Learning Complementary Maps for Light Field Salient Object Detection

Zeyu Xiao, Jiateng Shou, and Zhiwei Xiong(✉)

MoE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition,
University of Science and Technology of China
{zeyuxiao,shoujt}@mail.ustc.edu.cn zwxiong@ustc.edu.cn

Abstract. Light field imaging presents a promising avenue for advancing salient object detection (SOD). However, existing light field SOD (LFSOD) methods grapple with challenges related to effectively aggregating features from all-in-focus (AiF) images and focal slices. These methods often under-utilize the complementary nature of salient and non-saliency maps, leading to inaccurate predictions, particularly at fine boundaries. To tackle these limitations, in this paper, we introduce a novel method for LFSOD. Our method incorporates a Cross-Modality Aggregation (CMA) module at multiple levels, facilitating the efficient fusion of AiF image and focal slice features. This progressive aggregation capitalizes on global and local dependencies to harness implicit geometric information in an LF. Based on the observation that, salient regions and non-salient counterparts are complementary to each other, thus a better estimation on one side leads to an improved estimation on the other, and vice versa, we introduce the Complementary Saliency Map Generator (CSMG). The CSMG generates both saliency and non-saliency maps interactively to leverage the inherent complementary relationship between salient regions and their non-salient counterparts. Through extensive experiments conducted on benchmark datasets, we have demonstrated that our proposed method achieves superior performance in LFSOD.

Keywords: Light field · Salient object detecction · Complementary map.

1 Introduction

Salient object detection (SOD) [1, 2, 13] constitutes a foundational task in computer vision, focusing on identifying and segmenting prominent regions or objects within a scene. In parallel, the domain of light field SOD [37, 38] addresses the challenge of accomplishing SOD using light field data. The applications of SOD span a diverse spectrum, including object detection and recognition [11, 48, 57, 58, 86], semantic segmentation [70, 74, 75], unsupervised video object segmentation [59, 69], multimedia compression [31, 46, 47, 50], non-photorealist rendering [27], re-targeting [61], and human-robot interaction [3, 60]. Due to the prosperity of deep learning techniques, convolutional neural network (CNN) based methods and vision Transformer based method have demonstrated promising performance for SOD [6, 8, 14, 16, 17, 20, 41, 44, 45, 68, 71, 73, 76, 84].

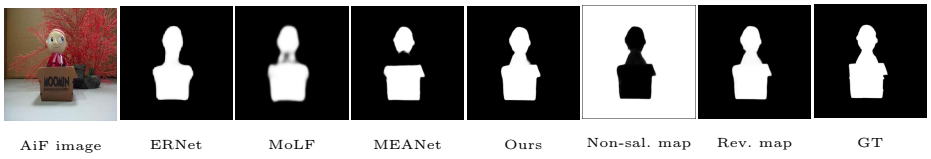

Fig. 1: Examples of LFSOD. We present saliency maps computed by state-of-the-art LFSOD methods in the initial row. Additionally, we showcase the saliency map generated by our methodology, along with the accompanying non-saliency map (non-sal. map) and the reversed non-saliency map (rev. map). This visual representation highlights the inherent complementarity between the salient region and the non-salient region. The salient regions within a light field and their non-salient counterparts possess a mutually complementary relationship. An improved estimation of one side directly contributes to enhancing the estimation of the other side and vice versa. This observation serves as a foundational principle in our method's design, enabling us to harness this symbiotic relationship for LFSOD.

In recent times, the domain of light field (LF) salient object detection (LFSOD) has garnered increasing attention, owing to its capability to significantly enhance results in challenging scenarios through the utilization of abundant LF cues [22–24, 79–81]. Compared with the RGB images captured by a regular camera or depth maps acquired by a depth sensor, the LF data acquired by a plenoptic camera (*e.g.*, Lytro ILLUM) records more comprehensive and complete information of natural scenes, covering, for example, depth information [32, 49, 53, 56, 62, 63, 65], focusness cues [37, 49] as well as angular changes [49, 54]. Diverging from conventional 2D SOD approaches that rely on RGB images as input or RGB-D SOD methods that incorporate both RGB images and depth maps for saliency map generation, LFSOD methods harness all-in-focus (AiF) images alongside auxiliary inputs like focal stacks (FSs) [93], depth maps [88], and epipolar plane images [64] to further enhance performance. Remarkably, models such as MEANet [33] revolving around focal stacks have significantly progressed in LFSOD, showcasing superior performance. This is attributed to the inherent depth information embedded in focal stacks in LF.

However, despite their notable advantages, these methods exhibit two primary limitations. First, they explore little about the long-term cross-modal information aggregation between AiF images and FSs. Typically, they rely on CNNs to capture local visual features from both modalities and simply concatenate them. This simplistic approach often leads to sub-optimal results, as it does not fully exploit the potential for cross-modal information fusion. Second, existing methods overlook the inherent complementary information available within this specific task. They typically focus on predicting saliency maps directly but do not consider the complementary information provided by their non-salient counterparts, often referred to as complementary saliency maps. This oversight limits their overall performance potential. To illustrate these limitations and the need for improvement, we provide a compelling example in Figure 1.

To tackle the challenges mentioned above, we introduce a novel method for LFSOD. To facilitate effective cross-modal aggregation of two critical modalities, namely AiF images and FSs, we introduce the CMA module. This module enables the simultaneous utilization of global implicit geometric information and intricate local cues in a coarse-to-fine manner. Within the CMA module, we employ a multi-head attention operation to coarsely aggregate features from AiF images and FSs. These aggregated features are further combined using a dual-branch structure, leveraging both global and local cross-modality dependencies. Furthermore, recognizing the inherent complementarity between salient regions and their non-salient counterparts, we propose to explicitly exploit this complementarity for LFSOD. To achieve this, we introduce a CSMG that concurrently produces both the saliency map and the non-saliency map in an interactive manner. These innovative designs are supported by extensive experiments conducted on benchmark datasets, showcasing the superior performance of our proposed method. As illustrated in Figure 1, the saliency map and the non-saliency map generated by our approach exhibit a complementary relationship, demonstrating the effectiveness of our method in handling the task of LFSOD.

Our main contributions are summarized as follows. (1) We propose the CMA module, which applies global and local cross-modal information aggregation in a coarse-to-fine manner to adaptively aggregate FSs features with the AiF feature. (2) We observe that salient regions and non-salient counterparts are complementary and interrelated. To harness this inherent complementarity, we design the CSMG to generate both saliency and non-saliency maps interactively for accurate LFSOD. (3) We validate the effectiveness of our method through extensive experiments conducted on benchmark datasets. The results demonstrate that our method can achieve superior performance.

