Exploring Spatial Correlation for Light Field Saliency Detection: Expansion From a Single View

Miao Zhang, *Member, IEEE*, Shuang Xu, Yongri Piao, *Member, IEEE*, and Huchuan Lu, *Senior Member, IEEE*

***Abstract*— Previous 2D saliency detection methods extract salient cues from a single view and directly predict the expected results. Both traditional and deep-learning-based 2D methods do not consider geometric information of 3D scenes. Therefore the relationship between scene understanding and salient objects cannot be effectively established. This limits the performance of 2D saliency detection in challenging scenes. In this paper, we show for the first time that saliency detection problem can be reformulated as two sub-problems: light field synthesis from a single view and light-field-driven saliency detection. This paper first introduces a high-quality light field synthesis network to produce reliable 4D light field information. Then a novel light-field-driven saliency detection network is proposed, in which a Direction-specific Screening Unit (DSU) is tailored to exploit the spatial correlation among multiple viewpoints. The whole pipeline can be trained in an end-to-end fashion. Experimental results demonstrate that the proposed method outperforms the state-of-the-art 2D, 3D and 4D saliency detection methods. Our code is publicly available at https://github.com/OIPLab-DUT/ESCNet.**

***Index Terms*— Light field synthesis, salient object detection, spatial correlation.**

I. Introduction

THE goal in salient object detection is to identify and segment the most relevant parts that grab the attention of what human see. As a fundamental task, salient object detection has received increasing attention in recent years and plays an important role in a variety of computer vision applications, *e.g.*, object detection [2], [3], [4], semantic segmentation [5], visual tracking [6], [7], and robot navigation [8].

Manuscript received 29 September 2021; revised 18 July 2022; accepted 25 August 2022. Date of publication 16 September 2022; date of current version 22 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 62172070 and Grant 61976035, in part by the Central Government Guided Local Science and Technology Development Funds of Liaoning Province under Grant 2022JH6/100100028, and in part by the Natural Science Foundation of Liaoning Province under Grant 2021-MS-123. An earlier version of this article was presented at IJCAI [DOI: 10.24963/ijcai.2019/127]. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Mai Xu. *(Corresponding author: Yongri Piao.)*

Miao Zhang is with the DUT-RU International School of Information Science and Engineering and the Key Laboratory for Ubiquitous Network and Service Software of Liaoning Province, Dalian University of Technology, Dalian 116024, China (e-mail: miaozhang@dlut.edu.cn).

Shuang Xu is with the School of Software Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: sxu1997@mail.dlut.edu.cn).

Yongri Piao and Huchuan Lu are with the School of Information and Communication Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: yrpiao@dlut.edu.cn; lhchuan@dlut.edu.cn).

Digital Object Identifier 10.1109/TIP.2022.3205749

The existing saliency detection methods can be roughly divided into three categories based on the 2D, 3D and 4D input images. Earlier 2D saliency detection methods [9], [10], [11] mainly rely on hand-crafted features and prior knowledge. To overcome the limitations caused by human priors, learning-based methods [12], [13], [14] are reported to exploit more meaningful feature representation and achieve promising results. However, current 2D learning-based methods are empowered by the learning capability of CNNs, they directly relate multi-level features to the ground truth but the relationship between scene understanding and salient objects can not be effectively established. This may lead 2D saliency detection methods to appear fragile when confronted with challenging scenes shown in Figure 1.

Contrast to the 2D methods, the introduction of geometric information allows 3D and 4D saliency detection methods to better handle these challenging scenes. On the one hand, a rich set of 3D methods [15], [16], [17], [18] which concentrate on effectively fusing the RGB and depth are proposed. Benefiting from the abundant spatial information embedded in depth images, great promotions have been made in saliency detection. On the other hand, light field provides richer visual information of a scene, including a stack of focal slices, depth images and multi-view images. Light field data with valuable spatial and geometric information has been demonstrated in favor of saliency detection [19], [20], [21]. However, most of these methods focus on exploring the inter relationship among focal slices. Effective use of multi-view images in saliency detection is still under study. The multiple viewpoints which spread over the extent of the lens aperture provide abundant spatial parallax information as well as accurate geometric information about the objects in a scene. Consequently, sufficiently exploiting the correlation and disparity information of multi-view images can contribute to light field saliency detection.

Building on the above consideration, this paper aims to take full advantage of the multi-view images rich in spatial information to facilitate saliency detection when confronted with challenging scenes. The primary challenge towards our goal is designing a mechanism that is effective for processing high-dimensional light field data but also is able to achieve flexibility for testing. The second challenge is how to comprehensively and attentively exploit and model the spatial correlation between multiple viewpoints.

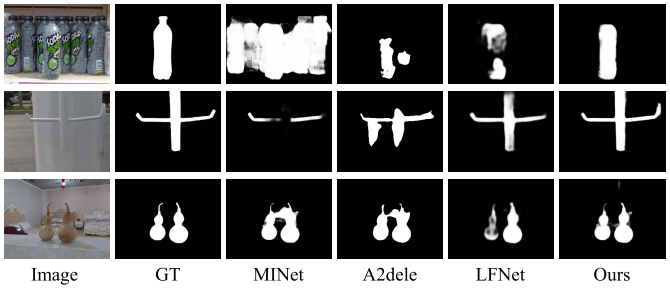


Fig. 1. Visual examples of challenging scenes in saliency detection (*e.g.*, similar foreground and background in row 1, low contrast in row 2, and multiple objects in row 3). Images and ground-truth masks (GT) are from DUTLF-V2 [22]. Compared with the state-of-the-art 2D method MINet [23], 3D method A2dele [24] and 4D method LFNet [20], the proposed approach can generate more accurate results when facing various challenging scenes.

In this paper, we strive to confront these challenges and shed light on light field saliency detection based on multi-view images. First, we propose an end-to-end CNN framework, from a novel perspective, that decomposes the saliency detection problem into the sub-problems: light field synthesis from a single view and light-field-driven saliency detection. The light field information (multi-view images and depth maps) can be automatically generated by the light field synthesis network. This allows the proposed network to be free of using high-dimensional light field data during testing.

Second, a novel light-field-driven saliency detection network is proposed. It consists of a spatial feature extractor, a Direction-specific Screening Unit (DSU) and a cascaded decoder. With the consideration of spatial parallax of the angular views with respect to the central view along the horizontal and vertical directions, the proposed DSU can effectively exploit the geometric information and fully incorporate the direction-discriminative features into the central view. The enhanced feature representation can further facilitate to achieve more accurate predictions for salient object detection.

Furthermore, experiments show that the proposed method achieves superior performance over state-of-the-art 2D, 3D and 4D methods on three light field saliency datasets.

A preliminary version of this work was published on [1]. This paper mainly makes the following modifications with significant improvements: 1) We rethink the selection of angular views and offer a more insightful glimpse into the balance of complexity and efficiency of utilizing multi-view images. Different from the conference version, this paper only synthesizes four side views along the horizontal and vertical directions, which are demonstrated to sufficiently represent the original light field data. 2) A more effective light-field-driven saliency detection subnetwork is proposed, in which a Direction-specific Screening Unit (DSU) is designed to comprehensively exploit and fully incorporate the spatial parallax information between the angular views and the central view. Extensive ablations have demonstrated the effectiveness of the proposed DSU. 3) The improved network makes a large margin of performance gain compared to the previous version. More comprehensive comparisons of the proposed method with currently advanced methods are conducted. Experimental results show that the proposed method performs favorably against the state-of-the-arts.

The rest of this paper is organized as follows. Related work is reviewed in Section II. Then details of the proposed network are reported in Section III. In Section IV, experimental results are analyzed and ablation studies are given. Section V presents the conclusion and future scope.

