We greatly appreciate the time and effort each of the reviewers have dedicated to providing insightful feedback on ways to strengthen our paper. Thus, it is with great pleasure that we make a point-by-point response to all comments.

To Reviewer1: **W1.1**: Thank you for your suggestion. We agree that a direct comparison between 12 RGB input-frames and 12 focal slices could highlight our spatial fusion module. 12 RGB input-frames containing the same information do not enable the SFM to learn the spatial correlation as well as the focal slices do. This is consistent with the results shown in the table below. We compare the results using two different inputs in terms of S-measure, F-measure and MAE. The results confirm the effectiveness of focusness information and our spatial fusion module. **W1.2**: We totally agree with you on that point. In our case, the 2D dataset (DUTS/MSRA10K) is 10 times larger than the proposed dataset. It is well understood that training dataset size is the most important factor determining the generalization ability regardless of model complexity. Based on this observation, we think it would be fairer to train different models on the same training dataset. Thus, it could provide an unbiased estimation for them tested on other datasets. The following table shows the results, tested on another dataset (LSFD), from the models we have retrained on our dataset. It is shown that the retrained models on our dataset do not perform as well as they are pre-trained on the much larger 2D dataset. This is the reason we choose more generalized models to compare with ours.

W2: Thank you for providing us with the constructive comments. In this paper, we adopted a simple model in Table 1(c) (concatenation followed by a 1×1 convolution layer) to replace the Mo-SFM for feature fusion. Experimental results indicates that the proposed Mo-SFM has superiority in learning spatial correlation of different features and performs better.

W3: We are sorry that the notation and display of results made the paper less readable than intended. We will carefully improve them in the final version according to your suggestion.

	Ours		PDNet[65]		DHS[35]		NLDF[37]		UCF[61]		Amulet[60]		R^3Net[11]	
Score on LFSD[29]	focal slices	RGB*12	w/o	with	w/o	with	w/o	with	w/o	with	w/o	with	w/o	with
$S_{measure}\uparrow$	0.830	0.757	0.786	0.746	0.770	0.741	0.745	0.731	0.762	0.748	0.773	0.757	0.789	0.769
$F_{measure}\uparrow$	0.819	0.740	0.780	0.734	0.761	0.732	0.748	0.720	0.710	0.702	0.757	0.738	0.781	0.759
$MAE\downarrow$	0.089	0.140	0.116	0.136	0.133	0.142	0.138	0.137	0.169	0.174	0.135	0.147	0.128	0.137

$*$ 'w/o' means no-retrain results, and 'with' means retrain results.

To Reviewer2: **About the originality:** [29] as the pioneering work is a traditional method in which various hand-crafted features and prior knowledges are used for light field saliency detection. However, deep-learning-based light field methods have been missing from contemporary studies in saliency detection. We first introduce the CNN framework for light field SOD and provide the largest light field dataset. The overall architecture of our paper differs substantially from [29] and other light field saliency detection methods. As to the proposed modules, the Mo-SFM resembles the memory mechanism to fuse information by emphasizing useful features and learning the spatial correlation between those light field features. This is the first attempt in light field SOD. Furthermore, different from other integration mechanisms for SOD, such as skip-connections [35,36], short-connections [21], and residual connections [11], our Mo-FIM is designed in a way high-level information is summarized as memory to guide low-level feature selection. To our best knowledge, this feature integration has never been explored in previous studies.

About the limitations: We appreciate the opportunity to include additional explanation about the limitation of our paper. First, our dataset is not big enough compared with other 2D datasets. This may limit the generalization ability of models training on it. Second, the use of the focal stack in the training process does require higher-end hardware devices or is a bit more time-consuming. We are currently working on these two limitations.

About the use of domains: We believe that there is a large space to explore our method in different fields, because Mo-SFM can be used to effectively fuse sequent features, and Mo-FIM can be applied to integrate multi-layer features for dense prediction tasks. However, deep learning based light field SOD is still in the initial stage. In the experiment, our main focus in this work lies on demonstrating the rationality and feasibility of the method. For the hyper spectral images, we suppose that our model can not be directly used on hyper spectral images due to different data types.

To Reviewer3: Thank you for the positive and constructive comments. We concur with you that a detailed description of our dataset would be important for NeurIPS readers and further strengthen the paper. To build this dataset, we first capture 3894 images under different lighting conditions in various environments around our daily life, e.g., parks, campus, streets and supermarkets and so on. Then we screen out blurred images or low-quality images. Acquisition of our ground truths follows annotation principle in the widely used saliency datasets, e.g., LFSD and DUTS. Because determining salient objects is considered highly subjective, three annotators are required to determine the salient objects in order to achieve annotation consensus. A free image annotation tool-GIMP is used manually to annotate the salient objects in a pixel-wise manner. Then 1462 light fields are selected as the final dataset, consisting of 900 indoor and 562 outdoor scenes. Moreover, it contains challenging scenes, e.g., multiple objects (95), transparent objects (28), low-contrast or similar foreground and background (108), low-intensity environment (9), and clutter background (31).

For the minor issues, we are sorry that they are not clearly described here. To be specific, we use the output of $Block5$ as the initial hidden state of ConvLSTM, i.e., $h_0 = F_5$. In the multi-level feature integration stage, h_{t-1} is considered as historical memory to learn a channel attention and update current light field features in SCIM. We will clarify the minor issues and incorporate a more clear description of our dataset in the final version.

Again, thank you all for giving us the opportunity to strengthen our paper with your valuable comments and queries. We will make the code and dataset publicly available. Hopefully this would maximize the contribution to the community.