2 Related Works

Image Salient Object Detection. In the early days of SOD, researchers heavily rely on hand-crafted features [26, 78] or incorporated saliency priors [12, 82] into their methods. However, these approaches have limitations in terms of their ability to represent complex features effectively. With the rise of CNNs, the landscape of SOD has significantly evolved. CNN-based methods have become dominant and are designed to harness both high-level semantic features and fine-grained low-level representations. Early CNN-based methods often operate by processing and classifying image regions to predict saliency [35, 98]. However, these methods discard the spatial layout of the input image. Later approaches introduce end-to-end multi-scale learning into SOD [10, 14, 17, 20, 28, 39, 66, 76, 77, 85, 90, 97], resulting in substantial performance improvements. For instance, EGNet [97] incorporate progressive fusion techniques to integrate salient edge information and global location information, effectively sharpening object boundaries. SCRN [77] tackle the challenge of over-reliance on edge information by refining multi-level features simultaneously for both salient object detection and edge detection, bridging the gap between these two tasks. More recently, vi-

sion Transformer-based methods have demonstrated their effectiveness in SOD. TCFNet [83] combines Transformer architectures with CNNs to leverage long-range relationships between image tokens and capture multi-scale local details. These advancements in SOD techniques have contributed to improved accuracy and robustness in identifying salient objects in complex scenes.

RGB-D Salient Object Detection. The RGB image provides essential appearance and texture information, while the depth map contributes crucial 3D layouts and spatial structural details. Effectively combining information from both modalities is a pivotal challenge in the RGB-D SOD task. Various fusion strategies have been explored in this context, including early fusion [7], middle fusion [34], and late fusion [9]. For instance, HDFNet [51] has introduced a depth-guided fusion scheme that significantly enhances accuracy. To achieve this, HDFNet leverages Densely Kernel Generation Units [30] and Kernel Transformation Units, which consist of dilated convolutions. These components play a crucial role in effectively fusing RGB and depth information. Another notable approach is JL-DCF [21], which adopts a joint-learning scheme to overcome the limitation of treating RGB and depth information independently. This method explores the intrinsic connections between RGB and depth data, leading to improved results. These breakthroughs in mining the internal connections between RGB and depth have significantly advanced the field of RGB-D SOD.

Light Field Salient Object Detection. The utilization of LF data in SOD is currently emerging as a promising trend due to its ability to capture comprehensive information beneficial for SOD tasks. Early LF-based methods leverage multiple visual cues such as depth, color contrast, LF flows, and boundary priors to detect saliency [37, 88, 89]. With the development of public LFSOD datasets [37, 54, 67, 87, 89], many CNN-based methods are proposed for LFSOD. These methods often take multi-view LF images as inputs and employ unified structures for synchronous salient object and edge detection [95]. Two-stream architectures, specifically middle fusion [92, 93] and late fusion [67, 95], are popular choices. Additionally, reconstruction-based methods like DLSD [54] consider geometric information for improved performance. A common approach in recent LFSOD methods is to use ConvLSTM and channel attention mechanisms during the decoding stage to detect salient objects from both AiF images and FSs [55, 67, 92, 93]. However, these existing methods tend to focus primarily on local dependencies between AiF images and FSs. They often predict saliency maps directly without fully considering the internal complementary information present in LF data. In contrast, our work places a strong emphasis on exploring long-range cross-modal information and effectively leveraging auxiliary information provided by saliency maps. Our proposed CMA module is designed to comprehensively extract both global and local cross-modal information to fully aggregate features from two modalities, AiF images and FSs. Additionally, the CSMG is introduced to simultaneously estimate both saliency regions and non-saliency counterparts, leading to a significant enhancement in LFSOD accuracy. This method effectively harnesses the intrinsic complementarity of LF data, setting it apart from previous methods.

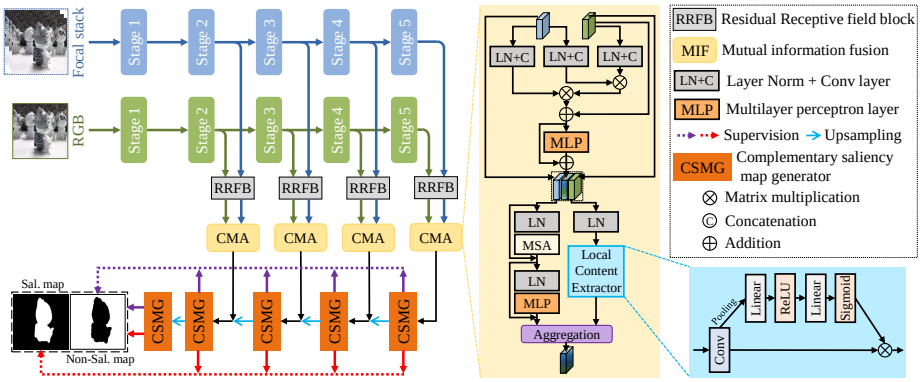

Fig. 2: The overall architecture of our proposed method. Our method adheres to a dual-stream structure to extract features and generate results. We employ the CMA modules to facilitate the aggregation of features and results across different modalities. Additionally, we integrate the CSMG to generate complementary saliency maps.

3 Method

3.1 Overview

The architecture of our method is depicted in Figure 2. Given an AiF image $I^A \in \mathbb{R}^{3\times H\times W}$ and Focal Stacks (FSs) $I^F \in \mathbb{R}^{S\times 3\times H\times W}$, our objective is to generate a saliency map $M \in \mathbb{R}^{H\times W}$. Here, H and W represent the height and width, respectively, and S signifies the number of focal slices. We set $S = 12$, and for FSs with fewer than 12 focal slices, we apply a zero-padding strategy [55,96].

We employ a dual-stream multi-scale encoder to extract features from both the AiF image and the FSs. The features extracted at five different levels are denoted as $\{\boldsymbol{F}_i^A\}_{i=1}^5$ for the AiF image and $\{\boldsymbol{F}_i^F\}_{i=1}^5$ for the FSs. To manage computational complexity efficiently, we exclude the features output from stage 1 ($\boldsymbol{F}_1^A$ and $\boldsymbol{F}_1^F$). These features, namely $\{\boldsymbol{F}_i^A\}_{i=2}^5$ and $\{\boldsymbol{F}_i^F\}_{i=2}^5$, are then processed by Residual Receptive Field Blocks (RRFBs) [42] to enhance global context information. This processing results in $\{\boldsymbol{F}_i^{A,R}\}_{i=2}^5$ and $\{\boldsymbol{F}_i^{F,R}\}_{i=2}^5$ with channel numbers aligned to 64, simplifying computations. Next, these extracted features are passed through CMA modules. These modules progressively aggregate cross-modal features, exploiting both global and local dependencies, and produce features denoted as $\{\boldsymbol{F}_i^M\}_{i=2}^5$. Subsequently, $\{\boldsymbol{F}_i^M\}_{i=2}^5$ are fed into the CSMG, which generates both salient and non-saliency maps. Finally, the hierarchically generated salient and non-saliency maps, along with their corresponding features, are bilinearly interpolated by a factor of $\times 2$ to match the lower layers optionally. These interpolated maps are concatenated and processed by residual channel attention blocks, forming a typical U-Net decoder. The final output of this decoder consists of the saliency and non-saliency predictions. The detailed structure of the decoder is in the supplementary material.

