Light Field Salient Object Detection with Sparse Views via Complementary and Discriminative Interaction Network

Yilei Chen, Gongyang Li, Ping An, *Member, IEEE*, Zhi Liu, *Senior Member, IEEE*, Xinpeng Huang, and Qiang Wu, *Senior Member, IEEE*

***Abstract*—4D light field data record the scene from multiple views, thus implicitly providing beneficial depth cue for salient object detection in challenging scenes. Existing light field salient object detection (LF SOD) methods usually use a large number of views to improve the detection accuracy. However, using so many views for LF SOD brings difficulties to its practical applications. Considering that adjacent views in a light field are actually with very similar contents, in this work, we propose defining a more efficient pattern of input views, *i.e.*, key sparse views, and design a network to effectively explore the depth cue from sparse views for LF SOD. Specifically, we firstly introduce a low rank-based statistical analysis to the existing LF SOD datasets, which allows us to conclude a fixed yet universal pattern for our key sparse views, including the number and positions of views. These views maintain the sufficient depth cue, but greatly lower the number of views to be captured and processed, facilitating practical applications. Then, we propose an effective solution with a key Complementary and Discriminative Interaction Module (CDIM) for LF SOD from key sparse views, named CDINet. The CDINet follows a two-stream structure to extract the depth cue from the light field stream (*i.e.*, sparse views) and the appearance cue from the RGB stream (*i.e.*, center view), generating features and initial saliency maps for each stream. The CDIM is tailored for inter-stream interaction of both these features and saliency maps, using the depth cue to complement the missing salient regions in RGB stream and discriminate the background distraction, to enhance the final saliency map further. Extensive experiments on three LF multi-view datasets demonstrate that our CDINet not only outperforms the state-of-the-art 2D methods, but also achieves competitive performance as compared with the state-of-the-art 3D and 4D methods. The code and results of our method are available at https://github.com/GilbertRC/LFSOD-CDINet.**

***Index Terms*—Light field, salient object detection, sparse views, complementary and discriminative interaction.**

I. Introduction

SALIENT object detection (SOD) is a fundamental task in computer vision, which aims to locate the most attractive objects in a scene. In the past years, great progress has been achieved in the SOD based on RGB images [1]–[6]. To further improve the detection accuracy in challenging scenes, some recent works [7]–[25] apply the emerging light field data to this task, referred to LF SOD. The light field data provide the depth cue of the scene [26]–[32], being a key supplement to the appearance cue of RGB images for SOD. Currently, there are two main forms of light field data used for LF SOD, *i.e.*, the focal stack [7]–[18] and the multi-view array [19]–[25]. In this paper, we focus on the further exploration of multi-view-based LF SOD.

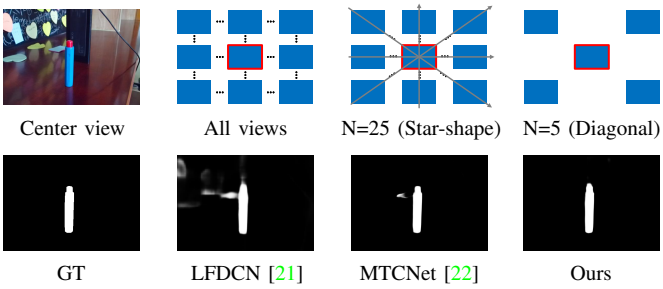


Fig. 1. The input pattern and the predicted saliency maps of LFDCN [21], MTCNet [22] and ours. Our method achieves more accurate result with sparse views (only with five diagonal views). GT represents ground truth, and the red box on each input pattern represents center view.

Different from a traditional RGB image, a multi-view array of light field comprises multiple RGB images (*i.e.*, views) that capture the scene from different perspectives [33]. The disparities among its views can implicitly provide the depth cue. However, for a multi-view-based LF SOD method, using a large number of views may present a challenge to its practical applications. That is, we need a very costly process to capture the required input views. The amount of data to be transmitted and stored will be also increased if we expect to compute saliency maps in cloud terminals. In Fig. 1, we show the two representative multi-view-based methods, *i.e.*, LFDCN [21] and MTCNet [22], including their required input views and the generated saliency maps. In LFDCN [21], all views are selected as their input. For example, for a 7×7 light field, 49 views must be captured. MTCNet [22] reduces the number of input views with a specific star-shaped pattern[1], but generates even better result than LFDCN. This inspires us that it is

This work was supported in part by the National Natural Science Foundation of China under Grant 62020106011, Grant 62171269, Grant 62001279 and Grant 62071287 and in part by the China Postdoctoral Science Foundation under Grant 2022M722037. *(Yilei Chen and Gongyang Li contributed equally to this work.) (Corresponding author: Ping An and Zhi Liu.)*

Yilei Chen, Gongyang Li, Ping An, Zhi Liu, and Xinpeng Huang are with Shanghai Institute for Advanced Communication and Data Science, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: yileichen@shu.edu.cn; ligongyang@shu.edu.cn; anping@shu.edu.cn; liuzhisjtu@163.com; xinpeng_huang@163.com).

Qiang Wu is with the University of Technology Sydney, Ultimo NSW 2007, Australia (e-mail: qiang.wu@uts.edu.au).

[1]This input pattern does not satisfy the definition of sparse views. Sparse views not only mean the fewer number of views, but also require the maximum disparity between the neighboring views larger than 1 pixel [33].

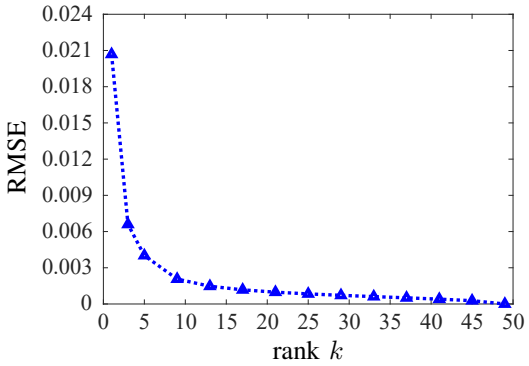


Fig. 2. The average RMSE between the low-rank light field and the original 7×7 light field in the mixed LF SOD dataset [21], [39]. The rank k can be considered as the number of key views for a rank-k light field.

possible to use much fewer views to address the LF SOD, as long as the appropriate views are selected.

In a light field, adjacent views are actually very similar to each other [34]–[36], as they capture the same scene only from slightly different perspectives. Given this fact, we argue that there is a pattern of *key sparse views* (*i.e.*, small number of selected views for providing sufficient depth cue) for LF SOD. To verify this assumption, we apply a low-rank model [37], [38] on the existing LF SOD datasets [21], [39]. It analyses the relationship between light field information and its number of views in the commonly used 7×7 light field. In Fig. 2, it shows the Root Mean Squared Error (RMSE) between the low-rank light field and the original one under different rank k. It reveals that the (approximate) rank of a light field is far less than the original number of views. For a 7×7 light field, selecting five to nine key views can represent most information of the whole forty-nine views. Furthermore, according to the analysis, the key views are located at the five diagonal positions for most light fields. Therefore, we define the views at these five fixed positions as our key sparse views. They greatly reduce the number of required input views but still can achieve expected performance for LF SOD. As shown in Fig. 1, with these key sparse views, we can still achieve accurate result.

Given the above key sparse views, another challenge is how to combine the depth cue[2] from them and the appearance cue from the RGB image[3] effectively for LF SOD. These two cues belong to different modalities, and most deep learning-based methods [14]–[17], [22] deal with the combination via the two-stream fusion strategy. To be specific, a two-stream network is adopted to extract the two cues separately. Then, some methods [15]–[17], [22] focus on fusing the extracted two-stream features (*i.e.*, feature fusion), while others [14] directly generate and fuse the two-stream saliency maps (*i.e.*, result fusion). However, the feature fusion ignores the prior information contained in initial two-stream saliency maps that can judge the validity of features, while the result fusion loses the rich texture and structure information of features. We take both fusion strategies into consideration, proposing comprehensive *feature-result fusion*. That is, we obtain the enhanced cross-modal features by exploiting both two-stream features and initial saliency maps. In this way, we can combine the two cues more effectively.

To this end, we propose our *Complementary and Discriminative Interaction Network* (CDINet) for LF SOD. It follows the two-stream structure, but both features and initial saliency maps are extracted from the light field stream and the RGB stream. For the effective two-stream feature-result fusion, we design a novel *Complementary and Discriminative Interaction Module* (CDIM), which has two ingenious units. **The first unit** is the Maximum-based Complementarity Unit. In this unit, by the pixel-wise maximum comparison between the two initial saliency maps, we can find the location of the missing salient regions in each stream. This location information provided by the two initial maps allows us to use the RGB and light field features in a guided way to highlight more complete salient regions. **The second unit** is the IOU-weighted Discrimination Unit. In this unit, the depth cue from the light field stream is used to eliminate the background distraction in RGB features. We further consider the non-ideal case for this unit, *i.e.*, the salient object is not close to the foreground (invalid depth cue). In this case, the light field-stream saliency map often has no intersection with the RGB-stream saliency map, which can be used as the prior information to assess the validity of the depth cue, and thereby mitigate the negative effects brought by the invalid depth cue on this unit. Using the features obtained by the above two units in CDIM together, our CDINet is able to use much fewer views in a light field and achieves competitive performance compared with 21 state-of-the-art 2D, 3D and 4D methods on three LF multi-view datasets.

Our main contributions are summarized as follows:

- We explore the LF SOD from sparse views for the first time, efficiently reducing the huge number of input views. Accordingly, we propose a novel CDINet to effectively explore the depth cue from sparse views with the comprehensive cross-modal feature-result interaction.
- We introduce a low-rank model in a thorough statistical analysis to select the key sparse views as the network input. Such efficient input pattern can preserve most of the light field information according to their specific number and positions.
- We propose a CDIM for the cross-modal feature-result interaction, which consists of a Maximum-based Complementarity Unit and an IOU-weighted Discrimination Unit. The former one maximizes the complementarity of cross-modal features to detect more complete salient regions, and the latter one discriminates the validity of depth cue before using it to eliminate the background distraction in RGB features.

II. Related Work

In this section, we firstly introduce the multiple representations of 4D light field. Then, we briefly review the existing SOD methods for 2D and 3D contents. Finally, we summarize the development history of 4D LF SOD community.

[2]We extract the depth cue in an unsupervised manner [29], [40] (*i.e.*, taking view synthesis as target) to avoid it being potentially supervised by low-quality depth maps. For more details, please refer to Sec. IV-D.

[3]In multi-view-based LF SOD, the center view is usually taken as a separate RGB image to provide the appearance cue. GT is aligned with it.

A. Light Field Representation

Light field, which is commonly captured by the plenoptic camera [41]–[43] or camera array [44], can simultaneously record the angular and spatial information of the light rays in the scene via its novel 4D representation $L(u, v, x, y) \in \mathbb{R}^{(U \cdot V) \times (X \cdot Y)}$, where (u, v) and (x, y) denote the angular and spatial domains respectively with the resolutions of $(U \cdot V)$ and $(X \cdot Y)$. In the first row of Fig. 3, we show examples of the captured light field, including the lenslet image (*i.e.*, macro-pixel array[4]) and the multi-view array. Lenslet image, which is the data form captured by the plenoptic camera, records the light field with the macro-pixel array. Each macro pixel $L_{x^*,y^*}(u, v)$ records the fixed spatial position (x^*, y^*) from $U \cdot V$ different views, providing the rich angular information of the light rays. Multi-view array, as the more familiar form of the light field, can be obtained by the camera array or split from the lenslet image. Each element $L_{u^*,v^*}(x, y)$ of this form is the normal image captured from a specific angular position (u^*, v^*), enabling the multi-view perception of the scene.

As shown in the second row of Fig. 3, we can subsequently generate the focal stack and depth map from the captured light field. The generating method has been described in many literatures [27], [29], [33], [46], and its core idea is the linear relationship between the spatial-angular position changes of light rays and their depth values. The focal stack comprises a series of images refocused on the specific depth planes, which can reflect the implicit depth cue via the blurriness on each plane. In contrast, the depth map can explicitly give the pixel-wise distance information of the scene.

B. Salient Object Detection for 2D and 3D

In the past decades, many efforts have been made to the 2D SOD methods due to the low-cost acquisition of input data, including traditional and deep learning-based methods [1]–[3]. The core idea of traditional methods is to exploit the prior knowledge like contrast prior [47] and objectness prior [48], [49], which is easily violated in complicated scenes. To alleviate these problems, the deep learning-based methods further integrate the high-level semantic features with diverse network strategies, *e.g.*, deep supervision and short connections [50], guidance of fixation prior [51], cascaded partial decoder [52], three-level hybrid loss [53], context-aware pyramid feature attention [54], interactive two-stream decoder for saliency and contour [55], content-aware guidance for feature learning [56], and spatial attenuation context [57]. However, the well-designed strategies cannot fully bridge the huge gap between 2D content and 4D scene.

Recently, deep learning-based methods on 3D RGB-D SOD have attracted the most research interest for its key supplementation of the depth cue [58], [59]. Considering the different modal information provided by the depth map and the RGB image, numerous strategies [60]–[74] are proposed to explore the complementarity between them, especially the two-stream

[4]The lenslet image as a macro-pixel array only applies to plenoptic 1.0 camera (*e.g.*, Lytro [42]). For plenoptic 2.0 camera (*e.g.*, Raytrix [43]), the captured lenslet image is formed by a micro-image array, but it can also be split into a multi-view array [45].

