Saliency Detection of Light Field Images by Fusing Focus Degree and GrabCut

Fuzhou Duan 1,2, Yanyan Wu 1,2, Hongliang Guan 1,2,3, and Chenbo Wu 1,2

1 Engineering Research Center of Spatial Information Technology, MOE, Capital Normal University, 105 West Third Ring North Road, Haidian District, Beijing 100048, China
2 Key Lab of 3D Information Acquisition and Application, MOE, Capital Normal University, 105 West Third Ring North Road, Haidian District, Beijing 100048, China
3 Academy for Multidisciplinary Studies, Capital Normal University, 105 West Third Ring North Road, Haidian District, Beijing 100048, China
* Correspondence: hlguan@cnu.edu.cn

Abstract: In the light field image saliency detection task, redundant cues are introduced due to computational methods. Inevitably, it leads to the inaccurate boundary segmentation of detection results and the problem of the chain block effect. To tackle this issue, we propose a method for salient object detection (SOD) in light field images that fuses focus and GrabCut. The method improves the light field focus calculation based on the spatial domain by performing secondary blurring processing on the focus image and effectively suppresses the focus information of out-of-focus areas in different focus images. Aiming at the redundancy of focus cues generated by multiple foreground images, we use the optimal single foreground image to generate focus cues. In addition, aiming at the fusion of various cues in the light field in complex scenes, the GrabCut algorithm is combined with the focus cue to guide the generation of color cues, which realizes the automatic saliency target segmentation of the image foreground. Extensive experiments are conducted on the light field dataset to demonstrate that our algorithm can effectively segment the salient target area and background area under the light field image, and the outline of the salient object is clear. Compared with the traditional GrabCut algorithm, the focus degree is used instead of artificial Interactively initialize GrabCut to achieve automatic saliency segmentation.

Keywords: salient object detection; light field; foreground matting; GrabCut

1. Introduction

Light field represents the propagation of light in all directions in three-dimensional space, and it is used to describe the radiative transfer characteristics of light in real space [1]. Since Gershun proposed it in 1936, researchers have never stopped exploring it [2–5]. The early way of recording the light field was through a single camera continuously changed its position through a mechanical arm. It was difficult to capture an entire light field with time inconsistency. Ren (2005) released the first handheld light field camera with the continuous development of imaging device technology. This light field camera’s feature of taking pictures first and then focusing attracted more and more scholars to join the light field technology research field [6]. Saliency object detection (SOD) is an algorithm that allows the fast acquisition of regions of interest in images [7]. Since Gershun proposed it in 1936, researchers have never stopped exploring it [2–5]. The early way of recording the light field was through a single camera continuously changed its position through a mechanical arm. It was difficult to capture an entire light field with time inconsistency. Ren (2005) released the first handheld light field camera with the continuous development of imaging device technology. This light field camera’s feature of taking pictures first and then focusing attracted more and more scholars to join the light field technology research field [6]. Saliency object detection (SOD) is an algorithm that allows the fast acquisition of regions of interest in images [7]. This algorithm can accelerate the process of other advanced computer vision processing tasks such as object recognition and image compression. Therefore, SOD based on light field cameras is highly significant in computer vision tasks.

Figure 1 shows that a light field camera can generate refocus images of different focal depth planes through digital refocusing technology. Based on this, a complete traditional RGB image can be obtained by using multifocal fusion technology [8,9]. At the same time, the depth map of the scene can be obtained through multi-view [10] and epipolar plane...
image (EPI) [11], which allows the use of traditional SOD methods on light field cameras. The focus-stacked images of the light field are used to find images focused on salient objects to generate focus clues. The foreground and background information of the image can be separated using the focus cues to generate a saliency map. The saliency map can provide feature information such as the depth associated with the salient object and constrain the salient areas of the scene [12,13].

![Figure 1](image1.png)

**Figure 1.** Two of the focus-stacked images: (a) Focus on the foreground; (b) Focus on the background.

GrabCut is a semi-automatic binary classification algorithm [14]. Most scenes only need to provide some seed points to achieve accurate segmentation, and the focus clue of the light field implies the seed point information of GrabCut. Combining the advantages of both, a SOD algorithm based on a light field camera is proposed. We compare the results of 2D, 3D, and light field SOD algorithms and the traditional GrabCut algorithm on existing light field datasets, demonstrating the accuracy of our method in focus calculation, foreground image extraction, and saliency results. The contribution of this work is twofold:

- Aiming at the problem of inaccurate focus area obtained by the focus degree calculation method based on the spatial domain in the existing light field SOD and the problem that the generated focus degree cues have redundant information, we propose a focus degree calculation method based on secondary blurring.
- In this paper, we use the foreground and background information contained in the focus cues of the light field to guide GrabCut to generate foreground color space models and background color space models, respectively. This method enhances the interaction between focus cues and colors and compensates for the drawback that GrabCut requires manual interaction.

The organization of this paper is as follows: Section 2 presents the related work of light-field SOD to summarize existing methods and suggest ideas for improvement. Section 3 introduces the focus calculation method and SOD algorithm proposed in this paper. The experimental setup, evaluation measures, experimental comparisons and experimental results are described in Section 4. Finally, Section 5 discusses the experimental results of this paper.