3.2 Learning Cross-Modality Aggregation

The significance of aggregating both modalities, the FSs and the AiF image, arises from the distinct characteristics they offer. The FSs provide implicit depth information as they focus on a specific depth within a scene, blurring other regions. Conversely, the AiF image contains rich texture details, allowing for the exploitation of global semantic information. To fully leverage these complementary characteristics, it is crucial to aggregate both modalities effectively. Furthermore, the effectiveness of simultaneously utilizing global semantic information and detailed local cues in aggregation has been demonstrated in previous work [45]. Therefore, we introduce the CMA module. This module is specifically designed to perform efficient cross-modal aggregation of the two modalities, FSs and AiF images, in a coarse-to-fine manner. Such a design ensures that we capture the benefits of both global and local information, leading to improved performance in LFSOD.

Drawing inspiration from the self-attention operation, which involves queries (Q), keys (K), and values (V) all originating from the same input, our CMA module follows a similar structure. In our case, the CMA module extracts the queries Q_A from the AiF image feature and the keys K_F and values V_F from the FSs feature, as visualized in Figure 2. To accomplish this, we first pass the features $\boldsymbol{F}_i^{A,R}$ and $\boldsymbol{F}_i^{F,R}$ through normalization and 1×1 convolutional layers. These convolutional layers produce features with C output channels. Subsequently, we apply cross-modal attention between vectorized features from the two modalities using a mechanism that involves the queries, keys, and values via

$$\text{Attention}\left(\mathrm{Q}_A, \mathrm{K}_F, \mathrm{V}_F\right) = \mathrm{V}_F \cdot \text{softmax}\left(\frac{\mathrm{Q}_A^T \mathrm{K}_F}{\sqrt{d_k}}\right). \tag{1}$$

It is important to note that the soft indexing of K_F by Q_A occurs at the channel dimension rather than the spatial dimensions. Consequently, the computational cost of the soft attention map in Equation 1 is $\mathcal{O}\left(C^2\right)$ rather than $\mathcal{O}\left(h_i^2 w_i^2\right)$, which makes the operation feasible even for features with high spatial resolution. Here, h_i and w_i represent the spatial resolution of features at the i-th layer. The output of the attention operation is added to the input AiF image features and then passed through a multi-layer perceptron (MLP) consisting of two fully connected layers with the GELU activation function in between. This process yields the coarse-aggregated feature, denoted as $\boldsymbol{F}_i^{ca}$.

Next, we reuse $\boldsymbol{F}_i^{A,R}$ and $\boldsymbol{F}_i^{F,R}$, along with $\boldsymbol{F}_i^{ca}$, for further aggregation by leveraging both global and local information. This involves concatenating $\boldsymbol{F}_i^{A,R}$, $\boldsymbol{F}_i^{F,R}$, and $\boldsymbol{F}_i^{ca}$ along the channel dimension to obtain $\boldsymbol{F}_i^{cat}$. We employ a window-based multi-head Transformer block [43] to capture global information. This block consists of a multi-head self-attention (MSA) operation and an MLP with a residual connection.

$$\begin{aligned} \boldsymbol{F}_i^{cat} &= \text{MSA}(\text{LN}(\boldsymbol{F}_i^{ca})) + \boldsymbol{F}_i^{ca}, \\ \boldsymbol{F}_i^{G} &= \text{MLP}(\text{LN}(\boldsymbol{F}_i^{cat})) + \boldsymbol{F}_i^{cat}, \end{aligned} \tag{2}$$

where LN denotes the LayerNorm operation, $\boldsymbol{F}_i^{cat}$ denotes the captured global contexts by multi-head self-attention (MSA) operation, and $\boldsymbol{F}_i^G$ denotes the final captured global contexts. We employ a local context extractor (LCE) based on the attention mechanism to extract local information $\boldsymbol{F}_i^L$ from neighboring pixels

$$\boldsymbol{F}_i^L = \mathrm{LCE}(\mathrm{LN}(\boldsymbol{F}_i^{cat})). \tag{3}$$

As a result, we can effectively exploit both long-range global and local information for further aggregation. Finally, we merge these contexts through element-wise addition to combine the global and local information.

$$\boldsymbol{F}_i^M = \boldsymbol{F}_i^G + \boldsymbol{F}_i^L. \tag{4}$$

Discussion. While existing methods for RGB-D SOD and LFSOD also emphasize the design of cross-modality modules, our CMA module differs from these methods in several key aspects. Firstly, our alignment process follows a coarse-to-fine approach. Specifically, other methods often design various modules that take RGB and another modality's information as direct inputs and obtain fused features and results directly. In contrast, our CMA module first coarsely aligns AiF and FSs features before performing a finer alignment, allowing for better fusion and enhancing the final performance. Secondly, we have introduced a fusion scheme based on the Vision Transformer. Unlike other methods that rely solely on CNNs for fusion, our approach can fully leverage both local and global information for efficient fusion. Our experimental results, as presented in the following sections, validate the effectiveness of our design.

3.3 Generating Complementary Saliency Maps

The relationship between saliency regions and non-saliency counterparts is inherently complementary. The saliency map M and the non-saliency map M_n are subject to the constraint $M + M_n = 1$. Consequently, improving the estimation of M contributes to an enhanced estimation of M_n, and vice versa. However, in practice, prior methods have predominantly focused on predicting saliency maps, often neglecting the potential offered by auxiliary information. To explicitly explore and harness this complementary interplay between saliency and non-saliency regions for enhanced performance, we introduce the CSMG. This module jointly generates the saliency map M and its non-saliency counterpart M_n in an interactive manner.

Figure 3 illustrates the architecture of the CSMG, which is composed of standard convolutional layers followed by rectified linear unit (ReLU) activation functions [25]. The ReLU activation serves the purpose of enhancing the salient regions while suppressing the non-salient regions. Considering the complementarity between salient and non-salient regions, it's beneficial to allow the suppressed saliency features to contribute to non-saliency estimation and vice versa. To achieve this interaction between the estimation of saliency and non-saliency maps at the feature level, we employ the ReLU^- technique [29]. The

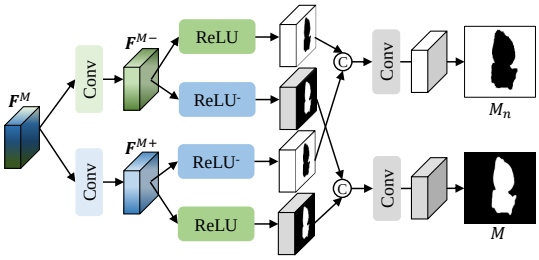

Fig. 3: The structure of CSMG. The saliency map M and non-saliency map M_n can be interactively estimated.