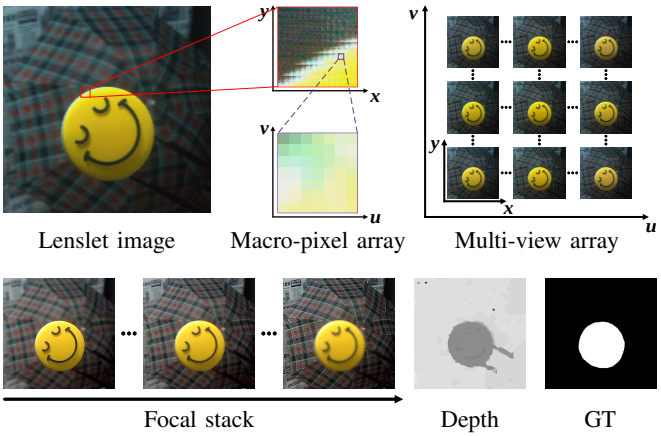


Fig. 3. Examples of light field in the HFUT-Lytro dataset [10]. The 1^{st} row shows the lenslet image (*i.e.*, macro-pixel array) captured by the plenoptic 1.0 camera, and the multi-view array. The 2^{nd} row shows the focal stack and depth map generated from light field, and the ground truth of saliency map.

fusion strategy [64]–[74]. Piao *et al.* [64] introduced a depth-induced multi-scale weighting module to combine the multi-scale context features with depth cue. Li *et al.* [65] designed an information conversion module to fuse the high-level cross-modal features. In [66], they further took the cross-scale features into consideration with a cross-modal weighting module. Fan *et al.* [67] used the initial saliency map generated from the high-level cross-modal features to refine the low-level ones. Li *et al.* [68] fused the cross-modal features progressively in an attention steered interweave manner. Huang *et al.* [69] used bilinear and linear fusion jointly to explore the complementarity between cross-modal features. Liu *et al.* [70] proposed a selective self-mutual attention module to propagate global contexts for both features. In [71], they propagated the global contexts further using multi-head attention in a cross modality transformer. Zhang *et al.* [72] paid attention to the quality of depth maps with the cross-modal attention unit.

Combined with the explicit depth cue provided by the depth map and the above fusion strategies, the SOD performance has achieved great gains. Most existing LF SOD methods [14]–[17], [22] including our CDINet are inspired by the two-stream fusion strategy of 3D SOD. In our solution, however, implicit depth cue is extracted from the light field stream, so that the quality of input depth maps is no longer a limitation, yielding richer perception of the scene.

C. 4D Salient Object Detection

Based on the multiple representation of light field, the existing LF SOD methods can be mainly classified into the focal stack-based methods [7]–[18] and the multi-view-based methods [19]–[25]. More details about the development of LF SOD can be found in the recently published survey paper [75].

Li *et al.* [7], [11] proposed the pioneering work to explore the focal stack for LF SOD. They calculated the focusness cue on each slice and fused them with the color cue from all-focus image (*i.e.*, RGB image) to predict the final saliency map. In [8], they further integrated this focusness cue into a unified saliency detection framework for the heterogeneous

data type. Zhang *et al*. [9] introduced the extra depth contrast saliency from light field to enhance the original color cue. Piao *et al*. [12] constructed an object-guided depth map to fuse the depth, focusness and color cues. Deep learning models were also applied to extract and integrate the focusness and color cues. Wang *et al*. [14] proposed the first deep learning-based method to explore the result fusion for two-stream cues. Zhang *et al*. [15] designed a memory-oriented spatial fusion module for the interaction of two-stream features. In [16], they proposed a light field refinement module to eliminate the homogeneity and refine the dissimilarities between cross-modal features. Liu *et al*. [17] first built dual local graphs to integrate focal stack features with the guidance of the all-focus feature. Then, they proposed fusing the two kinds of features in a recurrent guidance way. Different from the above methods, Piao *et al*. [18] distilled the features from focal stack to RGB image to improve the computational efficiency of LF SOD.

Focal stack simulates the refocusing ability of human vision system, which provides the natural advantage for distinguishing the salient objects from cluttered background. However, this data form loses partial angular information of light field. Zhang *et al*. [10] added the extra multi-view flow cue into their focal stack-based method. To directly explore the multi-view array for LF SOD, Piao *et al*. [19] proposed the first deep learning-based method to fuse the pseudo multi-view saliency maps from a single view. In [20], they extended their work to further exploit the spatial correlation among multi-view images synthesized from the single view. Zhang *et al*. [21] extracted potential depth cues by learning angular changes in each macro pixel of lenslet image. Zhang *et al*. [22] used a multi-task collaborative network to fuse the depth saliency features from star-shaped views with other saliency features. In [25], they explored the depth cue among views via graphs. Recently, two works [23], [24] exploit the depth cue in the multi-view array as a boundary cue for SOD. Jing *et al*. [23] predicted a salient edge map by analyzing the multi-view array's horizontal and vertical epipolar-plane images (EPIs), which was used to refine the RGB features. Wang *et al*. [24] extracted boundary features from macro-EPI forms of the multi-view array and designed a two-stream network with cascaded boundary interaction to fuse these features with the RGB features.

For its lossless representation of light field, in this work, we focus on the further exploration of the multi-view-based LF SOD. Different from [21]–[24], which require many neighboring views work together, we exploit a more sparse but efficient input pattern (*i.e*., the key sparse views) for our CDINet. Besides, our CDINet combines both feature fusion [15]–[17], [22], [24] and result fusion [14] for cross-modal interaction, which we call "feature-result fusion", to explore the implicit depth cue from the key sparse views more effectively.

III. Methodology

In this section, we firstly formulate the target problem in Sec. III-A. Then, we analyze our sparse views selection in Sec. III-B, give the overview of our CDINet in Sec. III-C, and elaborate its key module CDIM in Sec. III-D. Finally, we provide the implementation details in Sec. III-E.

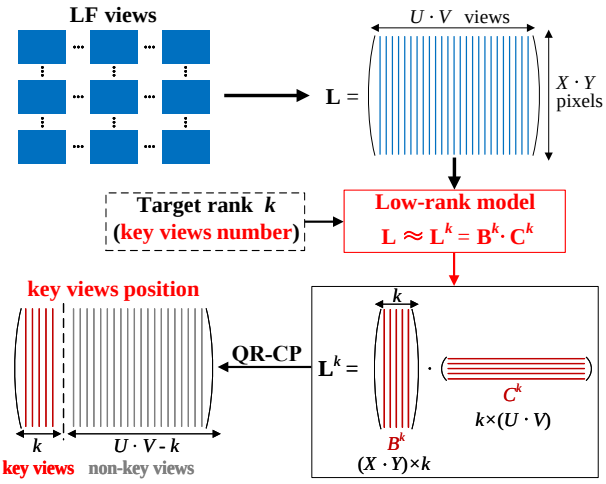


Fig. 4. The processing flow of low rank-based statistical analysis for each light field in the mixed LF SOD dataset [21], [39].

A. Problem Formulation

As described in Sec. II-A, a 4D light field $L(u, v, x, y) \in \mathbb{R}^{(U \cdot V) \times (X \cdot Y)}$ can be represented as a multi-view array:

$$L = \{L_{u,v} \mid 1 \leq u \leq U,\ 1 \leq v \leq V\}, \tag{1}$$

where each view $L_{u,v} \in \mathbb{R}^{X \times Y}$ captures the same scene from a specific angular position (u, v). The disparity between views provides the implicit depth cue of scene [27], [29], which can assist locating the foreground object for LF SOD.

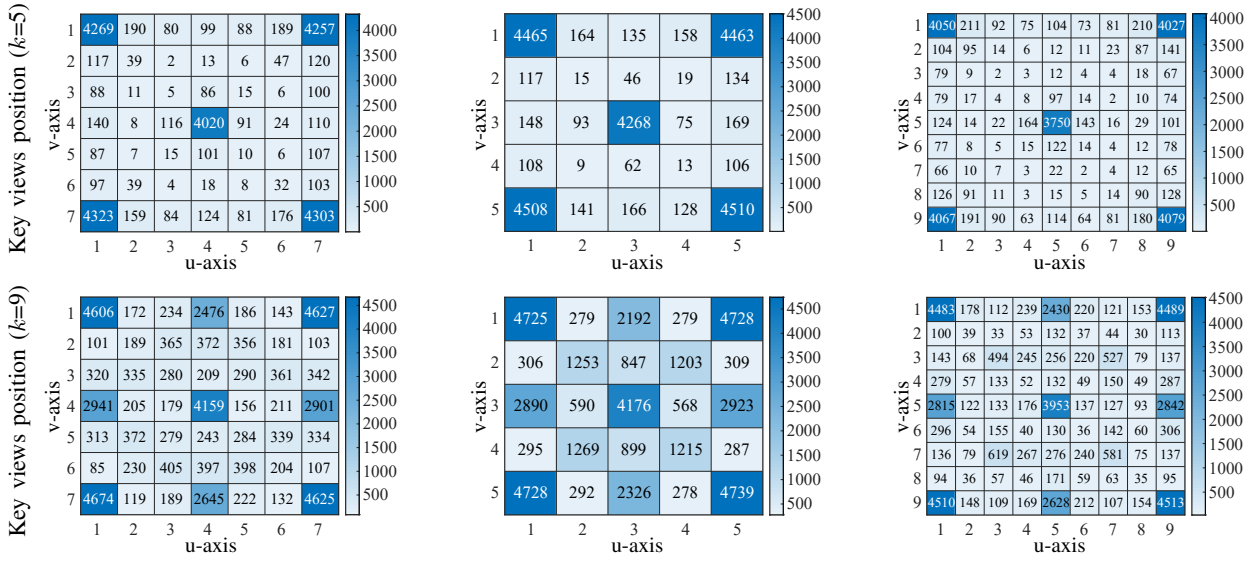
However, since the disparity between neighboring views is very small, the content of $L_{u,v}$ is very similar with each other. Using all $U \cdot V$ views is not only redundant for the extraction of depth cue but also costly for capturing process. Our goal is to find k most contributing views (*i.e*., the key sparse views) from original $U \cdot V$ views ($k \ll U \cdot V$), and design a network to explore the depth cue from these k views for LF SOD. This problem can be formulated as:

$$\boldsymbol{S} = \mathrm{g}(\{L_1, \cdots, L_k\}) \quad \text{s.t. } k \ll U \cdot V, \tag{2}$$

where $\boldsymbol{S}$ is the saliency map, $\{L_1, \cdots, L_k\}$ are the k key sparse views to be selected, and $\mathrm{g}(\cdot)$ is the network which predicts the saliency map from the key sparse views.

B. Key Sparse Views Selection

Considering the obvious similarities among the views in light field, we focus in this subsection on exploring a sparse but efficient input pattern (*i.e*., the key spare views), which enables the fewer input views while contains the sufficient implicit depth cue, for LF SOD. To determine the exact number and positions of these key views, we conduct a thorough low rank-based statistical analysis on a mixed LF SOD dataset, which includes HFUT-Lytro Illum [21] (640 scenes) and DUTLF-V2 [39] (4204 scenes), totaling 4844 scenes.



(a) Results on 7×7 light fields (b) Results on 5×5 light fields (c) Results on 9×9 light fields

Fig. 5. The appearance frequency of key views on each view position of the low-rank light field $\mathbf{L}^k$. We present results on 7×7 light fields, and also present results on 5×5 and 9×9 light fields for comprehensiveness.

As shown in Fig. 4, for each light field, we firstly vectorize its all views, and then stack them into a large light field matrix $\mathbf{L} \in \mathbb{R}^{(X \cdot Y) \times (U \cdot V)}$:

$$\mathbf{L} = \left[vec(L_{1,1}) \mid \cdots \mid vec(L_{u,v}) \mid \cdots \mid vec(L_{U,V}) \right], \quad (3)$$

where $vec(L_{u,v})$ denotes the vectorized form of view $L_{u,v}$, $X \cdot Y$ is the pixels number of each view, and $U \cdot V$ is the total views number in light field. In fact, many views in light field can be approximately represented by the weighted combination of some key views [76], which reveals that the full-rank $\mathbf{L}$ can be expressed in a low-rank way.

Thus, a low-rank model [37], [38] is introduced to complete the low-rank process of $\mathbf{L}$, which depends on the singular value decomposition (SVD) to $\mathbf{L}$:

$$\mathbf{L} = \mathcal{U}\Sigma\mathcal{V}^T = B \cdot C, \quad (4)$$

where we use $B \in \mathbb{R}^{(X \cdot Y) \times (U \cdot V)}$ and $C \in \mathbb{R}^{(U \cdot V) \times (U \cdot V)}$ to denote $\mathcal{U}\Sigma$ and $\mathcal{V}^T$, respectively. Given the target rank k, the low-rank light field $\mathbf{L}^k$ can be obtained as:

$$\mathbf{L} \approx \mathbf{L}^k = B^k \cdot C^k \quad \text{s.t. } rank(\mathbf{L}^k) = k, \quad (5)$$

where $B^k \in \mathbb{R}^{(X \cdot Y) \times k}$ is the first k columns of B, and $C^k \in \mathbb{R}^{k \times (U \cdot V)}$ is the first k rows of C. In this way, the low-rank light field $\mathbf{L}^k$ can be generated from the original full-rank one, which comprises k key views according to its rank.