II. Related Work

A. Salient Object Detection

Existing salient object detection models can be generally classified into three categories according to the types of input data: 2D RGB, 3D RGB-D and 4D light field based methods. Early traditional 2D methods [9], [10], [11] mainly rely on hand-crafted features and prior knowledge, these methods cannot well adapt to varying real-world scenarios. Benefiting from the powerful ability of CNNs in extracting feature representations, numerous learning-based 2D methods are proposed. Some methods focus on exploiting multi-layer feature integration for saliency detection. For example, Hou *et al.* [25] introduced short connections between shallower and deeper side-output layers, but the differences in features of different layers were ignored. Fang *et al.* [26] proposed a densely nested top-down flows-based framework and took advantage of the strong semantic information of high-level features. Wei *et al.* [27] designed a cross feature module to selectively integrate multi-level features and obtained finer details. Some methods propose to gradually refine the saliency map to obtain more accurate results. Deng *et al.* [28] developed a series of residual refinement modules to learn the residual between the intermediate saliency prediction and the ground truth. Wu *et al.* [29] introduced a cascaded partial decoder framework, which refined the features of backbone utilizing generated saliency map. These methods adopt intermediate predictions as guidance, but the incorrect primary predictions also introduced noises in learning. Besides multi-layer feature integration and saliency map refinement, some methods explore learning saliency detectors without pixel-level ground truths. Zhang *et al.* [30] proposed a novel supervision synthesis scheme that explored external and internal knowledge source, and avoided the expensive annotations. These 2D based methods without consideration of the relationship between scene understanding and salient objects are more likely to be compromised in challenging scenes (*e.g.*, multiple or transparent objects, similar foreground and background, complex background).

To alleviate this issue in RGB saliency detection, some researchers attempt to introduce depth cues and exploit geometric and structural information. For example, Han *et al.* [15] utilized two CNNs to extract both RGB features and depth representations, and made fusion automatically to generate final saliency prediction. Piao *et al.* [16] designed a depth refinement block to effectively extract and fuse multi-level paired cues from the RGB stream and the depth stream. Zhang *et al.* [17] proposed different fusion strategies for fusing features in high level and low level from two modalities

of RGB and depth. These methods focus on integrating cross-modality complementarities of RGB and depth features without the consideration of depth quality. Recently, some studies have noted that the saliency detection results may be affected by the noise and ambiguity in raw depth images and propose methods to tackle this problem. Li *et al.* [31] designed an alternate RGB-depth-RGB interaction framework to mitigate distractors in depth maps and highlight salient objects in RGB images. Ji *et al.* [32] designed a learning strategy to calibrate the latent bias in the original depth maps and a cross reference module to fuse features from RGB and depth modalities. Due to the embedded spatial information of depth cues, the RGB-D based methods have achieved great performance gains in saliency detection.

Light field data with rich spatial information of the scene has shown promising prospects in saliency detection. Previous work [33], [34], [35], [36] for light field saliency detection focus on developing hand-crafted features. For example, Li *et al.* [33] proposed to compute the focusness and objectness of the focal stack and integrate focusness-based saliency candidates with other contrast cues. They subsequently developed a weighted sparse coding framework [34] to handle the heterogenous types of input data. Zhang *et al.* [35] utilized both background prior and location prior and introduced an additional depth cues into the contrast computation. They further integrated multiple saliency cues extracted from light field images by a random-serach-based weighting method [36]. However, these methods strongly rely on low-level cues and are less capable of extracting semantic information. Recently, some learning-based methods [19], [20], [21] are proposed and make a further step towards light field saliency detection. These methods mainly design architectures to explore the correlation among focal sclices and all-focus images. For example, Zhang *et al.* [19] proposed a memory-oriented decoder with feature fusion mechanism for accurate prediction by exploiting internal correlation of focal slices. Zhang *et al.* [20] developed a light field refinement module to refine complementary information from each focal slice and the all-focus image, then integrated the refined features through a tailored light field integration module. Piao *et al.* [21] proposed an asymmetrical two-stream architecture which consists of a focal stream and a RGB stream to achieve versatility for both desktop computer and mobile devices. However, above learning-based methods typically focus on the internal discrimination property of focal slices, while ignoring the valuable spatial parallax and structural information embedded in multi-view images provided by light field.

In this paper, we factorize the light field saliency detection into light field synthesis and light-field driven saliency detection, further build the relationship between scene understanding and salient objects by exploiting the spatial correlation between the synthesized multi-view images.

B. View Synthesis From Light Fields

Over the past decades, there are fewer works on light field rendering. Levoy and Hanrahan [37] captured densely-sampled 4D light field images of a scene and interpreted the input images as 2D slices of a 4D light field. Gortler *et al.* [38] utilized the silhouette information to compute the approximate geometry and improved the quality of the rendered images. Buehler *et al.* [39] designed an unstructured lumigraph rendering framework to render convincing new images using a set of unstructured 2D slices of light field. Recently, Srinivasan *et al.* [40] proposed a learning-based light field synthesis method, which estimates depth maps and uses the geometry to render high-quality 4D light fields. However, the huge-amount and high-dimensional light field data make the processing of light field information computational-expensive and time-consuming. In this paper, we only synthesize four side views along the horizontal and vertical directions for efficient learning, which are sufficient to represent the original light field.

III. The Proposed Method

In this paper, the saliency detection problem is formulated as two sub-problems, namely light field rendering and light-field-driven saliency detection. The whole pipeline is described in Section III-A, then the light field synthesis network and the light-field-driven saliency detection network are introduced in Section III-B and Section III-C, respectively.

A. The Whole Pipeline

Light field contains both spatial and angular information of the light rays which benefits many tasks in computer vision, such as scene flow estimation [41], lens aberrations correction [42] and refocusing [43]. Inspired by this, we design a light field rendering network to facilitate saliency detection with rich geometric information of multi-view images. Then a light-field-driven saliency detection network is proposed to build the relationship between salient objects and scene understanding.

As shown in Figure 2, the central view is first expanded to an array of angular views. However, it should be noted that the adjacent views share small differences due to the subtle parallax. Furthermore, the high-dimensional light field data leads to high computational complexity. It is necessary to balance the computational complexity and prediction accuracy. Therefore, only four side views along the horizontal and vertical directions through the central view are synthesized, which are experimentally demonstrated efficient and effective to represent the original light field in the ablation study in Section IV. The light field rendering network is inspired by the recent view synthesis method [40], which is a learning-based method to generate the light field by warping a single image using the corresponding depth information. For light field synthesis, we train the depth CNNs to estimate scene depths, then render a Lambertian approximation of light field based on a physically-based warping layer, and finally generate total four side views along the horizontal and vertical directions regarding the central view as shown in Figure 3.

The generated multi-view images equipped with rich geometric information show great potential in detecting salient objects. However, due to the high-dimensional data structure and redundant information in light field, the exploitation of relation between multi-view images is limited in the previous

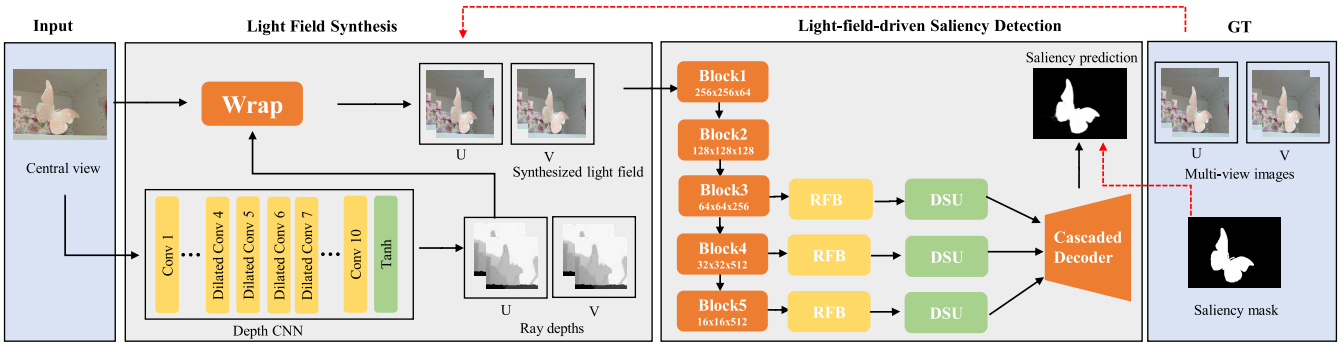


Fig. 2. Whole pipeline consists of two sub-networks: light field synthesis network and light-field-driven saliency detection network. The first network is designed to generate light field data. The second network contains three parts: spatial feature extractor, direction-specific screening unit and cascaded decoder.