2. Related Works

This section briefly summarizes the SOD algorithms available for light field images based on three image representations: all-in-focus images, depth maps and focus-stacked images.

2.1. SOD Based on All-in-Focus Images

A light field camera can generate an all-in-focus image of the scene by taking a photo at one time, while a traditional single-lens reflex (SLR) camera needs to be adjusted to
different focal planes multiple times before it can be fused to generate an all-in-focus image of the scene [15,16]. However, the all-focus image is easy to obtain, so the early SOD is mostly based on the 2D SOD algorithm. Itti (1998) first proposed a 2D SOD algorithm based on contrasting features, such as color, brightness and orientation, of RGB images [17]. Later, 2D SOD algorithms were mostly based on the improvement of global contrast or local contrast [18–20]. The saliency model that only uses two-dimensional information, such as color, treats the background with the same or similar color as the target object, resulting in the incomplete detection of the target. In order to solve the complex background problem, Yang (2013) and Zhu (2014) used the background before separating the foreground and background in the scene, reducing the background interference [21,22]. Although 2D SOD algorithms are constantly innovating, 2D images only contain light intensity information, so a single clue makes these 2D SOD algorithms require much prior knowledge. At the same time, it is difficult for these algorithms to obtain clean saliency maps in scenes with cluttered backgrounds and similar foreground and background.

2.2. SOD Based on Depth Images

The study by Poggio indicated that 2D-based RGB images ignore the human eye’s perception of scene depth [23]. The depth map contains depth information at any point in the scene, making up for the structural information between objects lost in the RGB image [14]. With the development of 3D cameras, such as Kinect, many depth map-based 3D saliency algorithms have emerged [24,25]. Jost believed that the smaller the depth value, the closer to the target, and using the depth information to post-process the 2D saliency results can distinguish the multiple saliency targets of the scene [26]. Ran considered the global depth structure and generated saliency maps based on the differences between the center and surrounding items of the depth map [27]. Peng proposed a multi-level RGBD saliency detection algorithm based on color and depth cues [28]. The RGBD saliency model proposed by Ren utilizes different global priors, such as background priors and depth priors [22]. These 3D saliency detection methods show that the spatial structure information implied in the depth information can distinguish objects at different depth layers in the scene, improving the results of saliency detection to a certain extent. These 3D SOD algorithms are mainly guided by depth, so when the depth map quality is poor, or the depths of the foreground and background are relatively close, 3D SOD cannot be used to generate accurate saliency maps.

2.3. SOD Based on Focus Stack Images

A focus-stacked image is a collection of images of different focal planes in a scene generated from light field data by digital refocusing techniques. If you can find an image that is focused near a salient area, it is easier to measure its salience based on the degree of focus. Therefore, Li first proposed a light-field SOD algorithm utilizing the focal feature of focused-stack images, but simply linearly fusing color cues and focality cues could not provide full play to their interaction [29]. Wang (2017) used the gradient value as the focus result based on the spatial domain, and the generated saliency map was close to the result of the method based on the frequency domain, which reduced the calculation time of the focus cue, but the gradient was sensitive to the edge information of the image, which was difficult to detect. The focus information of the background is suppressed [30]. Later, he (2017) proposed saliency maps generated by a probabilistic fusion of different cues using a Bayesian framework [31]. To a certain extent, this makes up for the inaccurate detection in different scenarios caused by the fixed weight value when the previous algorithm adopts the linear weighted fusion method. Zhou (2018) adopted the method of multiple K-Means clustering to fuse focus cues and color cues, and the focus message is required in this method [32]. Piao (2019) guided a cellular automata model based on depth information to generate saliency maps combined with color cues [33]. These light field SOD methods demonstrate the effectiveness of the focus cues provided by light field focus-stacked images to help light field cameras to generate high-quality saliency maps. However, these
methods generate focus cues by fusing multiple foregrounds or background images. In the foreground and background with similar depths, the fusion of the focus results of multiple images is prone to clue redundancy, which makes the generated saliency map contours indistinguishable. The chain block effect is pronounced. In addition, the frequency domain-based focus calculation method in the existing light field saliency model has a high time complexity, while the spatial domain-based focus calculation method is calculated by the gradient. For the out-of-focus region of the focal image, the detection result produces a certain error if the edge information is rich.

3. Methodology

This section introduces our fused focus degree with GrabCut for light field image SOD. The input images to our algorithm were focus-stacked images (depth numbers \( d = 1, 2, ..., L \)) and all-in-focus images. The focus-stacked image contains depth information, and the foreground and background images located at different depth layers are extracted according to the background connectivity detection. Therefore, corresponding to the foreground and background area in the all-focus image, respectively, the color Gaussian mixture model of the foreground and background of GrabCut was constructed. Next, the saliency map was iteratively calculated based on the maximum flow minimum cut algorithm of graph cuts.

