A Thorough Benchmark and a New Model for Light Field Saliency Detection

Wei Gao, *Senior Member, IEEE*, Songlin Fan, Ge Li, *Member, IEEE*, and Weisi Lin, *Fellow, IEEE*

***Abstract*—Compared with current RGB or RGB-D saliency detection datasets, those for light field saliency detection often suffer from many defects, e.g., insufficient data amount and diversity, incomplete data formats, and rough annotations, thus impeding the prosperity of this field. To settle these issues, we elaborately build a large-scale light field dataset, dubbed *PKU-LF*, comprising 5,000 light fields and covering diverse indoor and outdoor scenes. Our PKU-LF provides all-inclusive representation formats of light fields and offers a unified platform for comparing algorithms utilizing different input formats. For sparking new vitality in saliency detection tasks, we present many unexplored scenarios (such as underwater and high-resolution scenes) and the richest annotations (such as scribble annotations, bounding boxes, object-/instance-level annotations, and edge annotations), on which many potential attention modeling tasks can be investigated. To facilitate the development of saliency detection, we systematically evaluate and analyze 16 representative 2D, 3D, and 4D methods on four existing datasets and the proposed dataset, furnishing a thorough benchmark. Furthermore, tailored to the distinct structural characteristics of light fields, a novel symmetric two-stream architecture (*STSA*) network is proposed to predict the saliency of light fields more accurately. Specifically, our *STSA* incorporates a focalness interweavement module (FIM) and three partial decoder modules (PDM). The former is designed to efficiently establish long-range dependencies across focal slices, while the latter aims to effectively aggregate the extracted coadjutant features in a mutual-enhancement way. Extensive experiments demonstrate that our method can significantly outperform the competitors.**

***Index Terms*—Benchmark, focal stack, light field, salient object detection.**

I. Introduction

SALIENT object detection (SOD) [1], [2], [3], [4], [5], [6] mimics the human visual system to pop visually attractive objects out and has been widely applied in many vision tasks [7]. Regarding the inputs, the research community tends to categorize SOD tasks into 2D (RGB) [8], [9], [10], [11], [12], [13], [14], 3D (RGB-D) [1], [5], [15], [16], [17], and 4D (light field) [18], [19], [20], [21], [22], [23]. As an emerging media, light fields are increasingly popular for their powerful capability to reconstruct virtual reality and 3D immersive scenes. However, the visual attention modeling for light fields is immature compared with RGB or RGB-D images. As shown in Fig. 1, while light fields contain much more spatial information than the alternatives, their potential in SOD is not fully exploited.

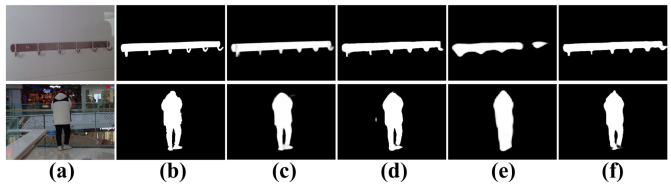


Fig. 1. Visual results from high-ranking 2D, 3D, and 4D methods. (a) All-focus images. (b) Ground truths. (c) Results from the 2D method GCPA [14]. (d) Results from the 3D method ATSA [16]. (e) and (f) Results from 4D methods ERNet [23] and ours. Though 2D and 3D images can be derived from 4D light fields, and light fields provide more spatial information, previous high-ranking 4D methods still perform worse than corresponding 2D or 3D cases. The proposed method can unleash the potential of light fields in SOD and achieve the best accuracy.

As convolutional neural networks (CNNs) revolutionize the 2D and 3D SOD fields, the lack of a high-quality light field dataset becomes a troublesome obstacle hindering the application of CNNs in 4D SOD. By simplifying the 7-dimensional plenoptic function, light fields parameterize spatial light rays by two parallel planes $L(u, v, x, y)$, thus whose directions can be recorded. As shown in Fig. 2(a), (u, v) and (x, y) encode the angular and spatial variation, respectively. Since the collected raw light data is obscure for analysis, post-processing [7] for data conversion often occurs. One can sample certain angular values (u^*, v^*) or spatial values (x^*, y^*) (see Fig. 2(b)) to obtain sub-aperture images and micro-lens images, respectively. Others may decode raw data into stacks of focal slices (focal stacks), all-focus images, or depth maps via digital refocusing techniques [24]. Focal slices share the viewpoint but with different focused depths (pixels), while all-focus images compose every focused pixel in focal slices. Thus, the acquisition of light field data often necessitates a vast workload. As summarized in Table I, there are four light field datasets, i.e., LFSD [18], HFUT [25], DUT-LF [20], and Lytro Illum [22]. But all of them

Manuscript received 4 October 2021; revised 15 May 2022; accepted 29 December 2022. Date of publication 9 January 2023; date of current version 5 June 2023. This work was supported in part by the Natural Science Foundation of China under Grants 62271013 and 62031013, in part by Guangdong Basic and Applied Basic Research Foundation under Grant 2019A1515012031, in part by Shenzhen Fundamental Research Program under Grant GXWD20201231165807007-20200806163656003, in part by Shenzhen Science and Technology Plan Basic Research Project under Grant JCYJ20190808161805519, and in part by The Major Key Project of PCL under Grant PCL2021A06. Recommended for acceptance by G. Pons-Moll. *(Corresponding author: Wei Gao.)*

Wei Gao, Songlin Fan, and Ge Li are with the School of Electronic and Computer Engineering, Peking University, Shenzhen 518055, China, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: gaowei262@pku.edu.cn; slfan@pku.edu.cn; geli@pku.edu.cn).

Weisi Lin is with the School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798 (e-mail: wslin@ntu.edu.sg).

Our dataset will be available at https://openi.pcl.ac.cn/OpenDatasets.

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TPAMI.2023.3235415.

Digital Object Identifier 10.1109/TPAMI.2023.3235415

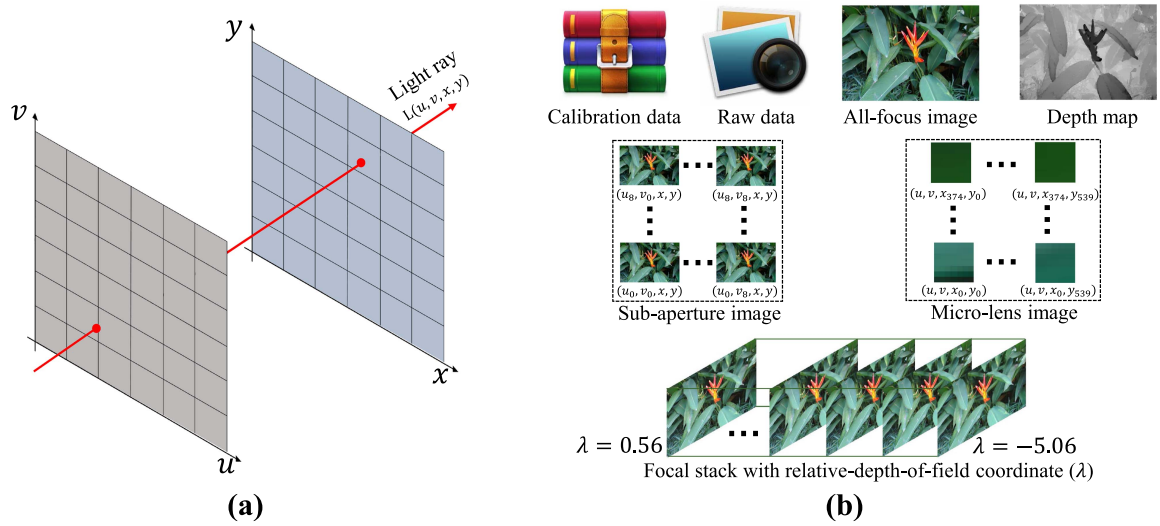


Fig. 2. Illustration of light fields. (a) Theory of light fields where a light ray is represented by the two-plane parameterization $L(u, v, x, y)$. (b) Light field data in our PKU-LF including the calibration data of our camera, raw light field data, all-focus images, depth maps, sub-aperture images, micro-lens images, and focal stacks with relative-depth-of-field coordinates (λ) [24]. Please refer to [7] for a detailed description of these light field representations.

TABLE I
SUMMARY OF FOUR EXISTING DATASETS AND THE PROPOSED PKU-LF DATASET

Dataset	Year	#Scale	Resulution	Light Field								Annotation					Device
				Cal.	Raw.	Foc.	Dep.	Sub.	Mic.	Rel.	Spl.	Scr.	Bou.	Obj.	Ins.	Edg.	
LFSD [18]	2014	100	360 × 360		✓	✓	✓							✓			Lytro
HFUT [25]	2017	255	328 × 328			✓	✓	✓						✓			Lytro
DUT-LF [20]	2019	1,462	600 × 400			✓	✓				✓			✓			Lytro Illum
Lytro Illum [22]	2019	640	400 × 590		✓				✓					✓			Lytro Illum
PKU-LF	2021	5,000	2022 × 1404	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	Lytro Illum

Cal.: Calibration data of the camera. Raw.: Raw light fields. Foc.: Focal stacks. Dep.: Depth maps. Sub.: Sub-aperture images. Mic.: Micro-lens images. Rel.: Relative-depth-of-field coordinates. Spl.: Official training/testing set split. Scr.: Scribble annotations. Bou.: Bounding boxes. Obj.: Object-level annot tions. Ins.: Instance-level annotations. Edg.: Edge annotations.

have detrimental shortcomings: *i) Insufficient data amount and diversity.* Compared to the RGB dataset [26] with over 10,000 images, the total sample amount of these four light field datasets is merely 2,457. Besides, the samples in existing datasets mainly involve specialized or artificially constructed scenes. The limited data amount and diversity of existing datasets restrict the generalization of algorithms in real-world scenes and lead to overfitting. *ii) Incomplete data formats.* The four existing datasets only provide parts of light field formats with significant inconsistency. For instance, HFUT contains sub-aperture images while those in DUT-LF are unavailable, making it inconvenient to evaluate algorithms that use different light field formats. *iii) Rough images and annotations.* The two early datasets, LFSD and HFUT, are established using the first generation Lytro camera [24] in poor collection conditions. Besides, the samples in existing datasets have a relatively low resolution (less than 600×400), on which the annotations are labeled. Therefore, the structural details of salient objects are lost, which is not conducive to utilization. The research community desperately demands a well-established light field dataset.

As depicted in Fig. 2(b), distinct from traditional RGB or RGB-D images, light fields record spatial geometry information in the form of image arrays. Thus, light field processing requires us to develop more efficient feature extraction and aggregation approaches according to the structural characteristics of light fields. Previous works [20], [21], [23] tend to employ ConvLSTM [27] to process features from all-focus images and focal stacks. However, ConvLSTM is initially proposed for video sequence modeling. Ignoring the unique structural characteristics of light fields will result in not only a low detection accuracy but also a slow inference speed (see Section V). This can also explain Fig. 1, where current 2D and 3D SOD methods outperform 4D ones while light fields provide more spatial information.

Overall, we present four contributions to address the issues above. *First*, we carefully construct PKU-LF, a high-quality light field SOD dataset. Compared with previous datasets, ours has the following merits:

- Our dataset comprises 5,000 light fields, covering over one hundred object categories. The entire data collection phase spans about one year and covers a wide range of real-world scenes from dozens of urban and rural areas.
- Our dataset includes many challenging scenes absent from existing datasets (see Fig. 3). For example, it is the first time to explore SOD in underwater scenes.