output of ReLU$^-$ can be defined as

$$\boldsymbol{F}^- = \text{ReLU}^-(\boldsymbol{F}) = \boldsymbol{F} - \text{ReLU}(\boldsymbol{F}) = \min\{\boldsymbol{F}, 0\}, \tag{5}$$

where $\boldsymbol{F}$ denotes the input feature. As shown in Figure 3, we feed $\boldsymbol{F}_i^M$ to convolutional layers to extract initial saliency feature $\boldsymbol{F}_i^{M^+}$ and non-saliency feature $\boldsymbol{F}_i^{M^-}$. We generate the saliency feature $\boldsymbol{F}_i^+$ and non-saliency feature $\boldsymbol{F}_i^-$ via

$$\begin{aligned} \boldsymbol{F}_i^+ &= [\text{ReLU}(\boldsymbol{F}_i^{M^+}), \text{ReLU}^-(\boldsymbol{F}_i^{M^-})], \\ \boldsymbol{F}_i^- &= [\text{ReLU}(\boldsymbol{F}_i^{M^-}), \text{ReLU}^-(\boldsymbol{F}_i^{M^+})], \end{aligned} \tag{6}$$

where $[\cdot, \cdot]$ is the concatenation operation. We finally estimate the saliency map M_i and the non-saliency map $M_{n,i}$ by feeding features into convolutional layers

$$M_i = \text{Conv}(\boldsymbol{F}_i^+), M_{n,i} = \text{Conv}(\boldsymbol{F}_i^-), \tag{7}$$

where Conv$(\cdot)$ is the convolutional layer. Upon feeding the hierarchically generated salient and non-saliency maps at different levels and their corresponding features into the decoder, we obtain the final saliency prediction denoted as M, and the non-saliency prediction denoted as M_n. These predictions possess the same spatial resolution as the input AiF image.

Discussion. Our CSMG draws inspiration from the methodology introduced in [29], where ReLU$^-$ is proposed for the image reflection removal task. However, unlike the approach presented in [29], we diverge by not utilizing a cascade of multiple ReLU and ReLU$^-$ layers. Significantly, our work pioneers the application of this concept specifically to LFSOD. Given the complementary nature of saliency and non-saliency regions, we introduce the complementary map for the first time in the context of LFSOD.

3.4 Loss Functions

As illustrated in Figure 2, our method predicts five saliency maps: $M \in \mathbb{R}^{H \times W}$ and $\{M_i\}_{i=2}^5 \in \mathbb{R}^{h_i \times w_i}$, along with five non-saliency maps: $M_n \in \mathbb{R}^{H \times W}$ and

$\{M_{n,i}\}_{i=2}^{5} \in \mathbb{R}^{h_i \times w_i}$. To train our method, let $M^G \in \{0,1\}$ represent the ground-truth saliency map, and $M_n^G = 1 - M^G$ denote the ground-truth non-saliency map. We employ a hybrid loss function for the training process.

We utilize $\mathcal{L}_s$ to optimize the estimated saliency maps

$$\begin{aligned}\mathcal{L}_s(M, M^G) = \mathcal{L}_B(M, M^G) + \mathcal{L}_I(M, M^G) + \mathcal{L}_E(M, M^G) + \\ \sum_{i=2}^{5} \left(\mathcal{L}_B(M_i{}^{\uparrow}, M^G) + \mathcal{L}_I(M_i{}^{\uparrow}, M^G) + \mathcal{L}_E(M_i{}^{\uparrow}, M^G)\right),\end{aligned} \tag{8}$$

where $M_i{}^{\uparrow}$ denotes M_i after bilinear upsampling. $\mathcal{L}_B(\cdot,\cdot)$ and $\mathcal{L}_I(\cdot,\cdot)$ denote the Binary Cross Entropy (BCE) loss and the Intersection over Union (IoU) loss, respectively; the loss $\mathcal{L}_E(\cdot,\cdot) = 1 - E_\phi$ with E_ϕ denoting E-Measure [16]. Similarly, we utilize $\mathcal{L}_n(M_n, M_n^G)$ to optimize the estimated non-saliency maps.

Therefore, the total loss function is denoted as

$$\mathcal{L} = \mathcal{L}_s + \lambda \mathcal{L}_n, \tag{9}$$

where λ is the weighting factor.

4 Experiments

4.1 Experimental Settings

Benchmark Datasets. We conduct experiments on three well-established benchmark datasets, following the protocol outlined in [22, 96]. (1) DUT-LFSD [93]: This is the largest dataset in our evaluation, consisting of 1,462 scenes. Each scene includes an AiF image, 12 FSs focused at different depths, and corresponding manually labeled ground truth. (2) HFUT-LFSD [88]: This dataset contains 255 scenes, each with similar components as DUT-LFSD. (3) LFSD [37]: LFSD is composed of 100 scenes, each with AiF images, FSs, and ground truth labels. For our training data, we select 1,000 scenes from the DUT-LFSD dataset as done in [22]. The remaining scenes from DUT-LFSD, as well as HFUT-LFSD and LFSD datasets, are used for testing our method.

Evaluation Metrics. We adopt S-Measure (S_α) [15], and E-Measure (E_ϕ) [16], F-Measure (F_β) [4] and Mean Absolute Error (M) [52] as evaluation metrics for the quantitative comparison between benchmark models and our method. These four evaluation metrics can provide comprehensive and reliable results and have been well explained in many kinds of literature.

(1) S-measure (S_α) [15] is proposed to measure the spatial structure similarities between the saliency map and ground-truth. It is defined as:

$$S_\alpha = \alpha * S_o + (1 - \alpha) * S_r, \tag{10}$$

where S_o and S_r denote the object-aware and region-aware structural similarity, respectively, and α balances S_o and S_r. We set $\alpha = 0.5$, as recommended in [15].

Table 1: Quantitative comparisons of different SOD methods on DUTLF-FS, LFSD, and HFUT-Lytro. The best results are marked in **bold**.