3.1. Extracting Focusness Information from Focus Stack

As seen in Figure 2, in the sequence of focus-stacked images, the depth position of the clear region moves from front to back as the sequence number of the focal slice increases. For optical field saliency detection, the focal degree of focus stack images is mostly calculated from the frequency domain [31–33]. Although the frequency domain-based focal degree calculation method can obtain more accurate high-frequency components of the images, it has a relatively high time complexity. In a recent light field-based SOD work [30], the focus was measured in the spatial domain by using a gradient operator, which is much faster to process. However, the gradient is sensitive to the grayscale changes between adjacent pixels in the image, and even in the defocused area, certain gradient information is generated. In order to suppress the gradient information in the defocused regions, we propose to perform secondary (multiple) blurring on the original image. It makes the difference of gradient change before and after the focused region of the original image greater and the difference of gradient change in the bokeh region smaller, which makes it easier to enhance and distinguish the results of the focused region and the defocused region in comparison.

![Figure 2. (a) Focus-stacked images. (b) All-in-focus images.](image)

Before using the gradient operator to measure the sharpness of the focus-stacked image, we needed to degrade it. Let \((i, j)\) denote the pixel in the image from the focus...
where \( \mu \) represents the result of the sharpness matrix filtered along the coordinate axis \( x \). One-dimensional sharpness values along the horizontal and vertical directions:

\[
\sigma_i = \left| \frac{\sum_{j=1}^{w} G(i, j)}{\sum_{j=1}^{w}} - \mu \right|
\]

Then, \( \text{Gray}(i, j) \) is the grayscale value of the pixel \((i, j)\) calculated by weighting the RGB value of \((i, j)\):

\[
\text{Gray}(i, j) = R(i, j) \times 0.299 + G(i, j) \times 0.587 + B(i, j) \times 0.114
\]

Next, the gradient value \( G(i, j) \) of any pixel \((i, j)\) of the focus stack image is calculated as:

\[
\begin{align*}
G_x(i, j) &= \text{Gray}(i + 1, j) - \text{Gray}(i, j) \\
G_y(i, j) &= \text{Gray}(i, j + 1) - \text{Gray}(i, j) \\
G(i, j) &= |G_x(i, j)| + |G_y(i, j)|
\end{align*}
\]

The sharpness matrix of each slice in the focus-stacked image can be defined as:

\[
F(i, j) = G_{\text{original}}(i, j) - G_{\text{Reblur}}(i, j)
\]

where \( G_{\text{original}}(i, j) \) and \( G_{\text{reblur}}(i, j) \) are denoted as the original image slice and the degraded image slice in the focus-stacked image, respectively. Then, we contrasted and enhanced the obtained \( F(i, j) \) to suppress the sharpness value of the defocused area.

### 3.2. Filtering Foreground/Background Images with Object Priors

Now, we filtered each image in the focus-stacked image to find the focal images that focus on the foreground and background regions’ focus stack. In recent work based on light field SOD, the position of the focal point is measured by the image’s center point as the filter center to pick out the focal image close to the center point of the image. However, salient objects are not always at the center point of the image [17–19]. In this paper, we used background connectivity detection [22] to measure the background region of the all-in-focus image, and we also obtained the object region through the inversion operation. We took the center point of the object region as our Gaussian filtered object center \((\mu_x, \mu_y)\). \( \text{Gauss}(x, \mu_x) \) represents the result of the sharpness matrix filtered along the coordinate axis \( x \):

\[
\text{Gauss}(x, \mu_x) = \exp \left( -\frac{(x - \mu_x)^2}{2\sigma^2} \right)
\]

where \( \mu_x \) is the center point of the one-dimensional Gaussian filter on the coordinate axis \( x \), and \( \sigma \) controls the band width of the filter. Next, we decomposed the sharpness into two one-dimensional sharpness values along the horizontal and vertical directions:

\[
\begin{align*}
D_x &= \frac{\sum_{i=1}^{m} F(i, j)}{\sum_{i=1}^{m} F(i, j)} \\
D_y &= \frac{\sum_{j=1}^{n} F(i, j)}{\sum_{j=1}^{n} F(i, j)}
\end{align*}
\]
Therefore, we computed the foreground measurement:

$$OS\left( I^k \right) = \exp \left( \frac{\lambda \cdot (-k)}{K} \right) \cdot \left[ \sum_{x=1}^{w} D_x^k \cdot G(x, \mu_x) + \sum_{y=1}^{h} D_y^k \cdot G(y, \mu_y) \right]$$  (8)

where \( k \) represents the serial number of the focus image and \( K \) is the total number of focus stack images. The focus-stacked images are sorted by depth from small to large, so the larger the value of \(-k/K\), the smaller the depth value of the focus image. \( \lambda \) represents the weighting factor of depth. The image with the highest foreground measurement value was selected as the foreground layer, and the sharpness matrix of the foreground layer was denoted as \( G_F \). The image with the smallest foreground measurement was selected as the background layer, and the sharpness matrix of the background layer was denoted as \( G_B \). However, when the depth of the scene is close, the focus area of the background layer may still contain foreground information; therefore, if \( OS_{\text{min}} \geq 0.7 \cdot OS_{\text{max}} \), the sharpness matrix \( G_B \) of the background layer is obtained by inverting the sharpness matrix \( G_F \) of the foreground layer.