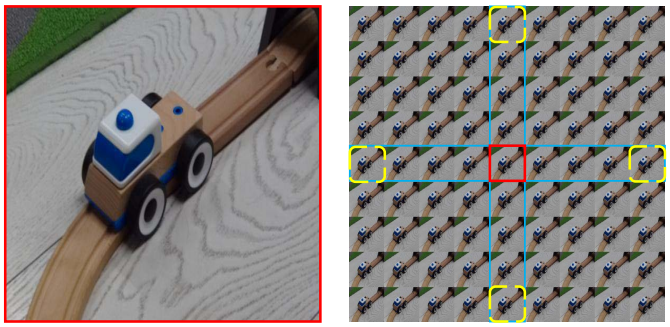


Fig. 3. Illustration of angular view selection in DUTLF-V2. Blue solid rectangles represent all 16 angular views along the horizontal and vertical directions, yellow dashed rectangles represent 4 side views.

method [1], which integrates the multi-view saliency maps in a view-wise attention fashion. A more efficient strategy aiming at exploring and modeling the spatial correlations between multi-view images should be considered. To solve this issue, a light-field-driven saliency detection network that consists of three major components is proposed. First, a spatial feature extractor is proposed for extracting spatial saliency features from the central image and the generated side views. Then a Direction-specific Screening Unit (DSU) is designed to effectively exploit the spatial correlation between the central view and side views along the horizontal and vertical directions. Finally, a cascaded decoder is adopted to integrate the multi-level features and make final prediction.

B. Light Field Synthesis Network

In what follows, we introduce the design of the light field synthesis network, as illustrated in Figure 2. First, the depth CNNs, represented by $d_u(\cdot)$ and $d_v(\cdot)$, are applied to estimate depth maps $D_u(x, y, u)$ and $D_v(x, y, v)$ along horizontal and vertical directions, respectively:

$$\begin{aligned} D_u(x, y, u) &= d_u(I(x, y); \theta_u) \\ D_v(x, y, v) &= d_v(I(x, y); \theta_v) \end{aligned} \tag{1}$$

where (x, y) represents spatial coordinates and $I(x, y)$ is the central view of the light field. (u, v) represents angular coordinates, and the values of $u, v \in \{-1, 1\}$ to generate depth maps corresponding to side views along the horizontal and vertical directions, respectively. The two depth CNNs have the same structure but different parameters θ_u and θ_v. The depth CNNs are developed based on the view synthesis network [40], in which there are ten convolutional layers including four dilated convolutional layers. The dilated convolution is used to obtain a large receptive field. The output channels of the last two convolutional layers are modified to the number of the views in horizontal and vertical directions. The detailed network architecture can be found in [40].

Then the central image and the predicted depth maps are fed into the warping layer to render other viewpoints of light field. The physically-based warping layer is the core component of the light field synthesis network. This process can be expressed as follows:

$$\begin{aligned} L_u(x, y, u) &= I(x + uD_u(x, y, u), y) \\ L_v(x, y, v) &= I(x, y + vD_v(x, y, v)) \end{aligned} \tag{2}$$

where $L_u(x, y, u)$ and $L_v(x, y, v)$ are the predicted multi-view images. After rendering the new viewpoints, the reconstruction error between the predicted images and the ground truths is calculated. Here, a simple L_1 loss is used to supervise the reconstruction quality:

$$\ell_{re} = \left\| L_u(x, y, u) - \hat{L}_u(x, y, u) \right\|_1 + \left\| L_v(x, y, v) - \hat{L}_v(x, y, v) \right\|_1 \tag{3}$$

where $\hat{L}_u(x, y, u)$ and $\hat{L}_v(x, y, v)$ represent the ground truths.

To further improve the quality of depth maps, the consistency regularization loss ℓ_c and the total variation regularization loss ℓ_{tv} proposed by [40] are applied in the light field rendering network. The parameters of the depth CNNs are updated by minimizing the final loss:

$$\min_{\theta_u, \theta_v} \sum_{T} (\ell_{re} + \lambda_c \ell_c + \lambda_{tv} \ell_{tv}) \tag{4}$$

where T denotes the training set. λ_c and λ_{tv} are the weight of consistency regularization loss and the weight of total variation regularization loss, respectively.

C. Light-Field-Driven Saliency Detection Network

The proposed light-field-driven saliency detection network can be divided into three parts: spatial feature extractor, direction-specific screening unit and cascaded decoder. More details are described as follows.

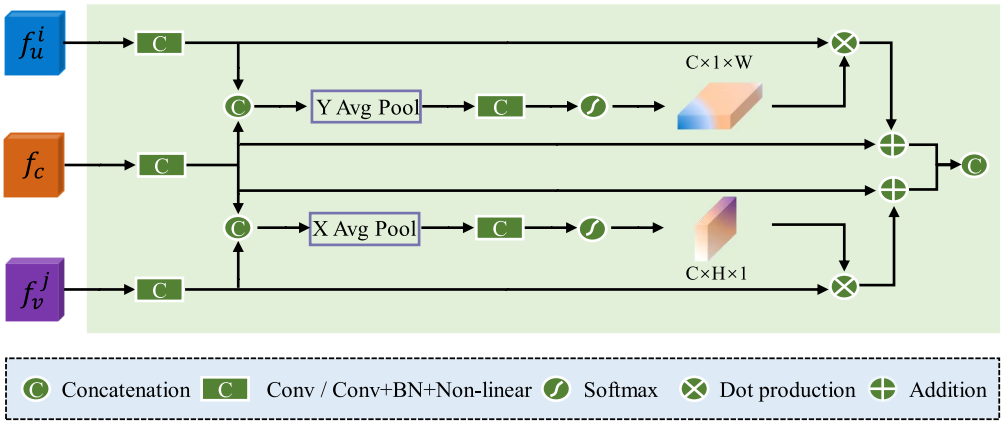


Fig. 4. Detailed structure of the Direction-specific Screening Unit (DSU). The DSU consists of two branches, each of which respectively exploits the spatial correlation along the horizontal and vertical directions, and supplements the central view with direction-specific screened information.

1) Spatial Feature Extractor: The spatial feature extractor takes an array of multi-view images $I = \{I_c, I_u^i, I_v^j\}$ as input, including the central view I_c, and the generated side views I_u^i, I_v^j, where $i, j \in \{-1, 1\}$ corresponding to the two side views in horizontal and vertical directions, respectively. The widely-used VGG-16 [44] is adopted as the backbone architecture. Specifically, we drop the last pooling layer and fully-connected layers, and preserve the five convolutional blocks for better fitting for the task of saliency detection. Then the high-level features (F_{conv}^3, F_{conv}^4, and F_{conv}^5) are selected to detect salient objects. For facilitating capturing parallax between different views, receptive field block (RFB) [45] is attached in each level to incorporate more global contrast information.

2) Direction-Specific Screening Unit (DSU): To explore the intrinsic geometric characteristics of multi-view images in a deeper insight, a Direction-specific Screening Unit (DSU) is proposed. DSU exploits the spatial correlation between the central view and the side views, and further supplements the central view with the screened direction-specific information which can facilitate to locate and segment the salient objects more precisely. The illustration of the proposed DSU is shown in Figure 4.

Considering the parallax of the synthesized side views with respect to the central view along the horizontal and vertical directions, the DSU is designed in a dual-branch manner. Specifically, 1×1 convolutions are first applied on high-level feature $F\{f_c, f_u^i, f_v^j\}$ extracted from different views:

$$\hat{f}_c = Conv_1(f_c),\ \hat{f}_u^i = Conv_1(f_u^i),\ \hat{f_v^j} = Conv_1(f_v^j) \quad (5)$$

It is noteworthy that the convolution operations for different views do not share parameters.