Types	Models	DUTLF-FS				LFSD				HFUT-Lytro			
		$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$F_\beta\uparrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$M\downarrow$
4D	Ours	**0.931**	**0.934**	**0.959**	**0.028**	**0.871**	**0.882**	**0.909**	**0.067**	0.723	**0.819**	**0.841**	**0.068**
	MEANet	0.927	0.933	**0.959**	0.031	0.851	0.850	0.891	0.075	0.759	0.701	0.824	0.084
	LFNet	0.878	0.891	0.930	0.054	0.820	0.824	0.885	0.092	0.736	0.657	0.799	0.092
	MoLF	0.887	0.903	0.939	0.051	0.825	0.824	0.880	0.092	0.742	0.662	0.812	0.094
	MAC	0.804	0.792	0.864	0.103	0.789	0.788	0.836	0.118	0.731	0.667	0.797	0.107
	DLSD	–	–	–	–	0.786	0.784	0.859	0.117	0.711	0.624	0.784	0.111
	RDFD	0.658	0.599	0.774	0.191	0.786	0.802	0.851	0.136	0.619	0.533	0.712	0.214
	DILF	0.725	0.671	0.802	0.156	0.811	0.811	0.861	0.136	0.672	0.601	0.748	0.150
	WSC	0.656	0.617	0.788	0.151	0.702	0.743	0.789	0.150	0.613	0.508	0.695	0.154
	LFS	0.585	0.533	0.711	0.228	0.681	0.744	0.809	0.205	0.565	0.427	0.637	0.221
3D	BBS	0.873	0.870	0.919	0.051	0.739	0.738	0.812	0.123	0.708	0.622	0.773	0.102
	SSF	0.881	0.889	0.930	0.050	0.790	0.793	0.861	0.097	0.687	0.612	0.781	0.099
	ATSA	0.880	0.892	0.936	0.045	0.816	0.823	0.873	0.087	0.727	0.673	0.805	0.094
	S2MA	0.894	0.893	0.934	0.046	0.827	0.829	0.873	0.086	0.672	0.572	0.735	0.120
	D3Net	0.906	0.911	0.947	0.039	0.827	0.821	0.877	0.086	0.720	0.645	0.801	0.092
	HDFNet	0.922	0.931	0.955	0.030	0.849	0.850	0.891	0.073	0.747	0.673	0.801	0.085
	JLDCF	0.924	0.931	0.958	0.030	0.850	0.860	0.900	0.071	0.755	0.694	0.823	0.086
2D	CPD	0.889	0.894	0.928	0.045	0.772	0.760	0.840	0.108	0.712	0.634	0.786	0.098
	GCPANet	0.867	0.863	0.914	0.055	0.727	0.704	0.794	0.131	0.693	0.585	0.742	0.107
	PoolNet	0.895	0.903	0.939	0.041	0.806	0.798	0.858	0.093	0.721	0.639	0.788	0.081
	EGNet	0.907	0.912	0.940	0.039	0.802	0.782	0.849	0.095	0.722	0.620	0.764	0.092
	SCRN	0.920	0.924	0.951	0.037	0.838	0.823	0.878	0.081	**0.762**	0.688	0.819	0.085

(2) E-measure (E_ϕ) [16] considers both the local and global similarity between the prediction and ground-truth. It is defined as:

$$E_\phi = \frac{1}{w * h} \sum_{i=1}^{w} \sum_{j=1}^{h} \phi(i, j), \tag{11}$$

where $\phi(\cdot)$ denotes the enhanced alignment matrix [16]. w and h are the width and height of the ground-truth map, while (i, j) are pixel indexes. Since E_ϕ also performs a comparison between two binary maps, we treat it similarly to the F-measure, thresholding a saliency map with all possible values and reporting the maximum and mean E_ϕ, denoted as $E_\phi^{\max}$ and E_ϕ^{mean}.
(3) F-measure (F_β) [13] is defined as the harmonic-mean of precision and recall:

$$F_\beta = \frac{(1 + \beta^2) Precision \cdot Recall}{\beta^2 \cdot Precision + Recall}, \tag{12}$$

where β is the weight between $Precision$ and $Recall$, and β^2 is often set to 0.3 to emphasize more on precision.
(4) Mean Absolute Error (M) [52] is defined as:

$$M = \frac{1}{N} \sum_{i=1}^{N} |S_i - G_i|, \tag{13}$$

where S_i and G_i denote the values at the i-th pixel in the saliency map and ground-truth map. N is the total number of pixels in both map.

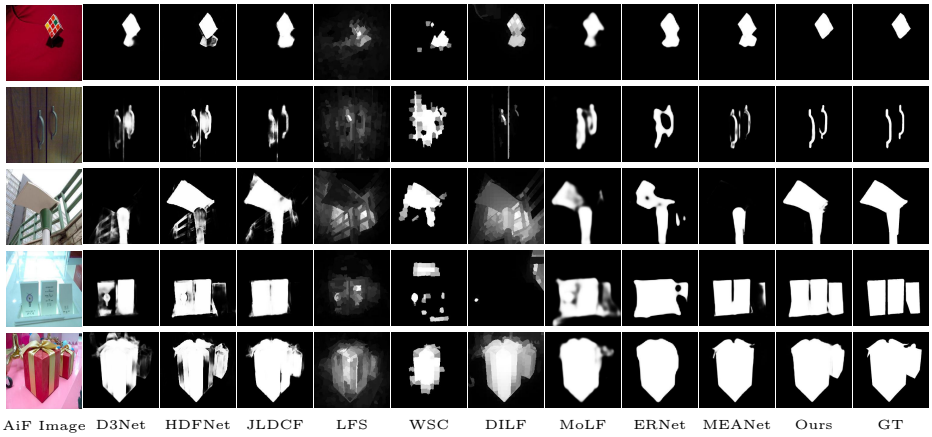

Fig. 4: Visual comparisons of different models on the DUTLF-FS testset.

Implementation Details. In our training process, we employ the Adam optimizer with the following parameters: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For the multi-scale encoder, we utilize an ImageNet-pretrained ResNet-50 for processing AiF images and an inflated 3D ResNet-50 for FSs. These networks are initialized with a weight transfer strategy from pre-trained models [5]. The learning rate is initially set to $1e-5$ and decreased by 0.5 when the training loss reaches a flat. Our implementation is based on the PyTorch toolbox. During training, we utilize 2 NVIDIA GeForce GTX1080Ti GPUs, with each mini-batch consisting of 2 samples. We apply data augmentation techniques such as rotation and flipping.

4.2 Quantitative and Qualitative Comparisons

To rigorously assess the efficacy of our proposed method, we conduct a comprehensive quantitative comparison against a range of state-of-the-art LFSOD methods, both deep-learning-based and traditional, as well as other advanced RGB-D SOD and RGB SOD techniques. The deep-learning-based LFSOD methods included in our evaluation are MEANet [33], LFNet [92], DLSD [54], MoLF [93], and MAC [87]. Additionally, we consider traditional LFSOD methods, such as RDFD [72], DILF [88], WSC [36], and LFS [37]. For RGB-D SOD methods (3D SOD), we evaluate BBS [19], SSF [94], ATSA [91], S2MA [40], D3Net [18], HDFNet [51], and JLDCF [21]. Furthermore, we include advanced RGB SOD methods (2D SOD) in the comparison, comprising CPD [76], GCPANet [10], PoolNet [39], EGNet [97], and SCRN [77]. To ensure fairness in comparisons, we retrain the RGB-D and RGB SOD methods using both RGB and depth data from the DUTLF-FS dataset.

Quantitative Evaluations. Table 1 presents comprehensive quantitative comparisons across various evaluation metrics, including F_β, S_α, E_ϕ, and M. Remarkably, our method consistently outperforms existing approaches on all four

metrics. To illustrate, our method exhibits a substantial superiority over MEANet, the current state-of-the-art LFSOD technique, by margins of 0.020, 0.032, 0.018, and 0.012 in terms of F_β, S_α, E_ϕ, and M on the LFSD dataset. Similarly, compared to JLDCF, an advanced RGB-D SOD method, our approach demonstrates significant improvements of 0.021, 0.022, 0.009, and 0.004 for the same four metrics on LFSD.