3.3. Detecting the Foreground/Background Regions Coarsely

We used the selected sharpness measures of the foreground and background layers to distinguish the foreground and background regions in the all-in-focus image, respectively. The focus degree of this paper is not obvious in the area where the adjacent pixels of the focus area does not change much. Through the simple linear iterative cluster (SLIC) superpixel segmentation algorithm [34], the all-focus image is divided into \( N \) superpixels \( r_i, i = 1, 2, \ldots, N \). We considered the focus degree of each pixel in the superpixel region to be the same. Therefore, we took the average value of the focus of all pixels in the region as its focus degree, and the focus degree for a superpixel region \( r \) can be expressed as:

$$F(r) = \sum_{(x,y) \in r} \frac{F(x,y)}{A_r}$$  (9)

where \( A_r \) represents the number of pixels in the superpixel. Here, we utilized an object Gaussian model to analyze each pixel of the foreground layer. Therefore, foreground cues are calculated as:

$$FC(r) = F_F(r) \cdot \left( \exp \left( \frac{-(r-r_o)^2}{2\sigma^2} \right) \right)$$  (10)

where \( r_o \) represents the center of the superpixel where the center point \((\mu_x, \mu_y)\) of the object region is located. Similarly, we analyzed the background layer according to the object biased Gaussian model. Therefore, background cues are calculated as:

$$BC(r) = F_B(r) \cdot \left( 1 - \exp \left( \frac{-(r-r_o)^2}{2\sigma^2} \right) \right)$$  (11)

where \( F_F(r) \) and \( F_B(r) \) denote the focus degree of the foreground and background superpixel regions, respectively. Thereafter, we threshold the foreground cues and background cues, respectively, to obtain the foreground and background regions of the all-focus image, whose pixel values are binary and can be regarded as rough foreground saliency map and background saliency map.

3.4. Fusioning Focus Cues and GrabCut to Generate Saliency Map

GrabCut is a binary image segmentation algorithm. However, GrabCut requires human interaction to provide foreground or background seed point information to initialize the Gaussian mixture model [14]. In Section 3.3, we obtained the foreground and background regions of the image through focus cues, which provides a new way for the
GrabCut algorithm to provide initialization information automatically. Therefore, accurate foreground and background seed point information can make the saliency segmentation result more accurate. We performed morphological operations on images of FC and BC to remove isolated points while optimizing edge information using a conditional random field model.

In GrabCut, the foreground and background are modeled using two fully covariance Gaussian mixture models (GMM) with K components, one for the foreground and one for the background. Then, there are additional vectors \( k = \{k_1, \ldots, k_n, \ldots, k_N \} \), where \( k_N \in \{1, \ldots, K\} \). Assign a unique GMM component to each pixel, according to \( a_n = 0 \) or 1, determine whether it is the GMM component of the background model or the GMM component of the foreground model. Therefore, the Gibbs energy of the whole image can be expressed as:

\[
E(\hat{\alpha}, k, \hat{\theta}, z) = U(\hat{\alpha}, k, \hat{\theta}, z) + V(\hat{\alpha}, z)
\] (12)

where the region term \( U(\hat{\alpha}, k, \hat{\theta}, z) \) is related to the color Gaussian mixture model and is defined as follows:

\[
U(\hat{\alpha}, k, \hat{\theta}, z) = -\log \pi(\alpha_n, k_n) + \frac{1}{2} \log det \sum(\alpha_n, k_n) + \frac{1}{2} |z_n - \mu(\alpha_n, k_n)|^T \sum(\alpha_n, k_n)^{-1} [z_n - \mu(\alpha_n, k_n)]
\] (13)

where the initial value of \( \alpha_n \) is determined by FC and BC that belongs to the GMM of the foreground or background. \( \pi \) is the weight of the Gaussian component, \( \mu \) is the mean vector of the Gaussian component, and \( \sum \) is the covariance of the Gaussian components. According to the RGB value of the pixel and the parameters of the three Gaussian components, the probability of its foreground or background can be obtained. The boundary term \( V(\hat{\alpha}, z) \) is calculated by the Euclidean distance of RGB to measure the similarity of two neighboring pixels:

\[
V(\hat{\alpha}, z) = \gamma \sum_{(m,n) \in E} [\hat{\alpha}_m \neq \hat{\alpha}_n] \exp -\beta \|z_m - z_n\|^2
\] (14)

where \( \gamma \) is the constant 50 and \( \beta \) is the contrast factor. The total energy \( E(\hat{\alpha}, k, \hat{\theta}, z) \) is segmented by the max-flow/min-cut algorithm, and the foreground and background region item data of the next iteration are obtained, until the \( E(\hat{\alpha}, k, \hat{\theta}, z) \) iteration converges to obtain the saliency segmentation result.

4. Experiments

In this section, we conduct experiments to evaluate the performance of our algorithm. First, the experimental setup was outlined, including datasets, evaluation metrics, and parameter settings. Next, a quantitative evaluation of our light-field SOD method was compared with the 2D SOD method \[22,34,35\], the 3D SOD method \[36\], and the light field SOD method \[15\]. Finally, the experimental comparison with the traditional GrabCut algorithm was carried out to prove the effectiveness of the focus results in this paper for the interaction-free GrabCut algorithm.