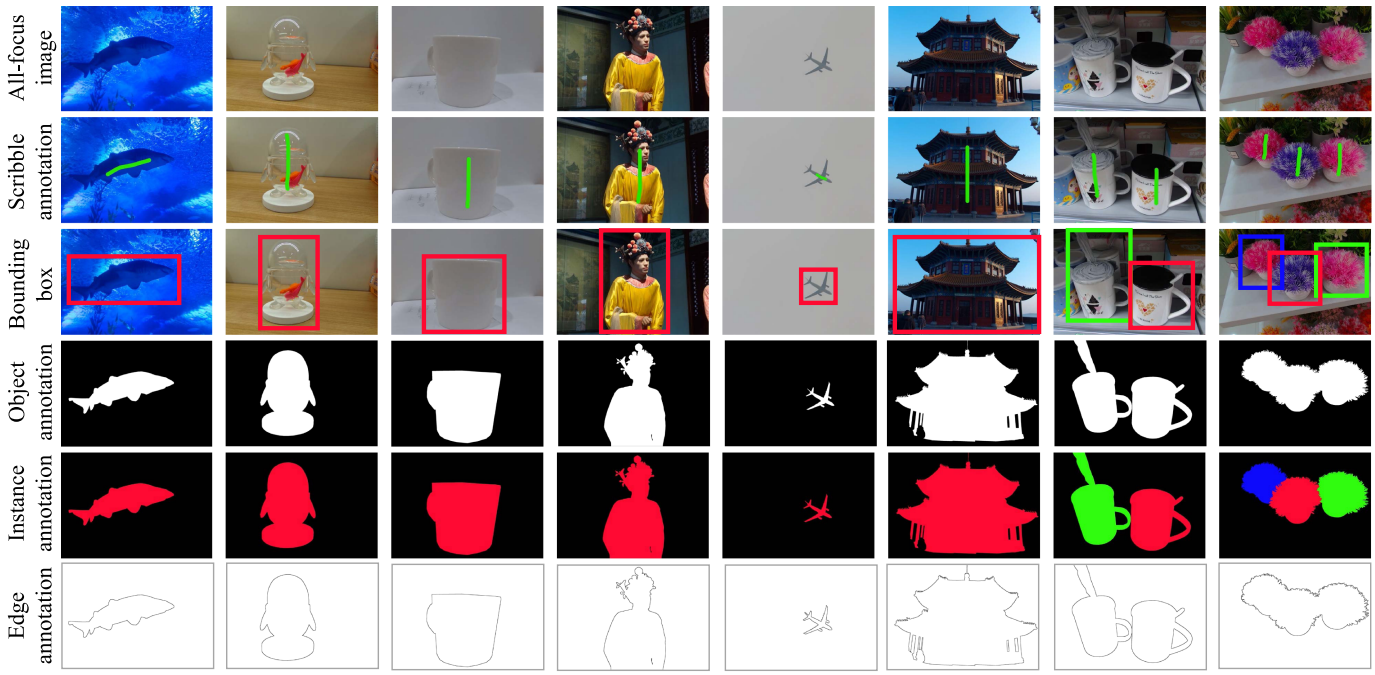


Fig. 3. Rich annotations of the proposed PKU-LF dataset including scribble annotations, bounding boxes, object-/instance-level annotations, and edge annotations. It is worth mentioning that the samples also indicate the complexity of our dataset, e.g., underwater scenes (1st column), transparent objects (2nd column), similar foregrounds and backgrounds (3rd column), low illumination (4th column), extremely small/large objects (5th and 6th columns), and multiple objects (7th and 8th columns).

- Our dataset has complete light field formats (see Table I), offering a unified platform for studying SOD methods with various inputs. Besides, our high-resolution images ensure the labeling quality of our annotations.
- Our dataset consists of rich annotations (see Fig. 3) that are expected to catalyze the advance of many new tasks, e.g., weakly supervised SOD, high-resolution SOD, salient instance detection, to name a few.

Second, to tap the superiority of light fields over RGB and RGB-D images, we develop a novel symmetric two-stream architecture (*STSA*) network, whose designs are tailored to the distinct structural characteristics of light fields. Concretely, our *STSA* incorporates a focalness interweavement module (FIM) and three partial decoder modules (PDM). Based on the unique structural characteristics of focal stacks where salient objects in different focal slices are spatially aligned, the proposed FIM adopts an efficient local interweavement operation to establish long-range dependencies across focal slices. Compared with the ConvLSTM in previous works, our FIM significantly improves the ability and efficiency of extracting the spatial geometry information of light fields. After extracting the task-benefited features of light fields, we further develop the PDM to fuse and decode the extracted coadjutant multi-modal and multi-scale features. Different from our previous works [28], [29], because there exist homogeneous and heterogeneous components in multi-modal features, as well as multi-scale features, our PDM adopts a mutual-enhancement strategy to emphasize the homogeneous components and suppress the heterogeneous ones. Convincing ablation experiments show that the proposed two modules are very effective.

Third, we demonstrate the performance superiority of 4D SOD over the 2D and 3D cases by achieving a new state-of-the-art performance, revealing that providing more spatial information is indeed beneficial for accurately detecting salient objects, especially for some complex scenes. *Finally*, we conduct a thorough evaluation and an in-depth analysis of 16 representative 2D, 3D, and 4D methods on four existing and our proposed datasets, paving the way for later researchers interested in this field.

Our four contributions constitute an entirety to facilitate the progress of light field saliency detection. The rest paper is organized as follows. Section II reviews the related work on existing light field datasets and representative 2D, 3D, and 4D saliency detection methods. Section III introduces the construction and statistics of the proposed high-quality light field dataset. Our *STSA* is described in Section IV. Extensive experiments are conducted in Section V to verify the effectiveness of the proposed method. Finally, we draw the conclusion in Section VI.

II. Related Work

In this section, we first perform a review of existing light field datasets. Then, the related 2D, 3D, and 4D saliency detection methods are discussed.

A. Light Field SOD Datasets

Table I lists the four existing light field datasets proposed for SOD tasks, i.e., LFSD [18], HFUT [25], DUT-LF [20], and Lytro Illum [22]. LFSD is the first light field dataset that consists of 60 indoor and 40 outdoor light fields with 360×360 spatial

resolution. Each light field is decoded into an all-focus image, a focal stack, and a depth map. Most images in this dataset have a single object with simple structures, and the salient objects usually appear at the image center. The limitation on diversity and complexity of this dataset causes its low generalization. Zhang et al. [25] construct the HFUT light field dataset that contains 255 light fields. Compared with LFSD, all-focus images, focal stacks, sub-aperture images, and depth maps are available. However, this dataset introduces color distortion [30] in the decoding phase. Apart from the deficient data amount, both LFSD and HFUT are established via the first generation Lytro camera [24], whose configuration is lagging. DUT-LF and Lytro Illum are collected by the advanced Lytro Illum camera. Lytro Illum has improved light field sensors and can thus capture larger spatial and angular resolution images compared with the first generation camera. DUT-LF, the largest light field SOD dataset, includes 1,462 light fields with 600×400 spatial resolution. The authors split them into 1,000 training samples and 462 testing samples. Nevertheless, this dataset does not provide the sub-aperture images and micro-lens images. It contains lots of specialized and artificially constructed scenes, restricting the generalization in real-world scenes. Lytro Illum incorporates 640 high-quality light fields, including several complex scenes, e.g., small objects, cluttered backgrounds, and texture variation. However, the common focal stack is absent from this dataset, and its data amount is also a severe limitation. The four existing datasets merely provide low-resolution (less than 600×400) images on which corresponding annotations are labeled. Consequently, their annotations lose many structural details of objects, eventually decreasing the performance of models.

B. Salient Object Detection

Since 4D SOD is the extension of 2D and 3D SOD, 2D and 3D SOD methods can also be applied to detect salient objects in light fields. The following paragraphs would like to review the recent advances in light field saliency detection, as well as those for 2D and 3D SOD.

2D SOD. Early 2D saliency detection methods [8], [9], [31], [32], [33], [34], [35], [36], [37], [38] mainly focus on hand-crafted priors, such as color contrasts and locations, with very limited feature expression. Recently, benefitting from the powerful feature extraction capability of neural networks, these traditional methods have been gradually surpassed. Zhao et al. [39] employ two independent CNNs to extract both local and global context information. Some works [40], [41] introduce the fully convolutional network (FCN) to promote the accuracy of SOD. A simple pooling-based network with a U shape architecture is designed by Liu et al. [42], which can reduce the aliasing effect of upsampling, especially with a sizeable upsampling rate. Chen et al. [14] propose a model that can integrate multi-level and multi-scale features and obtain satisfactory saliency maps. Inspired by mutual learning [43], aggregate interaction modules and self-interaction modules are utilized by Pang et al. [13] to fuse multi-level and multi-scale features and finally infer high-quality saliency maps.

3D SOD. Despite the gratifying achievements of 2D SOD, it still cannot make correct predictions when encountering complex scenes. Thanks to the advancement of 3D imaging technologies, depth maps are available and utilized by [44], [45], [46], [47], [48], [49] as complementary cues for saliency detection. Huang et al. [50] adopt an end-to-end model and a practical loss function to detect salient objects. Their proposed model achieves a visible performance promotion over the methods based on hand-crafted features. Chen et al. [51] aggregate different modality information through a multi-scale and multi-path network. A complementary-aware RGB-D saliency detection model developed by Chen et al. [52] uses a complementary-aware fusion block to integrate features from the same stage of each modality and conduct saliency detection. Chen et al. [53] design a novel cross-level combination block for multi-modality fusion, while Piao et al. [54] present a depth-induced multi-scale RGB-D saliency detection network. Zhang et al. [1] imitate the dataset labeling process to generate multiple saliency maps. These saliency maps are then aggregated into an accurate saliency map through the consensus process. Lately, we observed Fan et al. [17] and Zhang [16] identify the visually interesting region through a bifurcated backbone strategy and an asymmetric two-stream architecture, respectively.

4D SOD. At present, light field SOD is still at an early stage. Directly applying the above 2D and 3D SOD methods cannot effectively exploit the structural information in light fields. Hence, properly devising efficient mechanisms for light field saliency detection has attracted much attention. Li et al. [18] are the first to detect salient objects using light fields. A weighted sparse coding framework is proposed by Li et al. [55] to handle 2D, 3D, and 4D SOD problems simultaneously. Zhang et al. [56] calculate a contrast saliency map, then use background priors to eliminate the background distraction and obtain the final result. Recently, a depth-induced cellular automata is proposed by Piao et al. [57] for 4D SOD. In addition to the above traditional methods, Piao et al. [58] first attempt to introduce CNNs to process light fields and obtain corresponding saliency maps, while Zhang et al. [22] detect attractive objects on micro-lens images. Based on ConvLSTM [27], some recent works [20], [21], [23] process all-focus images and focal stacks separately; then, the extracted saliency cues are exploited to produce saliency maps. Though these deep learning-based 4D SOD methods can outperform their traditional counterparts, they still perform worse than some cutting-edge 2D and 3D methods because ConvLSTM fails to distinguish the valuable focal slices and is easily influenced by non-salient backgrounds. Combining the distinct structural characteristics of focal stacks with network designs is still unexplored.

III. Proposed Dataset

To meet the demand for a high-quality dataset and overcome the issues triggered by the shortcomings of existing datasets, we elaborately construct a large-scale light field dataset, PKU-LF. This section will detail the construction process and statistical properties of our dataset.

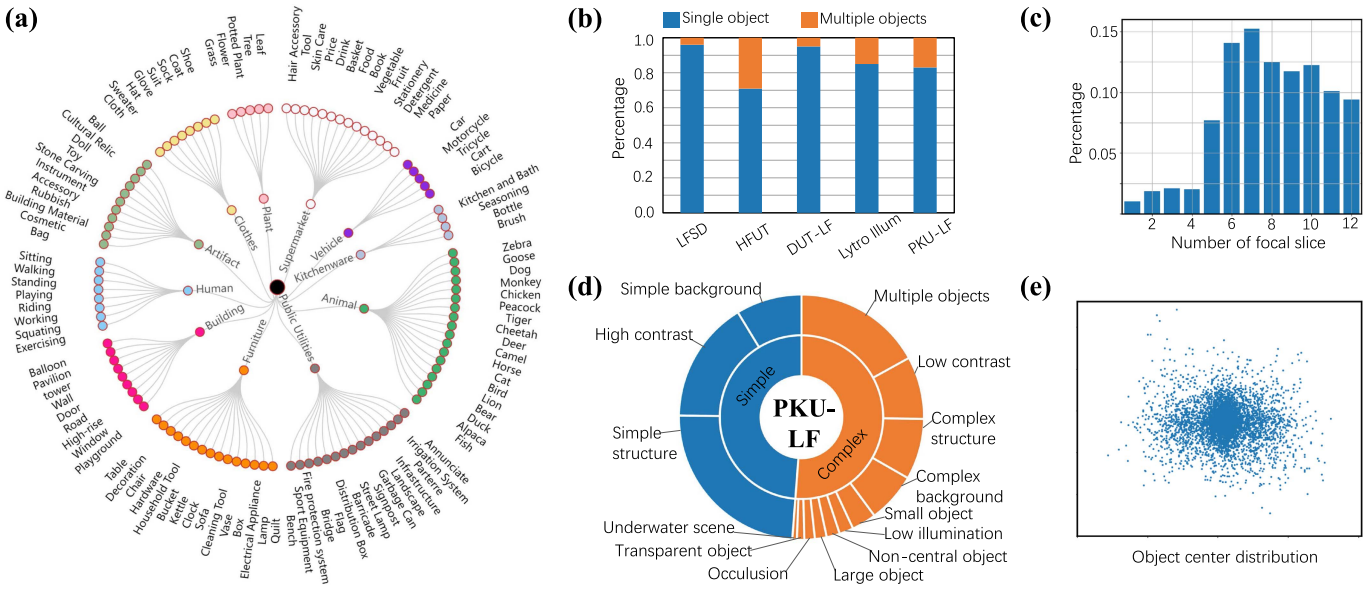


Fig. 4. Statistics of our proposed dataset. (a) Object categories including 11 superclasses and 112 subclasses. (b) Statistics on object numbers. (c) Number of focal slices in a focal stack. (d) Statistics on scene complexity. (e) Scatter of object centers.