Then the feature maps of the center view and the side views along the horizontal and vertical directions are concatenated. To take advantage of the parallax information, a pooling operation with band shape pooling window [46] $(1, W)$ or $(H, 1)$ is used to encode each channel along the vertical or horizontal dimension, respectively. Formally, this process can be written as:

$$\begin{aligned} \tilde{f}_u^i &= AvgPool_y(Cat(\hat{f}_c, \hat{f}_u^i)) \\ \tilde{f_v^j} &= AvgPool_x(Cat(\hat{f}_c, \hat{f_v^j})) \end{aligned} \quad (6)$$

where $Cat(\cdot)$ is the concatenation operation, $AvgPool_y$ and $AvgPool_x$ represent $AvgPooling$ operation performed by a pooling window shaped as $(1, W)$ or $(H, 1)$, respectively. Thus the aggregated feature maps regarding specific directions are obtained (separately denoted as '$\tilde{f}_u^i$' and '$\tilde{f_v^j}$').

In order to screen the spatial-discriminative and semantic-effective information, each of the dual branch conducts a series of convolution and activation operations:

$$\begin{aligned} S_u^i &= \sigma(Conv(\tilde{f}_u^i)) \\ S_v^j &= \sigma(Conv(\tilde{f_v^j})) \end{aligned} \quad (7)$$

where $Conv(\cdot)$ represents a sequential of operations including convolution, batch normalization, and activation. And a 1×1 convolution is applied to transorm the channel numbers. σ denotes the sigmoid function.

The enhanced feature representation of side views regarding horizontal and vertical directions is supplemented into the central view, which helps the network highlight the objects of interest more precisely. Lastly the generated feature maps are concatenated. This process can be listed as follows:

$$Z_u^i = S_u^i * \hat{f}_u^i + \hat{f}_c \quad (8)$$

$$Z_v^j = S_v^j * \hat{f_v^j} + \hat{f}_c$$

$$F_d = Cat(Z_u^i, Z_v^j) \quad (9)$$

where '$*$' represents Hadamard product, Z_u^i and Z_v^j denote the central view supplemented by side views with corresponding directions. F_d denotes the final output of DSU.

3) Cascaded Decoder: After obtaining the aggregated feature maps from the DSU, we need to hierarchically incorporate the spatial attentive context and the multi-scale context. This motivates us to apply a cascaded decoder [29] to fully integrate the features in multiple levels and make the dense prediction. Specifically, the cascaded decoder takes the three levels of features form DSU (F_d^3, F_d^4 and F_d^5) as input. Then a pyramid multiplication streategy is conducted to alleviate the gap between different levels, yielding

$$\begin{aligned} \tilde{F}_d^l &= F_d^l, l = 5 \\ \tilde{F}_d^l &= F_d^l \odot \prod_{k=l+1}^{K} Conv_3(UP(F_d^k)), l \in \{3, 4\} \end{aligned} \quad (10)$$

where $K = 5$, $UP(\cdot)$ represents up-sampling operation, $Conv_3(\cdot)$ represents 3×3 convolution layer, '$\odot$' represents the element-wise multiplication. Then these features are combined with the improved feature representation by a progressive concatenation operation. The final prediction is obtained by a 1×1 concovolation and up-sampling operation. The loss function for the light-field-driven saliency detection network is the cross entropy loss [47] which can be formulated as:

$$L_{CE} = Slog\hat{S} + (1 - S)log(1 - \hat{S}) \tag{11}$$

where $\hat{S}$ and S denote the saliency prediction and the ground truth, respectively.

IV. Experiments

A. Experimental Setup

1) Datasets: Our originally proposed light field dataset DUT-MV [1] consists of 1580 samples. Each light field consists of a RGB image, a corresponding ground truth and multi-view images along horizontal and vertical directions. However, the focal stacks are not provided. Currently the DUT-MV dataset has been extended to a more unified benchmark DUTLF-V2 [22], which includes rich data types (RGB images, corresponding manually labeled ground truths, depth maps, stacks of focal slices, and multi-view images). In order to provide a comprehensive comparison with current 2D, 3D and 4D state-of-the-arts, experiments are conducted on DUTLF-V2, as well as other two light fiels datasets, the LYTRO ILLUM [48] and the HFUT-LFSD [36].

DUTLF-V2 offers a generic benchmark for salient object detection. It contains 4204 samples and is officially split into 2957 for training and 1247 for testing. Each light field consists of a RGB image, a corresponding ground truth, a depth image, a stack of focal slices and an array of multi-view images with 9×9 angular resolution.

LYTRO ILLUM is consist of 640 light fields. This dataset provides the central images, depth maps, the corresponding per-pixel ground-truth saliency maps, and micro-lens image array which can be transformed into sub-aperture images with 9×9 resolution. The samples of LYTRO ILLUM are officially divided into five groups, four groups (512 samples) for training, and one group (128 samples) for test.

HFUT-LFSD contains 255 light fields captured by Lytro camera and is divided into 100 training samples and 155 testing samples. This dataset provides all-in-focus images, focal stacks, multi-view images with 7×7 angular resolution, coarse depth maps and ground truth masks. It is also a challenging dataset with real-life scenarios.

We selected 2957 samples from the DUTLF-V2 dataset and 512 samples from the LYTRO ILLUM dataset as the training set following their official divisions. The test set of DULF-V2 (1247 samples), LYTRO ILLUM (128 samples) and HFUT-LFSD (155 samples) are adopted for model evaluation. To prevent overfitting, the training set is extended by general data augmentation schemes (*e.g.,* flipping, cropping and rotating).

2) Implementation Details: The proposed method is implemented with Pytorch toolbox and trained on a PC with GTX 2080Ti GPU. All training images are uniformly resized to 256×256. The whole network is trained in an end-to-end manner using the Adam optimization algorithm. For light field synthesis, the learning rate is set to 1e-04. The weights of consistency regularization loss λ_c and total variation regularization loss λ_{tv} are experimentally set to 1e-02 and 1e-03. More analyses about the settings of λ_c and λ_{tv} are listed in Section IV-C. For saliency detection, the initial learning rate is set to 1e-04. The whole training time is about 10 hours. For inference, it takes 0.087 seconds for an image on the aforementioned GPU (about 11 FPS). The size of the proposed model is 133.0 MB.

3) Evaluation Metrics: To comprehensively evaluate the performance of various salient object detection methods, five widely-used evaluation metrics are adopted, including F-measure (F_β) [49], weighted F-measure (F_β^w) [50], E-measure (E_s) [51], S-measure (S_α) [52], and Mean Absolute Error (MAE). Specifically, F-measure is computed by the weighted harmonic mean of the precision and recall:

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

where β^2 is set to 0.3 as suggested in [53]. Weighted F-measure is adopted to overcome the interpolation flaw, dependency flaw and equal-importance flaw for a fair comparison,

$$F_\beta^w = \frac{(1 + \beta^2) \times Precision^w \times Recall^w}{\beta^2 \times Precision^w + Recall^w} \tag{13}$$

where w is a weighting function based on the Euclidean distance. E-measure can jointly capture image-level statistics and local pixel matching information:

$$E_s = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_s(i, j) \tag{14}$$

where $\phi_s(\cdot)$ is the enhanced alignment matrix. S-measure evaluates the structural similarity between the real-valued saliency map and the binary ground truth:

$$S_\alpha = \alpha * S_o + (1 - \alpha) * S_r \tag{15}$$

where S_o and S_r represent the object-aware and region-aware similarities, respectively. α is the balance parameter and is set to 0.5. MAE computes the average absolute per-pixel difference between the saliency map and the corresponding ground truth:

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |\hat{S}(i, j) - S(i, j)| \tag{16}$$

where $\hat{S}$ and S denote the normalized saliency map and the ground truth, respectively. The above evaluation metrics can provide standard and reliable evaluation results.