4.1. Experimental Setup

4.1.1. Datasets

We conducted SOD experiments on the publicly available light field dataset provided by Li et al. \[29\]. The dataset was captured by a Lytro light field camera and includes 60 images of indoor scenes and 40 images of outdoor scenes. Each scene contains fully focused RGB images, a focus-stacked image focused at different depths, and a depth image to facilitate testing different saliency models. One part of the dataset has complex background colors and similar foreground colors, and the other contains multiple salient objects.
4.1.2. Evaluation Metrics

In order to quantitatively illustrate the performance of the algorithm, we used three popular SOD quantitative indicators: PR curve, F-Measure, and Mean Absolute Error (MAE) \[37\], which are introduced in detail next. In the PR curve, P represents the precision rate (Precision) and R represents the recall rate (Recall). Precision and Recall are calculated as:

\[
\begin{align*}
\text{Precision} &= \frac{\text{sum}(S, GT)}{\text{sum}(S)} \\
\text{Recall} &= \frac{\text{sum}(S, GT)}{\text{sum}(GT)}
\end{align*}
\]

(15)

where \( \text{sum} \) represents the sum of pixels with a value of 1. \( S \) means saliency map \( GT \) means ground truth. If the PR curve is closer to the (1,1) position, the higher the precision rate under the same recall rate, the better the algorithm performance.

F-Measure is a weighted harmonic evaluation of precision rate \( P \) and recall rate \( R \). F-Measure can be used to evaluate the performance of the algorithm by combining the two indicators, which is an overall measurement method. The formula of F-Measure is calculated as:

\[
F_\beta = \frac{(1 + \beta^2)P \times R}{\beta^2 \cdot P + R}
\]

(16)

According to the description in the literature \[37\], by setting the weight \( \beta^2 \) to avoid the problem of excessive recall caused by expanding the SOD area, this paper set \( \beta^2 \) to 0.3. The larger the F-Measure, the more stable the algorithm performance.

MAE stands for mean absolute error. The smaller the value of MAE, the smaller the error with the true value, and the more stable the algorithm is. It is defined as the pixel-averaged absolute difference between the saliency map \( S \) and the ground-truth map \( GT \):

\[
\text{MAE} = \frac{1}{w \times h} \sum_{x=1}^{w} \sum_{y=1}^{h} |S(x, y) - GT(x, y)|
\]

(17)

4.2. Comparison with Different Methods

For the SOD of light field images, we compared the SOD algorithms of different input data, including three 2D SOD algorithms based on all-focus images, Geodesic Saliency (GS \[36\]), Manifold Ranking (MR \[21\]), Saliency Optimization (wCtr \[22\]); a 3D SOD algorithm based on depth maps and all-focus images, Local Background Enclosure (LBE \[38\]); and a light-field SOD algorithm based on focus stacks and all-focus images, Light Field Saliency (LFS \[29\]). To ensure fairness, all saliency results from competing methods were produced by running the source codes or pre-computed by the authors.

4.2.1. Quantitative Evaluation

Figure 3 shows the quantitative evaluation of the method in this paper and other SOD algorithms on the PR curve. Figures 4 and 5 are the comparison of the corresponding F-Measure and MAE values, respectively. According to these three images, it is clear that the algorithm’s performance in this paper is higher than that of other SOD methods on the three evaluation indicators, which shows the superiority of the SOD based on the focus of the light field in this paper. Compared with the SOD algorithm, the fusion method of focus and color cues can better exert their interaction and obtain more accurate saliency results.
4.2.2. Qualitative Evaluation

Figure 6 shows some of the SOD results on the dataset. Most of these are complex scenes with similar foreground and background and cluttered backgrounds. Comparing
the detection results with the true value, we can see the algorithm performance’s pros and cons. The 2D SOD algorithm is more susceptible to complex background interference, salient objects cannot be well separated from the background, and the detection results are better when the scene is relatively simple. The 3D SOD algorithms cannot effectively detect salient objects when their depths are close to the background. The light field SOD algorithm achieves a certain separation between the salient objects and the background in the scene according to the focus cue. Compared with other saliency algorithms, the saliency map obtained by this algorithm has a clear outline. Although the algorithm in this paper only recognizes the label on the transparent object (see the sixth row in Figure 6), there are some misjudgments. However, for most salient regions, regardless of whether the foreground and background are similar or whether the scene is complex, this paper’s algorithm can highlight the salient area better. At the same time, with the advantage of GrabCut binary classification, the algorithm in this paper is a complete binary image compared with the saliency map generated by other algorithms.

| Images | 2D | 3D | Light Field |
|--------|----|----|-------------|
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |
| ![Image](image.png) | ![Image](image.png) | ![Image](image.png) | ![Image](image.png) |

**Figure 6.** The saliency map obtained by different salient object detection methods.

### 4.3. Validity Analysis of Focus Degrees

In this part, we analyzed the focus degree results generated by the direct gradient method and the gradient method in this paper, including the degree of suppression of the defocused area and the results generated by the fusion GrabCut. At the same time, we also compared with the results generated using the traditional GrabCut algorithm.