To determine the required number of key views, we statistically analyze the Root Mean Squared Error (RMSE) value between $\mathbf{L}^k$ and $\mathbf{L}$ with different k in the mixed dataset:

$$\text{RMSE}(\mathbf{L}^k, \mathbf{L}) = \sqrt{\frac{1}{U \cdot V \cdot X \cdot Y} (\mathbf{L} - \mathbf{L}^k)^2}. \quad (6)$$

Then, for each k, we turn to find the positions of k key views in $\mathbf{L}^k$. Mathematically, this is the subset selection problem in a rank-deficient matrix. One solution is to perform the QR decomposition with column pivoting (QR-CP) [77], [78] on the $\mathbf{L}^k$ (*i.e.*, $\mathbf{L}^k\Pi = QR$) to make the diagonal elements in R in a descending order. In this way, the permutation matrix Π moves the k views that are as linearly independent as possible in $\mathbf{L}^k$ to the first k columns of $\mathbf{L}^k\Pi$, and these views are able to represent the rest of $U \cdot V - k$ views better, *i.e.*, they are the desired key views. According to the first k columns of Π, we can obtain their positions in $\mathbf{L}^k$. Note that, for each light field in the mixed dataset, these positions may be different. We count the number of times that each view position is identified as a key view position, denoting it as the appearance frequency of key views at that position, so as to find generalized positions for our key sparse views.

Here, the total views number $U \cdot V$ of each light field is selected as 7×7, and thus rank k varies from 1 to 49. As the RMSE-k curve shown in Fig. 2, $\mathbf{L}^k$ gradually approaches the original matrix $\mathbf{L}$ as target rank k increases. Notably, the RMSE value decreases sharply when rank k increases from 1 to 5, and it tends to level off after k reaches 9. This reveals that five to nine key views are sufficient to represent the original 7×7 light field in a quasi-lossless way. As shown in Fig. 5(a), when rank $k = 5$, although the position of key views varies in each light field according to the scene content, noise, *etc*, the appearance frequency of them on five diagonal positions exceeds that on other positions by a large margin. When $k = 9$, the nine concentrated positions contain the above five diagonal positions, while the extra four positions have a relatively lower frequency. In Fig. 5(b) and 5(c), we also present the statistical results on light fields with different view numbers, *i.e.*, 5×5 and 9×9 light fields. They show a similar pattern to the results obtained from the 7×7 light fields.

According to the above statistical analysis, we observe that the five diagonal positions have the prominent ability to represent the original light field. Therefore, we extract these five diagonal views from the original light field as the explored key sparse views, which form a concise yet efficient input

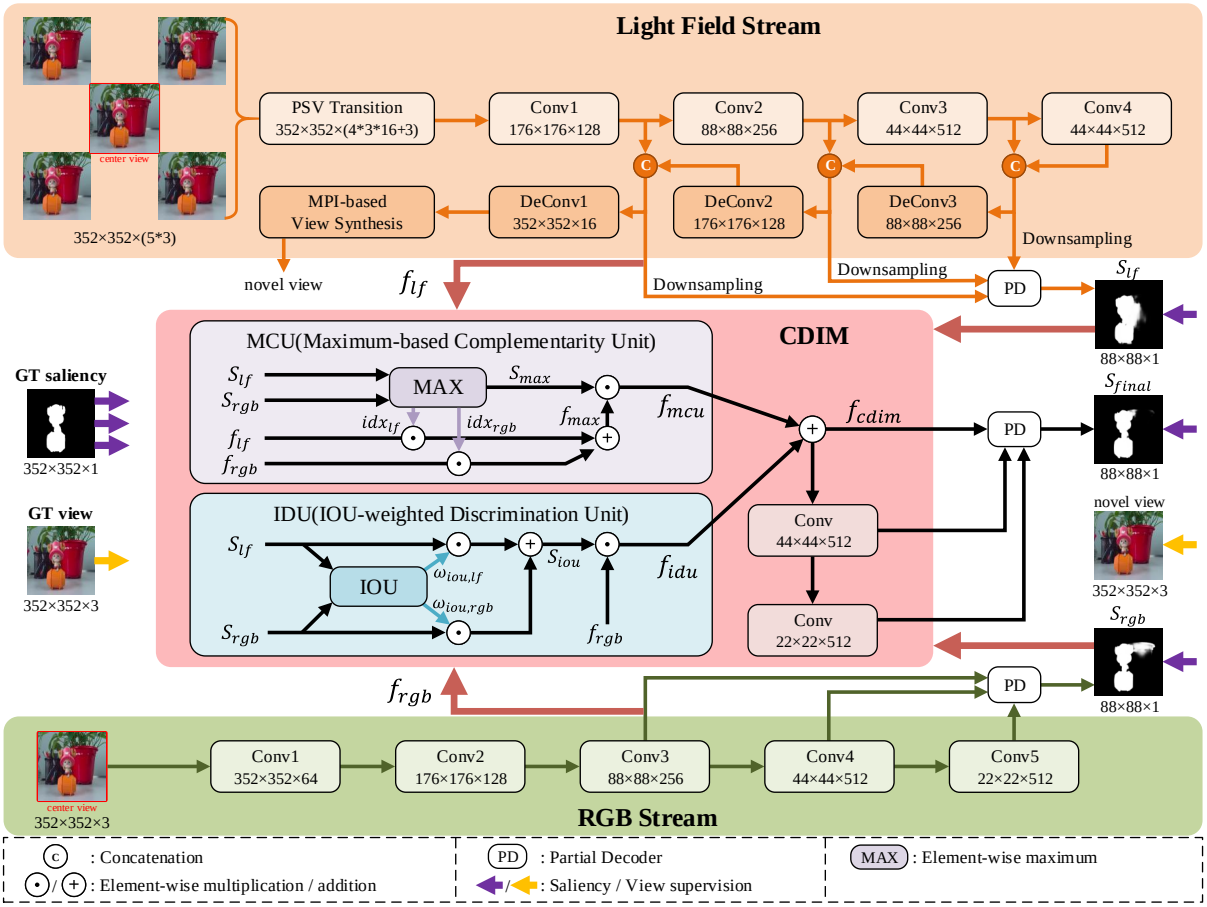


Fig. 6. Pipeline of the proposed CDINet. Our CDINet follows the two-stream structure for features and results extraction, and comprises a Complementary and Discriminative Interaction Module (CDIM) for cross-modal feature-result interaction. First, the light field stream and the RGB stream extract features and initial saliency maps from the sparse views and the center view, respectively. Then, these features and two initial saliency maps are fed to CDIM for complementing cross-modal features with each other via the maximum-based result index and enhancing the RGB features with discriminative depth cue from IOU-weighted (*i.e.*, validity-weighted) result. In our CDINet, all saliency maps are generated by the partial decoder and supervised by the GT. Besides, the light field stream adopts the PSV transition of input views and the extra supervision of novel view synthesis to assist the extraction of light field features. GT view denotes the ground truth of the synthesized novel view.

pattern for our method.

C. Overview of CDINet

As illustrated in Fig. 6, our CDINet comprises the light field stream, RGB stream and the key Complementary and Discriminative Interaction Module (CDIM). Light field stream explores the selected key sparse views via the multiplane-image (MPI) network [40] to extract light field features, and RGB stream employs the VGG-16 [79] backbone on the center view to extract RGB features. The extracted features of light field stream and RGB stream are fed to two partial decoders [52] to generate two initial saliency maps $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$. Subsequently, the extracted features and two saliency maps are fed to CDIM, which contains the parallel Maximum-based Complementarity Unit and IOU-weighted Discrimination Unit, for cross-modal feature-result interaction. Finally, the output features of CDIM are sent to the third partial decoder for generating the final saliency map $\boldsymbol{S}_{final}$.

Specifically, the MPI network adopted in light field stream is used to extract the implicit depth cue from the key sparse views. This network was originally developed to generate the MPI (*i.e.*, a set of fronto-parallel RGBα images at multiple depth planes) for scene representation, and it is supervised by view synthesis. The characteristic of scene representation makes the features from the MPI network contain the depth cue required by LF SOD. In our implementation, we follow its pipeline to project each non-center sparse view to the center view position under multiple depth planes, forming a 3D tensor input, *i.e.*, the plane-sweep-volume (PSV). Then, in MPI-based view synthesis, we synthesize a randomly selected non-key view (*i.e.*, novel view) with the MPI representation and use the ground truth of that novel view for supervision. More details about the MPI network can refer to [40].

D. Complementary and Discriminative Interaction Module

In CDIM, the sizes of input features (*i.e.*, $\boldsymbol{f}_{lf}$ and $\boldsymbol{f}_{rgb}$) and saliency maps (*i.e.*, $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$) are defined as $\mathbb{R}^{h\times w\times c}$ and $\mathbb{R}^{h\times w\times 1}$, respectively. A naive way for the cross-modal feature-result interaction is the direct concatenation-convolution operation, which ignores the explicit guidance of two-stream results for features. To fully explore the prior of initial results, we propose the CDIM, which contains two novel

units, *i.e.*, Maximum-based Complementarity Unit (MCU) and IOU-weighted Discrimination Unit (IDU). The details of CDIM are shown in the middle of Fig. 6, and we elaborate it as follows.

Maximum-based Complementarity Unit. A typical case in natural scene is that the salient object is composed of different colors, denoted as the different-looking regions. In this case, RGB features $\boldsymbol{f}_{rgb}$ may cause incomplete detection in $\boldsymbol{S}_{rgb}$. Thus, in the MCU, we aim at complementing these missing regions by $\boldsymbol{f}_{lf}$ and $\boldsymbol{S}_{lf}$ based on their depth cue, which is promising to generate more complete result.

To highlight all detected regions from two initial saliency maps, we first take the maximum of $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$ to generate the stitched map $\boldsymbol{S}_{max} \in \mathbb{R}^{h\times w\times 1}$, denoted as follows:

$$\boldsymbol{S}_{max} = \mathrm{MAX}(\boldsymbol{S}_{lf}, \boldsymbol{S}_{rgb}), \tag{7}$$

where $\mathrm{MAX}(\cdot)$ is the element-wise maximum. This operation is used to reveal the pixel-level contributions of each modal features, which can in turn guide the complementarity of them.

Then, we backtrack the source index (*i.e.*, $\boldsymbol{S}_{lf}$ or $\boldsymbol{S}_{rgb}$) of each value in $\boldsymbol{S}_{max}$, getting two binary index matrices $\{\boldsymbol{idx}_{lf}, \boldsymbol{idx}_{rgb}\} \in \mathbb{R}^{h\times w\times 1}$. For each pixel (x, y) in $\boldsymbol{idx}_{lf}(x, y)$, the value is defined as:

$$\boldsymbol{idx}_{lf}(x,y) = \begin{cases} 0, & \boldsymbol{S}_{max}(x,y) = \boldsymbol{S}_{rgb}(x,y) \\ 1, & \boldsymbol{S}_{max}(x,y) = \boldsymbol{S}_{lf}(x,y) \end{cases}. \tag{8}$$

Notably, $\boldsymbol{idx}_{lf}(x, y)$ is set to 0 if $\boldsymbol{S}_{rgb}(x, y) = \boldsymbol{S}_{lf}(x, y)$, and $\boldsymbol{idx}_{rgb} = 1-\boldsymbol{idx}_{lf}$. The obtained $\boldsymbol{idx}_{lf}$/$\boldsymbol{idx}_{rgb}$ provides the explicit index information of the larger saliency response at each pixel position, making up for the missing regions in $\boldsymbol{S}_{lf}$/$\boldsymbol{S}_{rgb}$.

According to $\boldsymbol{idx}_{lf}$ and $\boldsymbol{idx}_{rgb}$, we achieve the cross-modal complementarity feature $\boldsymbol{f}_{max} \in \mathbb{R}^{h\times w\times c}$, which can be defined as follows:

$$\boldsymbol{f}_{max} = \left(\boldsymbol{idx}_{lf} \odot \boldsymbol{f}_{lf}\right) \oplus \left(\boldsymbol{idx}_{rgb} \odot \boldsymbol{f}_{rgb}\right), \tag{9}$$

where $\odot$ is the element-wise multiplication, and $\oplus$ is the element-wise addition. This operation transfers the information of maximum index from two-stream saliency maps to cross-modal features, enhancing the feature-result interaction.

To further suppress the background distraction in $\boldsymbol{f}_{max}$, we adopt the stitched map $\boldsymbol{S}_{max}$ to purify $\boldsymbol{f}_{max}$, and obtain the output feature of MCU (*i.e.*, $\boldsymbol{f}_{mcu} \in \mathbb{R}^{h\times w\times c}$) as follows:

$$\boldsymbol{f}_{mcu} = \boldsymbol{S}_{max} \odot \boldsymbol{f}_{max}. \tag{10}$$

With the above maximum-based feature-result complementarity, $\boldsymbol{f}_{mcu}$ can complement the missing salient regions, containing the whole saliency information from $\boldsymbol{f}_{lf}$ and $\boldsymbol{f}_{rgb}$.