B. Comparison With State-of-the-Arts

The proposed method is compared with the conference version, DLFS [1] and other 22 state-of-the-art 2D, 3D and 4D methods, including both learning-based methods and

TABLE I

QUANTITATIVE COMPARISONS OF E-MEASURE, S-MEASURE, WEIGHTED F-MEASURE, F-MEASURE AND MAE SCORES ON THREE LIGHT FIELD DATASETS. BOLD: BEST, UNDERLINE: SECOND BEST (∗ REPRESENTS CONVENTIONAL METHODS, - MEANS NO AVAILABLE RESULTS)

Type	Methods	Years	DUTLF-V2					LYTRO ILLUM					HFUT-LFSD				
			$E_s\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE↓	$E_s\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE↓	$E_s\uparrow$	$S_\alpha\uparrow$	$F_\beta^w\uparrow$	$F_\beta\uparrow$	MAE↓
4D	DLFS	-	.868	.810	.676	.739	.080	.876	.830	.712	.774	.072	.799	.759	.611	.650	.097
4D	Ours	-	**.931**	**.882**	**.818**	**.852**	**.041**	**.936**	**.906**	**.858**	**.884**	**.038**	.848	.789	.672	.724	**.061**
	△gains		*↑.063*	*↑.072*	*↑.142*	*↑.113*	*↑.039*	*↑.060*	*↑.076*	*↑.146*	*↑.110*	*↑.034*	*↑.049*	*↑.030*	*↑.061*	*↑.074*	*↑.036*
4D	DLGLRG	ICCV'21	.908	.861	.780	.816	.046	-	-	-	-	-	.843	.765	.634	.709	.071
	LFNet	TIP'20	.915	.873	.799	.819	.047	-	-	-	-	-	.852	**.807**	.693	.718	.062
	ERNet	AAAI'20	.924	.852	.792	**.852**	.050	-	-	-	-	-	**.858**	.778	.687	**.753**	.069
	MoLF	NIPS'19	.915	.877	.803	.821	.047	-	-	-	-	-	.851	.795	.684	.722	.068
	DILF*	IJCAI'15	.733	.648	.388	.504	.187	.785	.731	.500	.608	.149	.736	.695	.458	.555	.131
	LFS*	CVPR'14	-	-	-	-	-	-	-	-	-	-	.686	.579	.264	.430	.205
3D	DCFNet	CVPR'21	.878	.821	.732	.772	.065	.908	.879	.823	.845	.046	.822	.779	.670	.703	.077
	HAINet	TIP'21	.886	.845	.760	.794	.060	.911	.882	.818	.848	.049	.792	.755	.628	.672	.097
	S2MA	CVPR'20	.844	.803	.679	.729	.087	.883	.868	.776	.802	.061	.770	.729	.573	.616	.112
	A2dele	CVPR'20	.888	.836	.771	.807	.048	.898	.854	.800	.834	.050	.833	.782	.688	.715	.069
	SSF	CVPR'20	.917	.869	.804	.831	.043	.903	.874	.810	.836	.049	.835	.781	.673	.714	.068
	BBSNet	ECCV'20	.893	.851	.754	.794	.059	.902	.874	.802	.834	.050	.806	.759	.612	.675	.086
	DMRA	ICCV'19	.897	.822	.740	.800	.060	.903	.845	.781	.837	.059	.844	.765	.646	.705	.073
2D	MSFNet	ACMMM'21	.888	.848	.784	.815	.050	.906	.882	.835	.857	.049	.809	.776	.684	.712	.083
	PurNet	TIP'21	.886	.849	.771	.802	.059	.918	.897	.846	.862	.046	.825	.798	**.703**	.721	.084
	MINet	CVPR'20	.870	.828	.736	.781	.065	.890	.864	.795	.825	.057	.800	.777	.670	.704	.090
	GCPANet	AAAI'20	.869	.838	.743	.782	.071	.888	.874	.800	.826	.057	.788	.772	.663	.683	.108
	F^3Net	AAAI'20	.878	.841	.756	.803	.063	.901	.876	.812	.843	.052	.810	.776	.673	.707	.094
	EGNet	ICCV'19	.855	.821	.710	.746	.078	.902	.884	.820	.843	.052	.794	.772	.634	.672	.094
	CPD	CVPR'19	.886	.836	.753	.794	.062	.895	.874	.802	.831	.052	.811	.766	.652	.691	.097
	PoolNet	CVPR'19	.876	.832	.732	.774	.069	.889	.867	.785	.820	.059	.803	.777	.652	.685	.092
	R^3Net	IJCAI'18	.842	.767	.665	.712	.083	.901	.857	.807	.837	.051	.741	.726	.622	.663	.136

TABLE II

QUANTITATIVE COMPARISONS OF CROSS ENTROPY ON THREE LIGHT FIELD DATASETS. BOLD: BEST

Type	Methods	DUTLF-V2	LYTRO ILLUM	HFUT-LFSD
4D	DLFS	.082	.132	.142
4D	Ours	**.062**	**.106**	.104
4D	DLGLRG	.079	-	.154
	LFNet	.066	-	**.098**
	ERNet	.079	-	.136
	MoLF	.083	-	.112
3D	DCFNet	.098	.110	.235
	HAINet	.066	.152	.167
	S2MA	.077	.111	.108
	A2dele	.086	.136	.215
	SSF	.065	.124	.143
	BBSNet	.074	.111	.183
	DMRA	.067	.124	.105
2D	MSFNet	.116	.144	.280
	PurNet	.072	.119	.143
	MINet	.076	.117	.130
	GCPANet	.110	.109	.238
	F^3Net	.101	.109	.169
	EGNet	.076	.116	.117
	CPD	.072	.112	.183
	PoolNet	.077	.182	.114
	R^3Net	.179	.185	.466

conventional ones (marked with '*'). There are six 4D light field methods: DLGLRG [54], LFNet [20], ERNet [21], MoLF [19], DILF [35] and LFS [33]; seven 3D RGB-D methods: DCFNet [32], HAINet [31], S2MA [55], A2dele [24], SSF [56], BBSNet [57], and DMRA [16]; and nine 2D RGB methods: MSFNet [58], PurNet [59], MINet [23], GCPANet [60], F^3Net [27], EGNet [47], CPD [29], PoolNet [61] and R^3Net [28].

1) Comparison With the Conference Version: For comparison of the improved method with the previous version DLFS [1], we retrain the DLFS on the DUTLF-V2 dataset. Comparison results are presented in Table I. The consistent improvements on all metrics across three light field datasets powerfully verify the superiority of the improved method. Specifically, compared with the conference version, the performance of the proposed method represents a 48.7%, 47.2%, and 37,1% increase towards MAE on DUTLF-V2, LYTRO ILLUM, and HFUT-LFSD dataset, respectively. It is worth noting that the performance improvement of the proposed method on the DUTLF-V2 and LYTRO ILLUM datasets is obviously higher than that on the HFUT-LFSD dataset. The main reason is that the three light field datasets vary in the quality of depth. The DUTLF-V2 and LYTRO ILLUM datasets provide high-quality scene depth, which facilities the utilization of the spatial geometry information. While in the HFUT-LFSD dataset, there exist many scenes with poor depth quality, which hardly provide reliable information and even a negative impact on the performance.

Meanwhile, qualitative comparisons are provided in Figure 5. It can be easily seen that the improved method can produce more accurate and complete saliency maps with high spatial consistency. Note that we obtain such significant improvements by only adopting four side views along the