#### 4.3.1. Analysis of the Defocused Area Suppression Degree of the Focus Image

The quality of the focus results affects the accuracy of the focus cues. Therefore, to verify the feasibility of using the secondary blurring method to calculate the gradient change proposed in this paper, we compared the approach of this paper with that of Wang [28]. We calculated the average number of pixels in the defocused area in the foreground of some different scenes with $F(i,j) > 0$. As shown in Table 1, we can see that the average
number of pixels in the defocused area obtained by the secondary blurring method in this paper is reduced by 30% compared with the average number of pixels in the defocused area calculated by the direct gradient method. The experimental comparison verifies that the method based on secondary blurring to calculate the focus degree in this paper has a good suppression effect on the number of pixels in the defocused area compared with the method of gradient directly as the focus degree. This method has certain advantages in calculating the focus degree clues.

Table 1. The number of pixels in some foreground images of different focus calculation methods.

| Focus Method | Focused Area Pixels | Defocused Area Pixels |
|--------------|---------------------|----------------------|
| gradient     | 8439                | 2492                 |
| ours         | 5697                | 1583                 |

4.3.2. Accuracy Analysis of Extracted Foreground Images

A foreground image is a target region of interest extracted from an image, and the target region is generally the focused region of the image, while the defocused region can be regarded as the background image. If accurate foreground and background images can be obtained, the accuracy can be significantly improved in the subsequent SOD tasks.

Suppose $I^k$ is a focal image in the focus stack image. Assume $I^k$ is a focal image in the focus-stacked image. We computed the foreground score values for the focus-stacked image sets $I_1$ to $I_{11}$ according to Equation (8) using the focus degree algorithm in the existing light field saliency model and our focus degree algorithm, respectively, and the results are shown in Table 2.

Table 2. The calculation result of the foreground score value of the focus image.

|                  | OS($I^1$) | OS($I^2$) | OS($I^3$) | OS($I^4$) | OS($I^5$) | OS($I^6$) | OS($I^7$) | OS($I^8$) | OS($I^9$) | OS($I^{10}$) | OS($I^{11}$) | Time  |
|------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|--------------|--------------|-------|
| frequency        | 0.826     | 0.876     | 0.855     | 0.804     | 0.781     | 0.758     | 0.736     | 0.715     | 0.695     | 0.675        | 0.656        | 8.758 |
| domain-based     | 0.796     | 0.849     | 0.845     | 0.680     | 0.646     | 0.578     | 0.510     | 0.464     | 0.431     | 0.418        | 0.425        | 6.303 |
| gradient-based   | 0.961     | 0.975     | 0.977     | 0.850     | 0.772     | 0.563     | 0.436     | 0.362     | 0.333     | 0.324        | 0.325        | 6.430 |
| ours             |           |           |           |           |           |           |           |           |           |              |              |       |

From Table 2, we can see that the time consumption of this method and that of the gradient-based focus calculation method are close, but both are faster than the frequency-domain-based focus calculation method. The frequency domain-based focus calculation has a higher time complexity because the image is converted to the frequency domain for DCT transformation. Finally, in the foreground scoring value, the frequency domain-based method does not vary significantly in the scoring value of each focus image. This is because the frequency domain is more sensitive to high frequencies, and in complex scenes, there may be high-frequency responses even in the scattered focus region, so the focus images of different depths may produce high-frequency responses in the target region. However, the gradient-based and our method significantly differ in the focus images at different depths.

4.3.3. SOD Results Using GrabCut with Different Focus Methods

We calculated its two evaluation indicators, F-Measure and MAE, on the LFSD dataset, as shown in Figure 7. According to the F-Measure and MAE values, it can be seen that the results generated by the GrabCut calculation method based on the focus degree in this paper are higher than the calculation method based on the gradient as the focus degree in both evaluation indicators.

Figure 8 shows the partial results of the two focusing degree calculation methods using GrabCut to calculate the saliency. It can be seen from the first and second rows that, in a simple scene, the saliency results obtained by the two focusing degrees can accurately highlight objects. In the complex scene where the third to fifth rows are located, the difference in the focus calculation method generates different focus cues, making the seed points of the color space model in the GrabCut algorithm different, resulting in the
unclear segmentation of significant objects. To sum up, the focus calculation method proposed in this paper suppresses the focus information of the defocused area and reduces the interference of the background area to the salient objects.

![Figure 7. F-Measure and MAE obtained by GrabCut for two focus calculation methods.](image)

![Figure 8. The saliency map obtained using GrabCut for two focus calculation methods.](image)

4.3.4. Comparison with the Traditional GrabCut Algorithm

Using human interaction, we generated saliency maps for GrabCut annotating foreground and background regions. Figure 9 shows the comparison between the results of our method and the traditional GrabCut method. Although the boundary information of foreground and background is provided by manual interaction, it can be seen from the results in the third row that there are still some missed detections in the interior of the salient objects. At the same time, in the fifth row, from the root of the flower, it can be seen that, if the seed point is not fully labeled at the edge, the missed detection still occurs, and the focus clue provided in this paper is the focus result of the light field focus-stacked image at the foreground and background depths.