Qualitative Evaluations. Exemplar visual results from various representative methods on DUTLF-FS are presented in Figure 4. It is worth noting that our proposed method consistently produces saliency maps characterized by distinct and precise object boundaries, thereby facilitating accurate segmentation. Furthermore, our method excels at preserving the structural integrity of salient objects while demonstrating remarkable resilience to diverse background interferences across multiple scenarios.

4.3 Ablation Studies

Effectiveness of the CMA Module. We design several baselines to demonstrate the effectiveness of the CMA module. (a) We remove coarse aggregation and fine aggregation parts simultaneously. (b) We remove the fine part and keep the coarse part unchanged. (c) We keep the coarse part unchanged and remove the local branch in the fine part. (d) We keep the coarse part unchanged and remove the global branch in the fine part. (e) We remove the coarse part and keep the fine part unchanged. We replace different components with several residual blocks (RBs), leaving the total number of parameters unchanged. As shown in Table 2, the complete CMA module provides F_β 0.018 gain compared with (a) and obtains the best performance on the DUTLF-FS test set. The above results demonstrate the effectiveness of our CMA module with a coarse-to-fine strategy and a dual-branch feature extractor. The coarse aggregation part (b) and the fine aggregation part (e) provide about 0.009 and 0.010 F_β gains compared with (a) on the DUTLF-FS test set, which indicates that the CMA module benefits from various types of information in a coarse-to-fine manner. In addition, the global branch and the local branch in the fine aggregation part provide 0.005 and 0.004 F_β gains compared with (b) on the DUTLF-FS test set. It is in line with our common sense that exploiting both global semantic information and local texture details for SOD is significant.

Effectiveness of the CSMG. We design several baselines to demonstrate the effectiveness of the CSMG. (a) We only estimate the saliency map using a single-branch structure. (b) We estimate the saliency and non-saliency maps using a dual-branch structure without utilizing the interactive structure. (c) We estimate the saliency and non-saliency maps using a dual-branch structure without utilizing the ReLU^- operation. As shown in Table 3, estimating both the saliency and non-saliency maps using the ReLU^- operation and an interactive structure obtain 0.005 and 0.008 gain in terms of F_β compared with baseline (a). Utilizing both designs simultaneously obtains the best performance. The above results demonstrate the effectiveness of the CSMG. We further visualize the estimated saliency maps and their non-saliency counterparts in Figure 5. When we only

Table 2: Quantitative comparisons of the CMA module and its variants on the DUTLF-FS testset.

Method	Coarse	Fine		$F_\beta \uparrow$
		Global	Local	
(a)	✘	✘	✘	0.913
(b)	✓	✘	✘	0.922
(c)	✓	✓	✘	0.927
(d)	✓	✘	✓	0.926
(e)	✘	✓	✓	0.923
CMA	✓	✓	✓	0.931

Table 3: Quantitative comparisons of the CSMG and its variants.

Method	Dual Branch	ReLU^-	Interaction	$F_\beta \uparrow$
(a)	✘	✘	✘	0.919
(b)	✓	✓	✘	0.924
(c)	✓	✘	✓	0.927
CSMG	✓	✓	✓	0.931

Table 4: Quantitative comparisons of the estimated saliency maps and the reversed non-saliency maps.

Dataset	Result	$F_\beta \uparrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$M \downarrow$
DUTLF-FS	M	0.931	0.934	0.959	0.028
	$1-M_n$	0.928	0.935	0.960	0.028
LFSD	M	0.871	0.882	0.909	0.067
	$1-M_n$	0.869	0.882	0.909	0.068
HFUT-Lytro	M	0.723	0.819	0.841	0.068
	$1-M_n$	0.716	0.819	0.837	0.068

estimate the saliency map using a single branch structure as in existing methods, our estimated saliency map is quite different from the actual ones. When salient objects are similar in appearance to the background (row 1) or when certain parts of the salient objects are difficult to detect (row 2) or when the border of the salient objects blends with the background (row 3), some portions of the salient objects may be missing in our estimated saliency maps. When we estimate both saliency and non-saliency maps, we find that the estimated results are accurate, showing that they can promote each other when explicitly estimating both saliency and non-saliency maps. In Table 4, we list the numerical results of the estimated saliency maps and reversed non-saliency maps compared to ground-truth maps. It can be found that their results are similar, indicating that a better estimation of saliency maps leads to an improved estimation of non-saliency maps and vice versa.

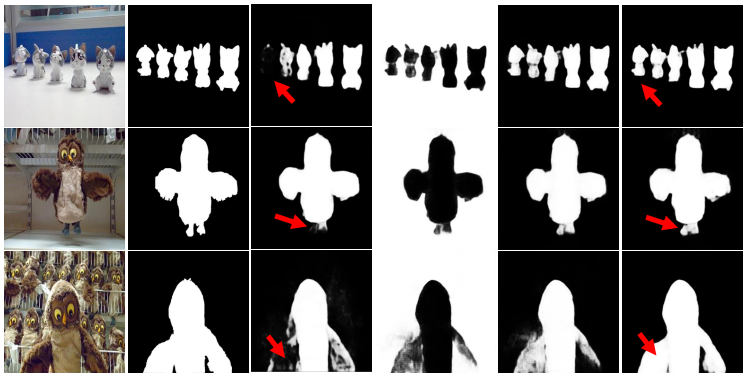

Fig. 5: Visual comparisons of the estimated saliency maps and the non-saliency maps. From left to right: AiF images, GT, the estimated saliency maps without predicting the non-saliency counterparts, the estimated non-saliency maps using CSMG, the reversed non-saliency maps, and the estimated saliency maps using CSMG.

Table 5: Quantitative comparisons of different layers.

Layer number	1	2	3	4	5
$F_\beta \uparrow$	0.901	0.912	0.921	0.931	0.932

Effectiveness of the Multi-Scale Architecture. We also conduct experiments on the multi-scale architecture of our method. We explore the performance of LFSOD using different layer features, and the results are shown in Table 5. As the number of layers increases, the performance gets better. Considering the computational overhead, we choose four layers in our method.

5 Conclusion

In this paper, we present an innovative approach tailored for LFSOD by introducing two meticulously designed components. Firstly, we introduce the CMA module, which aggregates the cross-modal AiF image and the focal stacks in a coarse-to-fine manner, leveraging both global and local dependencies within the feature domain. Secondly, considering the inherent complementarity between salient regions and their non-salient counterparts, we propose the CSMG to interactively generate both saliency and non-saliency maps, enhancing the accuracy of LFSOD. We believe that the observed complementarity between salient and non-salient regions holds the potential to advance tasks like 2D SOD and RGB-D SOD. Exploring these possibilities will be a focus of our future work.

Acknowledgement. We acknowledge funding from National Natural Science Foundation of China under Grants 62131003 and 62021001.