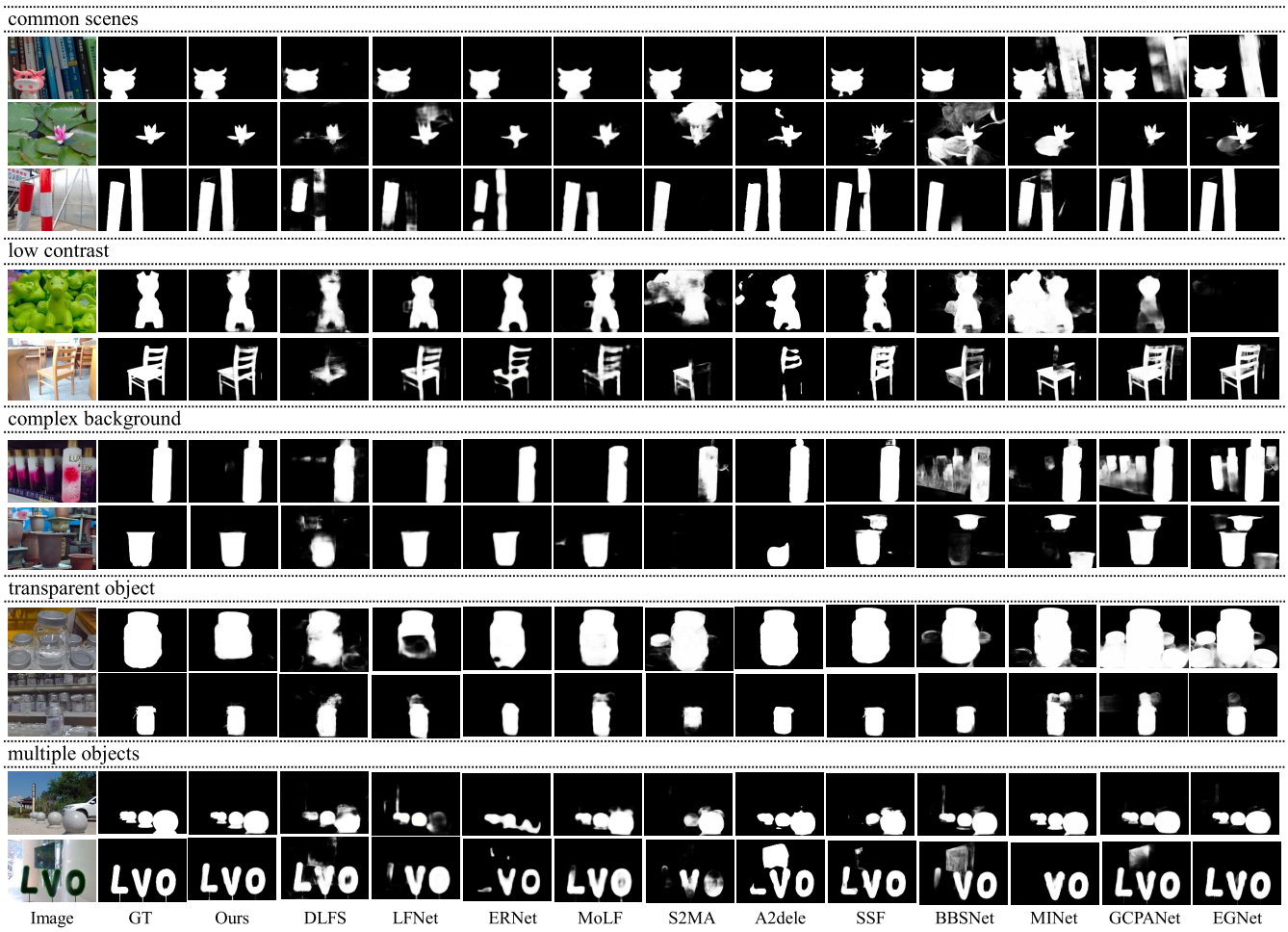


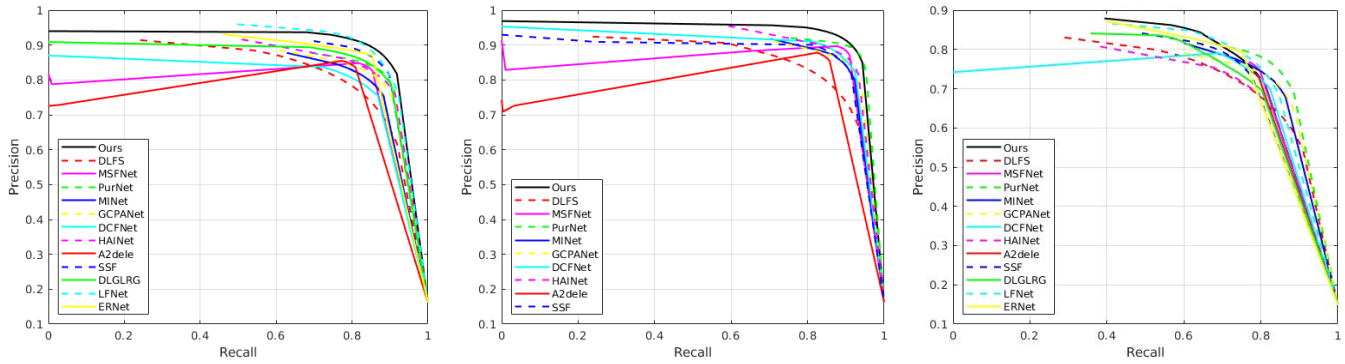
Fig. 5. Visual comparisons of the proposed method with top-ranking CNNs-based methods in some challenging scenes, including low contrast, complex background, transparent object and multiple objects.

horizontal and vertical directions, while the DLFS utilized all 16 views along the horizontal and vertical directions. This manifests that the proposed method can make full use of the spatial parallax information provided by multi-view images compared with the conference version that adopted a view-wise attention mechanism to integrate multi-view saliency maps.

2) Comparison With State-of-the-Arts: As the DUTLF-V2 dataset is an extended version, it can not be ignored that there is a small intersection between the training set of DUT-LFSD [19] and the test set of DUTLF-V2. To make a fair comparison, we retrain the methods on DUTLF-V2 that are originally trained on DUT-LFSD (*i.e.,* LFNet, ERNet and MoLF), following the released codes and the recommended parameter settings provided by the authors. Comparison results are listed in Table I. As illustrated in Table I, it is obvious that the proposed method achieves superior performance than current state-of-the-arts on DUTLF-V2 and LYTRO ILLUM dataset in terms of five metrics. Meanwhile, the proposed method achieves Top-2 F-measure and TOP-1 MAE on the HFUT-LFSD dataset. It is worth mentioning that compared with RGB methods that training with large-quantity datasets, the proposed method achieves significant advantages with a three times smaller training set (3469 *vs.* 10553). Moreover, the Cross Entropy (CE) are compared among state-of-the-art methods in order to evaluate the distribution difference between the predicted saliency maps and the ground truths. For CE, lower value is better. Corresponding results on three datasets are listed in Table II. It can be seen that the proposed method achieves TOP-1 result on the DUTLF-V2 and LYTRO ILLUM datasets and TOP-2 result on the HFUT-LFSD dataset. This more comprehensively demonstrates that the saliency maps predicted by the proposed method are closer to the ground truths.

In addition, the proposed method is compared with state-of-the-art models in terms of PR curves. PR curves describe the relationship between precision and recall, and the closer to the upper right, the better the performance of method. As illustrated in Figure 6, the proposed method outperforms other methods on the DUTLF-V2 and LYTRO ILLUM datasets, and achieves comparable performance on the HFUT-LFSD dataset. This comprehensively demonstrates that the proposed method achieves more accurate predictions.

For a more intuitive view, some representative results generated from the proposed method and other top-ranking CNNs-based approaches are visualized in Figure 5. It can be seen



(a) DUTLF-V2 (b) LYTRO ILLUM (c) HFUT-LFSD

Fig. 6. PR curves of the proposed method and other representative state-of-the-art methods on three light field saliency datasets.

TABLE III
QUANTITATIVE RESULTS OF THE ABLATION ANALYSIS FOR ANGULAR VIEW SELECTION. BOLD: BEST

Methods	Time(s)	DUTLF-V2		LYTRO ILLUM		HFUT-LFSD	
		$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$
Ours-2(H)	**0.049**	.829	.048	.859	.050	.707	.071
Ours-2(V)	**0.049**	.824	.050	.854	.047	.696	.077
Ours	0.087	.852	.041	**.884**	**.038**	**.724**	**.061**
Ours-8	0.094	**.854**	.041	.879	.039	.723	.065
Ours-16	0.189	.850	**.040**	.875	**.038**	.722	.067
Ours-20	0.193	.853	**.040**	.873	.039	**.724**	.064

TABLE IV
QUANTITATIVE RESULTS OF THE ABLATION ANALYSIS FOR THE PROPOSED DSU. BOLD: BEST

Methods	DUTLF-V2		LYTRO ILLUM		HFUT-LFSD	
	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$
Ours-add	.814	.055	.846	.051	.692	.078
Ours-H	.835	.045	.867	.041	.708	.071
Ours-V	.838	.044	.864	.042	.704	.070
Ours	**.852**	**.041**	**.884**	**.038**	**.724**	**.061**

that results of the proposed method are more consistent with the ground truths. When confronting with these challenging scenes including low contrast (row 4, 5), complex background (row 6, 7), transparent object (row 8, 9) and multiple objects (row 10, 11), most RGB-based and RGB-D based methods fail to detect the salient objects, while the proposed method can successfully generate accurate and robust saliency maps. When comparing with these CNNs-based light field methods, the proposed method also achieves more consistent prediction results with finer details. This further verifies the effectiveness of the proposed method and the superiority of exploiting spatial correlation in light field data.