Figure 10 compares the F-Measure and MAE values generated by the GrabCut algorithm with and without manual interaction. The results of this paper are slightly higher than those of the manual interaction method in the evaluation of the two indicators. It can be seen that the focus clue provided in this paper is significant for GrabCut interaction-free effectiveness.
5. Discussion

In this paper, we proposed a light field salient object detection algorithm that integrates focus and GrabCut, suppressing redundant information of focus cues and making the generated saliency map clear. In addition, in order to adapt to light field saliency detection in different complex scenes and maximize the complementarity of focus cues and color cues, the algorithm proposed in this paper combines the characteristics of the GrabCut algorithm to use focus cues to guide color cues to establish an initialized color space model. Interaction-free saliency detection for the foreground of light field images was achieved. The experimental results show that the proposed algorithm can achieve better saliency detection results in different light field scenarios compared with other algorithms by testing in the public light field dataset. Finally, through experimental comparison with the traditional manual interactive GrabCut algorithm, the seed point information provided by the focus of the light field is not much different from the saliency results generated by the manually marked seed point information, achieving the advantages of the focus of the light field and GrabCut.

However, our method still has certain shortcomings. The effect of the algorithm in this paper is more dependent on the accuracy of the focus cues. Although reliable seed
point information can be provided in most scenes, when the edge information of the focus area of the transparent object or the focus image is not rich, the focus result information of this paper is not enough. Some, although few, scenes will be missed. In Figure 6, the transparent objects make the model misjudge the background information and only detect the label area of the bottle. We believe that subsequent references to depth maps can help focus cues to determine thresholds.

**Author Contributions:** Conceptualization, F.D., C.W. and H.G.; Data curation, C.W. and Y.W.; Formal analysis, F.D., C.W. and Y.W.; Funding acquisition, H.G.; Investigation, C.W.; Methodology, F.D. and C.W.; Project administration, F.D. and H.G.; Resources, H.G.; Software, C.W. and Y.W.; Supervision, F.D. and H.G.; Writing—original draft, C.W.; Writing—review and editing, Y.W. All authors will be informed about each step of manuscript processing including submission, revision, revision reminder, etc. via emails from our system or assigned Assistant Editor. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work was supported by the National Key R&D Program of China (No. 2018YFC1800904); Capacity Building for Sci-Tech Innovation—Fundamental Scientific Research Funds (No. 20530290078).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data presented in this study are available on request from the corresponding author. The data are not publicly available due to privacy reason.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**

1. Photographique, A.R.A. [M. Gabriel Lippmann, dans son laboratoire de la Sorbonne]: [photographie de presse]/[Agence Rol]. 1908. Available online: https://gallica.bnf.fr/ark:/12148/btv1b69048248.item (accessed on 25 December 2021).
2. Gershun, A. The Light Field. *Stud. Appl. Math.* **1939**, *18*, 51–151. [CrossRef]
3. Adelson, E.H. The Plenoptic Function and the Element of Early Vision. Comput. Model. Vis. Process. **1991**, *1*, 8.
4. Mcmillan, G.L. Plenoptic modeling: An image-based rendering system. In Proceedings of the 22nd Annual Conference on Computer Graphics and Interactive Techniques, Los Angeles, CA, USA, 6–11 August 1995; pp. 39–46.
5. Levoy, P.M. Light Field Rendering. In Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH’96, New Orleans, LA, USA, 4–9 August 1996; pp. 31–42.
6. Ren, N.; Levoy, M.; Bredif, M.; Duval, G.; Horowitz, M.; Hanrahan, P. Light Field Photography with a Hand-Held Plenoptic Camera. Ph.D. Thesis, Stanford University, Palo Alto, CA, USA, 2005.
7. Ullah, I.; Jian, M.; Hussain, S.; Guo, J.; Yu, H.; Wang, X.; Yin, Y. A brief survey of visual saliency detection. *Multimed. Tools Appl.* **2020**, *79*, 34605–34645. [CrossRef]
8. Jiang, Y.; Zhou, T.; Ji, G.P.; Zhou, T.; Zhao, Q.; Fan, D.P. Light Field Salient Object Detection: A Review and Benchmark. *Comput. Vis. Media* **2020**, *8*, 509–534.
9. Zhang, X.; Li, C. Model of Light Field Imaging and its Application in Refocus and All-Focus. *Adv. Mater. Res.* **2014**, *1039*, 266–273. [CrossRef]
10. Ouazan, A.; Kovacs, P.T.; Balogh, T.; Barsi, A. Rendering multi-view plus depth data on light-field displays. In Proceedings of the 2011 3DTV Conference: The True Vision—Capture, Transmission and Display of 3D Video (3DTV-CON), Antalya, Turkey, 16–18 May 2011.
11. Zhang, S.; Sheng, H.; Li, C.; Zhang, J.; Xiong, Z. Robust depth estimation for light field via spinning parallelogram operator. *Comput. Vis. Image Underst.* **CVIU** **2016**, *145*, 148–159. [CrossRef]
12. Jiang, P.; Ling, H.; Yu, J.; Peng, J. Salient Region Detection by UFO: Uniqueness, Focusness and Objectness. In Proceedings of the 2013 International Conference on Computer Vision (ICCV), Sydney, NSW, Australia, 1–8 December 2013.
13. Gill, A.; Zerman, E.; Alain, M.; Le Pendu, M.; Smolic, A. Focus Guided Light Field Saliency Estimation. In Proceedings of the 2021 13th International Conference on Quality of Multimedia Experience (QoMEX), Montreal, QC, Canada, 14–17 June 2021.
14. Rother, C. GrabCut: Interactive foreground extraction using iterated graph cut. *ACM Trans. Graph.* **2004**, *23*, 309–318. [CrossRef]
15. Niu, N.N.; Han, X.; Ding, J.H. Light field image refocus method for Lytro camera. *Comput. Eng. Des.* **2016**, *7*, 1847–1850.
16. Takahashi, K.; Naemura, T. Layered light-field rendering with focus measurement. *EURASIP Signal Process. Image Commun.* **2006**, *21*, 519–530. [CrossRef]
17. Itti, L.; Koch, C.; Niebur, E. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [CrossRef]
18. Cheng, M.M.; Zhang, G.X.; Mitra, N.J.; Torr, P.H.; Hu, S.M. Global Contrast Based Salient Region Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR’11), NW Washington, DC, USA, 20–25 June 2011.
19. Duan, L.; Wu, C.; Miao, J.; Qing, L.; Fu, Y. Visual saliency detection by spatially weighted dissimilarity. In Proceedings of the CVPR 2011, Colorado Springs, CO, USA, 20–25 June 2011.

