang, L. Zhang, J. Qi, and H. Lu, "Depth-inspired label mining for unsupervised rgb-d salient object detection," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5669–5677.

[32] A. Li, Y. Mao, J. Zhang, and Y. Dai, "Mutual information regularization for weakly-supervised rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 1, pp. 397–410, 2024.

[33] Y. Xu, X. Yu, J. Zhang, L. Zhu, and D. Wang, "Weakly supervised rgb-d salient object detection with prediction consistency training and active scribble boosting," *IEEE Transactions on Image Processing*, vol. 31, pp. 2148–2161, 2022.

[34] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, vol. 7, pp. 37–69, 2021.

[35] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2014, pp. 2806–2813.

[36] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2015, pp. 5216–5223.

[37] M. Zhang, J. Li, J. Wei, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[38] Y. Piao, Y. Jiang, M. Zhang, J. Wang, and H. Lu, "Panet: Patch-aware network for light field salient object detection," *IEEE transactions on cybernetics*, 2021.

[39] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep learning for light field saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8838–8848.

[40] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen, "C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 989–998.

[41] N. Araslanov and S. Roth, "Single-stage semantic segmentation from image labels," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4253–4262.

[42] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[43] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[45] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[46] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2018, pp. 3089–3098.

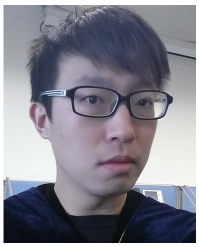
[47] Y. Piao, J. Wang, M. Zhang, and H. Lu, "Mfnet: Multi-filter directive network for weakly supervised salient object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 4136–4145.

[48] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7479–7489.

[49] Y. Piao, Z. Rong, S. Xu, M. Zhang, and H. Lu, "Dut-lfsaliency: Versatile dataset and light field-to-rgb saliency detection," *arXiv preprint arXiv:2012.15124*, 2020.
[50] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2017, pp. 4548–4557.
[51] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.
[52] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 9650–9660.
[53] R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," *Advances in neural information processing systems*, vol. 26, 2013.
[54] Z. Liu, Q. He, L. Wang, X. Fang, and B. Tang, "Lftransnet: Light field salient object detection via a learnable weight descriptor," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.
[55] D. Jing, S. Zhang, R. Cong, and Y. Lin, "Occlusion-aware bi-directional guided network for light field salient object detection," in *Proceedings of ACM International Conference on Multimedia*, 2021, p. 1692–1701.
[56] Y. Chen, S. Liu, and X. Wang, "Learning continuous image representation with local implicit image function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8628–8638.
[57] J. Lee and K. H. Jin, "Local texture estimator for implicit representation function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1929–1938.
[58] H. Zhang, J. Guo, J. Zhang, H. Qin, Z. Feng, M. Yang, and Y. Guo, "Deep fourier-based arbitrary-scale super-resolution for real-time rendering," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

YANFENG ZHENG was born in Liaoning, China,in 1998. He received the B.E. degree in Information Management and Information System in Dongbei University Of Finance And Economics. He is currently pursuing the master's degree in Computer Science and Technology with Dalian Minzu University. His primary research interests include computer vision and Salient Object Detection.

Zhong Luo was born in Liaoning, China, in 1998. He obtained a Bachelor's degree in engineering from Shenyang Normal University. He is currently pursuing the master's degree in Computer Science and Technology with Dalian Minzu University. His primary research interests include computer vision and Salient Object Detection.



Ying Cao received the B.Eng. and M.Sc. degrees in Software Engineering from Northeastern University, China, and a Ph.D. in Computer Science from the City University of Hong Kong. He is an assistant professor at ShanghaiTech University. His general research interests include computer graphics and computer vision, with a special interest in data-driven graphic design.

Xiaosong Yang is currently a Professor, Deputy Head of Department, Programme Leader of MSc AIM at the National Centre for Computer Animation, Bournemouth University. He is a member of the Peer Review College of the Arts and Humanities Research Council (AHRC) UK and the Computer Science Evaluation Group of the Natural Science and Engineering Research Council of Canada (NSERC). He is a member of the Program Committee for several international conferences (SIGGRAPH Asia, CGI, CASA etc), conference program chairs (CGI2012, CASA2020/2022), journal reviewer for many top journals (TVCG, ACM ToG, C&G, Signal Processing, Pattern Recognition, Neurocomputing, IEEE Access) and conferences (SIGGRAPH, Eurographics, ISMAR, PG, CGI etc). He has given several invited talks and keynote presentations internationally.

Weiwei Xu received the BS and master's degrees in computer science from Hohai university in 1996 and 1999, respectively, and the PhD degree in computer graphics from Zhejiang university. He is currently a professor with the state key laboratory of CAD&CG and the college of computer science and technology, Zhejiang university. He was awarded as Chang-Jiang scholar of the Ministry of Education of China in 2022. He was a distinguished professor of Qian-Jiang scholars of Zhejiang province in Hangzhou Normal university and a researcher in the internet graphics group with Microsoft Research Asia from 2005 to 2012. Before that, he was a post-doc researcher with Ritsumeikan university in Japan for over one year. His research interests include 3D reconstruction, physical simulation, 3D printing, and deep learning.

Zheng Lin is currently a postdoctoral researcher at Tsinghua University, under the supervision of Prof. Shi-Min Hu. He get the doctoral degree from Nankai University in 2023, under the supervision of Prof. Ming-Ming Cheng. His research interests includes computer vision and computer graphics.

Nan Yin is currently a postdoctoral fellow at Mohamed bin Zayed University of Artificial Intelligence, UAE. He received the B.S. degree from National University of Defense Technology, Changsha, China, in 2016 and the Ph.D. degree in School of Computer Science and Technology, National University of Defense Technology, Changsha, China. His current research interests includes transfer learning and graphs.

Pengjie Wang received his Ph.D. degree in computer science from Zhejiang University. He is currently a professor with School of Computer Science, Dalian Minzu University, Dalian, China. He is also with the National Centre for Computer Animation, Faculty of Media and Communication, Bournemouth University, U.K. His research interests include computer vision and computer graphics.