IOU-weighted Discrimination Unit. The depth cue implied in $\boldsymbol{S}_{lf}$ and $\boldsymbol{f}_{lf}$ plays a key role in suppressing the redundant background distraction in $\boldsymbol{S}_{rgb}$ detected by RGB features $\boldsymbol{f}_{rgb}$. And $\boldsymbol{S}_{lf}$ exhibits more explicit location of objects than $\boldsymbol{f}_{lf}$ according to its expression form, which is an intuitive probability map. Thus, a valid $\boldsymbol{S}_{lf}$ has the ability to highlight the foreground object in the final result. Similar to some work [72], [73] on 3D SOD, which focuses on the quality of depth map, the validity of $\boldsymbol{S}_{lf}$ needs to be carefully estimated now, avoiding its negative effects on the final results. Considering the fact that the valid $\boldsymbol{S}_{lf}$ always shares the common foreground object with $\boldsymbol{S}_{rgb}$, in the IDU, we exploit the intersection relationship between $\boldsymbol{S}_{rgb}$ and $\boldsymbol{S}_{lf}$ as a soft index to estimate the validity of $\boldsymbol{S}_{lf}$, and bring the valid depth cue from $\boldsymbol{S}_{lf}$ to enhance $\boldsymbol{f}_{rgb}$.

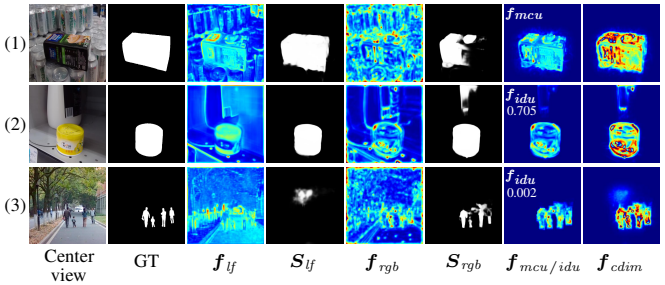


Fig. 7. Feature visualization in CDIM. $\boldsymbol{f}_{lf}$ and $\boldsymbol{f}_{rgb}$ are the input cross-modal features of CDIM, while $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$ are the input two-stream saliency maps. $\boldsymbol{f}_{mcu}$ and $\boldsymbol{f}_{idu}$ are the output features of MCU and IDU, respectively. $\boldsymbol{f}_{cdim}$ is the output feature of CDIM. The value in $\boldsymbol{f}_{idu}$ is the validity weight computed from the IOU between $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$.

Concretely, we novelly adopt the value of intersection-over-union (IOU) between $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$ as the validity weight $\boldsymbol{\omega}_{iou,lf} \in [0, 1]$ for $\boldsymbol{S}_{lf}$, computed as follows:

$$\boldsymbol{\omega}_{iou,lf} = \frac{|\boldsymbol{S}_{lf} \cap \boldsymbol{S}_{rgb}|}{|\boldsymbol{S}_{lf} \cup \boldsymbol{S}_{rgb}|}. \tag{11}$$

Notably, $1 - \boldsymbol{\omega}_{iou,lf}$ is denoted as $\boldsymbol{\omega}_{iou,rgb}$. In this way, the obtained $\boldsymbol{\omega}_{iou,lf}$/$\boldsymbol{\omega}_{iou,rgb}$ can confirm the validity weight of two-stream saliency maps $\boldsymbol{S}_{lf}$/$\boldsymbol{S}_{rgb}$, and facilitate fusing the valid depth cue from $\boldsymbol{S}_{lf}$ into the validity-weighted map $\boldsymbol{S}_{iou} \in \mathbb{R}^{h\times w\times 1}$, defined as follows:

$$\boldsymbol{S}_{iou} = \left(\boldsymbol{\omega}_{iou,lf} \odot \boldsymbol{S}_{lf}\right) \oplus \left(\boldsymbol{\omega}_{iou,rgb} \odot \boldsymbol{S}_{rgb}\right). \tag{12}$$

Specifically, if $\boldsymbol{S}_{lf}$ has the intersection region with $\boldsymbol{S}_{rgb}$ (*i.e.*, $\boldsymbol{\omega}_{iou,lf} > 0$), we consider it as the valid depth cue. In this time, in $\boldsymbol{S}_{iou}$, the value of the above intersection region is higher than other regions, which can distinguish the salient regions. Otherwise, if $\boldsymbol{S}_{lf}$ has no intersection with $\boldsymbol{S}_{rgb}$ (*i.e.*, $\boldsymbol{\omega}_{iou,lf} = 0$ and $\boldsymbol{\omega}_{iou,rgb} = 1$), we consider it as the invalid depth cue and Eq. 12 degenerates to $\boldsymbol{S}_{iou} = \boldsymbol{S}_{rgb}$. Generally, the obtained $\boldsymbol{S}_{iou}$ identifies the valid depth cue from $\boldsymbol{S}_{lf}$, and encourages the subsequent feature-result interaction.

Finally, to bring the valid depth cue to $\boldsymbol{f}_{rgb}$, we combine the validity-weighted map $\boldsymbol{S}_{iou}$ with $\boldsymbol{f}_{rgb}$, and obtain the output feature of IDU (*i.e.*, $\boldsymbol{f}_{idu} \in \mathbb{R}^{h\times w\times c}$) as follows:

$$\boldsymbol{f}_{idu} = \boldsymbol{S}_{iou} \odot \boldsymbol{f}_{rgb}. \tag{13}$$

With the above IOU-weighted (*i.e.*, validity-weighted) feature-result discrimination, $\boldsymbol{f}_{idu}$ can exploit the depth cue from $\boldsymbol{S}_{lf}$ without disturbance to locate the salient object accurately. As shown in Fig. 6, by adding the above $\boldsymbol{f}_{idu}$ with $\boldsymbol{f}_{mcu}$, we obtain the final output feature of CDIM (*i.e.*, $\boldsymbol{f}_{cdim} \in \mathbb{R}^{h\times w\times c}$), which comprehensively integrates appearance cue and depth cue from the RGB stream and the light field stream, to generate the final saliency map via the third partial decoder.

In Fig. 7, the penultimate column depicts features of MCU (1^{st} row) and of IDU (2^{nd} and 3^{rd} rows). Concretely, in the first row of Fig. 7, $\boldsymbol{f}_{rgb}$ detects the incomplete salient object in $\boldsymbol{S}_{rgb}$ due to different-looking regions. After the feature-result complementarity in MCU, these missing regions are successfully complemented with $\boldsymbol{f}_{lf}$ in $\boldsymbol{f}_{mcu}$. In the last two rows of Fig. 7, we show the case of valid and invalid $\boldsymbol{S}_{lf}$ separately. When $\boldsymbol{S}_{lf}$ provides valid depth cue, as shown in the second row, its validity weight is high, *i.e.*, 0.705, which can highlight the salient object from distracted background in $\boldsymbol{f}_{idu}$. When $\boldsymbol{S}_{lf}$ provides invalid depth cue, as shown in the third row, its validity weight is very low, *i.e.*, 0.002, which can be fully discriminated with $\boldsymbol{S}_{rgb}$ in $\boldsymbol{f}_{idu}$. As shown in the last column of Fig. 7, in $\boldsymbol{f}_{cdim}$, the salient regions of these three examples are accurately highlighted with depth cue through our novel cross-modal feature-result interaction in CDIM.

E. Implementation Details

1) Total Loss. As shown in Fig. 6, the generated two-stream saliency maps $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$ and final saliency map $\boldsymbol{S}_{final}$ are supervised by the GT saliency map $\mathbf{G}_s$ with Binary Cross-Entropy (BCE) loss. The synthesized view $\boldsymbol{L}_{novel}$ in light field stream is supervised by the GT view $\mathbf{G}_v$ with L_1 loss. The total loss function is defined as:

$$loss = \sum_i \left(\mathcal{L}_{bce}(up(\boldsymbol{S}_i), \mathbf{G}_s)\right) + \mathcal{L}_1(\boldsymbol{L}_{novel}, \mathbf{G}_v), \quad (14)$$

where $i \in \{lf, rgb, final\}$, $\mathcal{L}_{bce}(\cdot,\cdot)$ denotes the BCE loss, $\mathcal{L}_1(\cdot,\cdot)$ denotes the L_1 loss, and $up(\cdot)$ denotes the upsampling operation that all generated saliency maps are upsampled to the resolution of $\mathbf{G}_s$ by bilinear interpolation.

2) Training Setting. The proposed CDINet is implemented by TensorFlow with a Tesla P100 GPU. Following [21], [22], we choose 512 samples from HFUT-Lytro Illum [21] as our training set, and select the angular resolution of light field as 7×7. We augment them with rotation, flipping and brightness variations, and resize their spatial resolution to 352×352. Concretely, the augmenting process is implemented on the lenslet image, which can be considered as a 4D augmentation, and the rotation is limited to 90°, 180° and 270° to avoid leaking the information from one view into other views. For resizing of light field, we resize each view separately, and then adjust their angular positions with the same ratio. The parameters of light field stream and RGB stream are initialized from pre-trained MPI [40] and VGG-16 [79] model, respectively, and the newly added convolution layers are initialized by Xavier [80]. The number of depth planes in MPI network is set to 16. For other hyper parameters, we set the batch size to 4, and set the initial learning rate to 10^{-4}. The learning rate is divided by 10 every 30 epochs. We adopt the Adam optimizer [81] to train our CDINet for 40 epochs.

IV. Experimental Results

A. Datasets and Evaluation Metrics

1) Datasets. To evaluate the performance of the proposed method, we conduct experiments on three public LF multi-view SOD datasets: HFUT-Lytro Illum [21], HFUT-Lytro [10] and DUTLF-V2 [39]. The details of these datasets are described as follows.

HFUT-Lytro Illum[5] [21]: This dataset contains 640 high-quality light fields captured by a Lytro Illum camera. The provided light field form is the lenslet image, which can be split to the required multi-view array. We also generate the depth maps with the Lytro Power Tools[6]. In this dataset, 512 samples are used as the training set, and the remaining samples are used for test.

HFUT-Lytro[7] [10]: This dataset consists of 255 light fields captured by a Lytro first-generation camera, which provides all-focus images, focal stacks, depth maps, multi-view arrays as well as the corresponding ground truth. It is a challenging dataset, including small / multiple objects, various spatial distribution, *etc*. Some focal stack-based methods [15], [16] can be implemented on this dataset for comparison.

DUTLF-V2[8] [39]: This large-scale dataset is recently released for versatile 2D, 3D and 4D SOD, which is the updated version of DUTLF-FS [14] and DUTLF-MV [19]. It contains 102 classes and 4204 light fields captured by a Lytro Illum camera, and offers all-focus images, focal stacks, depth maps, multi-view arrays and the ground truth.

2) Evaluation Metrics. We employ five widely used metrics with the evaluation toolbox[9] to assess the predicted saliency maps. **E-measure** ($\mathcal{E}_\xi$) [82] is the metric proposed for binary map, which considers the local pixel-level matching and global image-level statistics information. **S-measure** ($\mathcal{S}_\lambda$, $\lambda = 0.5$) [83] is proposed for evaluating both region-aware and object-aware structural similarity between the predicted saliency maps and ground truth. **Weighted F-measure** ($\mathcal{F}_\beta^w$, $\beta^2 = 1$) [84] is the extended metric from **F-measure** ($\mathcal{F}_\beta$, $\beta^2 = 0.3$) [85]. F-measure is the harmonic mean of precision and recall for binary map, in which the precision is emphasized over recall. Weighted F-measure takes the location of errors into consideration, which provides the extra weight for precision and recall. **Mean Absolute Error** (MAE, $\mathcal{M}$) is the average pixel-wise errors. In this paper, we adopt the adaptive threshold for computing the E-measure and F-measure.

B. Ablation Analysis

To analyze the impact of our proposed components on the final performance of CDINet, we mainly investigate: 1) the number of key sparse views; 2) the positions of key sparse views; 3) the universality of key sparse views; 4) the effectiveness of view supervision and PSV transition in light field stream; 5) the importance of CDIM, MCU and IDU in CDINet; and 6) the rationality of MCU and IDU in CDIM. In this section, the ablation analyses are conducted on the challenging HFUT-Lytro [10], and all variants are retrained with the same setting in Sec. III-E.