A. Dataset Construction

Raw Data Collection. All light fields in our dataset are collected manually through the advanced Lytro Illum camera. We spent about one year in the entire data collection phase, and the scenes of our dataset are assembled from dozens of urban and rural areas. For a broad coverage of real-world scenes, we carefully change the taking conditions to capture a variety of light field samples, such as the site, time, environment, *etc*. Besides, the number, category, size, and location of salient objects are also well-considered for ensuring a rich dataset. A new dataset should not only solve existing issues but also pose new challenges. With this in mind, we collect ample complex samples (see Fig. 3), some of which are absent from existing datasets. In this way, we first obtain over 9,000 light fields. Following previous works [20], we check the quality of each light field and abandon repetitive, blurred, or over-exposed light field images. After the filter, 4,970 high-quality light fields are retained. To further diversify our dataset and avoid selection bias, 30 light fields without any salient object are appended into our dataset as negative samples. Finally, we utilize these 5,000 light fields to establish our dataset.

Raw Data Decoding. To provide a unified platform for studying 2D, 3D, and 4D SOD methods with different input formats, we use Lytro Power Tools [7] to decode raw light field data into multiple representation formats. As shown in Fig. 2(b), each light field is synthesized into an all-focus image, a depth map, sub-aperture images, micro-lens images, and a focal stack with relative-depth-of-field coordinates. The angular resolution of sub-aperture images is 9×9, while the focal stack contains from 1 to 12 focal slices selected randomly according to the relative-depth-of-field coordinates [24]. Fig. 4(c) shows that most focal stacks in our dataset have more than 4 focal slices. Without losing object details, we provide images with a high resolution of 2022×1404.

Professional Annotation. We employ thirty participants divided into ten groups to conduct our subjective experiments. Every three participants within a group jointly identify salient objects in light fields, which are then cross-validated by other groups. Only the object confirmed by over 80 percent of participants will be deemed the positive label to ensure a high consistency. The annotation quality is an essential factor influencing the lifespan of datasets. To this end, we resort to professional annotators to obtain diverse types of annotations (see Fig. 3), e.g., scribble annotations, bounding boxes, object-/instance-level annotations, and edge annotations. The rich annotations in our dataset are expected to facilitate many potential tasks, not limited to SOD. Besides, since our annotations are labeled on high-resolution (2022×1404) images, as shown in Fig. 5, every detail of salient objects can be traced, which benefits the performance improvement of existing SOD algorithms.

Data Split. Providing an official split is conducive to fairly observing and comparing the pros and cons of algorithms. Like the practice of previous works [20], [59], we randomly split the retained light fields into the training set and the testing set at the ratio of 7:3. Finally, our training set contains 3,500 samples, while the testing set has 1,500 samples.

B. Dataset Properties and Statistics

To help fully understand the proposed PKU-LF dataset, we present its several critical statistical properties.

Data Formats. As summarized in Table I, the proposed dataset has complete coverage of light field representations, e.g., all-focus images, depth maps, sub-aperture images, micro-lens images, and focal stacks with corresponding relative-depth-of-field coordinates. As a result, our dataset is expected to become

Fig. 5. Precise annotations. The annotations in our dataset are generated by accurate matting.

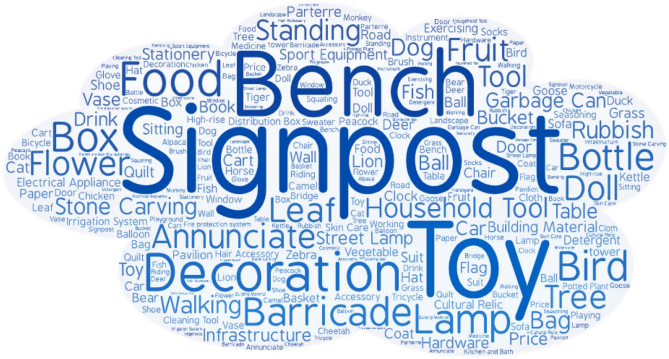


Fig. 6. Word cloud distribution of our PKU-LF. The size of a word is proportional to the word frequency.

a unified platform for studying 2D, 3D, and 4D SOD methods with different input formats. Besides, our high-resolution images enable the exploration of the fledgling high-resolution SOD [60]. To further promote the scalability and flexibility of our dataset in facilitating more exploration and optimization for light field processing and analysis tasks, we also publicize the raw light field data and calibration data of our camera.

Data Amounts. Compared with existing datasets, our dataset contains the most considerable light field amount, even exceeding the sum of four existing datasets (see Table I). This data amount can effectively boost the generalization of algorithms and prevent overfitting. Besides, with the introduction of our dataset, the emerging transformer-based methods [62] that demand large amounts of data can be investigated.

Object Categories. To demonstrate the object diversity of our dataset, we establish a taxonomic system [59] shown in Fig. 4(a). We first categorize our dataset into 11 superclasses, such as the animal, plant, human, vehicle, *etc*. Note that some rare samples are classified as "other" and not shown. Then, we can conclude 112 common subclasses that cover a variety of indoor and outdoor scenes. The word cloud distribution of these subclasses is illustrated in Fig. 6. The significant diversity of our dataset can virtually improve the performance of algorithms in real-world scenes.

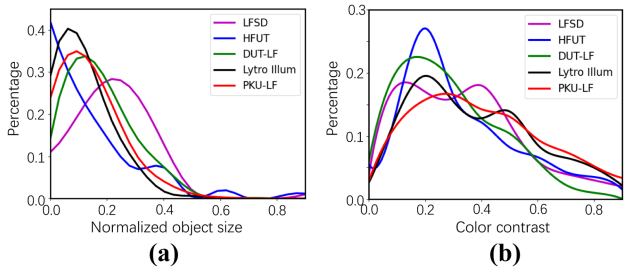


Fig. 7. Statistical illustration of existing datasets and ours. (a) Distribution of normalized object sizes [7]. (b) Distribution of color contrasts [61] between the foreground and background.

Object Numbers. We statistic the object numbers of existing datasets and ours in Fig. 4(b). It reveals that most samples in LFSD and DUT-LF contain a single object. Though HFUT and Lytro Illum promote the proportion of samples with multiple objects, their deficient data amounts restrict the practical applicability. However, our dataset shows a brilliant balance between the data amount and proportion of multi-object samples.

Object Sizes. The object size [7] is an important indicator reflecting the dataset complexity. We illustrate the distribution of normalized object sizes [7] in Fig. 7(a). It can be seen that the object size of our dataset has a significant variation from 0.00 to 0.74, averaging 0.19. Most objects in our dataset have a relatively small size, thus being challenging.

Scene Complexity. Our dataset involves numerous complex scenes. Several visual examples are shown in Fig. 3. Apart from the typical challenging scenes (such as transparent objects, similar foregrounds and backgrounds, low illumination, small/large objects, multiple objects, *etc*.), we also take into account some additional scenes, such as valuable underwater scenes. The detailed statistics in Fig. 4(d) indicate that samples with complex attributes in our dataset account for 51.2%.

Color Contrasts. The color contrast [61] between the foreground and background also expresses the diversity of a dataset. Similar to [61], we compute the color contrast and visualize the distribution in Fig. 7(b), from which we can learn that our dataset has a relatively balanced distribution of color contrasts.

Center Bias. To prevent the models from determining salient objects mainly from the center location priors, we allow the salient objects in an image to appear in a broader area, sometimes deviating from the image center. The scatter plot of object centers of our dataset in Fig. 4(e) reveals that our dataset suffers from less center bias.

Rich Annotations. As demonstrated in Table I, in addition to the common object-level annotations, our dataset first provides scribble annotations, bounding boxes, instance-level annotations, and edge annotations (see Fig. 3). These additional annotations not only help understand our dataset from different perspectives but also catalyze many new tasks, such as weakly supervised SOD, object proposal detection, salient instance detection, semantic edge detection, *etc*.

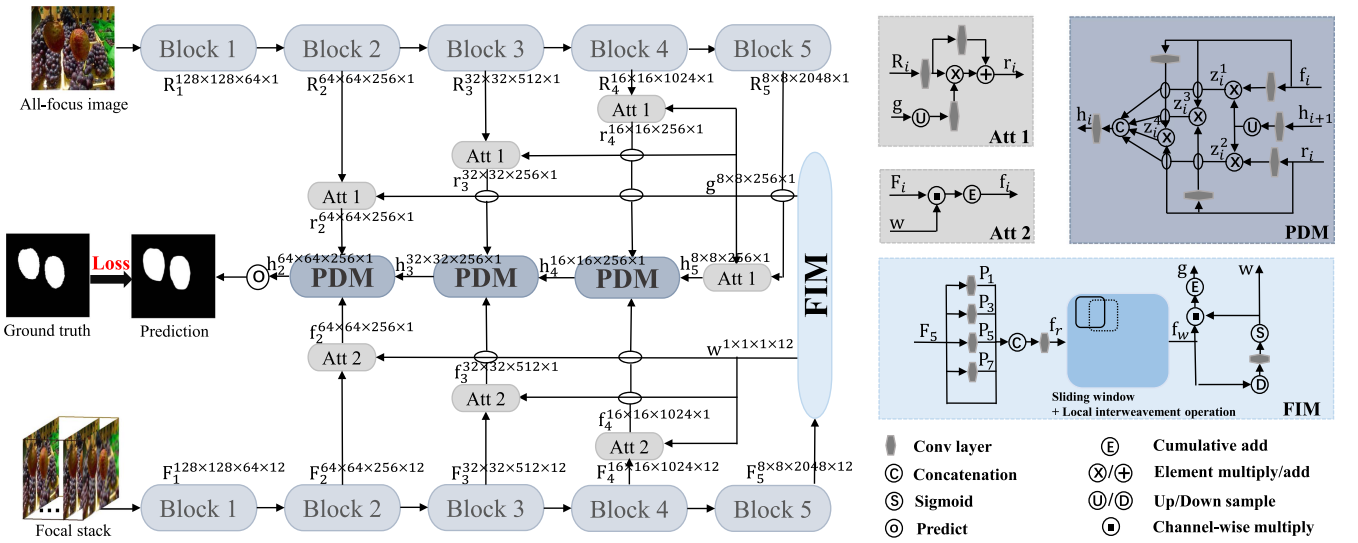


Fig. 8. Overall architecture of our symmetric two-stream architecture (*STSA*) network, which is developed in an encoder-decoder manner. Apart from a two-stream backbone ($\{\text{Block i}\}_{i=1}^{5}$), our encoder contains a focalness interweavement module (FIM) followed by several attention operations (Att 1/2), while our decoder includes three partial decoder modules (PDM) that can simultaneously aggregate and decode multi-modal (i.e., all-focus images and focal stacks) and multi-scale features.

IV. Proposed Method

A. Network Overview

As illustrated in Fig. 8, the overall architecture of the proposed symmetric two-stream architecture (*STSA*) network is developed in an encoder-decoder manner. Apart from a two-stream backbone [16], [21] that extracts features from all-focus images and focal stacks, respectively, our *STSA* mainly contains a focalness interweavement module (FIM) followed by several attention operations [62] and three partial decoder modules (PDM). The FIM is introduced to excavate and enhance the task-benefited features according to the unique structural characteristics of focal stacks, while the PDM is designed to aggregate and decode coadjutant multi-modal and multi-scale features into accurate saliency maps.