References

1. Borji, A., Cheng, M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Trans. Image Process. **24**, 5706–5722 (2015)
2. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A survey. Computational Visual Media **5**, 117–150 (2019)
3. Borji, A., Itti, L.: Defending yarbus: eye movements reveal observers' task. Journal of vision **14 3**, 29 (2014)
4. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: A benchmark. IEEE Trans. Image Process. **24**(12), 5706–5722 (2015)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR (2017)
6. Chen, G., Fu, H., Zhou, T., Xiao, G., Fu, K., Xia, Y., Zhang, Y.: Fusion-embedding siamese network for light field salient object detection. IEEE Transactions on Multimedia **26**, 984–994 (2023)
7. Chen, Q., Liu, Z., Zhang, Y., Fu, K., Zhao, Q., Du, H.: Rgb-d salient object detection via 3d convolutional neural networks. In: AAAI (2021)
8. Chen, Y., Li, G., An, P., Liu, Z., Huang, X., Wu, Q.: Light field salient object detection with sparse views via complementary and discriminative interaction network. IEEE Trans. Circ. Syst. Video Technol. (2023)
9. Chen, Z., Cong, R., Xu, Q., Huang, Q.: Dpanet: Depth potentiality-aware gated attention network for rgb-d salient object detection. IEEE Trans. Image Process. **30**, 7012–7024 (2020)
10. Chen, Z., Xu, Q., Cong, R., Huang, Q.: Global context-aware progressive aggregation network for salient object detection. In: AAAI (2020)
11. Cheng, M.M., Liu, Y., Lin, W.Y., Zhang, Z., Rosin, P.L., Torr, P.H.: Bing: Binarized normed gradients for objectness estimation at 300fps. Computational Visual Media **5**(1), 3–20 (2019)
12. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2014)
13. Cheng, M.M., Zhang, G.X., Mitra, N., Huang, X., Hu, S.: Global contrast based salient region detection. In: CVPR. pp. 409–416 (2011)
14. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: Bringing salient object detection to the foreground. In: ECCV. pp. 186–202 (2018)
15. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: ICCV (2017)
16. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421 (2018)
17. Fan, D.P., Li, T., Lin, Z., Ji, G.P., Zhang, D., Cheng, M.M., Fu, H., Shen, J.: Re-thinking co-salient object detection. IEEE Trans. Pattern Anal. Mach. Intell. (2021)
18. Fan, D.P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks. IEEE Trans. Neural Networks Learn. Syst. **32**(5), 2075–2089 (2020)
19. Fan, D.P., Zhai, Y., Borji, A., Yang, J., Shao, L.: Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In: ECCV (2020)

20. Fan, D.P., Zhang, J., Xu, G., Cheng, M.M., Shao, L.: Salient objects in clutter. IEEE Trans. Pattern Anal. Mach. Intell. (2022)
21. Fu, K., Fan, D.P., Ji, G.P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: CVPR (2020)
22. Fu, K., Jiang, Y., Ji, G.P., Zhou, T., Zhao, Q., Fan, D.P.: Light field salient object detection: A review and benchmark. Comput. Vis. Media pp. 1–26 (2022)
23. Gao, R., Liu, Y., Xiao, Z., Xiong, Z.: Diffusion-based light field synthesis. In: ECCVW (2024)
24. Gao, R., Xiao, Z., Xiong, Z.: Mamba-based light field super-resolution with efficient subspace scanning. In: ACCV (2024)
25. Glorot, X., Bordes, A., Bengio, Y.: Deep sparse rectifier neural networks. In: AISTATS (2011)
26. Gong, C., Tao, D., Liu, W., Maybank, S.J., Fang, M., Fu, K., Yang, J.: Saliency propagation from simple to difficult. In: CVPR (2015)
27. Han, J., Pauwels, E.J., de Zeeuw, P.M.: Fast saliency-aware multi-modality image fusion. Neurocomputing **111**, 70–80 (2013)
28. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply supervised salient object detection with short connections. In: CVPR (2017)
29. Hu, Q., Guo, X.: Trash or treasure? an interactive dual-stream strategy for single image reflection separation. NeurIPS (2021)
30. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
31. Itti, L.: Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans. Image Process. **13**, 1304–1318 (2004)
32. Jeon, H., Park, J., Choe, G., Park, J., Bok, Y., Tai, Y.W., Kweon, I.S.: Accurate depth map estimation from a lenslet light field camera. In: CVPR (2015)
33. Jiang, Y., Zhang, W., Fu, K., Zhao, Q.: Meanet: Multi-modal edge-aware network for light field salient object detection. Neurocomputing **491**, 78–90 (2022)
34. Li, C., Cong, R., Piao, Y., Xu, Q., Loy, C.C.: Rgb-d salient object detection with cross-modality modulation and selection. In: ECCV (2020)
35. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR (2015)
36. Li, N., Sun, B., Yu, J.: A weighted sparse coding framework for saliency detection. In: CVPR (2015)
37. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. In: CVPR (2014)
38. Li, N., Ye, J., Ji, Y., Ling, H., Yu, J.: Saliency detection on light field. IEEE Trans. Pattern Anal. Mach. Intell. **39**(8), 1605–1616 (2017)
39. Liu, J.J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR (2019)
40. Liu, N., Zhang, N., Han, J.: Learning selective self-mutual attention for rgb-d saliency detection. In: CVPR (2020)
41. Liu, N., Zhang, N., Wan, K., Shao, L., Han, J.: Visual saliency transformer. In: ICCV (2021)
42. Liu, S., Huang, D., et al.: Receptive field block net for accurate and fast object detection. In: ECCV (2018)
43. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021)
44. Liu, Z., He, Q., Wang, L., Fang, X., Tang, B.: Lftransnet: Light field salient object detection via a learnable weight descriptor. IEEE Trans. Circ. Syst. Video Technol. **33**(12), 7764–7773 (2023)