C. Ablation Study

1) Angular View Selection: In the conference version [1], we synthesized the angular views along the horizontal and vertical directions and supplemented the saliency detection network with the synthesized sixteen views. However, we find that the angular views share small differences due to the limited parallax, as shown in Figure 3. This brings redundant information and computational challenges for salient object detection. In this paper, only four side views along the horizontal and vertical directions are synthesized to facilitate the central view with spatial parallax information. To explore the impact of the number of selected views, detailed experiments are conducted under different settings: i) employing only two side views along the horizontal direction with respect to the central view (yellow dashed rectangles in the horizontal direction in Figure 3), denoted as 'Ours-2H'; ii) employing only two side views along the vertical direction with respect to the central view (yellow dashed rectangles in the vertical direction in Figure 3), denoted as 'Ours-2V'; iii) employing four side views along the horizontal and vertical directions with respect to the central view (yellow dashed rectangles in Figure 3), denoted as 'Ours'; iv) employing eight evenly spaced views along the horizontal and vertical directions with respect to the central view (the first, third, seventh, and ninth yellow dashed rectangles in the horizontal direction as well as the vertical direction in Figure 3), denoted as 'Ours-8'; v) employing all sixteen views along the horizontal and vertical directions with respect to the central view (blue solid rectangles in Figure 3), denoted as 'Ours-16'; vi) based on the setting of v, employing more views at four vertices in Figure 3, totally twenty angular views, denoted as 'Ours-20'. Comparison results are shown in Table III. It can be seen that 'Ours-2H' or 'Ours-2V' leads to an unsatisfactory result on three datasets, due to the absence of geometric information in the other direction. When adopting four side views, 'Ours' shows significant performance gains. It is worth noting that when the number of views further increases, the performance saturates and fluctuates in the thousandths. This demonstrates that the four side views along the horizontal and vertical directions are sufficient to represent the original light field data. Meanwhile, increasing the number of views inevitably brings burden for computation. For example, compared with 'Ours', 'Ours-16' tremendously increases the inference time by 117%. This further verifies that the angular view selection of 'Ours' achieves a better speed-accuracy trade off.

2) Effectiveness of the Direction-Specific Screening Unit (DSU): One of the core claims of this paper is that exploring the spatial correlation between side views and the central view

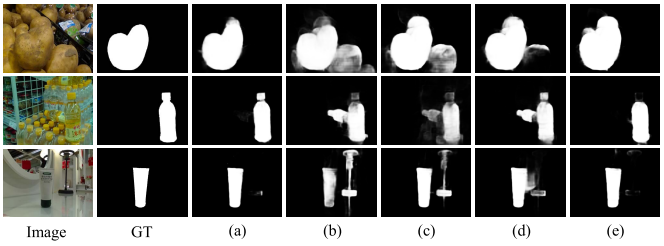


Fig. 7. Visual results of the ablation study. (a) represents the output saliency map of 'Ours'. (b), (c) and (d) correspond to the results of 'Ours-add', 'Ours-H' and 'Ours-V', respectively. (e) shows results of 'Ours-nl'.

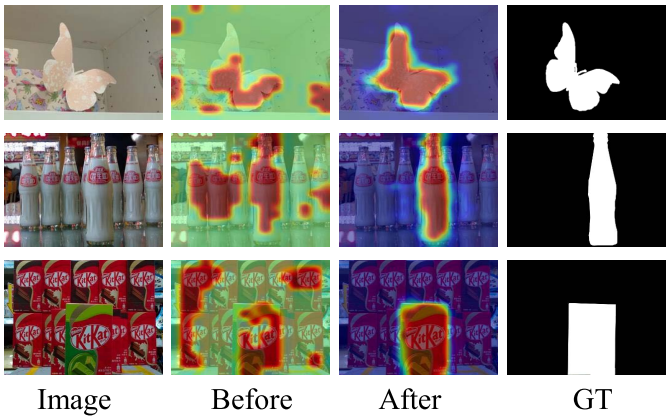


Fig. 8. Visualization for feature maps of several challenging scenes before and after being processed by DSU.

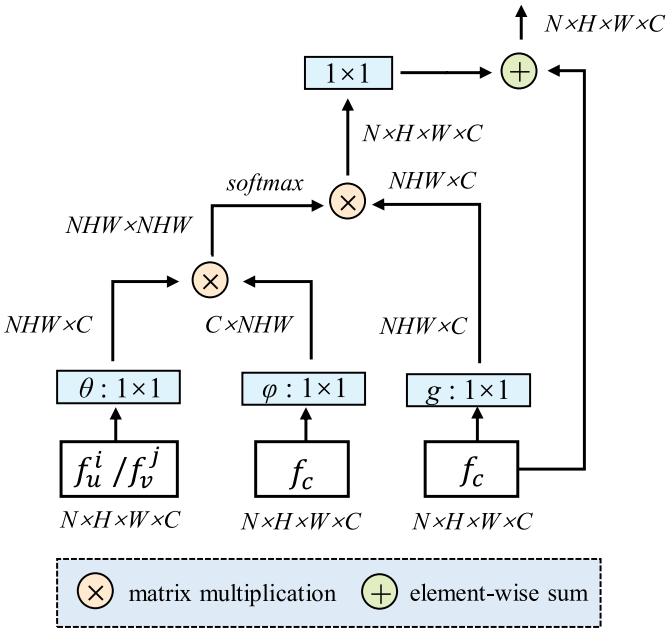


Fig. 9. Detailed structure of non-local block employed in the ablation study.

TABLE V
QUANTITATIVE RESULTS OF COMPARISON WITH NON-LOCAL MECHANISM. BOLD: BEST

Model	Time(s)	DUTLF-V2		LYTRO ILLUM		HFUT-LFSD	
		$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓	$F_\beta \uparrow$	MAE↓
Ours-nl	0.147	.845	**.041**	.870	.040	.718	.068
Ours	**0.087**	**.852**	**.041**	**.884**	**.038**	**.724**	**.061**

facilitates saliency detection to locate and segment the salient objects more precisely. In this section, a series of experiments are conducted to demonstrate the effectiveness of the DSU. Firstly, we simply combine the side views with the central view via simple addition, denoted as 'Ours-add'. It can be seen from the quantitative results in Table IV and the visual comparison results in Figure 7 that simple addition fails to take advantage of the spatial parallax information and leads to unsatisfactory results. Secondly, ablation studies are conducted under the settings that preserving the DSU in one direction and employing addition in another direction, denoted as 'Ours-H' and 'Ours-V' respectively. Specifically, in the setting of 'Ours-H', DSU is applied to combine the side views along the horizontal direction with respect to the central view, while the side views along the vertical direction are combined with the central view by addition. The setting of 'Our-V' is and vice versa. Comparison results in Table IV verify that even exploring spatial correlation in one direction shows impressive performance gains compared with 'Ours-add'. For example, compared to 'Ours-add', 'Ours-H' numerically reduces the MAE by 18.1%, 19.6%, 8.9% on DUTLF-V2, LYTRO ILLUM and HFUT-LFSD dataset, respectively. And 'Ours-V' reduces the MAE by 20.0%, 17.6%, 10.2% on DUTLF-V2, LYTRO ILLUM and HFUT-LFSD dataset, respectively. Finally, when applying the DSU to screen and supplement spatial information in a dual-branch manner (denoted as 'Ours'), we obtain the sweet point of performance across three light field datasets. Meanwhile, the corresponding visual results shown in Figure 7 also illustrate that the exploration of spatial information can help the central view to identify and segment the salient objects more accurately.

To further understand why DSU can achieve such improvements, we visualize the feature maps of three challenging scenes before and after being processed by the DSU, denoted as 'Before' and 'After' in Figure 8. It can be easily seen that the salient regions are emphasized and the background are suppressed after adopting DSU. This powerfully illustrates that the exploration of spatial correlation in DSU can significantly facilitate the saliency detection with finer prediction and more accurate results.