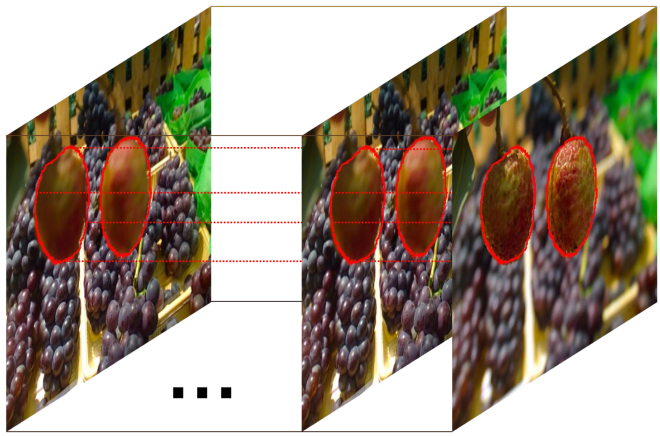


Fig. 9. Spatial alignment of salient objects across different focal slices. Left focal slices focus on the distant backgrounds, while right ones focus on the foregrounds. Salient objects in different focal slices are spatially aligned.

B. Proposed Focalness Interweavement Module (FIM)

As shown in Fig. 2, light fields record abundant spatial information in the form of image arrays, such as a stack of focal slices with different focused pixels. Properly excavating the SOD task-benefited features from the massive spatial information in light fields is still challenging for light field processing. Previous works [21], [23], [58] mainly rely on ConvLSTM [27], a module for video sequence modeling, to establish long-range dependencies across focal slices. However, focal stacks are essentially distinct from videos in their structural characteristics; the rigid application of ConvLSTM thus results in low accuracy and efficiency.

Fig. 9 illustrates an enlarged example of focal stacks, from which we can learn a helpful trait that salient objects in different focal slices are spatially aligned. According to this observation, we propose building connections in the local regions of a focal stack rather than the full when establishing long-range dependencies. This strategy can reduce the computational complexity and prevent the disturbance of non-salient backgrounds. To further clarify the roles of different focal slices in a focal stack, we repeat each focal slice to the number in a focal stack and predict corresponding salient objects using the proposed model. The visual results are shown in Fig. 10. It can be seen that the quality of saliency maps from these focal slices varies remarkably, indicating their unequal contributions to the SOD task. Specifically, focal slices focusing on salient objects have clear structural information of salient objects, thus benefitting the integrity of boundary segmentation, compared with those focusing on backgrounds. Since focal slices focusing on backgrounds as auxiliary information help models capture the spatial information of scenes, models can detect the most accurate salient objects from the complete focal stack.

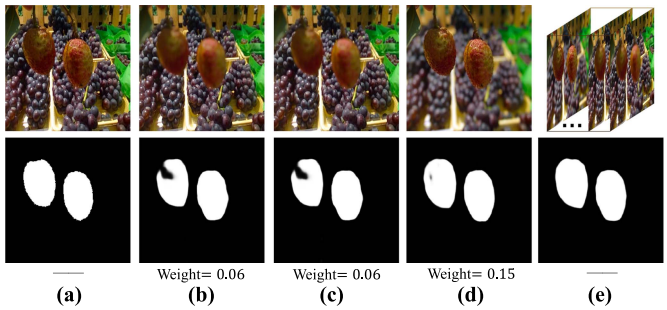


Fig. 10. Qualitative results of individual focal slices and focal stack. (a) All-focus image and ground truth. (b)-(d) Focal slices and predictions. (e) Focal stack and prediction. "Weight" denotes the weight coefficient of each focal slice learned by our FIM.

Local Interweavement Operation (LIO). To combine the powerful non-local operation [62], [63], [64] for establishing long-range dependencies with the unique structural characteristics of focal stacks, we define a local interweavement operation (LIO) $\mathcal{F}_l(\cdot)$ to establish local connections across focal slices. Concretely, as shown in Fig. 11, let $\mathbf{X} \in \mathbb{R}^{S\times H\times W\times C}$ represent a local feature patch from a focal stack, where S indicates the number of focal slices and H, W, C denote the height, width, and channel of the patch, respectively. We first map $\mathbf{X}$ to $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{S\times H\times W\times C}$ via three convolution operations with $1\times 1\times 1$ filters. Given the center vector $\mathbf{Q}_{s,h,w} \in \mathbb{R}^{C}$ at the location (s, h, w) of $\mathbf{Q}$, we calculate the similarity vector $\mathbf{S}_{s,h,w} \in \mathbb{R}^{S+H+W-2}$ between the vector $\mathbf{Q}_{s,h,w}$ and the vectors $\mathbf{K}_{s,h,\cdot} \cup \mathbf{K}_{s,\cdot,w} \cup \mathbf{K}_{\cdot,h,w}$ in $\mathbf{K}$ via inner product, where $\mathbf{K}_{s,h,\cdot} \cup \mathbf{K}_{s,\cdot,w} \cup \mathbf{K}_{\cdot,h,w}$ expresses the vectors in the same row, column, or image location with $\mathbf{Q}_{s,h,w}$. Then, we apply the softmax function to $\mathbf{S}$ and obtain the attention map $\mathbf{A} \in \mathbb{R}^{S+H+W-2}$. According to the attention map $\mathbf{A}$, vectors in $\mathbf{V}$ are weighted and added by their correlated ones measured by $\mathbf{A}$. Finally, to retain the integrity of the original features, a shortcut connection [65] is also introduced for residual learning.

To establish dense dependencies among adjacent areas that are not in the same row, column, or image location, we can recurrently apply the LIO T times on the feature patch. Besides, our LIO can establish dependencies with features that are farther away as T increases. Compared with the original non-local operation or ConvLSTM, the proposed method can significantly improve the efficiency of building connections across focal slices.

Focalness Interweavement Module. Based on the LIO, we develop the focalness interweavement module (FIM) to excavate and exploit the SOD task-benefited features. Formally, as shown in Fig. 8, let $\{\mathbf{R}_i\}_{i=1}^{5}$ and $\{\mathbf{F}_i\}_{i=1}^{5}$ denote the i_{th} layer features extracted from all-focus images and focal stacks, respectively, by the two-stream backbone. In our FIM, a receptive field block (RFB) [66] is first applied to the features $\mathbf{F}_5$ to expand the receptive field, which is conducive to accurately detecting large objects [17], [59]. Specifically, the RFB is achieved by arranging K parallel dilated convolution layers $\{\mathcal{D}_d^k\}_{k=1}^{K}$ with different dilated rates d, each of which outputs $\mathbf{P}_d$. Then, all $\mathbf{P}_d$ and the original $\mathbf{F}_5$ are concatenated followed by a 1×1 convolution layer to reduce the channel to C

$$\mathbf{f}_r = conv_{1\times 1}^{r}(cat(\mathbf{F}_5, \mathbf{P}_1, \mathbf{P}_3, \mathbf{P}_5, \ldots, \mathbf{P}_{2k-1})), \quad (1)$$

where $\mathbf{f}_r \in \mathbb{R}^{S\times H\times W\times C}$ and $\mathbf{P}_k = \{\mathcal{D}_{2k-1}^{k}\}(\mathbf{F}_5)$. cat denotes the concatenation operation, and $conv$ represents the convolution operation whose subscript indicates the size of its filter. Until this step, different focal slices in a focal stack are still independent, and there is still no information flow among focal slices. To establish connections across focal slices, we apply the LIO $\mathcal{F}_l(\cdot)$ on the patches of $\mathbf{f}_r$ via a sliding window with radius r. In this approach, the interweaved features $\mathbf{f}_w$ containing rich high-level spatial information, such as the amount and location of salient objects and the importance degree of each focal slice, can be obtained. Subsequently, we encode $\mathbf{f}_w$ into the semantic features $\mathbf{g} \in \mathbb{R}^{H\times W\times C\times 1}$ and the weight coefficients $\mathbf{w} \in \mathbb{R}^{1\times 1\times 1\times S}$ of focal slices. Finally, by introducing the attention mechanism [62], the semantic features are adopted to enhance the low-level features $\{\mathbf{R}_i\}_{i=2}^{5}$ from all-focus images, while the weight coefficients are utilized to compress the low-level features $\{\mathbf{F}_i\}_{i=2}^{4}$ from focal stacks

$$\mathbf{w} = \sigma(conv_{1\times 1}^{w}(Pool(cat(\mathbf{f}_w^1, \mathbf{f}_w^2, \ldots, \mathbf{f}_w^S))), \quad (2)$$

$$\mathbf{g} = conv_{1\times 1}^{g1}\left(\sum_{s=1}^{S}\mathbf{w}_i\mathbf{f}_w^s\right),\ \mathbf{f}_i = \sum_{s=1}^{S}\mathbf{w}_i\mathbf{F}_i^s, \quad (3)$$

$$\mathbf{r}_i = conv_{1\times 1}^{r1}(\mathbf{R}_i) \otimes conv_{3\times 3}^{g2}(\mathcal{U}(\mathbf{g})) + conv_{3\times 3}^{r2}(conv_{1\times 1}^{r1}(\mathbf{R}_i)), \quad (4)$$

where $\{\mathbf{r}_i\}_{i=2}^{5}$ and $\{\mathbf{f}_i\}_{i=2}^{4}$ denote the enhanced low-level features. σ, $Pool$, $\otimes$, and $\mathcal{U}$ indicate the sigmoid activation function, average pooling operation, element-wise multiplication, and bilinear upsampling operation, respectively.

C. Proposed Partial Decoder Module (PDM)

Features of different modalities or different levels usually contain noisy heterogeneous ingredients. Previous works [5], [28] merely use addition or concatenation for aggregating multi-modal and multi-scale features. The lack of effective interactions among these features makes it difficult to filter these heterogeneous ingredients and emphasize mutual consistency. Thus, the proper designs of decoders can further unleash the potential of light field SOD.

To take explicit feature interactions among multi-modal and multi-scale features, we propose an effective partial decoder module (PDM) to enhance and fuse the features from our encoder simultaneously. Specifically, the proposed PDM takes a mutual-enhancement manner to aggregate multi-modal and multi-scale features. The high-level features $\{\mathbf{h}_{i+1}\}_{i=3}^{5}$ (note that $\mathbf{h}_5 = \mathbf{r}_5$) are first utilized to enhance the multi-modal features $\{\mathbf{r}_i\}_{i=2}^{4}$ and $\{\mathbf{f}_i\}_{i=2}^{4}$ from the all-focus stream and focal stack stream, respectively. Mutual enhancements are then performed between the multi-modal features from the two streams. We concatenate all the enhanced features and introduce a 1×1 convolution layer to reduce the channel to C

$$\mathbf{z}_i^1 = \mathcal{U}(conv_{3\times 3}^{h1}(\mathbf{h}_{i+1})) \otimes conv_{3\times 3}^{f1}(\mathbf{f}_i),$$

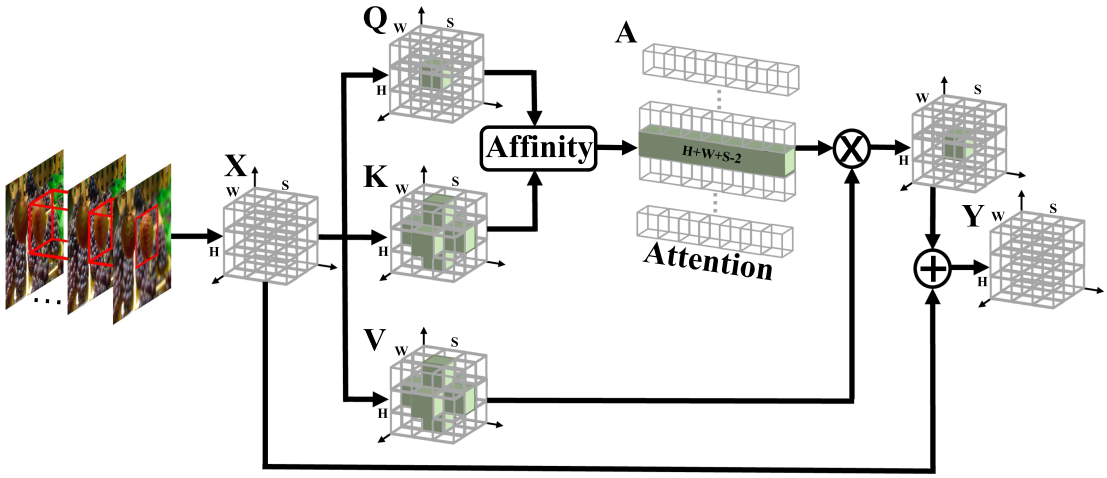


Fig. 11. Pipeline of our local interweavement operation described in Subsection IV-B. $\mathbf{X}$ denotes a local feature patch from a focal stack, which is subsequently mapped to $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ via convolution operations. We compute the attention map $\mathbf{A}$ between the center vector of $\mathbf{Q}$ (small solid cube) and the vectors in the same row, column, or image location in $\mathbf{K}$ (small solid cubes). According to the attention map, vectors in $\mathbf{V}$ (small solid cubes) are aggregated. We also introduce a residual connection to retain the original features.

$$\begin{aligned}
\mathbf{z}_i^2 &= \mathcal{U}(conv_{3\times3}^{h1}(\mathbf{h}_{i+1})) \otimes conv_{3\times3}^{r1}(\mathbf{r}_i),\\
\mathbf{z}_i^3 &= conv_{3\times3}^{f2}(\mathbf{f}_i) \otimes \mathbf{r}_i,\\
\mathbf{z}_i^4 &= conv_{3\times3}^{r2}(\mathbf{r}_i) \otimes \mathbf{f}_i,\\
\mathbf{h}_i &= conv_{1\times1}^{h2}(cat(\mathbf{z}_i^1, \mathbf{z}_i^2, \mathbf{z}_i^3, \mathbf{z}_i^4)).
\end{aligned} \tag{5}$$

Our decoder has three identical PDMs. Finally, a bilinear upsampling operation and a convolution operation, which constitute the prediction layer, are successively performed on the output of the last PDM $\mathbf{h}_2$ to produce the final result $\mathbf{P}$.