1) Number of Key Sparse Views. In Sec. III-B, we select five diagonal views as the explored key sparse views, which

[5] https://github.com/pencilzhang/MAC-light-field-saliency-net
[6] https://github.com/kmader/lytro-power-tools
[7] https://github.com/pencilzhang/HFUT-Lytro-dataset
[8] https://github.com/OIPLab-DUT/DUTLF-V2
[9] https://github.com/jiwei0921/Saliency-Evaluation-Toolbox

TABLE I
ABLATION ANALYSES FOR THE SELECTION OF KEY SPARSE VIEWS IN CDINET. WE PROVIDE EIGHT SOLUTIONS OF VIEW SELECTION WITH 1, 3, 5 (OURS), 9, 9-DIAGONAL, 5 (PATTERN-A), 5 (PATTERN-B) AND 5-RANDOM VIEWS. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

Views number	HFUT-Lytro [10]				
	$\mathcal{E}_\xi \uparrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$
1	0.7970	0.7504	0.6342	0.6823	0.0904
3	0.7974	0.7539	0.6366	0.6867	0.0870
5 (Ours)	**0.8187**	**0.7650**	**0.6587**	**0.7058**	**0.0837**
9	0.7993	0.7502	0.6388	0.6878	0.0907
9-diagonal	0.7963	0.7509	0.6280	0.6818	0.0883
5 (pattern-A)	0.8113	0.7552	0.6396	0.6894	0.0858
5 (pattern-B)	0.8047	0.7553	0.6391	0.6869	0.0870
5-random	0.8126	0.7564	0.6468	0.6958	0.0869

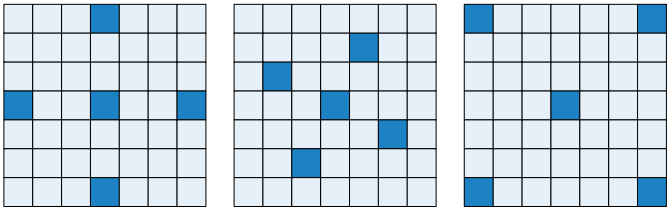


Fig. 8. Three input patterns of five views. 5 (pattern-A) and 5 (pattern-B) are the two input patterns as suggested in [86], and 5 (Ours) is the explored five diagonal views.

form a sparse but efficient input pattern for our CDINet. In this analysis, we change the key views number to "1", "3" and "9", respectively. Concretely, views number "1" denotes that we select the center view as input for both two streams, views number "3" denotes selecting the left-top, right-bottom and center views as the key sparse views, and views number "9" denotes the input pattern shown in Fig. 5(a) that we add the extra four most vertical and horizontal views. We also select nine diagonal views for comparison, named "9-diagonal".

As listed in the first five rows of Tab. I, we observe that the CDINet with our five diagonal views achieves the best performance among all selections. Notably, the performance is improved gradually when views number increases from "1" to "5", which is inline with the RMSE-k curve shown in Fig. 2. However, when we continue increasing the views number to "9", the performance is degraded. The reason is that the selected five diagonal views contain sufficient implicit depth cue for LF SOD, while adding more redundant views contributes less to the performance but brings more interference information for SOD. This ablation result indicates the appropriateness of selecting five views as our key sparse views.

2) Positions of Key Sparse Views. To validate the superiority of our explored pattern (*i.e.*, diagonal positions), we retrain our CDINet tailored for other two representative patterns of five views as suggested in [86]. The patterns of them are shown in Fig. 8, and denoted as "5 (pattern-A)" and "5 (pattern-B)", respectively. Besides, we also retrain the model to accept the

TABLE II
ABLATION ANALYSES FOR THE UNIVERSALITY OF KEY SPARSE VIEWS. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

Models	HFUT-Lytro [10]				
	$\mathcal{E}_\xi \uparrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$
MTCNet-25 [22]	**0.8178**	0.7502	0.6338	**0.7142**	0.0845
MTCNet-5	0.8121	**0.7625**	**0.6482**	0.7036	**0.0830**

TABLE III
ABLATION RESULTS OF NOVEL VIEW SUPERVISION, PSV TRANSITION AND PRE-TRAINED MPI PARAMETERS IN LIGHT FIELD STREAM. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

Models	HFUT-Lytro [10]				
	$\mathcal{E}_\xi \uparrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$
w/o VS	0.7981	0.7545	0.6375	0.6926	0.0870
w/o PSV	0.8048	0.7596	0.6496	0.7017	0.0862
w/o Pre_MPI	0.8091	0.7593	0.6453	0.6989	0.0851
Ours	**0.8187**	**0.7650**	**0.6587**	**0.7058**	**0.0837**

randomly selected input patterns, denoted as "5-random". To evaluate its performance, we randomly select 10 patterns for each sample, and report the average result of them.

As listed in the last three rows of Tab. I, the ablation results show the superiority of the explored diagonal positions. With the same number of views, the diagonal positions can provide more depth cue than "5 (pattern-A)" and "5 (pattern-B)" for LF SOD, and is also more effective than the random pattern. With both explored number and positions, our key sparse views form an efficient input pattern for the subsequent CDINet.

3) Universality of Key Sparse Views. To further verify the universality of our key sparse views, we apply them to the LF SOD method MTCNet [22]. We replace its original star-shaped 25 views with our key sparse views (only 5 views), denoted as "*MTCNet-5*". We also refer to the original MTCNet as "*MTCNet-25*". The results are listed in Tab. II. The two models exhibit comparable performance. It demonstrates that our key sparse views maintain the valuable information of the original light field for LF SOD, and they potentially have universality in other methods.

4) Effectiveness of View Supervision and PSV Transition in Light Field Stream. The MPI network in the light field stream extracts the depth cue by supervising the novel view synthesis. The PSV transition of input views also assists in this extraction process. To verify their effectiveness individually, we remove the novel view supervision (including MPI-based view synthesis and all DeConv layers), denoted as "*w/o VS*", and the PSV transition, denoted as "*w/o PSV*", respectively. Tab. III shows the results. Since both are critical to depth cue extraction, they contribute significantly to the SOD performance.

In this work, we initialize the parameters of the light field stream from pre-trained MPI [40]. To also validate this choice, we design a variant to train our light field stream from scratch, denoted as "*w/o Pre_MPI*". Tab. III lists the result. It indicates that the initialization from pre-trained MPI is important. It can

TABLE IV
ABLATION RESULTS ON CONFIRMING THE IMPORTANCE OF CDIM, MCU AND IDU IN CDINET. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

Models	HFUT-Lytro [10]				
	$\mathcal{E}_\xi \uparrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$
w/o CDIM	0.7852	0.7288	0.5993	0.6767	0.0935
w/o MCU	0.7981	0.7446	0.6278	0.6821	0.0907
w/o IDU	0.7919	0.7442	0.6139	0.6552	0.0940
Ours	**0.8187**	**0.7650**	**0.6587**	**0.7058**	**0.0837**

ease the training process for our light field stream in extracting depth cues from such sparse input views.

5) Importance of CDIM, MCU and IDU in CDINet. To evaluate the importance of the key module CDIM for our CDINet, we replace it with the concatenation-convolution operation for the cross-modal feature-result interaction, which is denoted as "*w/o CDIM*". Then, to further examine the importance of proposed MCU and IDU for CDINet, we remove each of them separately and represent these two variants as "*w/o MCU*" and "*w/o IDU*".

The comparison results are listed in Tab. IV. We can observe that "*w/o CDIM*" achieves the worst performance especially in the global metrics $\mathcal{S}_\lambda$ and $\mathcal{F}_\beta^w$. The reason is that the simple concatenation-convolution operation cannot actually perceive the guidance of initial two-stream results to features, causing inaccurate global structure of salient object. When removing the IDU, "*w/o IDU*" has no ability to suppress the background distraction, causing 4.5% loss in $\mathcal{F}_\beta^w$ and 5.1% loss in $\mathcal{F}_\beta$ compared with our complete CDINet. When removing the MCU, "*w/o MCU*" also causes 2.0% loss in $\mathcal{S}_\lambda$ and 3.1% loss in $\mathcal{F}_\beta^w$. The reason is that MCU plays a key role in highlighting all salient regions detected by two streams in the final result. With both MCU and IDU (*i.e.*, the complete CDIM), our method achieves the best performance.

6) Rationality of MCU and IDU in CDIM. To further validate the rationality of MCU and IDU in CDIM, we design three variants for MCU and three variants for IDU. The structure of these variants are shown in Fig. 9, and the ablation results are listed in Tab. V.

In MCU, the maximum-based operation is employed on the two-stream initial results to obtain the key index matrix, which is then used as the guidance for the cross-modal features complementarity. As shown in the first row of Fig. 9, in its first variant, we remove both the maximum-based operation and the guidance of index matrix, and replace them with the element-wise addition, which is denoted as "*w/o MAX-index-A*". Also, we replace the maximum-based operation with the element-wise multiplication, which is denoted as "*w/o MAX-index-M*". To further explore the rationality of index matrix, we remove this step separately, which is denoted as "*w/o index*".

The ablation results of these MCU variants are listed in the first three rows of Tab. V. We can observe that "*w/o MAX-index-A*" damages the performance especially in $\mathcal{S}_\lambda$ and $\mathcal{F}_\beta^w$. The first reason is that the direct addition between two-stream initial results actually emphasizes on their common

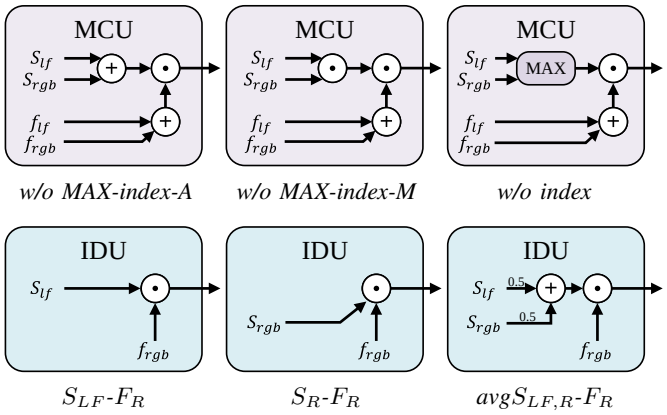


Fig. 9. Variants of MCU / IDU in CDIM.

TABLE V
ABLATION RESULTS ON VALIDATING THE RATIONALITY OF MCU AND IDU IN CDIM. THE BEST RESULT IN EACH COLUMN IS **BOLD**.

Models	HFUT-Lytro [10]				
	$\mathcal{E}_\xi \uparrow$	$\mathcal{S}_\lambda \uparrow$	$\mathcal{F}_\beta^w \uparrow$	$\mathcal{F}_\beta \uparrow$	$\mathcal{M} \downarrow$
w/o MAX-index-A	0.7845	0.7269	0.6019	0.6756	0.0961
w/o MAX-index-M	0.7756	0.6986	0.5645	0.6626	0.1070
w/o index	0.7934	0.7492	0.6320	0.6884	0.0907
S_{LF}-F_R	0.7781	0.6984	0.5491	0.6667	0.0946
S_R-F_R	0.8122	0.7599	0.6479	0.6963	0.0869
$avgS_{LF,R}$-F_R	0.8095	0.7523	0.6331	0.6876	0.0881
Ours	**0.8187**	**0.7650**	**0.6587**	**0.7058**	**0.0837**

regions but does not complement the missing regions. In this way, the salient response of the missing regions is lower than that of the common regions, degrading the complementary effect. Another reason is that the direct addition between cross-modal features may weaken the saliency information which is already extracted by one of them. "*w/o MAX-index-M*" shows that the direct multiplication between two-stream initial results is also worse than the maximum-based operation. When we add back the maximum-based operation in MCU, "*w/o index*" alleviates the performance loss, especially achieving 3.0% gain in $\mathcal{F}_\beta^w$ compared with "*w/o MAX-index-A*". However, due to the lack of index matrix transferred from the two-stream results, the cross-modal features complementarity cannot be fully explored. Compared to these three variants, our complete version of MCU effectively encourages the feature-result complementarity, complementing the missing salient regions to the maximum extent.

To also validate the rationality of IDU, we design three variants for its key validity weight $\boldsymbol{\omega}_{iou,lf}$ through changing its value to 1, 0 and 0.5, respectively. As shown in the second rows of Fig. 9, the first variant "S_{LF}-F_R" denotes $\boldsymbol{\omega}_{iou,lf} = 1$, which equals to enhancing RGB features $\boldsymbol{f}_{rgb}$ with $\boldsymbol{S}_{lf}$; the second variant "S_R-F_R" denotes $\boldsymbol{\omega}_{iou,lf} = 0$, which equals to enhancing $\boldsymbol{f}_{rgb}$ with $\boldsymbol{S}_{rgb}$; and the third variant "$avgS_{LF,R}$-F_R" denotes $\boldsymbol{\omega}_{iou,lf} = 0.5$, which equals to enhancing $\boldsymbol{f}_{rgb}$ with the average map of $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$.

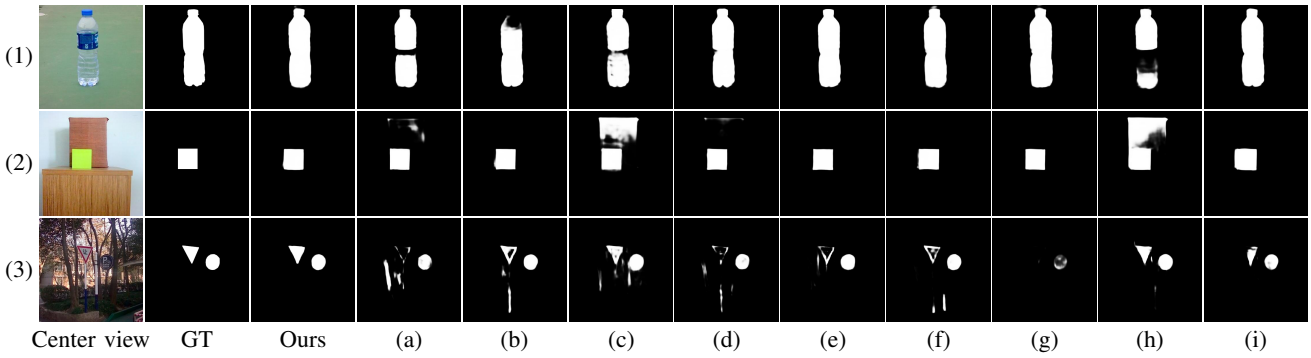


Fig. 10. Result visualization of ablation analysis. (a) *w/o CDIM*, (b) *w/o MCU*, (c) *w/o IDU*, (d) *w/o MAX-index-A*, (e) *w/o MAX-index-M*, (f) *w/o index*, (g) S_{LF}-F_R, (h) S_R-F_R and (i) $avgS_{LF,R}$-F_R.