Moreover, to further study the effectiveness of the DSU in dealing with subtle spatial parallax in multiple viewpoints, we perform comparison with non-local mechanism [62] which is demonstrated effective in capturing spatial dependencies. Specially, the DSU in the proposed method is removed, then the non-local mechanism is employed to build spatial correlation between the side views and the central view. The detailed structure is illustrated in Figure 9. Finally the supplemented central views are concatenated and fed into the decoder. As suggested in [62], different choices of affinity measuring function have similar performance, thus the embedded Gaussian version is adopted. For a fair comparison, the non-local block is adopted at three high levels, which is consistent with DSU. The quantitative comparison results are shown in Table V. It can be seen that 'Ours' outperforms 'Ours-nl' towards all

TABLE VI
QUANTITATIVE RESULTS UNDER DIFFERENT SETTINGS OF THE WEIGHTS FOR DEPTH REGULARIZATION IN THE LIGHT FIELD SYNTHESIS NETWORK. BOLD: BEST

Settings	DUTLF-V2		LYTRO ILLUM		HFUT-LFSD	
	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$	$F_\beta \uparrow$	MAE$\downarrow$
λ_c=5e-01, λ_{tv}=5e-01	.849	**.039**	.879	.039	.720	.064
λ_c=1e-04, λ_{tv}=1e-04	.844	.043	.873	**.038**	.713	.068
λ_c=1e-02, λ_{tv}=1e-03	**.852**	.041	**.884**	**.038**	**.724**	**.061**

the evaluation metrics across three datasets. More importantly, we calculate the inference time and find that 'Ours' minimizes the inference time by 40.8% compared with that of 'Ours-nl'. This demonstrates the superiority of the proposed method in learning spatial correlation of high-dimentional light field data.

3) Settings of the Weights for Depth Regularization: In the light field synthesis network, the consistency regularization loss and total variation regularization loss are regularizations for the predicted depths. These two losses encourage the predicted depths to be consistent across views and to be sparse in the spatial gradient domain, further improving the quality of depths. In this paper, the main target of the light field synthesis network is to synthesize the multi-view images, meanwhile, the quality of depths should also be considered because depths are involved in the process of view synthesis. Based on the above considerations, this paper follows the previous version [1] to set weights of consistency regularization loss and total variation regularization loss to be 1e-02 and 1e-03, respectively. To further explore the impact of the weights for depth regularization, experiments under different settings are conducted. Corresponding experimental results are listed in Table VI. It can be seen that if the weights are too large or too small, depths are not properly involved in the process of view synthesis, resulting in suboptimal performance. The setting of 1e-02 and 1e-03 achieves a better balance for the light field synthesis network and leads to better performance.

V. CONCLUSION AND FUTURE SCOPE

In this paper, a novel end-to-end framework is proposed to detect salient objects in challenging scenes. This paper shows for the first time that the saliency detection is decomposed into two sub-tasks: light field synthesis and light-field-driven saliency detection. The light field synthesis network generates high-quality 4D light fields from a single view. The light-field-driven saliency detection network exploits spatial correlation in multi-view images and builds the relationship between salient objects and scene understanding. Extensive quantitative and qualitative evaluations demonstrate that the proposed method outperforms the state-of-the-art 2D, 3D and 4D methods on three light field datasets and is capable of segmenting salient objects in challenging scenes.

Though the proposed method shows a 450% improvement on inference speed (11 FPS *vs.* 2 FPS) compared with the previous version [1], it still falls short of meeting real-time requirements like existing light filed methods (*e.g.*, 14 FPS for DLGLRG [54], 13 FPS for LFNet [20], 14 FPS for ERNet [21] and 5 FPS for MoLF [19]). The main reason is that extra data (*e.g.*, focal slices, multi-view images) needs to be processed. It is promising to introduce model compression technologies (*e.g.*, knowledge distillation) for light field salient object detection to improve the inference speed. We will devote to exploiting lightweight models for light filed salient object detection in the future work.

REFERENCES

[1] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu, "Deep light-field-driven saliency detection from a single view," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 904–911.

[2] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: A survey," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 84–100, Jan. 2018.

[3] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[4] J. Dai, Y. Li, K. He, and J. Sun, "R-FCN: Object detection via region-based fully convolutional networks," 2016, *arXiv:1605.06409*.

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

[6] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 597–606.

[7] A. W. M. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah, "Visual tracking: An experimental survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 7, pp. 1442–1468, Jul. 2014.

[8] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 2303–2309.

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

[10] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 110–119.

[11] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2334–2342.

[12] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

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

[14] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proc. CVPR*, 2017, pp. 5300–5309.

[15] J. Han, H. Chen, N. Liu, C. Yan, and X. Li, "CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion," *IEEE Trans. Cybern.*, vol. 48, no. 11, pp. 3171–3183, Nov. 2018.

[16] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7254–7263.

[17] M. Zhang, Y. Zhang, Y. Piao, B. Hu, and H. Lu, "Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 4107–4115.

[18] H. Chen, Y. Deng, Y. Li, T.-Y. Hung, and G. Lin, "RGBD salient object detection via disentangled cross-modal fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 8407–8416, 2020.

[19] M. Zhang, J. Li, W. Ji, Y. Piao, and H. Lu, "Memory-oriented decoder for light field salient object detection," in *Proc. NIPS*, vol. 32, 2019, pp. 898–908.

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

[21] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 11865–11873.

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

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

[24] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 9060–9069.

[25] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jul. 2017, pp. 3203–3212.

[26] C. Fang, H. Tian, D. Zhang, Q. Zhang, J. Han, and J. Han, "Densely nested top-down flows for salient object detection," 2021, *arXiv:2102.09133*.

[27] J. Wei, S. Wang, and Q. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 12321–12328.

[28] Z. Deng *et al.*, "R^3xNet: Recurrent residual refinement network for saliency detection," in *Proc. IJCAI*, 2018, pp. 684–690.

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

[30] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 7, pp. 1755–1769, Jul. 2020.

[31] 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, 2021.

[32] W. Ji *et al.*, "Calibrated RGB-D salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9471–9481.

[33] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *Proc. CVPR*, 2014, pp. 2806–2813.

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

[35] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu, "Saliency detection with a deeper investigation of light field," in *Proc. IJCAI*, 2015, pp. 2212–2218.

[36] 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. (TOMM)*, vol. 13, no. 3, p. 32, 2017.

[37] M. Levoy and P. Hanrahan, "Light field rendering," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 31–42.

[38] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proc. 23rd Annu. Conf. Comput. Graph. Interact. Techn.*, 1996, pp. 43–54.

[39] C. Buehler, M. Bosse, L. McMillan, S. J. Gortler, and M. F. Cohen, "Unstructured lumigraph rendering," in *Proc. 28th Annu. Conf. Comput. Graph. Interact. Techn.*, 2001, pp. 425–432.

[40] P. P. Srinivasan, T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng, "Learning to synthesize a 4D RGBD light field from a single image," in *Proc. ICCV*, 2017, pp. 2262–2270.

[41] P. P. Srinivasan, M. W. Tao, R. Ng, and R. Ramamoorthi, "Oriented light-field Windows for scene flow," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3496–3504.

[42] R. Ng and P. M. Hanrahan, "Digital correction of lens aberrations in light field photography," in *Proc. Int. Opt. Design Conf.*, 2006, p. WB2.

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

[44] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–14.

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

[46] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip pooling: Rethinking spatial pooling for scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4003–4012.

[47] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EgNet: Edge guidance network for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8779–8788.

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

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

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

[51] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[52] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. ICCV*, 2017, pp. 4558–4567.

[53] 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.

[54] 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/CVF Int. Conf. Comput. Vis.*, Oct. 2021, pp. 4712–4721.

[55] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for RGB-D saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 13756–13765.

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

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

[58] M. Zhang, T. Liu, Y. Piao, S. Yao, and H. Lu, "Auto-MSFNet: Search multi-scale fusion network for salient object detection," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 667–676.

[59] J. Li, J. Su, C. Xia, M. Ma, and Y. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Trans. Image Process.*, vol. 30, pp. 6855–6868, 2021.

[60] Z. Chen, Q. Xu, and R. Cong, "Global context-aware progressive aggregation network for salient object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 7, pp. 10599–10606.

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

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