It is worth noting that the decoders with the same name in existing works [59], [67], i.e., partial decoder module, have essentially different designs from ours. Their "partial" expresses that they only utilize the high-level features and neglect the intensive low-level ones for computation saving, thus sacrificing many low-level structural details. In contrast, ours represents the recurrent step of applying the proposed module in decoding high-/low-level features, avoiding losing low-level details.

D. Loss Function

Our loss is computed from the prediction $\mathbf{P}$ and corresponding ground truth $\mathbf{G}$. Previous works [21], [23] usually adopt the cross-entropy loss (ℓ_{ce}) to train their networks

$$\ell_{ce}(\mathbf{P}, \mathbf{G}) = -\frac{1}{N}\sum_{i=1}^{N}(\mathbf{G}_i \log \mathbf{P}_i + (1-\mathbf{G}_i)\log(1-\mathbf{P}_i)), \tag{6}$$

where N is the number of pixels in an image. When introducing the cross-entropy loss in our experiments, we find a huge gap between the mean F-measure (F_β^{mean}) and max F-measure (F_β^{max}) [68], indicating that the results are very susceptible to the threshold change [69]. It has been proved that FLoss (ℓ_f) [69] can alleviate the sensitiveness to the threshold because of its excellent gradient properties. Concretely, FLoss is reformulated on the relaxed precision ($\hat{prec}$) [70] and recall ($\hat{recall}$) [70]

$$\ell_f(\mathbf{P}, \mathbf{G}) = 1 - \frac{(1+\beta^2)\hat{prec}\cdot\hat{recall}}{\beta^2\hat{prec}+\hat{recall}}, \tag{7}$$

where $\beta^2 = 0.3$ is introduced to trade-off the precision and recall. However, since FLoss only resorts to the global matching information for optimizing and lacks the direct pixel-level optimization similar to that of the cross-entropy loss, the training processes of methods supervised by FLoss often appear to be unstable. Eventually, the performance on pixel-wise matching metrics, such as mean absolute error, becomes worse than before(see Table IV). If one can inherit the excellent gradient properties of FLoss and simultaneously contain the direct pixel-level optimization, the performance of SOD methods can be comprehensively improved. With this consideration, we develop a compound loss (ℓ_{cp}) by combining FLoss with pixel matching information

$$\ell_{cp}(\mathbf{P}, \mathbf{G}) = \ell_f(\mathbf{P}, \mathbf{G}) + \frac{1}{N}\sum_{i=1}^{N}|\mathbf{P}_i - \mathbf{G}_i|. \tag{8}$$

Extensive ablation experiments and detailed analysis in Subsection V-D evidence that this improvement can eliminate the sensitiveness to the threshold and visibly promote the accuracy of SOD.

V. Benchmark Evaluation Results

A. Implementation Details and Strategies

Implementation. We use the Pytorch toolbox to implement our *STSA* network, and our device is a single NVIDIA TiTan Xp GPU. The convolution blocks ($\{\text{Block i}\}_{i=1}^{5}$) of our two-stream backbone are borrowed from ResNet-50 [65] whose parameters are initialized from the model pre-trained on ImageNet. We set the network parameters as $C = 256$, $K = 4$, $T = 2$, and $r = 5$. Our model is trained by the SGD optimizer and warmed

up by the cross-entropy loss. The hyperparameters, including the batch size, maximum epoch, learning rate, momentum, and weight decay, are set to 1, 50, 1e-3, 9e-1, and 5e-4, respectively. Following previous works [20], [21], [23], each focal stack is expanded to contain $S = 12$ focal slices, and all images are resized to 256×256 with common data augmentation strategies, i.e., random flipping, rotating, and border clipping, to avoid overfitting in the training process.

Training and Testing Data. To disentangle the contributions of this work, i.e., the proposed PKU-LF dataset and *STSA* network, verify the generalization of our dataset, and clarify the effectiveness of our network designs, we set three training strategies: **i)** DUT-LF [20] + HFUT [25], **ii)** PKU-LF, and **iii)** DUT-LF + HFUT + PKU-LF, where the first strategy is widely used in existing works [23]. We evaluate our model on the whole LFSD [18] and Lytro Illum [22], the remaining samples of HFUT, and the testing datasets of DUT-LF and our PKU-LF.

Evaluation Metrics. To analyze the results of different methods, we employ eight popular evaluation metrics for quantitative performance benchmarking, including mean absolute error ($\mathcal{M}$) [75], the mean, maximum and adaptive scores of F-measure (F_β^{mean}, F_β^{max}, and F_β^{adapt}) [68], the mean, maximum and adaptive scores of E-measure (E_ϕ^{mean}, E_ϕ^{max} and E_ϕ^{adapt}) [76], and S-measure (S_α) [77]. Concretely, mean absolute error measures the pixel-wise difference between the prediction $\mathbf{P}$ and corresponding ground truth $\mathbf{G}$, i.e., $\mathcal{M} = \frac{1}{N}\sum_{i=1}^{N}|\mathbf{P}_i - \mathbf{G}_i|$, while F-measure is the harmonic mean value of the precision ($prec$) and recall ($recall$), i.e., $F_\beta = \frac{(1-\beta^2)prec \cdot recall}{\beta^2 prec + recall}$, where $\beta^2 = 0.3$ indicates that the precision is more dominant in the SOD task. Different F-measure scores, i.e., F_β^{mean}, F_β^{max}, and F_β^{adapt}, can be computed when we binarize the prediction by a varying threshold. S-measure is an indicator that emphasises the accurate detection of object structures. It is formulated as $S_\alpha = \alpha S_o + (1-\alpha)S_r$, where S_o and S_r denote the object-aware structural similarity and region-aware structural similarity, respectively. Like previous works [21], [23], [28], we set $\alpha = 0.5$ to balance the object and region similarity. E-measure is a new metric combining local matching information with image-level matching information for assessment. It is defined as $E_\phi = \frac{1}{N}\sum_{i=1}^{N}\phi_{FM}(i)$, where $\phi_{FM}(\cdot)$ is the enhanced alignment matrix. Similar to F-measure, a varying threshold results in different E-measure scores, i.e., E_ϕ^{mean}, E_ϕ^{max}, and E_ϕ^{adapt}. These eight evaluation metrics can provide convincing quality assessment results of saliency predictions.

B. Performance Benchmarking and Analysis

To have a thorough evaluation and analysis, we compare the proposed *STSA* network with 16 representative 2D, 3D, and 4D methods on four existing and the proposed datasets, among which there are 3 traditional methods and 13 deep learning-based methods. Specifically, the competitors include two state-of-the-art 2D methods, i.e., GCPA [14] and MINet [13], nine high-ranking 3D methods, i.e., BBS [17], JLDCF [5], SSF [71], UCNet [1], D3Net [15], S2MA [72], cmMS [73], HDF [74], and ATSA [16], and five cutting-edge 4D methods, i.e., LFS [18], WSC [55], DILF [56], MoLF [21], and ERNet [23]. For a fair comparison, the results of these models are generated by authorized codes with the default parameter settings or directly provided by the authors.

Traditional Methods versus Deep Methods. Table II lists the performance comparison of 3 traditional methods and 14 deep learning-based SOD methods, from which we can learn that learning-based methods consistently outperform the traditional ones by a clear margin. It reveals that the application of neural networks in light field SOD has excellent potential. Besides, advanced architectures benefit the performance improvements. We believe that introducing transformer/MLP-based architectures is a promising direction for light field SOD.

4D SOD versus 2D/3D SOD. Establishing a unified benchmark for 2D, 3D, and 4D SOD helps compare and analyze the pros and cons of various models, ultimately promoting their coordinated development. As shown in Table II, the performance of two 2D methods (GCPA and MINet) is comparable but slightly worse than that of top-ranking 3D or 4D methods, though this two are state-of-the-art methods in the 2D SOD field. The lack of spatial information hinders the performance improvement of 2D models, especially when confronting complex scenes. However, though 4D light fields provide more spatial information than the 3D alternatives, the best existing 4D method (ERNet) still performs worse than cutting-edge 3D methods (such as ATSA, JLDCF, and SSF). The underlying reason is that previous 4D methods using ConvLSTM ignore the unique structural characteristics of light fields, thus failing to exploit the rich spatial information in light fields. Since our network designs are tailored to light fields, the proposed method can significantly outperform all 2D, 3D, and 4D SOD methods, which indicates the potential of 4D SOD. Our best performance also demonstrates that tailored designs can unleash the potential of light fields in SOD.

Cross-Dataset Analysis. The generalizability of datasets determines their practical applicability. A new dataset should have a broad generalization for acquiring related performance goals. A reasonable approach to studying the dataset generalizability is to conduct a cross-dataset analysis [78]. Our cross-dataset analysis is performed between the largest two datasets, i.e., DUT-LF and our PKU-LF. Specifically, we alternatively train the proposed *STSA* on one dataset while testing it on other datasets and observing the performance change.

The S-measure scores for cross-dataset analysis are in Table III. Note that there are similar conclusions for other evaluation metrics, e.g., $\mathcal{M}$, F_β, and E_ϕ. We can observe a clear performance drop when training on DUT-LF whereas testing on our PKU-LF or other datasets, compared with testing on itself. This is because our dataset contains many scenes absent from DUT-LF. However, the model trained on our dataset consistently keeps the brilliant performance whatever the testing datasets are. This demonstrates the excellent generalizability of our dataset. The last column of Table III further lists the degree of performance drop when transferring from being tested on oneself to the