The ablation results of these variants are listed in the last three rows of Tab. V. Obviously, we can observe that directly exploring the depth cue from $\boldsymbol{S}_{lf}$ in "S_{LF}-F_R" severely damages the performance. The reason is that the invalid $\boldsymbol{S}_{lf}$ brings the negative effects on the salient regions. This result demonstrates the necessity of our validity discrimination for $\boldsymbol{S}_{lf}$ in IDU. Conversely, "S_R-F_R" discards the input $\boldsymbol{S}_{lf}$ and only relies on the appearance cue from $\boldsymbol{S}_{rgb}$ to enhance $\boldsymbol{f}_{rgb}$. Although this variant avoids the negative effects from invalid $\boldsymbol{S}_{lf}$, there is no depth cue that can be exploited to suppress the background distraction, inevitably causing 1.1% loss in $\mathcal{F}_\beta^w$ compared with our complete IDU. Different from the above two variants, "$avgS_{LF,R}$-F_R" fixes the validity weight to 0.5 as a kind of compromise. In this variant, the validity weight of $\boldsymbol{S}_{lf}$ cannot be flexibly adjusted according to the actual intersection relationship between $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$, thus the invalid $\boldsymbol{S}_{lf}$ still affects the final result, causing 1.3% loss in $\mathcal{S}_\lambda$ and 2.6% loss in $\mathcal{F}_\beta^w$. In contrast, our complete IDU novelly explores the IOU value between $\boldsymbol{S}_{lf}$ and $\boldsymbol{S}_{rgb}$ as the validity weight of $\boldsymbol{S}_{lf}$, which discriminates the validity of $\boldsymbol{S}_{lf}$ in a soft way, enhancing the subsequent feature-result interaction between $\boldsymbol{S}_{lf}$ and $\boldsymbol{f}_{rgb}$ without disturbance.

In Fig. 10, we show the saliency maps predicted by our CDINet and its nine variants on three examples, corresponding to different-looking regions (1^{st} row), valid depth cue from $\boldsymbol{S}_{lf}$ (2^{nd} row) and invalid depth cue from $\boldsymbol{S}_{lf}$ (3^{rd} row). For the first example, our CDINet with original MCU complements more missing regions for the bottle than "*w/o CDIM*", "*w/o MCU*" and "*w/o MAX-index-A*", and achieves better edge details than "*w/o index*". For the second example, we can obviously observe that our CDINet with original IDU can fully exploit the valid depth cue to highlight the foreground green box, while "*w/o CDIM*", "*w/o IDU*" and "S_R-F_R" fail to suppress the background box due to their lack or inadequate use of depth cue from $\boldsymbol{S}_{lf}$. For the third example, the salient traffic signs are not on the foreground, thus the depth cue from $\boldsymbol{S}_{lf}$ tends to be invalid. Our CDINet with original IDU can still highlight the salient regions with the appearance cue from the RGB stream, while "*w/o CDIM*", "S_{LF}-F_R" and "$avgS_{LF,R}$-F_R" make some of these regions disappear. In summary, our CDINet with original cross-modal feature-result interaction module CDIM, which complements the missing salient regions in MCU and discriminates valid depth cue from $\boldsymbol{S}_{lf}$ in IDU, achieves the best prediction results on these three examples compared with its nine variants.

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

1) Comparison Methods. We compare the proposed CDINet with 21 state-of-the-art CNN-based SOD methods, including 2D RGB methods, *i.e.*, CPD [52], BASNet [53], PFAN [54], ITSD [55] and CAGNet [56], 3D RGB-D methods, *i.e.*, IC-Net [65], S2MA [70], SSF [72], CMWNet [66], BBSNet [67], ASIF-Net [68], VST [71] and EBFS [69], and 4D LF methods, *i.e.*, DLFS [19], MoLF [15], LFNet [16], LFDCN [21], MTC-Net [22], DLGLRG [17], OBGNet [23] and ESCNet [20].

We first follow [21], [22] to train our CDINet with the training set of HFUT-Lytro Illum [21], and conduct the comparison on the test set of HFUT-Lytro Illum and all samples of HFUT-Lytro [10]. We also retrain OBGNet [23] and ESCNet [20] on this training set for comparison. For 2D methods [52]–[56], since they have very large-scale training set as compared with ours (10553 *vs.* 512), we retrain them on our training set for fair comparison. For other methods, the saliency maps are either provided by authors or produced with the released codes. Notably, we test focal stack-based 4D methods MoLF [15] and DLGLRG [17] on HFUT-Lytro Illum using the focal stack data supplemented by [75]. Since DLGLRG [17] has a training set that includes 100 samples of HFUT-Lytro, we do not test it on this dataset. LFNet [16] does not release the code and only provides the results on HFUT-Lytro.

To further conduct the comparison with these methods on the recently released dataset DUTLF-V2 [39], we retrain our CDINet and all other 4D methods[10] with the training set of DUTLF-V2 (*i.e.*, 2957 samples), and test them on the 1247 test samples of DUTLF-V2. For 2D and 3D methods, we directly test them on this dataset.

2) Quantitative Comparison. The quantitative comparison results are listed in Tab. VI. Specifically, when calculating the average ranking (*i.e.*, "AvgR") of each method, we exclude LFNet [16] and DLGLRG [17] because they cannot test on some datasets as explained before.

[10]Specifically, the performance of LFNet [16] on DUTLF-V2 is not reported since it does not release the code for training and test. For DLGLRG [17] and OBGNet [23], the former one only releases the test code and the latter one was originally trained on DUTLF-V2. We choose to directly test them on this dataset.

TABLE VI
QUANTITATIVE PERFORMANCE COMPARISON OF OUR METHOD AND OTHER 21 STATE-OF-THE-ART CNN-BASED 2D, 3D AND 4D METHODS ON LIGHT FIELD MULTI-VIEW DATASETS HFUT-LYTRO ILLUM [21], HFUT-LYTRO [10] AND DUTLF-V2 [39] WITH ADAPTIVE E-MEASURE, S-MEASURE, WEIGHTED F-MEASURE, ADAPTIVE F-MEASURE AND MAE. WE ALSO REPRESENT THE FRAMES PER SECOND (FPS) AND THE FLOATING POINT OPERATIONS (FLOPS) OF EACH METHOD, AND THE AVERAGE RANKING IN TERMS OF FIVE METRICS ON THREE DATASETS, DENOTED AS AVGR. ↑ AND ↓ INDICATE LARGER AND SMALLER IS BETTER, RESPECTIVELY. THE TOP THREE RESULTS ARE MARKED IN **RED**, **BLUE** AND **GREEN**.

Methods	FPS ↑	FLOPs ↓	HFUT-Lytro Illum [21]					HFUT-Lytro [10]					DUTLF-V2 [39]					AvgR
			$\mathcal{E}_\xi$ ↑	$\mathcal{S}_\lambda$ ↑	$\mathcal{F}_\beta^w$ ↑	$\mathcal{F}_\beta$ ↑	$\mathcal{M}$ ↓	$\mathcal{E}_\xi$ ↑	$\mathcal{S}_\lambda$ ↑	$\mathcal{F}_\beta^w$ ↑	$\mathcal{F}_\beta$ ↑	$\mathcal{M}$ ↓	$\mathcal{E}_\xi$ ↑	$\mathcal{S}_\lambda$ ↑	$\mathcal{F}_\beta^w$ ↑	$\mathcal{F}_\beta$ ↑	$\mathcal{M}$ ↓	
4D Light Field SOD Methods																		
Ours	15	307.85G	**.9088**	**.8801**	**.8265**	**.8540**	**.0479**	**.8187**	**.7650**	**.6587**	**.7058**	**.0837**	**.9399**	**.9053**	**.8617**	**.8796**	**.0328**	**1.4**
$ESCNet_{22}$ [20]	11	248.70G	**.9147**	.8762	**.8188**	.8498	**.0487**	**.8120**	.7499	.6253	.6842	.0866	.9286	.8806	.8190	.8493	.0422	4.9
$OBGNet_{21}$ [23]	11	73.38G	.9059	.8777	.8168	**.8518**	.0498	.7901	.7398	.6058	.6714	.0941	**.9362**	**.8963**	**.8459**	**.8687**	**.0368**	5.5
$DLGLRG_{21}$ [17]	14	353.85G	**.9107**	.8739	.8115	.8460	.0499	-	-	-	-	-	.9079	.8605	.7794	.8163	.0464	-
$MTCNet_{21}$ [22]	14	4431.66G	.9074	.8757	.8154	**.8515**	**.0485**	**.8178**	.7502	.6338	**.7142**	**.0845**	**.9315**	.8954	**.8373**	**.8661**	**.0390**	**3.9**
$LFDCN_{20}$ [21]	1	197.16G	.8852	.8613	.7500	.7973	.0622	.7419	.6759	.4398	.5661	.1299	.9004	.8672	.7453	.7994	.0576	13.5
$LFNet_{20}$ [16]	-	-	-	-	-	-	-	.7700	.7358	.5787	.6145	.0930	-	-	-	-	-	-
$MoLF_{19}$ [15]	10	530.90G	.8699	.8334	.7393	.7606	.0727	.7851	.7420	.5946	.6272	.0946	.9117	.8762	.8023	.8111	.0470	11.7
$DLFS_{19}$ [19]	2	967.15G	.8454	.8018	.6603	.7173	.0847	.7554	.7109	.5327	.5924	.1109	.8748	.8093	.6766	.7524	.0755	17.6
3D RGB-D SOD Methods																		
$EBFS_{22}$ [69]	10	147.34G	.8958	**.8784**	.8169	**.8515**	.0508	.7804	.7358	.5973	.6444	.0978	.8716	.8292	.7360	.7768	.0684	10.0
VST_{21} [71]	9	30.99G	.9022	**.8872**	**.8253**	.8408	.0532	.8069	**.7920**	**.6822**	**.7008**	.0874	.9233	**.8970**	.8363	.8488	.0394	**4.0**
$ASIF\text{-}Net_{21}$ [68]	31	209.34G	.8589	.8256	.7405	.7882	.0680	.7747	.7088	.5645	.6300	.0975	.8527	.7927	.6890	.7394	.0775	16.4
$BBSNet_{20}$ [67]	21	31.14G	.8604	.8256	.7200	.7830	.0820	.8029	.7510	.6036	.6648	.0892	.8927	.8514	.7538	.7942	.0592	11.5
$CMWNet_{20}$ [66]	7	322.34G	.8873	.8734	.7899	.8131	.0565	.7871	**.7661**	.6305	.6575	.0949	.8767	.8409	.7366	.7711	.0673	10.3
SSF_{20} [72]	21	46.53G	.8913	.8499	.7797	.8285	.0582	.7799	.7257	.5921	.6403	.1002	.8872	.8175	.7235	.7723	.0617	12.7
$S2MA_{20}$ [70]	29	141.06G	.8637	.8294	.7264	.7694	.0750	.7302	.7102	.5411	.5889	.1237	.8434	.8030	.6791	.7292	.0868	17.9
$ICNet_{20}$ [65]	13	125.72G	.9055	.8752	.7991	.8282	.0527	.7901	.7638	.6339	.6630	.0949	.8685	.8322	.7263	.7630	.0715	9.9
2D RGB SOD Methods																		
$CAGNet_{20}$ [56]	17	154.36G	.8991	.8483	.7939	.8335	.0550	.7940	.7315	.6227	.6790	.0927	.8551	.7780	.6902	.7457	.0760	12.3
$ITSD_{20}$ [55]	35	34.77G	.9004	.8722	.8084	.8325	.0499	.8031	.7609	.6517	.6855	.0896	.8790	.8190	.7336	.7722	.0676	8.4
$PFAN_{19}$ [54]	53	65.91G	.8517	.8323	.7045	.7402	.0750	.7457	.7199	.5479	.6107	.1100	.8308	.7907	.6272	.6810	.0905	18.1
$BASNet_{19}$ [53]	41	127.40G	.8949	.8648	.8002	.8287	.0560	.8041	.7643	**.6558**	.6902	**.0843**	.8530	.7876	.6907	.7343	.0797	10.5
CPD_{19} [52]	66	59.43G	.9042	.8693	.8094	.8412	.0495	.8064	.7541	.6376	.6931	.0875	.8574	.7869	.6894	.7532	.0740	9.2

Our method outperforms all these competitors on the large-scale dataset DUTLF-V2 in terms of five metrics and achieves competitive performance (mostly Top-2) on HFUT-Lytro Illum and HFUT-Lytro datasets. Across all fifteen metrics, it ranks first in ten and second in three. Notably, our method surpasses all 2D methods on these three datasets, demonstrating the critical role of implicit depth cues from light field data. As compared with the best 3D methods, *i.e.*, VST [71] (Transformer-based), our method achieves gains of 2.5% and 3.1% in $\mathcal{F}_\beta^w$ and $\mathcal{F}_\beta$, respectively, on the DUTLF-V2 dataset. As compared with the 4D methods, MTCNet [22] is the second-best method among all 2D, 3D and 4D competitors. Despite using a more sparse input pattern than MTCNet (25 star-shaped views→5 diagonal views), our method still outperforms MTCNet on HFUT-Lytro Illum and DUTLF-V2 datasets and achieves gains of 1.5% and 2.5% in $\mathcal{S}_\lambda$ and $\mathcal{F}_\beta^w$, respectively, on the HFUT-Lytro dataset. All these results indicate the effectiveness of our CDINet in addressing LF SOD, leveraging carefully selected key sparse views and novel feature-result interaction.