45. Liu, Z., Tan, Y., He, Q., Xiao, Y.: Swinnet: Swin transformer drives edge-aware rgb-d and rgb-t salient object detection. IEEE Trans. Circuits Syst. Video Technol. **32**(7), 4486–4497 (2021)
46. Ma, Y.F., Hua, X.S., Lu, L., Zhang, H.J.: A generic framework of user attention model and its application in video summarization. IEEE Transactions on Multimedia **7**, 907–919 (2005)
47. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: ACM MM (2002)
48. Moosmann, F., Larlus, D., Jurie, F.: Learning saliency maps for object categorization. In: International Workshop on The Representation and Use of Prior Knowledge in Vision. pp. 1–15 (2006)
49. Ng, R., Levoy, M., Brédif, M., Duval, G., Horowitz, M., Hanrahan, P.: Light field photography with a hand-held plenopic camera. Technical Report CTSR 2005-02 **CTSR** (01 2005)
50. Ouerhani, N., Bracamonte, J., Hugli, H., Ansorge, M., Pellandini, F.: Adaptive color image compression based on visual attention. In: Proceedings 11th International Conference on Image Analysis and Processing. pp. 416–421 (2001)
51. Pang, Y., Zhang, L., Zhao, X., Lu, H.: Hierarchical dynamic filtering network for rgb-d salient object detection. In: ECCV (2020)
52. Perazzi, F., Krähenbühl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR (2012)
53. Piao, Y., Ji, X., Zhang, M., Zhang, Y.: Learning multi-modal information for robust light field depth estimation. ArXiv **abs/2104.05971** (2021)
54. Piao, Y., Rong, Z., Zhang, M., Li, X., Lu, H.: Deep light-field-driven saliency detection from a single view. In: IJCAI (2019)
55. Piao, Y., Rong, Z., Zhang, M., Lu, H.: Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In: AAAI (2020)
56. Piao, Y., Zhang, Y., Zhang, M., Ji, X.: Dynamic fusion network for light field depth estimation. ArXiv **abs/2104.05969** (2021)
57. Ren, Z., Gao, S., Chia, L., Tsang, I.: Region-based saliency detection and its application in object recognition. IEEE Trans. Circ. Syst. Video Technol. **24**, 769–779 (2014)
58. Rutishauser, U., Walther, D., Koch, C., Perona, P.: Is bottom-up attention useful for object recognition? In: CVPR. vol. 2 (2004)
59. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: ECCV (2018)
60. Sugano, Y., Matsushita, Y., Sato, Y.: Calibration-free gaze sensing using saliency maps. In: CVPR (2010)
61. Sun, J., Ling, H.: Scale and object aware image retargeting for thumbnail browsing. In: ICCV (2011)
62. Tao, M.W., Hadap, S., Malik, J., Ramamoorthi, R.: Depth from combining defocus and correspondence using light-field cameras. In: ICCV (2013)
63. Tao, M.W., Srinivasan, P.P., Malik, J., Rusinkiewicz, S., Ramamoorthi, R.: Depth from shading, defocus, and correspondence using light-field angular coherence. In: CVPR (2015)
64. Wang, M., Shi, F., Cheng, X., Zhao, M., Zhang, Y., Jia, C., Tian, W., Chen, S.: Lfbcnet: Light field boundary-aware and cascaded interaction network for salient object detection. In: ACMMM (2022)
65. Wang, T., Efros, A.A., Ramamoorthi, R.: Occlusion-aware depth estimation using light-field cameras. In: ICCV (2015)

66. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: ICCV (2017)
67. Wang, T., Piao, Y., Li, X., Zhang, L., Lu, H.: Deep learning for light field saliency detection. In: ICCV (2019)
68. Wang, W., Lai, Q., Fu, H., Shen, J., Ling, H., Yang, R.: Salient object detection in the deep learning era: An in-depth survey. IEEE Trans. Pattern Anal. Mach. Intell. **44**(6), 3239–3259 (2021)
69. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **40**, 20–33 (2018)
70. Wang, X., You, S., Li, X., Ma, H.: Weakly-supervised semantic segmentation by iteratively mining common object features. In: CVPR (2018)
71. Wang, X., Chen, S., Wei, G., Liu, J.: Tenet: Accurate light-field salient object detection with a transformer embedding network. Image and Vision Computing **129**, 104595 (2023)
72. Wang, X., Dong, Y., Zhang, Q., Wang, Q.: Region-based depth feature descriptor for saliency detection on light field. Multim. Tools Appl. **80**(11), 16329–16346 (2021)
73. Wang, Z., Zhang, Y., Liu, Y., Wang, Z., Coleman, S., Kerr, D.: Tf-sod: a novel transformer framework for salient object detection. Neural Computing and Applications **34**(14), 11789–11806 (2022)
74. Wei, Y., Feng, J., Liang, X., Cheng, M.M., Zhao, Y., Yan, S.: Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: CVPR (2017)
75. Wei, Y., Liang, X., Chen, Y., Shen, X., Cheng, M.M., Zhao, Y., Yan, S.: Stc: A simple to complex framework for weakly-supervised semantic segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**, 2314–2320 (2017)
76. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR (2019)
77. Wu, Z., Su, L., Huang, Q.: Stacked cross refinement network for edge-aware salient object detection. In: ICCV (2019)
78. Xia, C., Li, J., Chen, X., Zheng, A., Zhang, Y.: What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors. In: CVPR (2017)
79. Xiao, Z., Cheng, Z., Xiong, Z.: Space-time super-resolution for light field videos. IEEE Trans. Image Process. (2023)
80. Xiao, Z., Liu, Y., Gao, R., Xiong, Z.: Cutmib: Boosting light field super-resolution via multi-view image blending. In: CVPR (2023)
81. Xiao, Z., Xiong, Z.: Incorporating degradation estimation in light field spatial super-resolution. arXiv preprint arXiv:2405.07012 (2024)
82. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR (2013)
83. Yao, C., Feng, L., Kong, Y., Xiao, L., Chen, T.: Transformers and cnns fusion network for salient object detection. Neurocomputing **520**, 342–355 (2023)
84. Yun, Y.K., Lin, W.: Selfreformer: Self-refined network with transformer for salient object detection. arXiv preprint arXiv:2205.11283 (2022)
85. Zeng, Y., Lu, H., Zhang, L., Feng, M., Borji, A.: Learning to promote saliency detectors. In: CVPR (2018)
86. Zhang, D., Meng, D., Zhao, L., Han, J.: Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In: IJCAI (2016)

87. Zhang, J., Liu, Y., Zhang, S., Poppe, R., Wang, M.: Light field saliency detection with deep convolutional networks. IEEE Trans. Image Process. **29**, 4421–4434 (2020)
88. Zhang, J., Wang, M., Gao, J., Wang, Y., Zhang, X., Wu, X.: Saliency detection with a deeper investigation of light field. In: IJCAI (2015)
89. Zhang, J., Wang, M., Lin, L., Yang, X., Gao, J., Rui, Y.: Saliency detection on light field: A multi-cue approach. ACM Trans. Multim. Comput. Commun. Appl. **13**(3), 1–22 (2017)
90. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: CVPR (2018)
91. Zhang, M., Fei, S.X., Liu, J., Xu, S., Piao, Y., Lu, H.: Asymmetric two-stream architecture for accurate rgb-d saliency detection. In: ECCV (2020)
92. Zhang, M., Ji, W., Piao, Y., Li, J., Zhang, Y., Xu, S., Lu, H.: Lfnet: Light field fusion network for salient object detection. IEEE Trans. Image Process. **29**, 6276–6287 (2020)
93. Zhang, M., Li, J., Wei, J., Piao, Y., Lu, H.: Memory-oriented decoder for light field salient object detection. NeurIPS (2019)
94. Zhang, M., Ren, W., Piao, Y., Rong, Z., Lu, H.: Select, supplement and focus for rgb-d saliency detection. In: CVPR (2020)
95. Zhang, Q., Wang, S., Wang, X., Sun, Z., Kwong, S., Jiang, J.: A multi-task collaborative network for light field salient object detection. IEEE Trans. Circuits Syst. Video Technol. **31**(5), 1849–1861 (2020)
96. Zhang, Y., Chen, G., Chen, Q., Sun, Y., Xia, Y., Deforges, O., Hamidouche, W., Zhang, L.: Learning synergistic attention for light field salient object detection. In: BMVC (2021)
97. Zhao, J.X., Liu, J.J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV (2019)
98. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: CVPR (2015)