TABLE II
BENCHMARKING RESULTS OF REPRESENTATIVE 2D, 3D, AND 4D MODELS ON FOUR EXISTING AND OUR PROPOSED DATASETS

		Traditional			Deep Learning-based															
	Metric	LFS‡ [18]	WSC‡ [55]	DILF‡ [56]	MoLF‡ [21]	ERNet‡ [23]	BBS† [17]	JLDCF† [5]	SSF† [71]	UCNet† [1]	D3Net† [15]	S2MA† [72]	cmMS† [73]	HDF† [74]	ATSA† [16]	MINet⋆ [13]	GCPA⋆ [14]	Ours‡$_1$	Ours‡$_2$	Ours‡$_3$
LFSD [18]	$S_\alpha \uparrow$	.681	.702	.811	.831	.835	.864	.862	.859	.858	.825	.837	.850	.846	.858	.815	.830	.859	.864	.871
	$F_\beta^{\max} \uparrow$	.744	.743	.811	.834	.850	.858	.867	.868	.859	.812	.835	.858	.837	.866	.790	.811	.868	.871	.877
	$F_\beta^{\text{mean}} \uparrow$	.513	.722	.719	.809	.836	.842	.848	.862	.848	.797	.806	.850	.818	.856	.781	.800	.853	.864	.868
	$F_\beta^{\text{adp}} \uparrow$	.735	.743	.795	.819	.839	.840	.827	.862	.838	.788	.803	.857	.818	.852	.810	.816	.860	.866	.872
	$E_\phi^{\max} \uparrow$	.809	.789	.861	.888	.888	.900	.902	.901	.898	.863	.873	.896	.880	.902	.840	.850	.905	.907	.909
	$E_\phi^{\text{mean}} \uparrow$	.567	.753	.764	.872	.883	.883	.894	.890	.893	.850	.855	.881	.869	.899	.834	.839	.902	.902	.905
	$E_\phi^{\text{adp}} \uparrow$	.773	.788	.846	.886	.887	.889	.882	.896	.890	.853	.863	.890	.872	.897	.864	.869	.906	.905	.910
	$\mathcal{M} \downarrow$	.205	.150	.136	.088	.082	.072	.070	.067	.072	.095	.094	.073	.086	.068	.096	.093	.067	.065	.062
HFUT [25]	$S_\alpha \uparrow$	.565	.613	.672	.742	.778	.751	.789	.725	.748	.749	.729	.723	.763	.772	.769	.775	.810	.789	.834
	$F_\beta^{\max} \uparrow$	.427	.508	.601	.662	.722	.676	.727	.647	.677	.671	.650	.626	.690	.729	.692	.701	.784	.779	.810
	$F_\beta^{\text{mean}} \uparrow$	.323	.493	.513	.639	.709	.654	.707	.639	.672	.651	.623	.617	.669	.706	.683	.682	.771	.770	.805
	$F_\beta^{\text{adp}} \uparrow$	.427	.485	.530	.627	.706	.654	.677	.636	.675	.647	.588	.636	.653	.689	.691	.687	.769	.767	.804
	$E_\phi^{\max} \uparrow$	.637	.695	.748	.812	.841	.801	.844	.778	.804	.797	.777	.784	.801	.833	.804	.812	.868	.841	.884
	$E_\phi^{\text{mean}} \uparrow$	.524	.684	.657	.790	.832	.765	.825	.763	.793	.773	.756	.746	.788	.819	.787	.794	.864	.836	.879
	$E_\phi^{\text{adp}} \uparrow$	.666	.680	.693	.785	.831	.804	.811	.781	.810	.789	.744	.779	.789	.810	.816	.822	.865	.838	.880
	$\mathcal{M} \downarrow$	.221	.154	.150	.094	.082	.089	.075	.100	.090	.091	.112	.097	.095	.084	.088	.094	.067	.072	.057
DUT-LF [20]	$S_\alpha \uparrow$	.585	.657	.654	.887	.899	.865	.877	.879	.831	.822	.787	.804	.822	.901	.870	.881	.911	.915	.928
	$F_\beta^{\max} \uparrow$	.533	.621	.585	.903	.908	.852	.878	.887	.816	.797	.754	.803	.801	.915	.855	.866	.928	.929	.941
	$F_\beta^{\text{mean}} \uparrow$	.358	.610	.492	.855	.891	.834	.846	.879	.806	.776	.733	.774	.776	.900	.845	.849	.906	.911	.928
	$F_\beta^{\text{adapt}} \uparrow$	.525	.619	.597	.843	.885	.848	.835	.885	.803	.784	.735	.819	.778	.898	.862	.865	.906	.912	.929
	$E_\phi^{\max} \uparrow$	.711	.789	.757	.939	.949	.900	.925	.923	.876	.860	.839	.879	.864	.941	.895	.904	.959	.962	.965
	$E_\phi^{\text{mean}} \uparrow$	.511	.762	.635	.921	.943	.879	.911	.907	.870	.841	.817	.817	.848	.937	.886	.894	.954	.958	.961
	$E_\phi^{\text{adapt}} \uparrow$	.742	.789	.784	.923	.943	.908	.910	.918	.878	.869	.842	.870	.867	.938	.915	.922	.956	.961	.964
	$\mathcal{M} \downarrow$	.227	.149	.165	.051	.039	.066	.058	.050	.081	.083	.102	.079	.091	.041	.060	.061	.033	.030	.027
Lytro Illum [22]	$S_\alpha \uparrow$	.619	.708	.751	.834	.843	.876	.881	.822	.852	.860	.856	.872	.856	.882	.862	.874	.878	.884	.890
	$F_\beta^{\max} \uparrow$	.545	.663	.688	.820	.827	.848	.868	.787	.827	.836	.832	.849	.835	.875	.829	.845	.868	.873	.888
	$F_\beta^{\text{mean}} \uparrow$	.385	.646	.599	.766	.800	.830	.840	.776	.817	.809	.795	.832	.806	.848	.812	.823	.850	.856	.877
	$F_\beta^{\text{adapt}} \uparrow$	.547	.640	.666	.747	.796	.830	.826	.780	.821	.801	.788	.833	.809	.842	.818	.825	.846	.855	.878
	$E_\phi^{\max} \uparrow$	.721	.804	.827	.908	.911	.909	.926	.877	.899	.905	.903	.907	.901	.929	.888	.898	.928	.928	.936
	$E_\phi^{\text{mean}} \uparrow$	.546	.792	.721	.882	.900	.896	.914	.865	.893	.889	.882	.897	.887	.919	.881	.887	.916	.922	.932
	$E_\phi^{\text{adapt}} \uparrow$	.771	.798	.817	.876	.900	.911	.914	.885	.905	.901	.886	.912	.894	.917	.900	.908	.917	.925	.936
	$\mathcal{M} \downarrow$	.197	.115	.127	.065	.056	.047	.044	.066	.053	.055	.060	.045	.056	.042	.051	.050	.045	.043	.037
PKU-LF	$S_\alpha \uparrow$	.579	.641	.618	.809	.826	.847	.854	.841	.792	.802	.765	.846	.822	.860	.841	.852	.862	.876	.887
	$F_\beta^{\max} \uparrow$	.424	.540	.529	.776	.781	.802	.811	.798	.736	.742	.683	.797	.775	.839	.787	.804	.843	.863	.878
	$F_\beta^{\text{mean}} \uparrow$	.325	.519	.507	.716	.761	.775	.785	.786	.730	.717	.660	.782	.753	.814	.776	.785	.820	.854	.870
	$F_\beta^{\text{adapt}} \uparrow$	.435	.505	.498	.683	.754	.761	.778	.787	.737	.715	.651	.777	.744	.800	.784	.784	.809	.852	.870
	$E_\phi^{\max} \uparrow$	.685	.750	.745	.883	.883	.882	.889	.884	.852	.858	.811	.880	.873	.909	.870	.876	.918	.919	.930
	$E_\phi^{\text{mean}} \uparrow$	.546	.730	.731	.850	.870	.864	.878	.876	.842	.828	.791	.869	.858	.898	.858	.862	.909	.914	.927
	$E_\phi^{\text{adapt}} \uparrow$	.693	.722	.727	.832	.867	.876	.880	.889	.864	.855	.808	.881	.861	.891	.885	.888	.909	.917	.930
	$\mathcal{M} \downarrow$	.214	.132	.143	.066	.059	.056	.049	.052	.070	.067	.100	.052	.065	.045	.050	.055	.047	.042	.035

Note that the 2D, 3D, and 4D SOD models are marked with "⋆", "†", and "‡", respectively. "↑ "/" ↓ " indicates that larger/smaller is better. The top-performance methods are successively highlighted in red, blue, cyan, and orange, respectively. The last three columns represent the three training strategies: i) DUT-LF + HFUT, ii) PKU-LF, and iii) DUT-LF + HFUT + PKU-LF.

TABLE III
S-MEASURE [77] SCORES FOR CROSS-DATASET ANALYSIS

Trained on: \ Tested on:	DUT-LF [20]	PKU-LF	Self	Mean Others	Drop ↓
DUT-LF [20]	.913	.841	.913	.846	7.34%
PKU-LF	.915	.876	.876	.863	1.48%

Our STSA model is trained on the dataset in the column and tested on the dataset in the row. "Self" means training and testing on the same dataset. "Mean Others" means averaged testing scores on all the other datasets. "Drop" equals the difference of "Self" and "Mean Others" divided by "Self".

others. We can learn that the cross-dataset performance drops for DUT-LF and our PKU-LF are 7.34% and 1.48%, respectively.

C. Comparison With State-of-The-Arts

In this subsection, we compare our method with existing methods and verify the superiority of our approach.

Quantitative Comparisons. The quantitative results in Table II show that the best performance of our proposed *STSA* network significantly outperforms 16 state-of-the-art 2D, 3D, and 4D methods across four existing and the proposed datasets on eight evaluation metrics. As previous methods of the best performance are inconsistent on different evaluation metrics or datasets, we bag previous methods of the best performance in each row into an integrated one for comparisons. Note that every previous method actually performs worse than this integrated method. The proposed method outperforms the integrated one by a clear

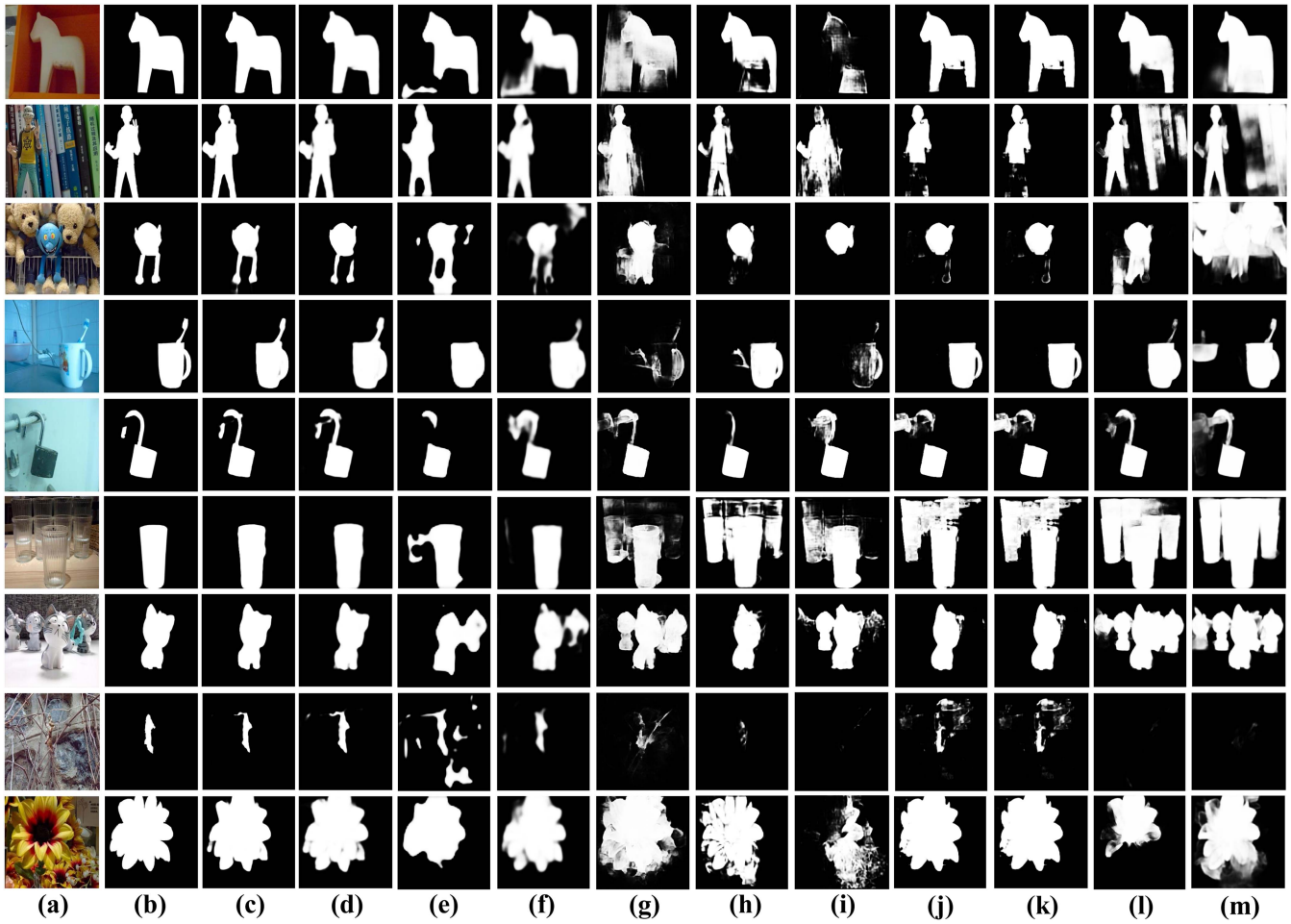


Fig. 12. Qualitative comparisons of the proposed *STSA* network with nine top-ranking methods. (a) All-focus images. (b) Ground truths. (c) Ours with the compound loss. (d) Ours with the cross-entropy loss. (e) ERNet [23]. (f) MoLF [21]. (g) BBS [17]. (h) SSF [71]. (i) UCNet [1]. (j) HDF [74]. (k) ATSA [16]. (l) MINet [13]. (m) GCPA [14].

margin. Concretely, our method has averaged performance improvements of 2.08%, 4.22%, 5.53%, 6.30%, 2.05%, 2.57%, 3.20%, and 24.79% on S_α, F_β^{max}, F_β^{mean}, F_β^{adapt}, E_ϕ^{max}, E_ϕ^{mean}, E_ϕ^{adapt}, and $\mathcal{M}$, respectively.