3) Visual Comparison. To further illustrate the superiority of our method, we show the visual comparisons with top-ranking methods among 2D, 3D and 4D SOD methods in Fig. 11. We select several representative images and divide them into four challenging scenes, including (1) different-looking regions (1^{st} and 2^{nd} rows), *i.e.*, the same salient object comprises different color regions, (2) different-depth planes (3^{rd} and 4^{th} rows), *i.e.*, similar salient objects are located at different distances, (3) complicated background (5^{th} and 6^{th} rows) and (4) non-salient foreground (7^{th} and 8^{th} rows).

(1) **Different-looking regions.** Our method detects more complete salient object which is split by some different-looking regions. For example, the panel and bracket of the warning sign (1^{st} row) and the handle of the transparent cup (2^{nd} row) cannot be completely detected by most competitors. In this case, the proposed MCU plays a key role to the final results. It explores the depth cue from the light field stream to successfully complement these missing regions.

(2) **Different-depth planes.** Our method can highlight the similar salient objects distributed on different-depth planes. For example, most competitors only detect the closest price card (3^{rd} row) and billboard (4^{th} row), while the farther one disappears in the results. In our MCU, the appearance cue

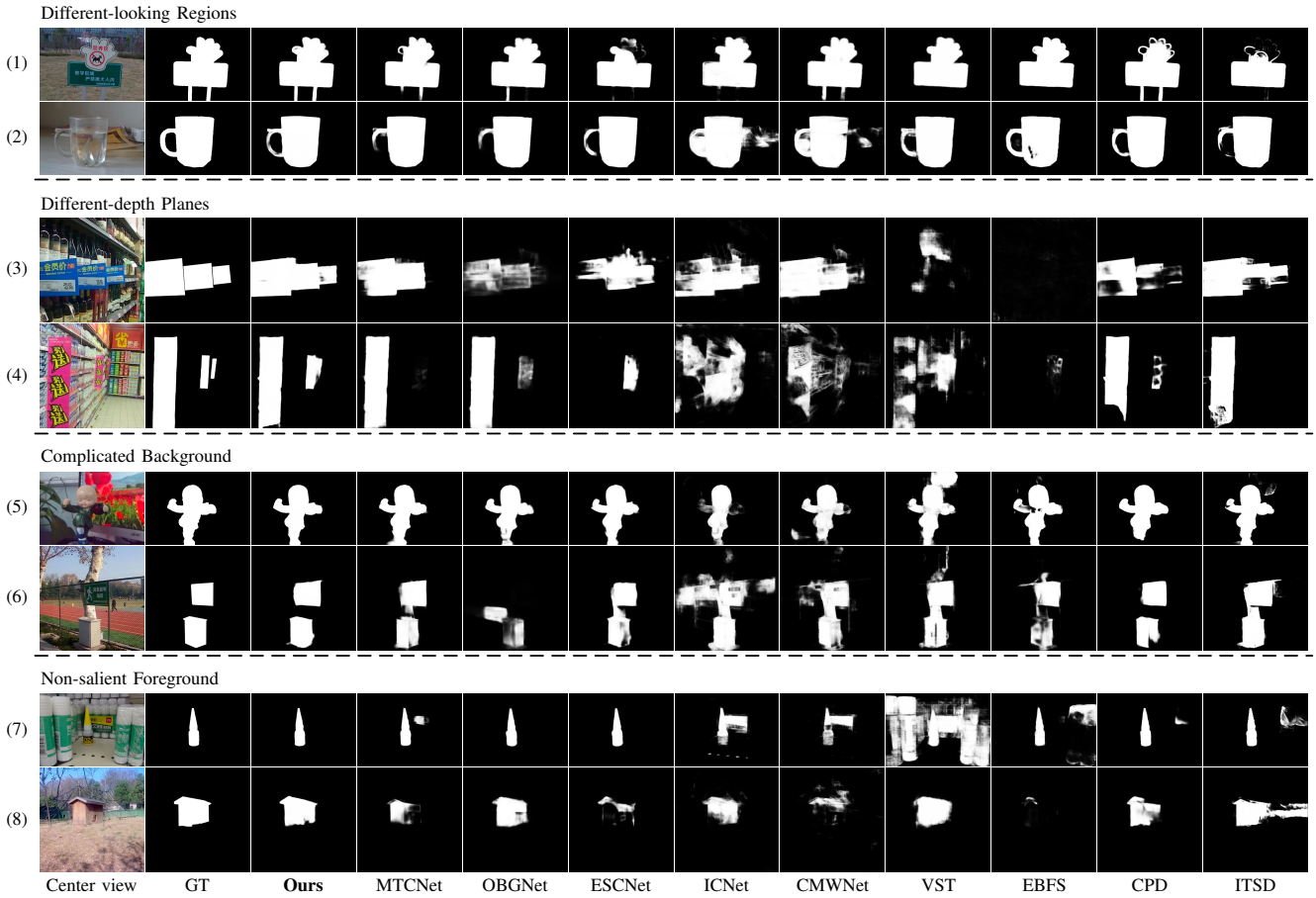


Fig. 11. Visual comparisons with nine top-ranking CNN-based SOD methods, including 4D LF methods (MTCNet [22], OBGNet [23], ESCNet [20]), 3D RGB-D methods (ICNet [65], CMWNet [66], VST [71], EBFS [69]), and 2D RGB methods (CPD [52], ITSD [55]).

and the depth cue are actually complemented with each other, and thus our CDINet is the only method which can detect the farthest billboards in the 4^{th} row.

(3) **Complicated background.** Our method can accurately locate the foreground object from the complicated background with the valid depth cue of light field. For example, the bright-colored flowers behind the doll (5^{th} row) and the tree branch between the target objects (6^{th} row) are also detected as salient regions by some competitors. Thanks to the valid interaction between initial light field saliency map and RGB features in IDU, our CDINet can effectively suppress the redundant background distraction.

(4) **Non-salient foreground.** Our method can also produce the accurate results even when the salient object is not in the foreground. In this scene, the depth cue has less or none contribution to the final results. Some 3D and 4D methods, which rely heavily on the depth cue, either detect the non-salient foreground objects (7^{th} row) or make part of the salient object disappear (8^{th} row). The IDU not only exploits the depth cue from initial light field saliency map, but also flexibly discriminates its validity with the intersection relationship between two-stream initial results, improving the robustness of our CDINet in this challenging scene.

4) Complexity Evaluation. We also report the Frames Per Second (FPS) and the Floating Point Operations (FLOPs) of each method in Tab. VI. The FPS of all methods is calculated on a Tesla P100 GPU. Due to the high-dimensional data to be processed, 4D methods have the slower FPS and larger FLOPs than most of 2D and 3D methods. Since we use the key sparse views to reduce the redundancy among all views, our CDINet achieves the fastest FPS and the fourth smallest FLOPs among all 4D methods. Taking number of required input views, SOD performance, FPS and FLOPs together, we believe our CDINet has the potentiality for practical applications of LF SOD.

D. Discussion

1) Depth Cue in Light Field Stream. Our light field stream adopts the MPI network [40] to extract the depth cue. It is an unsupervised manner, which takes the view synthesis as the target and does not need ground truth depth maps. To visually demonstrate the extracted depth cue, we follow [87] to present some depth maps generated from the predicted MPI in Fig. 12. These scenes are from the DUTLF-V2 dataset [39].

Our depth maps are consistent with those from the dataset, indicating successful extraction of depth cues. Moreover, one advantage of using the unsupervised manner can be found in the 4^{th} scene, where the dataset's depth map has low quality. Since we do not use this map as supervision, we can avoid its negative impacts and extract more reliable depth cues (see our generated depth map).

In our CDINet, the light field features are from the above MPI network, so they contain the depth cue. Since the light

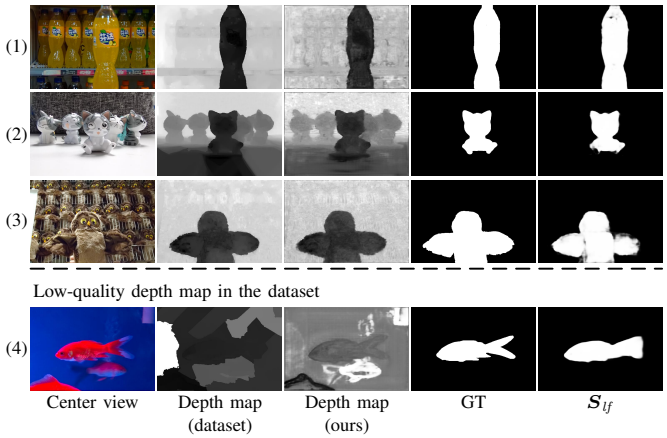


Fig. 12. Depth maps generated in the light field stream of our CDINet in the DUTLF-V2 dataset [39]. We compare them with depth maps provided in the dataset. We also show our initial light field-stream saliency maps S_{lf} and the corresponding ground truth.

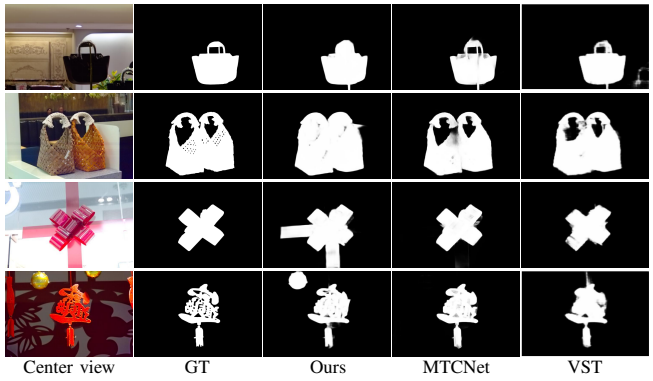


Fig. 13. Illustration of some failure cases.

field-stream saliency map is generated by these features, it also contains the depth cue, as shown in the last column of Fig. 12. Using these features and the saliency map, we explore the depth cue from key sparse views for LF SOD.

2) Key Sparse View Selection. In this work, the selected key sparse views follow a fixed pattern. According to the low-rank statistical analysis (refer to Sec. III-B), they are universal for most scenes. While finding key views for each scene may benefit SOD, it can be costly as it still requires capturing all views before selecting the key views. Our fixed pattern directly reduces the number of views to be captured while maintaining relatively reliable performance for LF SOD. Therefore, it is beneficial in practical applications.

3) Limitation and Future Work. Although our model outperforms others in most scenes, there are still some exceptions. Fig. 13 shows cases in which our model performs worse than MTCNet [22] and VST [71]. First, our model cannot preserve object boundaries well (1^{st} and 2^{nd} rows). This limitation is due to the lack of boundary prediction and supervision as in [22], [71], which can be exploited in the future. Second, when multiple bright objects are in the foreground, our model may highlight them without discrimination (3^{rd} and 4^{th} rows). The long-range context information [71] or deeper aggregation of RGB features [22] is required in such cases beyond using the appearance and depth cues solely.

Besides, exploring light-weighted networks is another potential future work. While our method has achieved the fastest FPS (15 *fps*) and relatively small FLOPs (307.85G) compared to existing 4D methods, it still faces challenges in real-time applications. One approach worth considering is knowledge distillation [88], which can retain the advantages of light field data while improving efficiency. We can also incorporate some high-efficiency modules [89] to reduce model complexity.

V. Conclusion

In this paper, we propose a novel CDINet to address LF SOD. To extricate our CDINet from the high number requirement of input views, we conclude the key sparse views (*i.e.*, five diagonal views), which are concise but efficient, as the input pattern through the low rank-based statistical analysis on the mixed LF SOD dataset for the first time. To better explore the implicit depth cue from the selected key sparse views for LF SOD, we propose a key module CDIM in our CDINet to encourage the comprehensive cross-modal feature-result interaction. In CDIM, the Maximum-based Complementarity Unit exploits the information of maximum index from two-stream initial results to complement cross-modal features with each other, and the IOU-weighted Discrimination Unit discriminates the actual intersection relationship between two-stream initial results to enhance the RGB features with the valid depth cue. Extensive experiments, including ablation and comparison analyses, demonstrate that our CDINet with only five diagonal views achieves competitive performance as compared with 21 state-of-the-art 2D, 3D and 4D methods under different evaluation metrics.

Acknowledgment

The authors would like to thank Dr. Anwa Zhou for valuable discussions about the key sparse views selection.