20. Yu, Z.; Guo, X.; Ling, H.; Lumsdaine, A.; Yu, J. Line Assisted Light Field Triangulation and Stereo Matching. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013.

21. Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; Yang, M.-H. Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 3166–3173.

22. Zhu, W.; Liang, S.; Wei, Y.; Sun, J. Saliency Optimization from Robust Background Detection. In Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 2814–2821.

23. Poggio, G.F.; Poggio, T. The Analysis of Stereopsis. *Annu. Rev. Neurosci.* 1984, 7, 379–412. [CrossRef] [PubMed]

24. Xia, L.; Chen, C.C.; Aggarwal, J.K. Human detection using depth information by Kinect. In Proceedings of the Computer Vision & Pattern Recognition Workshops, Colorado Springs, CO, USA, 20–25 June 2011.

25. Ren, J.; Gong, X.; Lu, Y.; Zhou, W.; Ying Yang, M. Exploiting global priors for RGB-D saliency detection. In Proceedings of the Computer Vision & Pattern Recognition Workshops, Boston, MA, USA, 7–12 June 2015.

26. Jost, T.; Ouerhani, N.; Wartburg, R.V.; Müri, R.; Hügli, H. Contribution of depth to visual attention: Comparison of a computer model and human behavior. *Proc. Early Cogn. Vis. Workshop* 2004, 1–4.

27. Ran, J.; Gong, L.; Geng, W.; Ren, T.; Wu, G. Depth saliency based on anisotropic center-surround difference. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2015.

28. Feng, H.; Bing, L.; Xiong, W.; Hu, W.; Ji, R. RGBD Salient Object Detection: A Benchmark and Algorithms. In *European Conference on Computer Vision*; Springer: Cham, Switzerland, 2014.

29. Li, N.; Ye, J.; Ji, Y.; Ling, H.; Yu, J. Saliency Detection on Light Field. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2014, Columbus, OH, USA, 23–28 June 2014.

30. Wang, H.; Yan, B.; Wang, X.; Zhang, Y.; Yang, Y. Accurate saliency detection based on depth feature of 3D images. *Multimed. Tools Appl.* 2017, 77, 14655–14672. [CrossRef]

31. Wang, A.; Wang, M.; Li, X.; Mi, Z.; Zhou, H. A Two-Stage Bayesian Integration Framework for Salient Object Detection on Light Field. *Neural Process. Lett.* 2017, 46, 1083–1094. [CrossRef]

32. Zhou, J.; Naren, T.; Chen, X.; Ma, Y.; Liu, J.; Dai, F. Light Field Foreground Matting Based on Defocus and Correspondence. In *International Conference on Multimedia Modeling*; Springer: Cham, Switzerland, 2018.

33. Piao, Y.; Li, X.; Zhang, M.; Yu, J.; Lu, H. Saliency Detection via Depth-Induced Cellular Automata on Light Field. *IEEE Trans. Image Process.* 2019, 28, 1879–1889. [CrossRef] [PubMed]

34. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Trans. Pattern Anal. Mach. Intell.* 2012, 34, 2274–2282. [CrossRef] [PubMed]

35. Créte-Roffet, F.; Dolmiere, T.; Ladret, P.; Nicolas, M. The Blur Effect: Perception and Estimation with a New No-Reference Perceptual Blur Metric. In Proceedings of the OAI, 2007:649201-649201-11, San Jose, CA, USA, 12 February 2007.

36. Wei, Y.; Fang, W.; Jian, S. Geodesic Saliency Using Background Priors. WO2015013908 A1, 4 July 2017.

37. Achanta, R.; Hemami, S.; Estrada, F.; Süssstrunk, S. Frequency-tuned salient region detection. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009.

38. Feng, D.; Barnes, N.; You, S.; McCarthy, C. Local Background Enclosure for RGB-D Salient Object Detection. In Proceedings of the Computer Vision & Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.