Qualitaive Comparisons. To observe the visual quality of predicted saliency maps from previous methods and ours, we illustrate several representative examples in Fig. 12. The methods selected for comparisons are nine top-ranking methods in Table II. Though great progress has been made by recent 2D and 3D SOD methods, they still cannot accurately locate and segment the salient objects in complex scenes, e.g., clustered backgrounds (2nd and 3rd rows), similar foregrounds and backgrounds (4th row), occlusion (5th row), transparent objects (6th row), multiple objects (7th row), and small/large objects (8th and 9th rows). In contrast, the three 4D methods, i.e., ours, ERNet, and MoLF, are capable of distinguishing the salient objects in these complex scenes, showing the superiority of light fields containing abundant spatial information. Further comparing our method with the state-of-the-art 4D SOD methods, we can learn that saliency maps produced by our method are more refined and satisfactory thanks to the tailored designs of our method.

Generalizability. To validate the generalizability of our method, we set three training strategies (see Subsection V-A) to conduct our experiments. We list the performance of our *STSA* in the last three columns of Table II. We can learn that our method always achieves the best performance whatever the training strategy is adopted, which proves its excellent generalizability. Because our PKU-LF dataset is by far the highest quality dataset, its applications (Ours$_2^\ddagger$ and Ours$_3^\ddagger$) can visibly improve the performance of models. Besides, even if we exclude our PKU-LF dataset, the proposed *STSA* network (Ours$_1^\ddagger$) still has superiority over its comparative counterparts, especially with previous state-of-the-art 4D SOD methods. This validates the effectiveness of our network designs tailored to light fields.

More Discussion. In our experiments, we find the performance of existing and our proposed methods are susceptible to one typical challenging scene, i.e., objects with complex fine-grained structures. As demonstrated in Fig. 13, though our method can accurately locate the lobster and segment its majority compared with previous top-performance 4D methods, fine-grained structures of the lobster are missing. The underlying reason is that it is challenging to extract sufficiently fine-grained

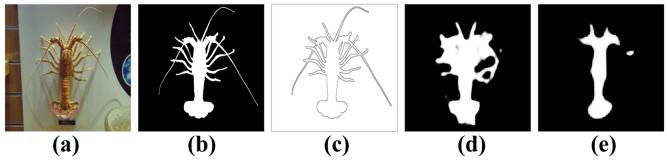


Fig. 13. Typical challenging scene of an object with complex fine-grained structures. (a) All-focus image. (b) Ground truth. (c) Edge annotation. (d) Result from our method. (e) Result from the top-performance 4D method ERNet [23].

TABLE IV
ABLATION ANALYSIS ON DUT-LF DATASET

No.	BaseR	BaseV	LSTM	FIM	PDM	ℓ_{ce}	ℓ_f	ℓ_{cp}	$\mathcal{M}\downarrow$	$F_\beta^{\text{adapt}}\uparrow$
1	✓					✓			.072	.802
2	✓		✓			✓			.061	.826
3	✓			✓		✓			.057	.831
4	✓			✓	✓	✓			.041	.864
5	✓			✓	✓		✓		.043	.883
6		✓		✓	✓			✓	.035	.899
7	✓			✓	✓			✓	.033	.906
8	✓			✓	✓			✓	.027	.929

"BaseR" denotes a baseline model with ResNet-50 [65] backbone, while "BaseV" expresses a baseline model with VGG-19 [80] backbone. "LSTM" denotes the widely used ConvLSTM in previous works. Note that models No.1-No.7 adopt the training strategy DUT-LF + HFUT, and No.8 uses the training strategy DUT + HFUT + PKU-LF.

features portraying objects with complex structures. Thanks to our additional edge annotations, properly devising approaches to utilizing the edge annotations can be a promising direction for segmenting fine-grained structures.

D. Ablation Study

To demonstrate the validity of our contributions, we conduct extensive ablation experiments on the DUT-LF dataset. Specifically, we first exclude all components described in Section IV to obtain an FPN-like network [79] as our baseline model. Then, we separately install the comparative components in previous works and ours into the baseline model to observe the performance change.

Effectiveness of FIM. Previous methods [20], [21], [23] consistently adopt ConvLSTM [27] when establishing long-range dependencies across focal slices. ConvLSTM is initially developed for video sequence modeling and appears to be inapplicable in light field processing. On the one hand, ignoring the alignment of salient objects across focal slices, ConvLSTM is easily influenced by non-salient backgrounds. On the other hand, ConvLSTM treats every focal slice equally and fails to distinguish the valuable focal slices, which contradicts our conclusions from Fig. 10. The proposed FIM is tailored to the unique structural characteristics of focal stacks and can thus effectively exploit SOD-benefited features in light fields. As shown in Table IV, compared with the baseline model (No.1), the proposed FIM (No.3) can bring 20.83% and 3.62% performance improvements on $\mathcal{M}$ and F_β^{adapt}, respectively. Our FIM also outperforms ConvLSTM (No.2), whose performance improvements on $\mathcal{M}$ and F_β^{adapt} are 15.28% and 2.99%. This indicates that our FIM has more outstanding ability to excavate SOD-benefited features than ConvLSTM.

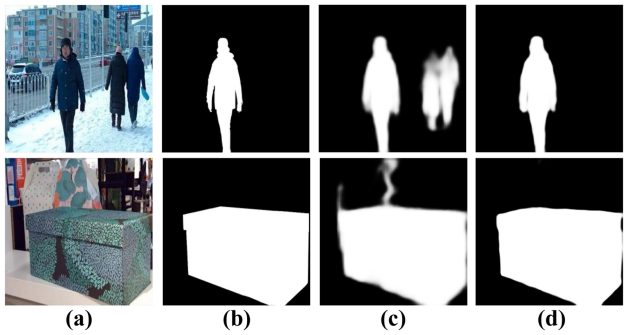


Fig. 14. Visual comparisons of our *STSA* network with/without the FIM. (a) All-focus images. (b) Ground truths. (c) Results from the model without our FIM. (d) Results from the model with our FIM.

TABLE V
ABLATIONS OF DIFFERENT FIM SETTINGS ON DUT-LF DATASET

Metrics	w/o FIM	FIM	Subcomponent		Radius			+All
			w/o RFB	w/o LIO	$r=1$	$r=3$	$r=7$	
$\mathcal{M}\downarrow$	.031	.027	.028	.029	.030	.028	.028	.031
$F_\beta^{\text{adapt}}\uparrow$	.912	.929	.924	.919	.921	.926	.930	.908

"+All" indicates our FIM simultaneously fuses features from focal stacks (F_5) and all-focus images (R_5).

To learn the impact of different configurations of our FIM on the final results, we study the effectiveness of the entire FIM and its subcomponents, respectively. As discussed in Subsection IV-B, the interweaved features from our FIM encode rich high-level information, such as the amount and location of salient objects. To verify this, we show some visual results of the model with/without our FIM in Fig. 14. We can learn that the model without our FIM has difficulty in determining the amount (1st row) and location (2nd row) of the salient objects. In contrast, the model with our FIM can segment the salient objects accurately, which evidences the benefits of our FIM to the final results. The structures of focal stacks cater to the refocusing ability of the human visual system (eye movements). Our FIM is tailored to exploit spatial information within the structural characteristics of focal stacks. If we simultaneously fuse the features from focal stacks and all-focus images via the proposed FIM, the intact structure of focal stacks may be destroyed. Eventually, as shown in Table V, the performance of this approach is even worse than the model without the FIM. Enlarging the radius r of the sliding window can enrich the extracted local contexts, but the computational complexity of subsequent LIO also increases dramatically. Therefore, properly setting the radius r is necessary for balancing related accuracy and efficiency goals. As listed in Table V, our setting $r=5$ can achieve the best balance between accuracy and efficiency.

The proposed FIM includes two key subcomponents, i.e., RFB and LIO. To validate the effectiveness of these two subcomponents, we separately exclude one of them and observe the performance change from another. Experimental results in Table V reveal that both the RFB and LIO benefit the accuracy promotion of SOD. Concretely, the introduction of RFB can

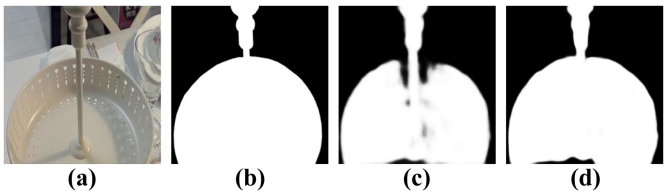


Fig. 15. Validity verification of the application of RFB in our FIM. (a) All-focus image. (b) Ground truth. (c) Result from the model without the RFB. (d) Result from the model with the RFB.

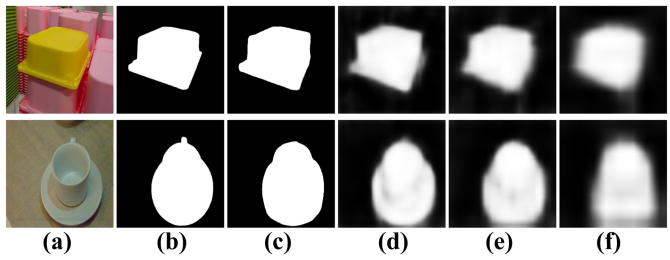


Fig. 16. Illustration of the effectiveness of our PDM. (a) All-focus images. (b) Ground truths. (c) Final results from $\mathbf{h}_2$. (d)-(f) Side outputs from $\{\mathbf{h}_i\}_{i=3}^{5}$.

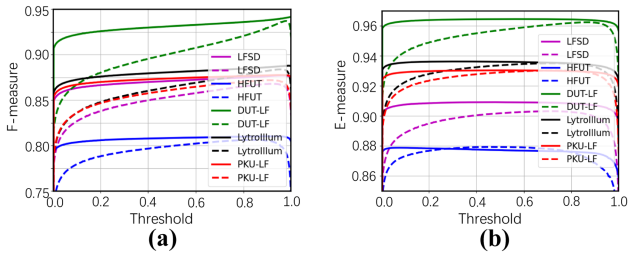


Fig. 17. F-measure and E-measure scores of our model under different thresholds across five datasets. The solid and dashed lines mean results from the cross-entropy loss and compound loss, respectively.

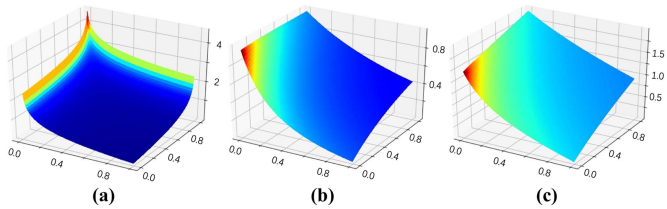


Fig. 18. Surface plots of loss functions in a two-point binary classification problem [69]. (a) Cross-entropy loss. (b) FLoss [69]. (c) Compound loss. Note that the ground truth is {1,0} here.

expand the receptive field of models, enhancing the perception for large objects. Fig. 15 illustrates a typical scene with a large object of the normalized size of 0.59. It can be seen that the model without the RFB fails to produce a satisfactory segmentation result due to the lack of sufficient receptive fields. Since the model with the RFB can capture adequate semantic information of the large object, it enables the production of a more complete saliency map. Furthermore, the proposed LIO can effectively establish long-range dependencies across focal slices, which helps models grasp the global information of the entire focus stack. According to the global information of focal stacks, our model can adaptively learn the weight coefficient of each focal slice, explicitly emphasizing the most task-related focal slice and suppressing the unrelated ones. An example of weight coefficients is visualized in Fig. 10. We can learn that our method can effectively distinguish the focal slices with unequal contributions to the SOD task.

Effectiveness of PDM. Compared with the decoders in previous works, the proposed PDM introduces explicit feature mutual-enhancement among multi-modal and multi-scale features whose mutual consistency can thus be raised. Besides, we recurrently apply PDM three times in the feature decoding phase so that the noisy heterogeneous ingredients among features can be effectively filtered and wrong predictions can be corrected.