References

[1] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[2] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Comput. Vis. Media*, vol. 5, no. 2, pp. 117–150, Jun. 2019.
[3] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 6, pp. 3239–3259, Jun. 2022.
[4] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 2941–2959, Oct. 2019.
[5] W. Wang *et al.*, "Revisiting video saliency prediction in the deep learning era," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 220–237, Jan. 2021.
[6] D.-P. Fan *et al.*, "Re-thinking co-salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4339–4354, Aug. 2022.
[7] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. IEEE CVPR*, Jun. 2014, pp. 2806–2813.
[8] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *Proc. IEEE CVPR*, Jun. 2015, pp. 5216–5223.
[9] J. Zhang *et al.*, "Saliency detection with a deeper investigation of light field," in *Proc. IJCAI*, Jul. 2015, pp. 2212–2218.
[10] 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, Jul. 2017.

[11] 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.

[12] 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, Oct. 2019.

[13] X. Wang, Y. Dong, Q. Zhang, and Q. Wang, "Region-based depth feature descriptor for saliency detection on light field," *Multimed. Tools Appl.*, vol. 80, pp. 16 329–16 346, May 2021.

[14] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, "Deep learning for light field saliency detection," in *Proc. IEEE ICCV*, Oct. 2019, pp. 8837–8847.

[15] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Proc. NeurIPS*, Dec. 2019, pp. 898–908.

[16] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu, "LFNet: Light field fusion network for salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 6276–6287, Apr. 2020.

[17] N. Liu, W. Zhao, D. Zhang, J. Han, and L. Shao, "Light field saliency detection with dual local graph learning and reciprocative guidance," in *Proc. IEEE ICCV*, Oct. 2021, pp. 4692–4701.

[18] 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*, Feb. 2020, pp. 11 865–11 873.

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

[20] M. Zhang, S. Xu, Y. Piao, and H. Lu, "Exploring spatial correlation for light field saliency detection: Expansion from a single view," *IEEE Trans. Image Process.*, vol. 31, pp. 6152–6163, Sep. 2022.

[21] 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, Feb. 2020.

[22] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "A multi-task collaborative network for light field salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1849–1861, May 2021.

[23] D. Jing, S. Zhang, R. Cong, and Y. Lin, "Occlusion-aware bi-directional guided network for light field salient object detection," in *Proc. ACM MM*, Oct. 2021, pp. 1692–1701.

[24] M. Wang, F. Shi, X. Cheng, M. Zhao, Y. Zhang, C. Jia, W. Tian, and S. Chen, "LFBCNet: Light field boundary-aware and cascaded interaction network for salient object detection," in *Proc. ACM MM*, Oct. 2022, pp. 3430–3439.

[25] Q. Zhang, S. Wang, X. Wang, Z. Sun, S. Kwong, and J. Jiang, "Geometry auxiliary salient object detection for light fields via graph neural networks," *IEEE Trans. Image Process.*, vol. 30, pp. 7578–7592, Sep. 2021.

[26] Y. Zhang *et al.*, "Light-field depth estimation via epipolar plane image analysis and locally linear embedding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 739–747, Apr. 2017.

[27] C. Shin *et al.*, "EPINET: A fully-convolutional neural network using epipolar geometry for depth from light field images," in *Proc. IEEE CVPR*, Jun. 2018, pp. 4748–4757.

[28] Y. Zhang *et al.*, "Depth estimation from light field using graph-based structure-aware analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 4269–4283, Nov. 2020.

[29] J. Jin and J. Hou, "Occlusion-aware unsupervised learning of depth from 4-D light fields," *IEEE Trans. Image Process.*, vol. 31, pp. 2216–2228, Mar. 2022.

[30] T. Wang *et al.*, "EPI-guided cost construction network for light field disparity estimation," in *Proc. IEEE CVPRW*, Jun. 2023, pp. 3437–3445.

[31] H. Sheng *et al.*, "UrbanLF: A comprehensive light field dataset for semantic segmentation of urban scenes," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7880–7893, Nov. 2022.

[32] R. Cong *et al.*, "Combining implicit-explicit view correlation for light field semantic segmentation," in *Proc. IEEE CVPR*, Jun. 2023, pp. 9172–9181.

[33] G. Wu *et al.*, "Light field image processing: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 926–954, Aug. 2017.

[34] C. Brites, J. Ascenso, and F. Pereira, "Lenslet light field image coding: Classifying, reviewing and evaluating," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 1, pp. 339–354, Jan. 2021.

[35] Y. Chen, P. An, X. Huang, C. Yang, D. Liu, and Q. Wu, "Light field compression using global multiplane representation and two-step prediction," *IEEE Signal Process. Lett.*, vol. 27, pp. 1135–1139, Jun. 2020.

[36] X. Huang, P. An, Y. Chen, D. Liu, and L. Shen, "Low bitrate light field compression with geometry and content consistency," *IEEE Trans. Multimedia*, vol. 24, pp. 152–165, Jan. 2022.

[37] Y. Peng *et al.*, "RASL: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2233–2246, Nov. 2012.

[38] X. Jiang, M. L. Pendu, R. A. Farrugia, and C. Guillemot, "Light field compression with homography-based low-rank approximation," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 7, pp. 1132–1145, Oct. 2017.

[39] Y. Piao, Z. Rong, S. Xu, M. Zhang, and H. Lu, "DUT-LFSaliency: Versatile dataset and light field-to-rgb saliency detection," *arXiv preprint arXiv:2012.15124*, 2020.

[40] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo Magnification: Learning view synthesis using multiplane images," *ACM Trans. Graph.*, vol. 37, no. 4, pp. 1–12, Jul. 2018.

[41] R. Ng *et al.*, "Light field photography with a hand-held plenopic camera," *Comput. Sci. Tech. Rep.*, vol. 2, no. 11, pp. 1–11, Jan. 2005.

[42] "Lytro," 2021. [Online]. Available: https://www.lytro.com/

[43] "Raytrix—3d light field camera technology," 2021. [Online]. Available: http://www.raytrix.de/

[44] B. Wilburn *et al.*, "High performance imaging using large camera arrays," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 765–776, Jul. 2005.

[45] T. Georgiev and A. Lumsdaine, "Superresolution with plenoptic camera 2.0," *Adobe Tech. Rep.*, pp. 1–9, Apr. 2009.

[46] R. Ng, "Fourier slice photography," *ACM Trans. Graph.*, vol. 24, no. 3, pp. 735–744, Jul. 2005.

[47] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. 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. 2015.

[48] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.

[49] L. Ye *et al.*, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, Aug. 2017.

[50] 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.

[51] 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.

[52] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3902–3911.

[53] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 7479–7489.

[54] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3080–3089.

[55] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 9138–9147.

[56] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelana, and M. Havaei, "CAGNet: Content-aware guidance for salient object detection," *Pattern Recognit.*, vol. 103, pp. 1–12, Jul. 2020.

[57] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng, "SAC-Net: Spatial attenuation context for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 3, pp. 1079–1090, Mar. 2021.

[58] D.-P. Fan *et al.*, "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.

[59] T. Zhou *et al.*, "RGB-D salient object detection: A survey," *Comput. Vis. Media*, vol. 7, pp. 37–69, Mar. 2021.

[60] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proc. IEEE CVPR*, Jun. 2019, pp. 3922–3931.

[61] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu, "A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 9057–9066.

[62] 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.

[63] 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.

[64] Y. Piao *et al.*, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE ICCV*, Oct. 2019, pp. 7253–7262.

[65] G. Li, Z. Liu, and H. Ling, "ICNet: Information conversion network for RGB-D based salient object detection," *IEEE Trans. Image Process.*, vol. 29, pp. 4873–4884, Mar. 2020.
[66] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 665–681.
[67] 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. ECCV*, Aug. 2020, pp. 275–292.
[68] C. Li *et al.*, "ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection," *IEEE Trans. Cybern.*, vol. 51, no. 1, pp. 88–100, Jan. 2021.
[69] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing bilinear fusion and saliency prior information for RGB-D salient object detection," *IEEE Trans. Multimedia*, vol. 24, pp. 1651–1664, Mar. 2022.
[70] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 13 753–13 762.
[71] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *Proc. IEEE ICCV*, Oct. 2021, pp. 4702–4712.
[72] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for RGB-D saliency detection," in *Proc. IEEE CVPR*, Jun. 2020, pp. 3469–3478.
[73] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, and H. Ling, "Hierarchical alternate interaction network for RGB-D salient object detection," *IEEE Trans. Image Process.*, vol. 30, pp. 3528–3542, Mar. 2021.
[74] Y. Yang *et al.*, "Bi-directional progressive guidance network for RGB-D salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5346–5360, Aug. 2022.
[75] K. Fu *et al.*, "Light field salient object detection: A review and benchmark," *Comput. Vis. Media*, vol. 8, pp. 509–534, May 2022.
[76] S. Zhao and Z. Chen, "Light field image coding via linear approximation prior," in *Proc. IEEE ICIP*, Sept. 2017, pp. 4562–4566.
[77] G. H. Golub, V. C. Klema, and G. W. Stewart, "Rank degeneracy and least squares problems," Univ. of Maryland, Tech. Rep., Aug. 1976.
[78] H. Engler, "The behavior of the QR-factorization algorithm with column pivoting," *Appl. Math. Lett.*, vol. 10, no. 6, pp. 7–11, Nov. 1997.
[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, May 2015, pp. 1–14.
[80] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. AISTATS*, May 2010, pp. 249–256.
[81] D. P. Kingma and J. L. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, May 2015, pp. 1–15.
[82] D.-P. Fan *et al.*, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. IJCAI*, Jul. 2018, pp. 698–704.
[83] 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 ICCV*, Oct. 2017, pp. 4548–4557.
[84] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE CVPR*, Jun. 2014, pp. 248–255.
[85] R. Achanta *et al.*, "Frequency-tuned salient region detection," in *Proc. IEEE CVPR*, Jun. 2009, pp. 1597–1604.
[86] J. Jin *et al.*, "Deep coarse-to-fine dense light field reconstruction with flexible sampling and geometry-aware fusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 4, pp. 1819–1836, Apr. 2022.
[87] R. Tucker and N. Snavely, "Single-view view synthesis with multiplane images," in *Proc. IEEE CVPR*, Jun. 2020, pp. 551–560.
[88] J. Shen *et al.*, "Distilled siamese networks for visual tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8896–8909, Dec. 2022.
[89] Z. Zhao *et al.*, "Real-time and light-weighted unsupervised video object segmentation network," *Pattern Recognit.*, vol. 120, pp. 1–10, Dec. 2021.

Yilei Chen received the B.E. degree, in 2018, from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, where he is currently working toward the Ph.D. degree in signals and information processing. His research interests include light field processing, compression, and its application.

Gongyang Li received the Ph.D. degree from Shanghai University, Shanghai, China, in 2022. From July 2021 to June 2022, he was a Visiting Ph.D. Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Postdoct with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include saliency detection, multi-modal image processing, and image/video segmentation.

Ping An (Member, IEEE) received the B.E. and M.E. degrees from the Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree from Shanghai University, Shanghai, China, in 2002. Since 1993, she has been affiliated with Shanghai University, where she currently holds the position of Professor in the Video Processing Group, School of Communication and Information Engineering. From 2011 to 2012, she joined the Communication Systems Group, Technische University at Berlin, Germany, as a visiting professor. She has completed more than 15 projects supported by the National Natural Science Foundation of China, the National Science and Technology Ministry, and the Science and Technology Commission of Shanghai Municipality. Her research interests include image and video processing, with a focus on immersive video processing. She was a recipient of the Second Prize of the Shanghai Municipal Science and Technology Progress Award, the Second Prize in Natural Sciences of the Ministry of Education, and the Second Prize in Natural Sciences of the Chinese Institute of Electronics.

Zhi Liu (M'07-SM'15) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 1999, 2002 and 2005, respectively. He is currently a Professor at the School of Communication and Information Engineering, Shanghai University, Shanghai, China. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He is an Area Editor of *Signal Processing: Image Communication* and served as a Guest Editor for the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*.

Xinpeng Huang received the B.S. and M.E. degrees from the Zhengzhou University of Light Industry in 2013 and 2016, respectively, and the Ph.D. degree from Shanghai University in 2019. He started his Postdoctoral career with Shanghai University as a Candidate for the 2020 Shanghai Super Postdoctoral Incentive Program. Since 2022, he has been affiliated with Shanghai University, where he is currently a Lecturer at the School of Communication and Information Engineering. His research interests include depth estimation and light field compression, quality assessment, and processing.

Qiang Wu (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees from the Harbin Institute of Technology, Harbin, China, in 1996 and 1998, respectively, and the Ph.D. degree from the University of Technology Sydney, Ultimo, NSW, Australia, in 2004. He is currently an Associate Professor and a Core Member with the Global Big Data Technologies Centre, University of Technology Sydney. His research interests include computer vision, image processing, pattern recognition, machine learning, and multimedia processing. The application fields where the research outcomes are applied span over video security surveillance, biometrics, video data analysis, and human–computer interaction. His research outcomes have been published in many premier international conferences, including ECCV, CVPR, ICCV, ICIP, and ICPR and the major international journals, such as the IEEE TIP, IEEE TSMC-B, IEEE TCSVT, IEEE TIFS, PR, PRL, and Signal Processing.