The validity verification of our PDM is shown in Table IV (see No.3 and No.4). We can learn that the proposed PDM can further bring significant performance improvements orthogonal to our FIM. Specifically, based on the performance improvements from our FIM, our PDM has 22.22% and 4.11% performance improvements on $\mathcal{M}$ and F_β^{adapt}, respectively. This reveals that the proposed PDM is capable of effectively aggregating the coadjutant multi-modal and multi-scale features. To qualitatively study the efficacy of our PDM, we introduce three additional prediction layers to predict saliency maps from the side outputs $\{\mathbf{h}_i\}_{i=3}^{5}$ and observe the gradual changes in the quality of results. As shown in Fig. 16, as the decoding phase goes on, low-level features with fine-grained structures are gradually utilized to refine the high-level coarse features, and the saliency maps become sharper. Besides, due to the mutual-enhancement among features of different modalities, wrong predictions can be gradually corrected, and the boundaries of salient objects become more accurate.

Effectiveness of Compound Loss. As depicted in Fig. 17, methods supervised by cross-entropy loss are very susceptible to the threshold as the threshold change brings about substantial performance fluctuations. Though this issue can be alleviated by introducing FLoss [69], the training process with FLoss often appears to be unstable because FLoss only considers the overall statistical information and lacks direct pixel-level optimization. Eventually, the performance of models with Floss measured by the pixel-wise matching metric, such as $\mathcal{M}$, worsens (see No.4 and No.5 in Table IV). To this end, we develop the compound loss by introducing direct pixel-level optimization to make up for the flaws of Floss while inheriting the excellent gradient properties of FLoss [69]. To clarify our designs, we illustrate the surface plots of cross-entropy loss, FLoss, and our compound loss in a two-point binary classification problem. As shown in Fig. 18, our compound loss has a similar gradient plot with FLoss. Experiments in Fig. 17 reveal that the performance of models with our compound loss always appears as an almost horizontal line on the upper side, indicating their strong robustness to the thresholds. Moreover, compared with cross-entropy loss and FLoss, our compound loss has the most significant gradient even when the training process is close to the saturation area. Hence, our compound loss can force the networks to achieve higher accuracy (see No. 4, No. 5, and No. 7 in Table IV).

TABLE VI
INFERENCE SPEED COMPARISONS BETWEEN EXISTING 4D SOD METHODS AND OURS

Metric	DLSD [58]	MoLF [21]	ERNet [23]	Ours
FPS	2	8	14	25

Fig. 12 further shows the qualitative result improvements of the proposed model with our compound loss, from which we can learn that our compound loss can help models generate more accurate segmentation results (3rd, 5th, and 8th rows).

Ablation Study of Backbone. Considering that about half of the deep methods in Subsection V-B adopt ResNet-50 as their backbone while the others employ VGG [80], we replace the ResNet-50 backbone in our *STSA* network with the VGG-19 backbone to study the impact of backbone change. As listed in Table IV, the change of backbones has a very slight influence on the performance of the proposed method, demonstrating that our designs are very effective.

E. Efficiency Analysis

Light fields contain abundant spatial information, and light field processing tends to take up unaffordable computing time. When establishing long-range dependencies across focal slices, our FIM merely connects the much smaller adjacent areas via a sliding window. Hence, our FIM significantly reduces the computational complexity and accelerates the inference process. As shown in Table VI, compared with the previous 4D SOD works based on ConvLSTM, the proposed method can achieve an inference speed of 25 frames per second (FPS).

VI. CONCLUSION

In this work, we have presented a comprehensive study on light field SOD by constructing a high-quality dataset, developing a novel *STSA* network, and having a thorough evaluation of 2D, 3D, and 4D methods. Concretely, the proposed dataset, namely PKU-LF, contains diverse indoor and outdoor scenes. We provide rich annotations that make our dataset adaptable to a variety of vision tasks. The proposed *STSA* network achieves a new state-of-the-art performance compared with existing SOD methods. The superiority of our method validates the potential of light field SOD in different application scenarios. To facilitate the progress of SOD, we benchmark 16 representative 2D, 3D, and 4D SOD methods and have an in-depth analysis of them, paving the way for further study.

REFERENCES

[1] J. Zhang et al., "Uncertainty inspired RGB-D saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5761–5779, Sep. 2022.

[2] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

[3] Y. Xu, S. Gao, J. Wu, N. Li, and J. Yu, "Personalized saliency and its prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 12, pp. 2975–2989, Dec. 2018.

[4] H. Zhang, Y. Zeng, H. Lu, L. Zhang, J. Li, and J. Qi, "Learning to detect salient object with multi-source weak supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3577–3589, Jul. 2021.

[5] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for RGB-D salient object detection and beyond," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 5541–5559, Sep. 2022.

[6] M.-M. Cheng, S. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, Nov. 2022.

[7] Y. Jiang, T. Zhou, G.-P. Ji, K. Fu, Q. Zhao, and D.-P. Fan, "Light field salient object detection: A review and benchmark," 2020, *arXiv:2010.04968*.

[8] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[9] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 2976–2983.

[10] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, Apr. 2019.

[11] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Salient object detection with recurrent fully convolutional networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1734–1746, Jul. 2019.

[12] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[13] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9413–9422.

[14] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 10599–10606.

[15] 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," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, May 2021.

[16] M. Zhang, S. X. Fei, J. Liu, S. Xu, Y. Piao, and H. Lu, "Asymmetric two-stream architecture for accurate RGB-D saliency detection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 374–390.

[17] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis.*. Springer, 2020, pp. 275–292.

[18] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 8, pp. 1605–1616, Aug. 2017.

[19] M. Zhang et al., "LFNet: Light field fusion network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6276–6287, 2020.

[20] T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu, "Deep learning for light field saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8838–8848.

[21] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 898–908.

[22] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang, "Light field saliency detection with deep convolutional networks," *IEEE Trans. Image Process.*, vol. 29, pp. 4421–4434, 2020.

[23] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, no. 07, 2020, pp. 11865–11873.

[24] R. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, and P. Hanrahan, "Light field photography with a hand-held plenoptic camera," Ph.D. dissertation, Stanford Univ., 2005.

[25] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui, "Saliency detection on light field: A multi-cue approach," *ACM Trans. Multimedia Comput., Commun. Appl.*, vol. 13, no. 3, pp. 1–22, 2017.

[26] L. Wang et al., "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 136–145.

[27] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," 2015, *arXiv:1506.04214*.

[28] G. Liao, W. Gao, Q. Jiang, R. Wang, and G. Li, "MMNet: Multi-stage and multi-scale fusion network for RGB-D salient object detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 2436–2444.

[29] W. Gao, G. Liao, S. Ma, G. Li, Y. Liang, and W. Lin, "Unified information fusion network for multi-modal RGB-D and RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 2091–2106, Apr. 2021.

[30] D. G. Dansereau, O. Pizarro, and S. B. Williams, "Decoding, calibration and rectification for lenselet-based plenoptic cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 1027–1034.

[31] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2083–2090.

[32] L. Zhang, C. Yang, H. Lu, X. Ruan, and M.-H. Yang, "Ranking saliency," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 9, pp. 1892–1904, Sep. 2017.

[33] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 29–42.

[34] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2014.

[35] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[36] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 447–456.

[37] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 473–480.

[38] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 883–890.

[39] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1265–1274.

[40] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3089–3098.

[41] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 714–722.

[42] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3917–3926.

[43] Y. Zhang, T. Xiang, T. M. Hospedales, and H. Lu, "Deep mutual learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4320–4328.

[44] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. Int. Conf. Internet Multimedia Comput. Service*, 2014, pp. 23–27.

[45] C. Zhu, G. Li, W. Wang, and R. Wang, "An innovative salient object detection using center-dark channel prior," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, 2017, pp. 1509–1515.

[46] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–11.

[47] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *Proc. 24th Brit. Mach. Vis. Conf.*, 2013, pp. 112.1–112.11.

[48] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, "HSCS: Hierarchical sparsity based co-saliency detection for RGBD images," *IEEE Trans. Multimedia*, vol. 21, no. 7, pp. 1660–1671, Jul. 2019.

[49] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[50] P. Huang, C.-H. Shen, and H.-F. Hsiao, "RGBD salient object detection using spatially coherent deep learning framework," in *Proc. IEEE 23rd Int. Conf. Digit. Signal Process.*, 2018, pp. 1–5.

[51] 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," *Pattern Recognit.*, vol. 86, pp. 376–385, 2019.

[52] H. Chen and Y. Li, "Progressively complementarity-aware fusion network for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3051–3060.

[53] H. Chen, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 28, no. 6, pp. 2825–2835, Jun. 2019.

[54] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7253–7262.

[55] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 5216–5223.

[56] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field," in *Proc. Int. Joint Conf. Artif. Intell.*, 2015, pp. 2212–2218.

[57] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu, "Saliency detection via depth-induced cellular automata on light field," *IEEE Trans. Image Process.*, vol. 29, pp. 1879–1889, 2019.

[58] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep light-field-driven saliency detection from a single view," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 904–911.

[59] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, Oct. 2021.

[60] L. Tang, B. Li, Y. Zhong, S. Ding, and M. Song, "Disentangled high quality salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3580–3590.

[61] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 280–287.

[62] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[63] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[64] Z. Huang et al., "CCNet: Criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jul. 03, 2020, doi: 10.1109/TPAMI.2020.3007032.

[65] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[66] S. Liu et al., "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 385–400.

[67] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3907–3916.

[68] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 1597–1604.

[69] K. Zhao, S. Gao, W. Wang, and M.-M. Cheng, "Optimizing the f-measure for threshold-free salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 8849–8857.

[70] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 248–255.

[71] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3472–3481.

[72] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13756–13765.

[73] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "RGB-D salient object detection with cross-modality modulation and selection," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 225–241.

[74] Y. Pang, L. Zhang, X. Zhao, and H. Lu, "Hierarchical dynamic filtering network for RGB-D salient object detection," 2020, *arXiv:2007.06227*.

[75] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 733–740.

[76] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," 2018, *arXiv:1805.10421*.

[77] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4548–4557.

[78] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 1521–1528.

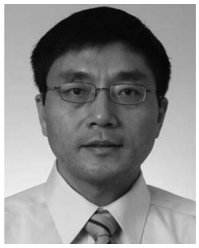
[79] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[80] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

Wei Gao (Senior Member, IEEE) received the PhD degree in computer science from the City University of Hong Kong, Kowloon, Hong Kong, February 2017. From 2012 to 2013, he was a camera ISP engineer with the OmniVision Technologies, Shanghai, China. In 2016, he was a visiting scholar with the University of California, Los Angeles (UCLA), CA, USA. From 2017 to 2019, he worked with the City University of Hong Kong, Hong Kong, and Nanyang Technological University, Singapore. Since 2019, he has been an Assistant Professor with the School of Electronic and Computer Engineering, Peking University, China, and also affiliated to Peng Cheng Laboratory, Shenzhen, China. His research interests include image and video processing, multimedia computing, and deep learning.

Songlin Fan received the BE degree from the School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China, in 2018. He is currently working toward the graduate degree with the School of Electronic and Computer Engineering, Peking University, China, and is also an intern with Peng Cheng Laboratory, Shenzhen, China. His research interests include object segmentation, optimization theory and multimodal learning.

Ge Li (Member, IEEE) received the PhD degree from the Department of Electrical Engineering, Auburn University, AL, USA, in 1999. He was a post-doctoral fellow with the Department of Electrical and Computer Engineering, University of California at Davis, from 2003 to 2004. He was a summer engineer with CSG Research Laboratories, Motorola Inc., Libertyville, IL, USA. After several years of research work in industry, he joined the School of Electronic and Computer Engineering, Peking University, China, in 2014, as a full professor. His general research interests include image/video processing and analysis, machine learning, and signal processing.



Weisi Lin (Fellow, IEEE) received the PhD degree from Kings College London, U.K. He served as the laboratory head of visual processing with Institute for Infocomm Research, Singapore. He is currently a professor with the School of Computer Engineering, Nanyang Technological University, Singapore. His research interests include image processing, perceptual signal modeling, video compression, and multimedia communication, in which he has published 200+ journal articles, more than 230 conference papers, filed 11 patents, and authored two books. He is a fellow of IET and an honorary fellow of the Singapore Institute of Engineering Technologists. He has been the Technical Program chair of the IEEE ICME 2013, the PCM 2012, the QoMEX 2014, and the IEEE VCIP 2017. He has been an associate editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Circuits and Systems for Video Technology*, *IEEE Transactions on Multimedia*, the *IEEE Signal Processing Letters*, and *Journal of Visual Communication and Image Representation*. He has been an invited/panelist/keynote/tutorial speaker for more than 20 international conferences. He was a distinguished lecturer of the IEEE Circuits and Systems Society from 2016 to 2017 and the Asia-Pacific Signal and Information Processing Association (APSIPA) from 2012 to 2013.